TREC 2012 Crowdsourcing: Text Relevance Assessing Task Guidelines

Author: Mark D. Smucker (mark.smucker@uwaterloo.ca)
Crowdsourcing Track Organizers: Gabriella Kazai (Microsoft Research), Matthew Lease (University of Texas at Austin), Panagiotis G. Ipeirotis (New York University), Mark D. Smucker (University of Waterloo)
Date: June 5, 2012
Version: 1.0

Introduction
The text relevance assessing task (TRAT) is one of the two TREC 2012 Crowdsourcing Track tasks. The goal of the task is to evaluate approaches to text relevance assessing. We assume that many participants will utilize crowdsourcing to do the relevance assessing, but the task is open to all approaches that follow the task's guidelines.

Participants in the task will simulate playing the relevance assessing role of NIST in the TREC 8 ad-hoc track with a subset of the TREC 8 topics. The TREC 8 Overview by Voorhees and Harman provides details about the TREC 8 ad hoc task: http://trec.nist.gov/pubs/trec8/papers/overview_8.pdf

Participating
To participate, groups must have registered to participate in TREC 2012:

Submissions will not be accepted from groups that have not registered to participate in TREC 2012.

Data
The data available for the track consists of the following:

List of topic-docno pairs to judge. We have randomly selected the following 10 topics for use in the TRAT: 411, 416, 417, 420, 427, 432, 438, 445, 446, 447. The topic-docno pairs to judge are provided in the file http://www.mansci.uwaterloo.ca/~msmucker/trec2012TRAT/trec2012-crowdsourcing-text-task.topic-docnos.txt. There are 18260 topic-docno pairs to be judged.

Topic descriptions. The NIST provided topics for TREC 8 ad-hoc: http://trec.nist.gov/data/topics_eng/topics.401-450.gz. The topic's title, description, and narrative when taken together define what the original NIST assessor considered to be relevant and non-relevant to the topic.

Documents. TREC 8 ad-hoc used the Text Research Collection Volumes 4 (May 1996) and 5 (April 1997) minus the Congressional Record (CR). These are two of the document disks that TREC makes available free to registered TREC participants.

The subcollections of vol 4 and 5 minus the Congressional Record are:

2. Federal Register, 1994 (FR94)
3. Foreign Broadcast Information Service (FBIS)
4. Los Angeles Times (LA).

It is important to note that redistribution of the documents is prohibited. The documents are copyrighted and your use of these documents is limited by your signed agreement with NIST. For example, you are forbidden from putting all of the documents up on a website and allowing the world to view them. NIST has approved the showing of a small number of documents to workers for the purpose of your experiments.

To obtain the TREC disks, please refer to the instructions in the "Welcome to TREC 2012" email from NIST.

Qrels before TREC 8. Participants may use any other TREC collection and its associated relevance judgments (qrels) for training of their system so long as that collection and the qrels predates TREC 8. Qrels for TREC 8 and the later tracks that used TREC 8 topics (#'s 401-450) are explicitly forbidden for use. Participants may have previously used TREC 8 topics in crowdsourcing or relevance assessing research. Any prior use of the TREC 8 topics and qrels should be disclosed by participants in their write-ups. Note that the TREC 7 ad-hoc topics and qrels are for the same document collection as TREC 8, and thus form a reasonable set of training data.

Past TREC 8 Ad Hoc Results. Since the idea behind the TRAT is to simulate being NIST for the TREC 8 ad-hoc task, a few groups may find it helpful to have knowledge of the submitted runs that NIST used to create the judging pools. To obtain these runs requires a separate application and permission from NIST beyond that required to participate in TREC 2012. See: http://trec.nist.gov/results.html

Relevance assessing guidelines. Participants may find it useful to look at the guidelines provided to NIST assessors. Copies are provided of the TREC 7 guidelines:
and the portion on relevance for the TREC 8 guidelines:
http://www.mansci.uwaterloo.ca/~msmucker/trec2012TRAT/trec-8-relevance-judging-guidelines.txt
Participants should note that NIST provides training, supervision, and their own relevance assessing platform to their paid assessors. In addition, the NIST assessors have been personally responsible for the development of their search topics, and while the topic narrative aims to capture a description of what is relevant, it is always possible that the NIST assessors made relevance judgments based on notions of relevance known only to them.

Task
As mentioned in the introduction, participants will simulate the relevance assessing role of NIST for 10 of the TREC 8 ad-hoc topics. A key difference is that NIST assessors created the search topics and then determined the relevance of documents.

For each of the 10 topics, participants will need to provide both a binary relevance decision and a probability of relevance. If a probability of relevance is not possible for a run, that run will only be evaluated based on its binary judgments.
While this is a crowdsourcing track task, there are no limits on the methods that can be employed to obtain a set of relevance judgments for the topic-docno pairs. The only restriction that must be strictly adhered to is that the NIST judgments (qrels) for TREC 8 and later tracks that used the TREC 8 topics (401-450) are forbidden for use in any way shape or manner. For example, participants may not even use the qrels to help them understand the topic or train workers. Participants may not even make counts of the number of relevant documents etc. If participants want to use "gold" in their crowdsourcing, the participants must create this gold themselves. **The TREC 8 qrels and later qrels that used TREC 8 topics (#'s 401-450) may not be used for any purpose.**

Participants should do their best to keep track of their time and cost to produce their submissions so that these variables can be reported in their reports. Other researchers will want to understand the time and cost needed to achieve difference levels of performance.

**Evaluation**

Participants will be asked to designate one of their runs for use in a majority vote consensus process. The majority vote of the submitted runs will be compared to the NIST relevance judgments. When there is disagreement between the majority vote and NIST, the organizers will adjudicate the judgment and decide on the final relevance judgment for these documents. All such decisions will be recorded and made available to participants. Thus, a benefit to participating in the TRAT is access to a new set of relevance judgments over an existing TREC collection.

Evaluation of TRAT will be similar to that of the former Spam Track. For each topic, the performance of a submitted run will be judged on both its binary judgments and its probabilities of relevance. The binary judgments will be evaluated using the logistic average misclassification rate:

$$\text{LAM} = \logit^{-1}\left(\frac{\logit(fpr) + \logit(fnr)}{2}\right)$$

where \(fpr\) is the false positive rate, \(fnr\) is the false negative rate, and

$$\logit(p) = \log\frac{p}{1-p}$$

and

$$\logit^{-1} = \frac{e^x}{1+e^x}$$

We will smooth the calculation of \(fpr\) and \(fnr\):

$$fpr = \frac{|FP| + 0.5}{|FP| + |TN| + 1}$$

$$fnr = \frac{|FN| + 0.5}{|FN| + |TP| + 1}$$
where |FP| is the number of false positives, etc, as given in the table below:

<table>
<thead>
<tr>
<th>Run Submission</th>
<th>Relevant (Positive)</th>
<th>Non-Relevant (Negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant</td>
<td>TP = True Positive</td>
<td>FP = False Positive</td>
</tr>
<tr>
<td>Non-Relevant</td>
<td>FN = False Negative</td>
<td>TN = True Negative</td>
</tr>
</tbody>
</table>

We will measure the performance of the runs given their probabilities of relevance with Area under the ROC curve (AUC).

**Submission Format**

Participating groups will be allowed to submitted as many runs as they like, but must ask permission before submitting more than 10 runs.

One of the submitted runs must be designated as the run to be used for group majority voting.

Each run will be submitted as an ASCII text file with one judgment per line. There will only be one judgment per topic-docno pair, i.e. the topic-docno pairs must be unique. Binary judgments are required of all submissions. Each line will consist of the following whitespace separated columns:

1. topic
2. docno
3. binary judgment, 0 for non-relevant and 1 for relevant.
4. probability of relevance (a value in the inclusive range [0,1] with 1 being relevant). If the run’s method does not produce probabilities of relevance, this column should have a value of NA for every topic-docno.
5. runtag that uniquely identifies your group and the method used to produce the run. Run tags should be unique across years (don’t use a tag you have used in previous years). Tags must be 12 of fewer characters and made up of letters and numbers only.

For example:

411 FBIS3-10248 0 0.3 UWatMDSvg
411 FBIS3-10319 1 0.952 UWatMDSvg
411 FBIS3-10334 1 1.0 UWatMDSvg
etc.

**Timeline**

- June 5, 2012: Release of these guidelines for discussion.
- August 31, 2012: Submissions due.
- October 1, 2012: Results available.
- November 6-9, 2012: TREC Conference.