The TREC-9 Query Track

Chris Buckley
Sabir Research, Inc
chrisb@sabir.com

1 Introduction

The Query Track in TREC-9 is unlike the other tracks in TREC. The other tracks attempt to compare systems to determine the best approaches to solve the particular track problem. This comparison is normally done over a given set of topics, with a single query per topic. The Query Track, on the other hand, compares multiple queries on a single topic to determine which queries perform best with which systems. There is no emphasis on system-system comparisons: none of the participating systems were even the most advanced system from that particular participating group. Instead, the goal is to try and understand how the statement (query) of the user's information need (topic) affects retrieval.

Information Retrieval is a somewhat odd discipline. It's one where a human can do much better than any IR system, given infinite time and patience. Given any particular information need and some representative relevant documents, a user can often find an automatic retrieval strategy that does much better than an IR system. But, as any experienced IR system designer knows, implementing such a strategy may improve performance on this query and/or topic, but end up hurting performance on other types of queries and/or topics. Humans are remarkably adept at finding different ways to express similar ideas in both queries and documents; this variability is the heart of the difficulty of the information retrieval task.

The Query Track is an attempt to isolate some of the issues dealing with query and topic variability. Automatic IR systems perform tremendously differently across a typical IR task such as in the Web Track, but much of this variability is concealed by the evaluation averages. What is often quite surprising, especially to people just starting to look at IR, is the large variability in system performance across topics as compared to other systems. In a typical TREC task, no system is the best for all the topics in the task. It is extremely rare for any system to be above average for all the topics. Instead, the best system is normally above average for most of the topics, and best for maybe 5%-10% of the topics. It very often happens that quite below-average systems are also best for 5%-10% of the topics, but do poorly on the other topics. The Average Precision Histograms presented on the TREC evaluation result pages are an attempt to show what is happening at the individual topic level.

One of the major purposes of the Query Track is to try to understand how much of the system variability is due to issues of *how* the user's information need is being expressed

(the query syntax), and how much is due to *what* the information need is (topic semantics).

1.1 Query vs Topic

For the purposes of this track, a *topic* is considered an information need of a user. It includes a full statement of what information is wanted as well as information the user knows that pertains to the request. A *query* is what the user actually types to a retrieval system. It is much shorter than a topic, but is the only direct information from the user that the system has. Topic 51 (the first topic used in the Query Track) is given below. A query corresponding to Topic 51 might be something as simple as "Airbus subsidies".

TOPIC 51

<top>
<head> Tipster Topic Description
<num> Number: 051
<dom> Domain: International Economics
<title> Topic: Airbus Subsidies
<desc> Description:Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<smry> Summary:Document will discuss government assistance to Airbus Industrie, or mention atrade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.
<narr> Narrative:A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S.government, over federal subsidies to Airbus.
<con> Concept(s):
1. Airbus Industrie
2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.
3. federal subsidies, government assistance, aid, loan, financing
4. trade dispute, trade controversy, trade tension
5. General Agreement on Tariffs and Trade (GATT) aircraft code
6. Trade Policy Review Group (TPRG)
7. complaint, objection
8. retaliation, anti-dumping duty petition, countervailing duty petition, sanctions

<def> Definition(s): ...

1.2 Issues to Examine

There are a number of issues that we wish to examine in both last year's and this year's Query Track data, and in the future with the NIST Query Station. They include

- Can we distinguish between easy and hard queries/topics?
  - Are queries hard or are topics hard?
  - Even if we can distinguish this from the results, can NLP analysis of a query distinguish this before-hand?
- What categories of queries can potentially yield performance differences?
- Where do query performance differences come from?
  - Examine system vs topic vs query.

- Can we easily create test collections with large numbers of queries with judgments?

If we can answer these questions, then we may make it possible to improve retrieval systems dramatically.

2 Query Track Test Collection Creation

The construction of the Query Track test collection consists of 2 sub-tasks. In the first sub-task, groups take each of topics 51-100 from TREC 1 and create one or more queries based on the topic. In the second sub-task, each group runs one or more versions of their system on all the queries from all the groups. The results are then evaluated and analysis can begin!

2.1 Query Creation Sub-Task

Groups create one or more versions of each of TREC topics 51-100 in categories

- Very short: 2-4 words based on the topic and possibly a few relevant documents from TREC disk 2.
- Sentence: 1-2 sentences using topic and relevant documents.
- Sentence-Feedback only: 1-2 sentences using only the relevant documents. The aim is to increase vocabulary variability.

This is the second (and final) year of the Query Track. Last year there were five participating groups who produced 23 Query Sets. Each query set consisted of 50 queries corresponding to topics 51-100. Two of the Query Sets were not natural language (lists of weighted terms) and were not re-used. The other 21 Query Sets were used again this year. To this we added another 22 Query Sets, giving us a total of 43 Query Sets from 6 groups.

APL	INQ	Sab	Acs	Pir	Uof M
Johns Hopkins	Umass	Sabir	Acsys	Queens	Melbourne
Expert	Students	Expert	Expert	Expert	Expert
1 short 1 sent.	10 short 10 sent 10 fdbk	4 short 1 sent 1 fdbk	1 short	1 short	2 short 1 sent

Several versions of queries for topic 51 are given below. It was quite surprising how few duplicate queries there were. There were 2150 original queries. Of those, 1982 were unique after removing spaces, extra punctuation, and capitalization. After that, if hyphens were removed there were 1973 unique queries left. Every topic had at least 33 unique queries (out of the 43 possible.)

Sample of queries for Topic 51

•51 01 recent airbus issues
•51 02 Airbus subsidies dispute
•51 03 Airbus subsidy battle
•51 04 Airbus subsidies dispute
•51 05 U.S. Airbus subsidies
•51 06 What are the reactions of American companies to the trade dispute and how the dispute progresses?
•51 07 What are the issues being debated regarding complaints against Airbus Industrie?
•51 08 News related to the Airbus subsidy battle.
•51 09 U.S. and Europe dispute over Airbus subsidies
•51 10 Is European government risking trade conflicts over issue of Airbus subsidies?

2.2 Retrieval Sub-Task

After the Query Sets were constructed, they were distributed to all the groups to run one or more retrieval runs on the TREC Disk 1 document collection (about 510,000 documents). Six groups performed 18 retrieval runs:

- INQ: 3 runs
  - only query terms
  - query terms plus structure
  - query terms plus structure plus blind feedback
- SUN: 2 runs
  - Used two slightly different versions of their Question Answering Track engine
- Sab: 3 runs
  - query terms plus adjacency phrases
  - query terms plus phrases plus 7 terms expansion from blind feedback
  - query terms plus phrases plus 60 terms expansion
- UoM: 2 runs - no expansion
- hum: 7 runs
  - baseline, linguistic morphology
  - spelling correction (for words occurring in less than 10 documents)
  - no keywords in documents
  - varying idf weight (squared normally, but not here)
  - keep high frequency terms (normally dropped)
  - old version of software
- ok7: 1 run - no expansion, base run

The groups submitted the results (top 1000 documents retrieved for each query) to NIST for evaluation. There were a total of 774 runs: 18 system variants times 43 queries.

The runs were evaluated at NIST using trec_eval, concentrating on Mean Average Precision. The results of the initial evaluation were given to the six groups. This included

- Rankings of all documents (1.7 Gbytes in size)
- MAPs of all groups on all queries
- Various averages and standard deviations

These results are now publicly available at NIST on the TREC web site.

We can now compare systems on 2000 queries, making a qualitative difference in possible investigations. It has proven to be great tool for analyzing systems. Some of the differences among queries of a single topic pinpoint weaknesses in stemming, phrasing, hyphenation, and spelling correction. Other differences show that some systems are able to handle an entire topic better than other systems, while being worse on other topics. This comparison of differences due to syntax (queries) and semantics (topics) should prove very interesting.

The short-term goal of the Query Track has been to gather raw data for analysis. The long-term depends on you, the members of the community. You can both contribute more data, submitting runs of your system to the Query Station, and contribute your analysis.