TREC 2010 Relevance Feedback Track

[[Changes from Preliminary Version]]

This is the final version of the 2010 Relevance Feedback Guidelines. The only important change from the Preliminary Version of June 4 is that an additional optional task has been added, allowing more categories of relevant documents to be explored (explicitly including spam relevant documents.)

Introduction

This is the third year of the TREC relevance feedback track. The first year concentrated on the RF algorithm itself. All participants were given the same sets of judged docs, and used their own algorithms to retrieve a new set of docs. In the second year, the concentration shifted to finding good sets of docs to base their retrieval on. Each participant submitted one or two sets of 5 docs for each topic, and 3-5 other participants ran with those docs, thus getting a non-system dependent score on how good those docs were.

One disappointment of the first two year results, is that there was little analysis of what made a relevant document relevant. Unlike a query, there is a lot of well-formed text indicating relevance in a typical relevant document, but groups in general did not attempt to get a deeper understanding of the query from a syntactic or semantic analysis of the documents. The format of the first two years (possibly many relevant documents) made it much easier to simply apply statistical models of relevance.

Task

For 2010, the track aims to examine what makes an individual document good or bad for feedback. The focus on single document relevance feedback should encourage groups not to just look at a document as an unordered bag of words, but to examine the document and language structures to focus on the essence of relevance. The limited amount of relevance information known should also make failure analysis much easier: “This other group, with the same basic performance as my group, was able to get much better results using this document than my group. Why?”

The basic user scenario envisioned for this track is that a user has submitted a short (title only in TREC parlance) query to a retrieval system that has returned one or more documents to the user that it thinks are relevant/useful. The user looks at the first document and decides that it is indeed relevant/useful. Based upon this document and the original query, the system reconstructs the list of documents to be given the user. This new list is then evaluated for both accuracy (top documents are relevant) and completeness (all relevant documents are retrieved, not just those that have the same aspect as the known document). Like almost all relevance feedback tasks, we assume the user has a need for several or many relevant documents – otherwise there's no need for relevance feedback once the user has one relevant document!

This scenario will be investigated within the track as follows: There will be a set of 100 topics for which we already have some known relevant and non-relevant documents. For each topic, each participating group will receive 5 documents. For each of these documents, the group is to make a relevance feedback retrieval run using the original short query plus that document, which should be assumed to be relevant. The group will submit the returned documents of this run to NIST. The document list will be judged and evaluated, and the groups will report the results at the 2010 TREC workshop.

As well as the single document run submissions, the groups will be asked to supply a baseline run with no relevance information. In addition, groups can optionally supply their opinion, in the form of a relative ranking, of how well the 5 documents will perform. Can systems pick out those documents that they feel will provide good relevance information?

Please note that this, like most relevance feedback experiments, is emphatically not an attempt to evaluate the entire relevance feedback process for a system (we hope to do that next year). For instance, the systems are not necessarily operating upon the original document that that system would have given the user, and there is only one round of feedback, where in a real environment there would be many. Instead, the experiment is evaluating and comparing approaches for one vital subtask that most real systems would be doing: given an original query and a good document, extract as much relevance related information as possible out of the document, while still maintaining a focus on all aspects of the original query.

Data

Corpus

The track will use again the ClueWeb09 (http://boston.lti.cs.cmu.edu/Data/clueweb09/) corpus. We will use only the English portion of ClueWeb09 (English[1-10]). In 2009, a distinction was made between the Category B subset (CatB is English_1) and the full English portion. Participants may use the standard subset of full English[1-10], CatB, but we strongly encourage use of the full dataset. On submission of runs, we will ask groups whether they used all of English, or CatB. Documents for feedback will be drawn from English_[1-10] – we will make available the text of all those documents for those groups who only have access to the CatB documents.

There are resources available for the ClueWeb09 collection. For instance Gord Cormack and his group has done a nice job at spam detection, with results available at http://durum0.uwaterloo.ca/clueweb09spam . The ClueWeb09 Wiki has other resources.

Topics

Topics will be selected from those run in the TREC 2009 Million Query track (which includes those queries run in both web and relevance feedback tracks). Groups may train on odd numbered topics from the MQ track – all selected topics will be even-numbered topics. Thus groups should NOT do any training on any full set of 2009 track topics, since those topics will contain both odd and even numbered topics. Just train on the odd topics!

Documents for Feedback

There will be 5 documents chosen for each topic, with all groups doing a run for each document. One of the 5 documents will be randomly chosen from among that topic's known relevant documents. This will be the standard relevant document. The other 4 documents will be chosen according to some property of the document. The criteria for choosing documents might include

- How commonly retrieved the document was in last year's runs
- The level of relevance (highly relevant, relevant, or even possibly commonly retrieved but non-relevant)
- Document length

For instance, the documents for each topic might include the most commonly highly retrieved relevant document for that topic in last year's tracks. The documents will not be marked with the criteria used to choose them, and the documents chosen by any one criteria will be randomly mixed among the sets of documents given the participants.

The goal here is to start an investigation into some of the properties of documents and how they affect relevance feedback performance, and determine whether those affects are general or system-dependent.

Run Input and Submissions

The participants will be given 5 sets of documents per topic in standard TREC results text format with the only important fields being the topic id and the ClueWeb09 document id. The document sets will be called "10-1" through "10-5". The participants will then submit a retrieval run for each of these document sets. Each group should also submit a base case run, B, with no relevance information used. Runs will be submitted to the NIST submission system

All runs should be submitted in standard TREC results format. The run name should consist of a common basename identifying the group, followed by a period and the document set used for that run. The basename may not exceed 8 characters. Thus Sabir Research might submit the following runs: Sab10rf.B, Sab10rf.10-1, Sab10rf.10-2, Sab10rf.10-3, Sab10rf.10-4, Sab10rf.10-5

Participants should submit a ranked list of 2500 documents per topic. Only the top 1000 will be used for evaluation in this track; the submission of the extra documents allows for flexibility in meta-evaluation experiments, and allows for future residual collection evaluation where researchers may remove documents from your run for fair comparisons. The result should not include the relevant document used as input for feedback.

Optional Sub-task Submission

As an optional task, groups can submit a ranking giving the group's prediction of the relative effectiveness of each input document for relevance feedback. This will be in the form of a standard TREC results format, 5 documents per topic. The score can be on any scale the participants desire – it does not have to be based on trying to predict MAP or Prec@5 scores. The predictive scores will be used only to rank the input documents for a particular topic.

Submission of Optional Second Set of Runs

Groups may submit two complete sets of runs instead of just one if they wish to. Each of the 5 runs in the second set should be named as the first set were, except a suffix of '.2' should be added (e.g., 'Sab10rf.10-3.2'). This allows a rough within system comparisons. Note that the second set of runs will not contribute to the judgment pool. Given last year's results, there are some worries that a bias may exist in favor of groups submitting more runs than other groups. That will be explored this year (and the optional second set of runs will help the organizers explore this), but the official evaluation of the second runs will be based on judgments determined by the first set of runs only.

Optional Second Set of Factors

If they have completed and submitted the required sets of runs (6 runs, base plus 10-1 through 10-5), groups may optionally submit an additional set of runs using documents selected using a different set of factors. The input documents for this second set of factors are included in the general topic and input file distribution and are labeled 10-6 through 10-10. Groups doing only the required task will not be using these files at all. The same 100 topics will be used.

Groups should use exactly the same procedure and system on this second set of factors as they did on the required task. There should be one run for each of the new 5 input files, with runs again being labeled with the suffix of the input set (e.g. Sab10rf.10-10). The results from input documents of both the first set and second set of factors will be combined – it is important that groups do not change their systems when doing the two sets.

Assessments

Amazon's Mechanical Turk will be used for judging the relevance of documents. That has implications for the timing of run submissions, the timing of results being returned to the participants, and the evaluation procedures. We expect to have several unofficial evaluation measures as we experiment with the Mechanical Turk judgments, but the official measures will be given as below. We expect there will be at least two partial evaluation releases to enable the participants to get a start on failure analysis, and their own evaluations, but the final numbers will not be released until quite close to the TREC conference.

Evaluation

The primary official evaluation measure will be a recall oriented measure like MAP, statMAP (an approximation to MAP that handles missing judgments better).MAPjudged (MAP, but eliminating all unjudged documents from the top 1000 before the evaluation) or Prec@1000. All recall-oriented measures have problems on CLueWeb09 that are still being investigated. There will be a secondary official measure of Precision at 10 docs. These measures will be calculated using binary relevance judgments – possibly conflicting Mechanical Turk judgments will be coerced into binary judgments.

There will also be unofficial measures reported which will explicitly model the probability of relevance of a document given the Mechanical Turk judging process. However, since these measures have not yet been fully developed or tested, and there are as yet no test collections that participants can train their systems with using these measures, they will not be official measures, but only used as guides for future evaluations.

The official measures were chosen to reflect a task where the user is interested not just a single relevant document (since they already have one!), but all varieties of relevant documents for this topic. Thus a recall oriented measure like MAP is desired. In addition, we would like some measure that gives an indication of just performance at the top of the rankings. Ideally, we would like that measure to have a diversity component as in the Web Track 09, but given the Mechanical Turk environment, we don't want to commit to being able to do that.

A final single score for each participating group will be the average over all 5 runs over all the topics, for each of the measures. In addition, the input documents will be broken down into sets based on the criteria used to select them, one per topic. Thus there might be a set for "long documents" where each topic has one long relevant document in it. There will be a set for the "standard" randomly selected relevant documents, one per topic. Thus, scores will be reported for each of the 5 criteria sets.

The predictive relative effectiveness task for the track will be evaluated for each topic, comparing the predicted ordering of the 5 input documents against the actually recall system performance ordering the 5 documents. Average Kendall tau over all topics will be reported.

The second optional task will be evaluated the same as the required task. There will be results reported for the 5 additional categories of the second task.

Research Goals

The overall focus for this year's track is to study the notion of what is a good document to use for relevance feedback, how can a system recognize a good document, and what is the expected gain from using a good document as opposed to poorer document for relevance feedback. In addition, different types of relevant documents will be used, and the performance of systems using those types will be studied. The design above allows conclusions to be reached about

1. What is the level of performance using different relevant documents (for each topic, max score – min score among the input docs), both for a system and between systems. How important is the choice of a good relevant document?
2. Do systems agree about the notion of goodness of a relevant document (Is the ordering of system performance of the input docs generally the same for all systems for a topic?)
3. Can systems predict the ordering of system performance of documents? (Is the ordering of actual system performance for a topic's input docs generally the same as the ordering induced by the predictive score of an input doc?)
4. Are there characteristics of good relevant documents that make them easier to use (Among the input doc sets determined by criteria, are some doc sets better than others. and is that consistent among all systems?).
5. If these characteristics have an effect, is that effect system independent?

Important Dates

June 28: Topics and documents for RF released.

August 12: Run Submissions due (tentative date)

September: Relevance judgments and scores to groups.

November 16-19: TREC 2010

Coordinators

Chris Buckley (chrisb@sabir.com)

Mark Smucker

Matt Lease