QUERY PROCESSING FOR RETRIEVAL FROM LARGE TEXT BASES

John Broglio and W. Bruce Croft

Computer Science Department
University of Massachusetts
Amherst, MA 01003

ABSTRACT

Natural language experiments in information retrieval have often been inconclusive due to the lack of large text bases with associated queries and relevance judgments. This paper describes experiments in incremental query processing and indexing with the INQUERY information retrieval system on the TIPSTER queries and document collection. The results measure the value of processing tailored for different query styles, use of syntactic tags to produce search phrases, recognition and application of generic concepts, and automatic concept extraction based on interword associations in a large text base.

1. INTRODUCTION: TIPSTER AND INQUERY

Previous research has suggested that retrieval effectiveness might be enhanced by the use of multiple representations and by automated language processing techniques. Techniques include automatic or interactive introduction of synonyms [Har88], forms-based interfaces [CD90], automatic recognition of phrases [CTL91], and relevance feedback [SB90]. The recent development of the TIPSTER corpus with associated queries and relevance judgments has provided new opportunities for judging the effectiveness of these techniques on large heterogenous document collections.

1.1. TIPSTER Text Base and Query Topics

The TIPSTER documents comprise two volumes of text, of approximately one gigabyte each, from sources such as newspaper and magazine articles and government publications (Federal Register). Accompanying the collections are two sets of fifty *topics*. Each topic is a full text description, in a specific format, of an information need. (Figure 1).

Each TIPSTER topic offers several representations of the same information need. The *Topic* and *Description* fields are similar to what might be entered as a query in a traditional information retrieval system. The *Narrative* field expands on the information need, giving an overview of the classes of documents which would or would not be considered satisfactory, and describes facts that must be present in relevant documents, for example, *the location of the company*. The *Concepts* field lists words and phrases which are pertinent to the query. The *Factors* field lists constraints on the geographic and/or time frame of the query. All of these fields offer opportunities for different kinds of natural language processing.

<top>
<dom> Domain: International Economics
<Title> Topic: Satellite Launch Contracts
<desc> Description:
Document will cite the signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite.

<narr> Narrative:
A relevant document will mention the signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite.

<con> Concept(s):
1. contract, agreement
2. launch vehicle, rocket, payload, satellite
3. launch services, commercial space industry, commercial launch industry
4. Arianespace, Martin Marietta, General Dynamics, McDonnell Douglas
5. Titan, Delta II, Atlas, Ariane, Proton

</top>

Figure 1: A TIPSTER topic.

1.2. The INQUERY Information Retrieval System

INQUERY is a probabilistic information retrieval system based upon a Bayesian inference network model [TC91, Tur91]. The object network consists of object nodes (documents) (o_j's) and concept representation nodes (r_m's). In a typical network information retrieval system, the text representation nodes will correspond to

words extracted from the text [SM83], although representations based on more sophisticated language analysis are possible. The estimation of the probabilities $P(r_m|o_j)$ is based on the occurrence frequencies of concepts in both individual objects and large collections of objects. In the INQUERY system, representation nodes are the word stems and numbers that occur in the text, after stopwords are discarded.

2. QUERY PROCESSING EXPERIMENTS

Our current set of natural language techniques for query enhancement are:

- deletion of potentially misleading text;
- grouping of proper names and interrelated noun phrase concepts;
- automatic concept expansion;
- simple rule-based interactive query modification.

Future experiments will use more extensive automatic noun phrase processing and paragraph level retrieval.

In addition to the traditional recall/precision table, we show tables of the precision for the top n documents retrieved, for 5 values of n. The recall/precision table measures the ability of the system to retrieve all of the documents known to be relevant. The precision for the top n documents gives a better measure of what a person would experience in using the system.

2.1. Deletion processes.

Table 1 illustrates an incremental query treatment. The (**Words**) column shows results from the unprocessed words of the query alone. (Formatting information, such as field markers, has been removed.) The first active processing (**Del1**) removes words and phrases which refer to the information retrieval processes rather than the information need, for example, *A relevant document will describe* .... We further remove words and phrases which are discursive, like *point of view*, *sort of*, *discuss*, *mention* as well as expressions which would require deep inference to process, such as *effects of* or *purpose of* (Figure 2). Some of these expressions would be useful in other retrieval contexts and different lists would be appropriate in different domains. An interactive user is given feedback regarding deletions and could have the capability of selectively preventing deletion.

In the experiment in the fourth column (**-NARR**) the Narrative field has been deleted from each query. Since the Narrative field is usually a very abstract discussion of the criteria for document relevance, it is not well-suited to a system like INQUERY, which relies on matching words from the query to words in the document. New terms introduced by the Narrative field are rarely useful as retrieval terms (but note the small loss in precision at the very lowest level of recall).

2.2. Grouping Noun Phrases and Recognizing Concepts

The simplest phrasing or grouping techniques are recognition of proper noun groups (**Caps** in Table 1) and recognition of multiple spellings for common concepts such as *United States*.

Proximity and phrase operators for noun phrases. Simple noun phrase processing is done in two ways. Sequences of proper nouns are recognized as names and grouped as arguments to a proximity operator. The proximity operator requires that its arguments appear in strict order in a document, but allows an interword distance of three or less. Thus a query such as *George Bush* matches *George Herbert Walker Bush* in a document.

Secondly, the query is passed through a syntactic part of speech tagger [Chu88], and rules are used rules to identify noun phrases (Figure 2). Experiments showed that very simple noun phrase rules work better than longer, more complex, noun phrases. We believe this is because the semantic relationships expressed in associated groups of noun phrases in a query may be expressed in a document as a compound noun group, a noun phrase with prepositional phrase arguments, a complex sentence, or a sequence of sentences linked by anaphora. This hypothesis is supported by the success of the unordered text window operator used in the interactive query modification experiments (Table 4).

On the other hand, there are verbal "red herrings" in some query noun phrases due to overprecise expression. For example, the phrase *U.S. House of Representatives* would be more effective for retrieval without the *U.S.* component (*Congress* might be even nicer).

2.3. Concept Recognition

Controlled vocabulary. The INQUERY system has been designed so that it is easy to add optional object types to implement a controlled indexing vocabulary [CCH92]. For example, when a document refers to a company by name, the document is indexed both by the the company name (words in the text) and the object type (*#company*). The standard INQUERY document parsers recognize the names of companies [Rau91], coun-

Table 1: Precision and recall tables for experiments starting with *words-only* queries (**Words**) through phrase (**Del1**) and word (**Del2**) deletion to proper noun (*Caps*) and noun phrase (**NP**) grouping. The queries were evaluated on Volume 1 of the TIPSTER document collection, using relevance judgements from the 1992 Text Retrieval and Evaluation Conference (TREC).

	Precision (% change) – 50 queries											
Recall	Words	Del1		Del2		-Narr		Caps		NP		
0	71.6	73.5	(+ 2.7)	76.2	(+ 6.4)	83.2	(+16.2)	81.9	(+14.4)	83.5	(+16.6)	
10	49.2	52.7	(+ 7.0)	54.7	(+11.0)	59.6	(+21.1)	60.0	(+21.9)	62.9	(+27.8)	
20	41.2	44.2	(+ 7.5)	46.1	(+12.1)	50.6	(+22.9)	51.3	(+24.6)	54.5	(+32.4)	
30	35.3	38.9	(+10.4)	40.5	(+14.8)	45.2	(+28.2)	45.9	(+30.1)	48.8	(+38.5)	
40	30.7	34.6	(+12.6)	35.9	(+17.1)	39.9	(+30.0)	40.5	(+32.1)	43.6	(+42.1)	
50	26.2	30.3	(+15.6)	31.7	(+21.1)	35.9	(+37.1)	35.6	(+36.0)	37.8	(+44.1)	
60	22.1	25.5	(+15.5)	26.9	(+21.8)	31.0	(+40.4)	30.9	(+40.3)	32.6	(+47.9)	
70	18.7	21.1	(+12.9)	22.0	(+17.9)	26.1	(+40.0)	25.8	(+38.2)	27.2	(+46.1)	
80	15.0	17.0	(+13.4)	17.8	(+18.4)	20.5	(+36.6)	19.9	(+32.8)	21.4	(+42.6)	
90	9.2	10.5	(+13.7)	11.1	(+20.0)	12.7	(+37.3)	12.3	(+33.4)	12.9	(+39.8)	
100	2.4	2.8	(+19.9)	3.2	(+33.8)	2.6	(+10.2)	2.5	(+ 5.2)	2.9	(+23.2)	
avg	29.2	31.9	(+ 9.2)	33.3	(+13.9)	37.0	(+26.7)	37.0	(+26.5)	38.9	(+33.2)	

	Precision (% change) – 50 queries											
Recall	Words	Del1		Del2		-Narr		Caps		NP		
5	54.4	57.2	(+ 5.1)	58.4	(+ 7.4)	66.4	(+22.1)	65.6	(+20.6)	66.8	(+22.8)	
15	46.4	49.7	(+ 7.1)	50.9	(+ 9.7)	57.1	(+23.1)	57.5	(+23.9)	62.8	(+35.3)	
30	44.2	47.2	(+ 6.8)	49.3	(+11.5)	53.6	(+21.3)	53.3	(+20.6)	56.3	(+27.4)	
100	33.9	37.0	(+ 9.1)	38.7	(+14.2)	43.0	(+26.8)	43.2	(+27.4)	45.0	(+32.7)	
200	27.5	30.1	(+ 9.5)	31.5	(+14.5)	35.4	(+28.7)	35.2	(+28.0)	37.2	(+35.3)	

tries, and cities in the United States.

With wide-ranging queries like the TIPSTER topics, we have had some success with adding *#city* (and *#foreign-country*) concepts to queries that request information on the *location* of an event (Table 2). But the terms *#company* and *#usa* have not yet proved consistently useful. The *#company* concept may be used to good effect to restrict other operators. For example, looking for the terms *machine, translation*, and *#company* in an *n*-word text window would give good results with respect to companies working on or marketing machine translation products. But, the current implementation of the *#company* concept recognizer has some shortcomings which are exposed by this set of queries. Our next version of the recognizer will be more precise and complete[1], and we expect significant improvement from these it.

The *#usa* term tends to have unexpected effects, because a large part of the collection consists of articles from U.S. publications. In these documents U.S. nationality is often taken for granted (term frequency of *#usa*=294408, *#foreigncountry*=472021), and it is likely that it may be mentioned explicitly only when that presupposition is violated, or when both U.S. and non-U.S. issues are being discussed together in the same document. Therefore, because focussing on the *#usa* concept will bring in otherwise irrelevant documents, it is more effective to put negative weight on the *#foreign-country* concept where the query interest is restricted to U.S. matters. For the same reason, in a query focussed only on non-U.S. interests, we would expect the opposite: using *#foreigncountry* should give better performance than *#NOT(#usa)*.

Research continues on the 'right' mix of concept recognizers for a document collection. In situations where text and queries are more predictable, such as commercial customer support environments, an expanded set of special terms and recognizers is appropriate. Names of products and typical operations and objects can be recognized and treated specially both at indexing and at query time. Our work in this area reveals a significant improvement due to domain-specific concept recognizers, however, standardized queries and relevance judgments are still being developed.

[1] Ralph Weischedel's group at BBN have been generous in sharing their company database for this purpose.

Original:
Document will cite the signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite.

Discourse phrase and word deletion:
the signing of a contract or preliminary agreement, or the making of a tentative reservation, to launch a commercial satellite.

Proper noun group recognition (Concept field):
#3(Martin Marietta) #3(General Dynamics) #3(McDonnell Douglas) #3(Delta II)

Noun phrase grouping (and stopword deletion):
#PHRASE (signing contract)
#PHRASE (preliminary agreement)
#PHRASE (making tentative reservation)
#PHRASE (commercial satellite)

Figure 2: Progressive changes in the *Description* field of the Topic.

Automatic concept expansion. We have promising preliminary results for experiments in automatic concept expansion. The **Expand** results in Table 3 were produced by adding five additional concepts to each query. The concepts were selected based on their preponderant association with the query terms in text of the 1987 Wall Street Journal articles from Volume 1 of the TIPSTER corpus. The improvement is modest, and we anticipate better results from refinements in the selection techniques and a larger and more heterogenous sample of the corpus.

2.4. Semi-Automatic query processing.

In the following experiments in interactive query processing, human intervention was used to modify the output of the best automatic query processing. The person making the modifications was permitted to

1. Add words from the Narrative field;
2. Delete words or phrases from the query;
3. Specify a text window size for the occurrence of words or phrases in the query.

The third restriction simulates a paragraph-based retrieval.

Table 4 summarizes the results of the interactive query modification techniques compared with the best automatic query processing **Q-1** (similar to **NP** in the other tables). The **Q-M** query-set was created with rules (1) and (2) only. The **Q-O** query-set used all three rules.

Table 2: The effect of replacing the query word *location* with the concepts *#us-city* and *#foreigncountry*. (We do not yet have a *#foreigncity* recognizer).

	Precision (8 queries)				
Recall	NoCity	— City —		– City+FC –	
25	45.8	46.7	(+2.0)	46.8	(+2.3)
50	30.3	30.4	(+0.2)	30.7	(+1.2)
75	15.0	14.9	(−1.2)	15.2	(+1.4)
avg	30.4	30.6	(+0.9)	30.9	(+1.8)

The improvement over the results from automatically generated queries demonstrates the effectiveness of simple user modifications after automatic query processing has been performed. The most dramatic improvement comes at the top end of the recall scale, which is a highly desirable behavior in an interactive system. The results also suggest that, based on the text window simulation, paragraph-based retrieval can significantly improve effectiveness.

3. CONCLUSION

The availability of the large TIPSTER text base and query sets has enabled us to undertake a series of experiments in natural language processing of documents and queries for information retrieval. We have seen steady improvements due to lexical and phrase-level processing of natural language queries. Our experiments with interactive modification of the resulting queries indicate how much potential gain there is in this area, provided we can refine our phrasing and selection criteria, and provided actual paragraph retrieval is at least as good as our text window simulation of it. Refinement of our recognition and use of controlled indexing vocabulary is already showing benefits in more predictable domains, and we expect to see improvement in the results in the TIPSTER queries as well.

The experiments in automatic concept expansion based on cooccurrence behavior in large corpora are extremely interesting. Although the effects shown here are very preliminary, it is reassuring that they are positive even at this early stage.

It is clear that incremental application of local (word and phrase-level) natural language processing is beneficial in information retrieval. At this stage, the only expected limits to this approach are represented by the improvement achieved with the experiments in interactive query modification.

Table 3: Automatic concept expansion (**Expand**) compared with the automatic query baseline (**NP**).

	Precision (50 queries)		
Recall	NP	– Expand –	
0	77.1	75.2	(−2.4)
10	55.2	56.1	(+1.7)
20	48.3	49.0	(+1.4)
30	41.5	43.0	(+3.4)
40	36.7	37.7	(+2.8)
50	32.0	32.9	(+3.0)
60	27.9	27.9	(+0.3)
70	22.1	22.9	(+3.5)
80	17.5	18.0	(+2.8)
90	12.5	12.8	(+2.7)
100	2.4	2.7	(+12.1)
avg	33.9	34.4	(+1.4)

	Precision (50 queries)		
Recall (#Docs)	NP	– Expand –	
5	58.4	58.0	(−0.7)
15	51.5	53.5	(+3.9)
30	48.7	50.1	(+2.9)
100	34.6	35.5	(+2.6)
200	26.3	26.9	(+2.3)

Table 4: A comparison of two semi-automatic methods of constructing adhoc queries. The methods were evaluated on Volume 1 of the TIPSTER document collection, using relevance judgements from the 1992 Text Retrieval and Evaluation Conference (TREC).

	Precision (50 queries)				
Recall	Q-1	– – Q-M – –		– – Q-O – –	
0	83.9	83.8	(−0.2)	93.0	(+10.8)
10	60.5	64.1	(+6.0)	71.6	(+18.3)
20	52.7	55.4	(+5.1)	63.4	(+20.3)
30	46.6	48.6	(+4.3)	54.2	(+16.3)
40	40.5	42.1	(+3.9)	46.8	(+15.5)
50	35.0	36.4	(+4.1)	40.4	(+15.6)
60	30.5	30.9	(+1.5)	34.1	(+11.8)
70	25.4	25.0	(−1.4)	28.4	(+11.6)
80	19.9	18.3	(−7.8)	21.7	(+ 9.1)
90	12.1	11.8	(−3.0)	13.4	(+10.3)
100	2.5	2.3	(−6.5)	2.4	(− 2.5)
avg	37.2	38.1	(+2.3)	42.7	(+14.6)

	Precision (50 queries)				
Recall (#Docs)	Q-1	– – Q-M – –		– – Q-O – –	
5	64.8	67.2	(+3.7)	76.4	(+17.9)
15	59.2	63.9	(+7.9)	72.4	(+11.7)
30	54.1	57.5	(+6.3)	64.9	(+20.0)
100	42.4	45.5	(+7.3)	49.4	(+16.5)
200	35.6	36.7	(+3.1)	39.2	(+10.1)

References

[CCH92] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag, 1992.

[Chu88] Kenneth Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, 1988.

[CD90] W. B. Croft and R. Das. Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 349–368, 1990.

[CTL91] W. B. Croft, H.R. Turtle, and D.D. Lewis. The use of phrases and structured queries in information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 32–45, 1991.

[Har88] D. Harman. Towards interactive query expansion. In Y. Chiaramella, editor, *Proceedings of the 11th International Conference on Research and Development in Information Retrieval*, pages 321–332. ACM, June 1988.

[Rau91] Lisa F. Rau. Extracting company names from text. In *Proceedings of the Sixth IEEE Conference on Artificial Intelligence Applications*, 1991.

[SM83] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[SB90] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41:288–297, 1990.

[TC91] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), July 1991.

[Tur91] Howard Robert Turtle. *Inference networks for document retrieval.* PhD thesis, Department of Computer and Information Science, University of Massachusetts, Amherst, 1991.