The TREC 7 Query Track

Chris Buckley - SabIR Research, Inc.

Introduction

General IR research is being held up because we don't have enough queries of various types to investigate advanced retrieval techniques that are query dependent. There's no way we can get enough relevance judgements on new queries to form a good query pool. The Query track looks at multiple query variations of past TREC topics to get a large number of query formulations.

The track guideline states four goals:

1. Start investigating the split between query formation/analysis and back-end engines. Evaluating what makes a good general query formation approach.
2. Get many variations of the same topic so we can start analyzing (including with strong NLP approaches) queries, and determining what sorts of things we want to pull out of queries.
3. Get a collection of mixed fact/content queries. For decades we've had systems (eg Pnorm) that can handle these, but haven't been able to evaluate and compare due to lack of a query collection.
4. Get a collection of reasonable very short queries, more typical of real-life ad-hoc queries.

Query Track Task

Each group forms variations of each of the 50 topics in some subsets of the following categories (as defined in the guidelines):

1. Very short: (2-3 words) based on topic.
2. Sentence: NL (natural Language), based on topic and judgements
3. Manual Feedback: Manual NL sentence based on reading 5 or so relevant documents without reference to the topic (done by someone who doesn't have the topics memorized and who might use different vocabulary than the topic). An attempt to get a sentence which might use different vocabulary than the topic.
4. Manual structured query: based on topics and judgements. Perhaps mixed fact and content queries. Perhaps result of manual NL analysis.
5. Automatic structured query: based on topics and judgements (Note that "structure" could be just a list of words, or could be very complicated based on semantics.) Perhaps the result of automatic NL analysis.

Then all groups run everybody's queries for some subset of the categories above (whatever categories their system can be made to support). The names of the submitted runs consist of 7-8 letters/digits. The first 3 letters identify the group running the query. The last 4-5 letters are the queryset id, including category. Thus, "CorAPL5a" would be Cornell running the first Category 5 query set that was constructed by APL.

Query Track Runs

This was the first year for the query track. As it ended up, only two groups participated in the track. Thus it is impossible to come up with as many conclusions as we had wanted.

The two groups are Cornell/SabIR and the APL Labs at Johns Hopkins. Cornell constructed one set of queries in each of the 5 categories; pretty much directly using the definitions of the categories. APL constructed 4 query sets, skipping category 3 and 4, but having two versions of category 5. For the first two categories, APL deliberately tried to construct different queries than the obvious choice of words. This increased query variability, though at a cost of overall effectiveness as we will see later.

All 5 sets of queries were reasonably easy to construct. Cornell's category 4 queries do not have much detailed structure; they are basically a weighted sum of a vector query and a pnorm query. Cornell's category 5 queries are straight weighted relevance feedback vectors.

The queries were all constructed in DN2 format. DN2 is a quite complicated query language, but luckily very few features needed to be known for the queries the two groups constructed. We did not run directly on the DN2 queries but translated them back and forth from normal TREC queries. In the future, it is clear we should use standard TREC form as much as possible. The DN2 format scared several groups away who might have participated.

Query Examples

As an example of variability of the queries, here are all the different forms of topic 4, expressed in DN2 format.

```
<DN2 ID=4 QUERYSET=APL1a>
  "Foreign debt reorganization"
</DN2>

<DN2 ID=4 QUERYSET=Cor1>
"debt rescheduling agreements"
</DN2>

<DN2 ID=4 QUERYSET=APL2a>
  "What countries have received assistance in the form of a
reduction in the rate at which they must repay their loans?"
</DN2>

<DN2 ID=4 QUERYSET=Cor2>
"debt rescheduling agreements and loan restructuring
accords between debtor countries and the EC, Paris Club
and creditor banks"
</DN2>

<DN2 ID=4 QUERYSET=Cor3>
"What restructuring of debt repayment by third-world
countries have creditor nations accepted?"
</DN2>

<DN2 ID=004 QUERYSET=APL5a>
  <INDEPENDENT>
```

```
    <FULL_TERM WEIGHT=1.000000> "creditor"  </FULL_TERM>
    <FULL_TERM WEIGHT=0.839627> "debtor"    </FULL_TERM>
    <FULL_TERM WEIGHT=0.695880> "rescheduling"</FULL_TERM>
    <FULL_TERM WEIGHT=0.692837> "debt"      </FULL_TERM>
    ...
  </INDEPENDENT>
</DN2>

<DN2 ID=004 QUERYSET=APL5b>
  <INDEPENDENT>
    <FULL_TERM WEIGHT=1.000000> "creditor"  </FULL_TERM>
    <FULL_TERM WEIGHT=0.639725> "debt"      </FULL_TERM>
    <FULL_TERM WEIGHT=0.596154> "billion"   </FULL_TERM>
    <FULL_TERM WEIGHT=0.556568> "nobrega"   </FULL_TERM>
    <FULL_TERM WEIGHT=0.556568> "mailson"   </FULL_TERM>
    ...
  </INDEPENDENT>
</DN2>

<DN2 ID=4 QUERYSET=Cor5>
  <INDEPENDENT>
    <FULL_TERM weight=0.1142> "repayers"  </FULL_TERM>
    <FULL_TERM weight=0.1311> "brazil"    </FULL_TERM>
    <FULL_TERM weight=0.2155> "paris"     </FULL_TERM>
    <FULL_TERM weight=0.2056> "accordance"</FULL_TERM>
    ...
  </INDEPENDENT>
</DN2>

<DN2 ID=4 QUERYSET=Cor4>
  <INDEPENDENT>
    <INDEPENDENT weight=0.7>
      <FULL_TERM weight=0.6756> "rescheduled"  </FULL_TERM>
      <FULL_TERM weight=0.2056> "accordance"   </FULL_TERM>
      <FULL_TERM weight=0.1844> "pact"         </FULL_TERM>
      <FULL_TERM weight=0.1764> "agreement"    </FULL_TERM>
      <FULL_TERM weight=0.1592> "debt"         </FULL_TERM>
      <FULL_TERM weight=0.1359> "restructuring"</FULL_TERM>
        ...
    </INDEPENDENT>
    <AND weight=0.3>
      <OR> "debt" "interest" "loan" "repayment" </OR>
      <OR> "rescheduling" "restructuring" </OR>
      <OR> "agreement" "accord" "settlement" "pact" "talks"
           "propose" "negotiate" "request" "grant" </OR>
    </AND>
  </INDEPENDENT>
</DN2>
```

Query Track Results

Table 1 gives results on running the 9 query set variations (5 variations from Cornell and 4 from APL) on the test document collection (TREC Disk 1 plus the AP subcollection from Disk 3). The runs all strongly differ from each other in results. In general, the Cornell queries performed better for Cornell than the APL queries. Part of that is that goals of the APL queries were explicitly to use different, possibly non-optimal, vocabulary. But part of it could be that Cornell constructed queries to suit Cornell's system. In particular, the query set Cor5 was constructed using relevance feedback based on Cornell document weights. How well these weights suit other systems remains to be seen. We didn't have enough participating systems to be able to conclude anything.

	APL		Cornell/SabIR	
Query Set	P(20)	Ave Prec	P(20)	Ave Prec
APL1a	.1460	.0559	.2350	.1051
APL2a	.1230	.0477	.2710	.1142
APL5a	.3010	.1627	.4010	.1971
APL5b	.5480	.2577	.6450	.3219
Cor1	.2730	.1055	.5030	.2457
Cor2	.4290	.1846	.6040	.3367
Cor3	.2330	.0917	.4560	.2020
Cor4	—	—	.6500	.3282
Cor5	.4540	.2296	.7760	.4586

Table 1: Results of Cornell and APL on Different Query Sets

As normal, even with the very strong overall differences in results between query sets, large numbers of individual queries of the weaker query set do better than the corresponding query in the stronger set. Table 2 gives the number of queries (out of 50) for which one query set beats another, keeping the system constant (Cornell's system was used). For instance, APL5b beat Cor2 on 28 out of 50 queries, despite having weaker overall evaluation averages.

>	Cor1	Cor2	Cor3	Cor4	Cor5	APL1a	APL2a	APL5a	APL5b
Cor1	0	7	32	11	2	43	39	30	18
Cor2	43	0	46	23	4	48	47	43	22
Cor3	18	4	0	5	1	38	36	26	12
Cor4	39	27	45	0	8	48	47	41	23
Cor5	48	46	49	42	0	50	49	48	46
APL1a	6	2	11	2	0	0	27	9	2
APL2a	11	3	14	3	1	22	0	16	3
APL5a	20	7	24	9	2	40	32	0	15
APL5b	32	28	38	27	4	48	47	35	0

Table 2: Comparative Query (row better than column for X queries)

There is a tremendous amount of query variability hidden in the comparative averages. We need to understand this variability. It is not clear that 9 query variations is enough to get a handle on variability; but at least it is a start.

Query Track Conclusions

It is impossible to conclude much from this initial track attempt since there were only two participants. We can verify what we already knew about queries:

- Different formulations of the same query can behave tremendously differently. In general, the more information included in the query, the better the results.
- Different queries behave very differently. There are significant numbers of queries where more information hurts.

We simply do not have enough information to look at how different systems interact with the various forms of the queries. Many interesting questions remain to be tackled in next year's track!