HARD Track Overview in TREC 2003
High Accuracy Retrieval from Documents

James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst

1 Introduction

The effectiveness of ad-hoc retrieval systems appears to have reached a plateau. After several years of 10% gains every year in TREC, improvements dwindled or even stopped. This lack of progress was undoubtedly one of the reasons behind abandoning suspending the ad-hoc TREC after TREC-9.

One plausible reason that document retrieval has been unable to improve is that the nature of the task requires that systems adopt "one size fits all" approaches. Given a query, a system will generally do best to return results that are good for an "average" user. Doing otherwise (i.e., targeting the results for a particular type of user) might result in substantial improvements on a query, but it is just as likely (in a TREC environment) to cause horrible degradation. By ignoring the user (or, more accurately, by treating all users identically), systems cannot possibly advance beyond a particular level of accuracy on average for a specific user.

The goal of this track is to bring the user out of hiding, making him or her an integral part of both the search process and the evaluation. Systems do not have just a query to chew on, but also have as much information as possible about the person making the request, ranging from biographical data, through information seeking context, to expected type of result.

The HARD track is a variant of the ad-hoc retrieval task from the past. It was a "pilot" track in 2003 because of the substantial extension on past evaluation—i.e., it is not clear how best to evaluate some of the aspects of the track, so at least for this year it was intended to be very open ended. HARD is also running as a track of TREC 2004.

The HARD 2003 track ran in three phases: baseline, clarifying, and final. In the first phase, sites received and ran topics that were essentially idential to a classic TREC topic: title, description, and narrative fields.

In the second phase, sites were able to acquire clarifying information about the topics. They had two means and could use either or both of them:

1. Biographical, contextual, preferred result format, and any information that disambiguates the query was captured when the topics were generated. This metadata about the query was made available for phase two.

2. Sites were permitted to generate a single "clarifying form" that the searcher would answer. For each topic, this form was a Web page that solicited useful information about the query or the searcher (e.g., disambiguating words in the query or finding out more information about what the searcher wants). The assumption was that the "clarification" would be generated automatically, but sites could have opted 0to generate manual clarification questions (can a librarian beat the best IR systems?). None did.

In the final phase of the track, sites used all user- and query-information that they acquired to construct a better and more accurate ranked list. That substantially improved (because it is more targeted) list was submitted to NIST for evaluation.

Because accurate retrieval could also just be pinpointed retrieval, an extension of the HARD track evaluated passage retrieval, a system's ability to select passages within documents that are relevant. Passage retrieval was an option available to sites, but could be ignored by returning full documents.

2 Participation

The following 14 sites participated in the HARD track. A summary of each group's activity is provided below. The summaries were written by the site (except for those that are in italics) and are listed in alphabetical order.

Clairvoyance [Shanahan et al., 2004]

The Clairvoyance team participated in the HARD Track, submitting fifteen runs. Our experiments focused primarily on exploiting user feedback through clarification forms for query expansion. We made limited use of the genre metadata. Within the clarification form feedback framework we explored the following hypothesis: could we organize the top retrieved documents for a query into intuitive groups (through clustering) that the user then selects as representative/relevant for the topic. Within this we explored two types of clarification forms: one was based upon representing each group using a list of terms, corresponding to typical terms for the group, and a list of documents, where each document was represented using its title and the information source (very similar to the forms used in Scatter-Gather [Hearst et al., 1995]); the second form represented each group using just a list of terms (forty), corresponding to typical terms for the group. The user was asked to judge the relevance of each group as being "On Topic", "Not on Topic", "Unsure", or "Unjudged". The group judgments were then used to expand (as feedback) the query. We explored various schemes for expansion. Overall, our results for the experiments suffered due to a bad baseline run that was used in the generation of both clarification forms. The mean average precision (MAP) for the top 1000 documents was 0.23 (about median for this track), which has since been improved to 0.31. Having said that, when we did incorporate feedback from the title-based form, the MAP was improved to 0.29 (from 0.23). The second term-based form did not yield any significant improvement. Follow-up experiments on our new baseline, where we choose a single group that yields the best performance for a topic (assume we have an oracle), has yielded an overall MAP of 0.37. We are currently regenerating the title-based forms using our new baseline run and will have a human re-evaluate them. We are also exploring how to automatically select these groups for feedback.

IIT Bombay [Ramakrishnan et al., 2004]

The description of the IIT Bombay system is minimal. They used the open source retrieval system Lucene to index the document collection and select passages for retrieval. The focus of their work in TREC 2003 was text summarization and they applied summarization techniques for several of the tracks.

Illinois Urbana-Champaign [Shen and Zhai, 2004]

For the clarification forms, we focused on studying the problem of "active feedback": Given that a user is willing to make relevance judgments on k documents, how do we choose k documents to present to the user so that we can learn *most* from the user's feedback on them?

The simplest baseline approach is to present the top k documents. However, this may not be the best strategy; for one thing, some of the top k documents may be redundant. Thus we proposed and tested three other methods: (1) "Gap-based methods": We sample k documents from the top ranked documents so that these k documents would form some "gaps" between them. E.g., we can pick k documents with ranks 1, 1+g, 1+2*g, 1+3*g, ..., 1+(k-1)*g. In this way, the gap between two adjacent documents in the ranked list is "g-1". If we assume that documents that are close in the ranked list are likely similar to each other, then this method would help reduce redundancy. (2) The Marximal marginal relevance (MMR) method: Here we select the k documents with a greedy algorithm. At each step, we try to pick a document that is both relevant and novel. (3) The clustering centroid method: We cluster the top N documents into k clusters, and pick the centroid document from each cluster.

To reduce the labor on the user side for judging documents, we presented the best (fixed window) passage, for each document in the clarification form, rather than the whole document. Our form contains essentially just 6 passages.

After obtaining the judgments, we explore two ways of using the judgments: (1) Using them as "passage judgments" and perform passage-based feedback (i.e., query expansion); (2) Using them to infer relevance status of the corresponding document – actually we simply take the judgment on a passage as if it were a judgment on the document that contains the passage.

Our basic retrieval approach is the KL-divergence retrieval formula with mixture model for feedback.

The results so far suggest that (1) All feedback methods perform better than the baseline no-feedback method. This isn't surprising at all, as it just shows that relevance feedback is effective. It does show that language model based feedback is effective. (2) Passage-based feedback performs *substantially* better than document-based feedback, which is also consistent with what others have seen (e.g., the local context analysis method?). But again, we did it with language models. (3) The gap-based method performs slightly better than the clustering method in terms of average precision, but the clustering method performs slightly better by pr@10 docs and R-precision. We are waiting for results of the MMR approach and of the baseline "top k" document approach. The comparison with the "top k" method would be most interesting, as it would indicate how effective our "active feedback methods" is as compared with the more standard way of presenting top k documents. (4) Compared with the group, our performances seem to be usually above medians.

Microsoft Research Cambridge [Robertson et al., 2004]

For the HARD task, we concentrated on the use of clarification forms (system-generated forms offered to the assessor who originated the topic for a one-pass user interaction). The primary intention was to obtain some data that could be used in a relevance feedback algorithm, The limitations of the clarification form (both screen space and assessor time) prevent the presentation of entire documents or even substantial passages, but it is possible to offer the user small, query-specific snippets. We used a form of passage retrieval, where the passages are pre-defined exclusive units at a little above the sentence level – each passage consisting of one or a few sentences, with no overlap between passages. The main focus was on an "active learning" approach to selecting the snippets to show the user: we wanted to choose items that would give us most information for relevance feedback purposes. Out of the top 30 documents, we attempted to choose a set of five which would maximise the expected change to the query after judgement (which might be positive or negative) by the user. The particular functions chosen to measure this effect will be described. The result of this selection process generally differed from the top 5 documents (our baseline run). The user was not asked specifically for a relevance judgement on each snippet, but rather for something that might correspond to click-through data. We also presented the user with some phrases selected from the top snippets, and invited them to select positive or negative phrases. Various possible ways to use this data will be discussed.

We incorporated minimal use of metadata. Since we are doing a form of relevance feedback with the clarification forms, we include the relt texts along with the rest of our relevant fragments (although the difference in length might be problematic). We did the obvious thing with the US Govt stuff.

For use of passages, see the description above. This is a step backwards from the kind of overlapping, any-size passages we used to do with Okapi, which we haven't yet reproduced in the new environment. We did some rather obvious matching of metadata granularity onto these passages.

Chinese Academy of Sciences [Wu et al., 2004]

They used natural language processing to restrict queries to just nouns and verbs and to classify them into positive and negative classes that could be treated differently. Their primarily emphasis was on how to train the relative weights of the positive and negative words in the query.

Queens College, CUNY [Grunfeld et al., 2004]

Basic retrieval was done using our PIRCS system with pseudo relevance feedback (expand using 20 docs and 60 terms). Three runs were submitted: pircHDBt1 and pircHDBt2 which are title runs but with slightly different parameters. pircHDBtd1 is another run using title and description.

Three clarification forms were submitted QCSU:1-3. The QCSU:2-3 forms were designed to display data produced by the BASIC retrieval for evaluation – essentially 20 feedback *terms* (out of 60) that have lowest document frequencies. In addition we displayed top 10 (out of 20) PRF documents with their title/first *sentence* for the user to judge. We also make available a window for the user to add whatever they want as *key* terms to improve the topic description. We believe the user can complete these in 3 minutes. The first form QCSU1 has synonym terms from WordNet based on the topic title only. Phrases were sent to WordNet first. If synonyms were found, their constituent single words were removed; else single words were used to find synonyms. This does not involve a BASIC retrieval, and is more efficient. We like to compare results to see if a BASIC retrieval is necessary.

For enhanced retrieval without metadata, *terms* and *keys* from the clarification form were added to enhance the raw query. A full retrieval was repeated including PRF. During PRF we make sure that the docs corresponding to the *sentences* marked relevant are included for feedback for QCSU:2-3. For QCSU1, no *sentence* information is available, just topic enhancement. These submissions are respectively; pircHDC1t1, pircHDC2t1 and pircHDC3td1. For QCSU:2-3, some clarification results lead to fewer than 3 relevant *sentences* (less than 3 relevant docs among the 10 displayed). This may signal that the BASIC retrieval is bad and the topic is hard. Even with enhanced *keys*, retrieval may only be mediocre. We regard this as suggestion that one should disallow PRF for these queries - 16 queries in QCSU2 and 20 in QCSU3. These submissions are pircHDC2t2, pircHDC3t2.

For enhanced retrieval using granularity metadata, starting with the results above, we further processed the 1000 retrieved docs of each topic by our QA system that returns 250 bytes as answer to a question. 250 bytes is like 40 words, close to a passage size; no effort was made to return sentences. Since such QA clue words as WHO, WHAT, HOW LONG, WHEN, etc are not available, the QA system Essentially defaults to finding text spans that contain most topic words and of higher weights. We believe answers could be words/phrases interspersed among these topic words in such a window. These runs are pircHDC1tp, pircHDC2tp and pircHDC3tdp.

Rutgers University [Belkin et al., 2004]

We were particularly concerned with such knowledge which could be gained through implicit sources of evidence, rather than explicit questioning of the information seeker. We therefore did not submit any clarification form, preferring to rely on the categories of supplied metadata concerning the user which we believed could, at least in principle, be inferred from user behavior, either in past or the current information seeking episode. We did not attempt to retrieve only passages. Below, we describe how we used the supplied metadata.

FAMILIARITY This we addressed by promoting the value of documents which score toward the unreadable end of readability scales for people highly familiar with the topic, and by promoting the value of documents which scored toward the easily readable end of the scales for people unfamiliar with the topic.

GENRE This we addressed in two ways. One was by constructing language models for all the retrieved documents for each training topic and for just the completely relevant documents for each topic. We then identified words which occurred with greater than expected probability, based on the entire topic language model, in the relevant documents, for all topics which had the same genre. These words were considered to be indicators of the genre. We added the words associated with a particular genre to queries for topics which requested that genre. The second way was to promote documents from certain sources to the top of the retrieved list for topics with some genres, by removing documents from some sources entirely from the retrieved list for topics with some genres, and by demoting the value of documents from some sources in the retrieved list for topics with some genres.

RELEVANT TEXTS We used relevant texts as the basis for automatic query expansion.

GRANULARITY If the desired granularity of the retrieval result was passage, we ranked documents on the basis of their best passage, rather than on the document as a whole.

Tsinghua University, CS IR Group [Ma et al., 2004]

Main idea: Though the HARD is new experimental track, our research work mainly focus on delivering a practical solution for applied search environment. Therefore, all the submitted results(including CF and runs) are constructed in a automatic way, for we think it is more feasible than manual mode.

1. Baseline run. We get the baseline run (only with document) using the initial query by a TF*IDF scoring schema (BM 25). For each topic, the initial query is constructed simply by the task description(The detail restriction for none-relevant document are ignored). For the search items, different weights are set according to their position and importance in the task description. No positive training documents are used to refine the query, because usually the training resource is unlikely to be provided for various immediate search requirements in Web IR.

2. Clarification form. In the form, all the potential search issues to be confirmed by user are listed with checkbox, together with a text field to fill if he/she find there are something we missed. The search issues are presented as keywords or phrases, which are automatically extracted by two methods: (1) the kernel words/phrase in topic description. (2) terms with high statistical weight in top-100 ranked documents in search result. To keep the search deviation under control, we limit the search items up to 10 issues. It is an efficient method for delivering clarification form to the user, while the accurate of the question seem not satisfied.

3. Final Run.

1. *refine the query term.* The resource to refine the query terms is from: (1) attached text field in the return CF form; (2) the searchitem field in metadata. The new terms are added by Rocchio-like style.

2. *focus probe and reconstruct the query.* From the CF result returned from LDC, we do the work in two ways: In first method, all the items in selected checkbox in return CF are thought as one search focus. Based on the kernel terms in initial query and the current search item, a sub-query is constructed for a specific search focus. Then the initial query is divided into several queries for different search focus. And the final result of the topic is the combine of the results from all the sub-query. In second method, all the search terms of the search focus are simply taken as new weighted terms to be added into the initial query. Then using the new refined query, we get the final run.

3. *return type detection.* There are three different types available. We return document if topic require so. For passage and sentence, we usually return the single paragraph(For sentence, it is nearly impossible to present an efficient result in such rough retrieval). If any type is welcomed, we analyze the topic description and decide the result should be passage or document.

UMass Amherst [AbdulJaleel et al., 2004]

The CIIR at UMass Amherst participated in all three aspects of the HARD task. First, we mapped query metadata values to document metadata values that we assigned. We then adjusted the ranking of documents depending on whether their metadata matched the query metadata.

We also generated clarification forms to tease more information out of the searcher. We tried several types of clarification forms, including providing a list of keywords that might appear in relevant documents, a list of top-ranking clusters that might contain relevant documents, and a list of passages that might appear in relevant documents.

Finally, we explored passage-level retrieval of documents to see if we could pinpoint the relevant portions of documents.

In the final analysis, all runs using metadata or clarification forms failed to outperform our best baseline run (which included query expansion). Further exploration is needed to understand why the adjustments did not help more. Passage retrieval provided a gain for a subset of queries, but there were just as any that did not improve or dropped in effectiveness.

University of Buffalo, CEDAR [Srikanth et al., 2004]

Metadata: We used the purpose, genre and granularity metadata items in our solution to the HARD problem. Documents were processed by InfoXtract - an information extraction engine from Cymfony Inc. InfoXtract parses the documents to tag named entities, semantic structures and discovers relations between entities. HARD queries are parsed by InfoXtract to identify question type. The occurrence of these features and query keywords in documents and text snippets (in the case of passage, sentence or phrase granularity) are used to model some of the metadata values of the query and rank the answers.

Passages: Document snippets are selected for all granularity values except 'Document'. For granularity of sentence/phrase, each sentence with at least one query keyword is shortlisted as candidate answer snippets. For passage granularity, contiguous sentences are selected based on keyword hits based on minimum (3 sentences) and maximum (6 sentences) window sizes.

Documents processed by InfoXtract are indexed (words and extracted features are used as index terms) by TAPIR toolkit. Document retrieval for HARD queries is based on Concept Language Models. Expected answers for a user query is modeled as a sequence of keyword and non-keyword features (e.g. passage position in document, answer-type match, occurrence and count of location/time/person/organization features). Document and/or text snippets are ranked based on the probability of their model *generating* the query features. adhoc weights were assigned to query features.

University of Helsinki

No information provided.

University of Maryland [He and Demner-Fushman, 2004]

The goal of University of Maryland team (UMD team) in this year's HARD experiment is to leverage existing theories, and models about information need negotiation in information science literature to design and implement an automated process of generating clarification questions and utilizing the answers to improve the ranked list of documents for a given query statement.

The clarification questions generated by UMD team came from four aspects of context information related to a given query, which was motivated by research work about information need negotiation. The four aspects are 1) characteristics of the subject area that the user is querying; 2) user's motivation/background, especially user's recent experience with searching on the subject area; 3) user's preference to sub-collections within the document collection; and 4) user's anticipation to result format. The questions were then further narrowed down to those whose answers can be utilized in an automated process. UMD team also preferred those questions that would probe complimentary information to the metadata provided. They applied three techniques in their automated process to utilize the extra information obtained from clarification forms and meta data. The three techniques are query expansion based on keyword extraction, document reranking within a ranked list, and ranked lists merging.

Not all metadata were used in UMD team's HARD experiment. The used metadata satisfied two conditions: 1) the data were able to apply in the automated process, and 2) the data were not covered by the answers from their clarification forms. Therefore, only genre and granularity information were used, where the first one was used in document reranking, and the latter was used to trigger passage retrieval.

UMD team designed their own simple passage retrieval module. Their passage retrieval module assumes that the relevance of a passage is related to how many query terms it contains, how important those query terms are, and how relevant the document containing the passage is. Among the three, they gave more emphasis to the document containing the passage. All passages are ranked according to their relevance, with the condition that only three passages were allowed from the same document. The final result is top 1000 passages for a given query.

University of Waterloo and Bilkent University [Vechtomova et al., 2004]

This year we decided to focus on developing techniques for eliciting additional search criteria from users, preparing the ground for the next year's participation, where we plan to focus on techniques that exploit genre, familiarity and purpose metadata. We developed two topic clarification techniques for this year's entry:

1) The first technique consists in selecting one representative sentence from each of the top-ranked documents retrieved by the terms taken from the topic titles. The selected sentences were presented in the clarification form, and the users were asked to choose those sentences that are likely to represent relevant documents. We selected sentences on the basis of the sum of idf of query term instances in the sentence. Sentences with equal scores were ranked by the sum of tf*idf of all terms in the sentence, normalised by the sentence length. The documents which contained sentences chosen by users were used for query expansion. We used word co-occurrence measure of Z-score to select the query expansion terms.

2) The second technique is to show to the users phrases, selected from the two top-scoring sentences in each document from the initially retrieved set, and asking them to choose those phrases that are likely to represent relevant documents. We used a POS tagger and a noun phrase chunker to identify noun phrases, which were ranked by the sum of idf weights of their constituents. Top-ranked phrases were included in the clarification form. Terms from user-selected phrases were then used for query expansion.

We used Okapi BM250 for document and passage retrieval. For the topics requiring retrieval of best sentences, we used the sentence selection method described above.

3 HARD Corpus

The evaluation corpus is a combination of newswire text from the 1999 portion of the AQUAINT corpus and of U.S. government documents. The following table provides details on the make-up of the corpus. All information is from only 1999 because only documents from that year are included:

NYT	APW	XIE	CR	FR	Totals
137,806	77,876	104,698	16,609	35,230	372,219
750Mb	245Mb	310Mb	147Mb	330Mb	1.7Gb
Jan-Dec	Jan-Nov	Jan-Dec	Jan-Dec	Jan-Dec	

The New York Times (NYT), Associated Press Worldstream (APW), and Xinghua English (XIE) articles are all available on the AQUAINT disks. Those disks were available free-of-charge to all TREC participants.

The Congressional Record (CR) and Federal Register (FR) data set was gathered by the LDC for this track. Particularly lengthy documents from either source were not included because they cause serious annotation problems. This set of data was provided free-of-charge to all participants in the HARD track.

4 Topics

Topics follow the basic TREC style, but are more richly annotated with metadata that describes the searcher and the context of the query. The format of a topic is:

<top>
<num> Number: HARD-*nnn*
<title> *Web-style description of topic*
<desc> Description: *Sentence-length description of topic*
<narr> Narrative: *Paragraph-length description of topic,*
indended primarily to help future relevance assessors
<hard> item=*label*, value=*value*
<hard> item=*label*, value=*value*
<hard> item=*label*, value=*value*
<hard> item=*label*, value=*value*
. . .
</top>

The following metadata items and values are provided for each topic:

1. *item=PURPOSE* represents why the user is searching for the information.
   - value=BACKGROUND indicates that the searcher wants to know where the topic came from.
   - value=DETAILS means the searcher wants to know the details of the topic.
   - value=ANSWER indicates the user is looking for an answer to a specific question. (This value is implicitly linked to some of the GRANULARITY values.)
   - value=ANY means that the user has no specific purpose in mind or, at least, has not specified one.
2. *item=GENRE* represents the type of material the searcher is interested in.
   - value=OVERVIEW means the searcher is interested in general news related to the topic.
   - value=REACTION indicates the searcher is looking for news commentary on the topic.
   - value=I-REACTION is like REACTION but is specifically about non-U.S. news commentary.
   - value=ADMINISTRATIVE means the search is interested in official US government documents.
   - value=ANY indicates that any genre is acceptable or none was indicated.
3. *item=FAMILIARITY* represents how familiar the searcher is with the topic. Presumably a user who is fully aware of the details of a topic would not be interested in background material, for example.
   - value=1, no prior knowledge
   - ...
   - value=5, know details of topic
   - value=UNKNOWN means that the user does not know his or her familiarity or has not specified one.
4. *item=GRANULARITY* captures the amount of text that the searcher is anticipating in a value response.
   - value=DOCUMENT means the searcher is expecting complete documents (one or more).
   - value=PASSAGE will be selected when the search expects extracts from documents that are on the paragraph or multi-paragraph level.
   - value=SENTENCE means that the retrieved units should be roughly at the sentence level.
   - value=PHRASE means that user is expecting a small number of words (including just one) as a response.
   - value=ANY means the user has no specific granularity in mind or did not specify one.
5. *item=RELATED-TEXT*. This item includes sample relevant text. It may be repeated if there are multiple sample texts to be included.
   - value="..." identified text that is known to be related to the topic being specified. This provides a kind of pre-query relevance feedback. The intent is that this text not come from the evaluation corpus.

During topic creation, the LDC made an effort to have topics vary across each of the indicated metadata items.

When the LDC created the *evaluation* topics, they also gathered additional metadata beyond what was required by the HARD track. That information is provided along with the HARD metadata after the baseline runs and may be used in any way a site likes. The items collected were:

- OCCUPATION

- SPECIAL TRAINING
- SPECIAL INTERESTS, where the annotator can candidly explain why he or she chose this topic
- LANGUAGES SPOKEN
- AGE
- SEX

5 Relevance judgments

For each topic, documents that are annotated get one of the following judgments:

- NON-RELEVANT means that the document is known not to be relevant to the topic. (As is common in TREC, a document without any judgment is assumed to be non relevant.)
- SOFT-REL means that the document is relevant to the topic but that it does not satisfy the appropriate metadata. Given the metadata items listed above, that means it either does not satisfy the PURPOSE, GENRE, or the FAMILIARITY items (the others are not document-level items).
- HARD-REL means that the document is relevant *and* it satisfies the appropriate metadata.

In addition, if the GRANULARITY value is not DOCUMENT, then each judgment will come with information that specifies which portion of the documents is relevant.

To specify passages, HARD used the same approach used by the question answering track. A passage is specified by its byte offset and length. The offset is from the "<" in the "<DOC>" tag of the original document (an offset of zero would mean include the "<" character). The length indicates the number of bytes that are included. If a document contains multiple relevant passages, the document is listed multiple times.

The HARD track used the standard TREC pooling approach to find possible relevant documents. The top 75 documents from one baseline and one final run from each submitted system were pooled (i.e., 75 times 14 times 2 documents). The LDC considered each of those documents as possibly relevant to the topic.

Judging was done in three passes:

1. Decide if the document contains relevant material (soft rel)
2. Decide if the document matches metadata restrictions (hard rel)
3. Select relevant passages within the document

Across all topics, the LDC annotated 42,016 documents, finding 5,123 that were HARD-REL and another 2,533 that were HARD-REL. Topics ranged from three HARD-REL documents to 400, and from 6 to 714 if SOFT-REL is also included.

6 Training data

The LDC provided 10 training topics. The topics incorporated a selection of metadata values and came with relevance judgments (though the relevance judgments were delayed).

In addition, the LDC provided a mechanism to allow sites to validate their clarification forms. Sites could send a form to the LDC and get back confirmation that the form was viewable and some "random" completion of the form. The resulting information was sent back to the site in the same format that was used in the evaluation.

7 Results format

Results were returned for evaluation in standard TREC format extended, though, to support passage-level submissions since it possible that the searcher's preferred response is the best passage (or sentence or phrase) of relevant documents. Results included the top 1000 documents (or top 1000 passages) for each topic, one line per document/passage per topic. Each line will have the format:

topic-id Q0 docno rank score tag psg-offset psg-length

where:

- *topic-id* represents the topic number from the topic (e.g., HARD-001)
- *"Q0"* is a constant provided for historical reasons
- *docno* represents the document that is being retrieved (or from which the passage is taken)
- *rank* is the rank number of the document/passage in the list. Rank should start with 1 for the document/passage that the system believes is most likely to be relevant and continue to 1000.
- *score* is a system-internal score that was assigned to the document/passages. High values of score are assumed to be better, so score should generally drop in value as rank increases.
- *tag* is a unique identifier for this run by the site.
- *psg-offset* indicates the byte-offset in document docno where the passage starts. A value of zero represents the "<" in "<DOC>" at the start of the document. A value of negative one (-1) means that no passage has been selected and the entire document is being retrieved.
- *psg-length* represents how many bytes of the document are included in the passage. A value of negative one (-1) must be supplied when psg-offset is negative one.

8 Evaluation approach

Results were evaluated at the document level, both in light of and ignoring the query metadata. Ranked lists were also evaluated incorporating passage-level judgments. We discuss each evaluation in this section.

Two of the 50 HARD topics (155 and 231) had no "HARD rel" documents. That is, although there were documents that matched the topics, no document in the pool matched the topic *and* the query metadata. Accordingly, those two topics were dropped from both the HARD and SOFT evaluations. (They could have been kept for the SOFT evaluation, but then the scores would not have been comparable.)

8.1 Document-level evaluation

In the absence of passage information, evaluation was done using standard mean average precision. There were two variants, one for HARD-REL judgments and one for SOFT-REL.

Some of the runs evaluated in this portion were actually passage-level runs and could therefore include a document at multiple points in the ranked list—i.e., because more than one passage was considered likely to be relevant. For the document-level evaluation, only the first occurrence of a document in the ranked list was considered. Subsequent occurrences were "deleted" from the ranked list.

Figure 1 shows a tradeoff between the number of relevant documents found in the first ten retrieved and the system's average precision. Each submitted run generates a point on the scatter plot, the main purpose of which is to show the range of scores that came in. Not surprisingly, there is a clear relationship between the two values. Figure 2 shows the same relationship for the "hard relevance" condition.

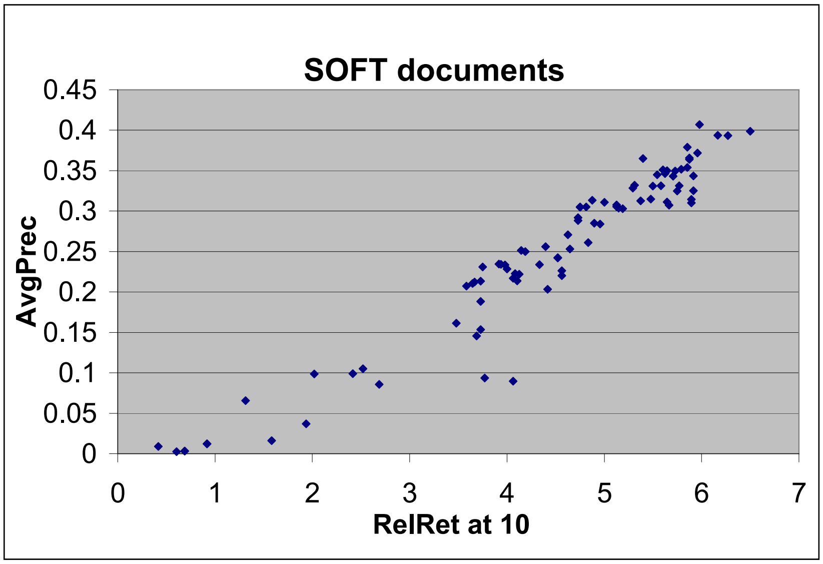

Figure 1: Scatter plot of number of relevant documents found in the first 10 compared to average precision, for "soft rel" documents.

8.2 Passage-level evaluation

The HARD track participants floated several passage evaluation measures. In the end, the track coordinator and NIST used one that was easy to implement and that attempted to match the goals of the community discussion.

The following operational description of passage recall and passage precision is provided by Ellen Voorhees of NIST to the HARD track participants.

> The passage level evaluation for a topic consists of values for passage recall, passage precision, and the F score at cutoff 5, 10, 15, 20, 30, 50, and 100, plus a R-precision score. As with standard document level evaluation, a cutoff is the rank within the result set such that passages at or above the cutoff are "retrieved" and all other passages are not retrieved. So, for example, if the cut-off is 5 the passage recall and precision are computed over the top 5 passages. R-precision is defined similarly to the document level counterpart: it is the passage precision after R passages have been retrieved where R is the number of relevant passages for that topic. We are using passage R-precision as the main evaluation measure reported for the track because it is a cutoff-based measure that tracks mean average precision extremely closely in document evaluations.
>
> For each relevant passage, allocate a string representing all of the character positions contained within the relevant passage (i.e., a relevant passage of length 100 has a string of length 100 allocated). Each passage in the retrieved set marks those character positions in the relevant passages that it overlaps with. A character position can be marked at most once, regardless of how many different retrieved passages contain it. (Retrieved passages may overlap, but relevant passages do not overlap.) The passage recall is then defined as the average over all relevant passages of the fraction of the passage that is marked. The passage precision is defined as the total number of marked character positions divided by the total number of characters in the retrieved set. The F score is defined in the same way as for documents, assigning equal weight

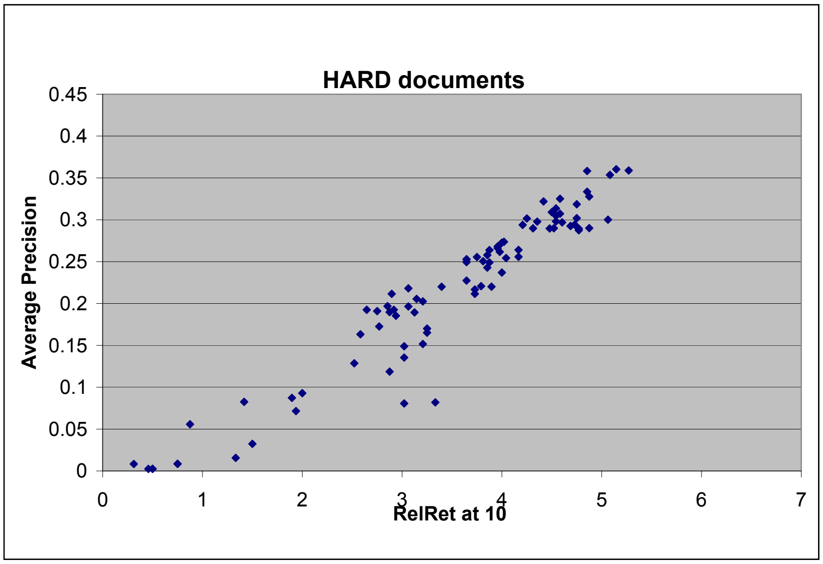

Figure 2: Scatter plot of number of relevant documents found in the first 10 compared to average precision, for "hard rel" documents.

to recall and precision:

$$F = (2 * \text{prec} * \text{recall})/(\text{prec} + \text{recall})$$

where F is defined to be 0 if prec+recall is 0. We included the F score because set-based recall and precision average extremely poorly but F averages well. R-precision also averages well.

In all of the above, a document is treated as a (potentially long) passage. That is, for topics where the granularity is "document" the relevant passage starts at the beginning of the document and is as long as the document. (These are represented in the judgment file as passages with -1 offset and -1 length, but are treated as described above.) For any topic, a retrieved document (i.e., where offset and length are negative one) is again just a passage with offset 0 and length the length of the document.

Using the above definition of passage recall, passage recall and standard document level recall are identical when both retrieved and relevant passages are whole documents. That is not true for this definition of passage precision. Passage precision will be greater when a shorter irrelevant document is retrieved as compared to when a longer irrelevant document is retrieved. This makes sense, but is different from standard document level precision.

Figure 3 shows the tradeoff between three measures for all submitted runs. Note that even runs that did not attempt any passage retrieval are included here; their "passages" are entire documents.

9 Conclusion

The HARD track is running again for TREC 2004. The experience of this track suggests the following changes, some of which have been adopted already:

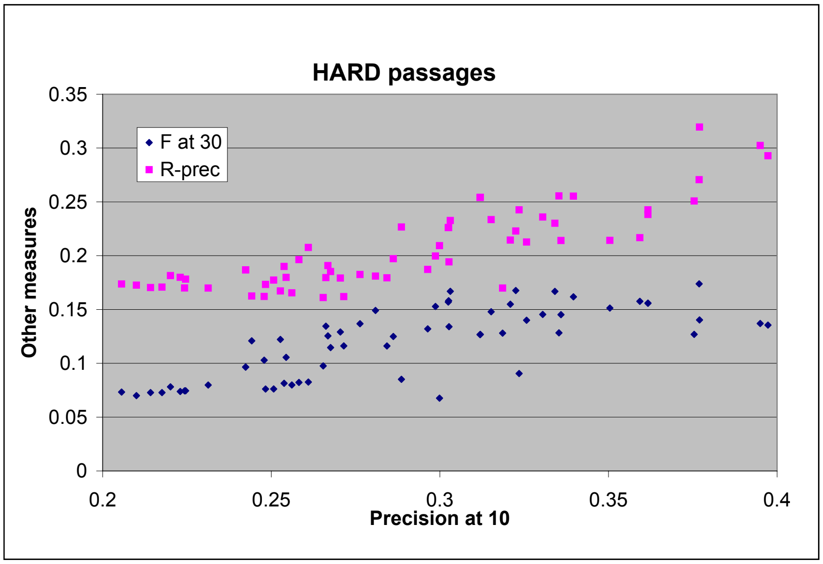

Figure 3: Scatter plot of precision at 10 documents retrieved compared to both the F measure at 30 documents retrieved and to R-precision, for “hard rel” passages.

- A corpus that permits a wider range of “interesting” metadata values would be useful. The current corpus was intended to provide a contrast between news and US government documents, but they were not different enough for metadata to be clearly useful.

  HARD 2004 will use a corpus of news from 2003. This does not provide the wide range we dream of, but it is a richer set of news than was used for HARD 2003. Also, because it is more recent news, it will be more pleasurable for the annotators to read.

- Passage-level judging was a terrifically difficult task for the LDC and needs to be revisited. The LDC has some thoughts on this but they have not been finalized at this time.

- As is typical with new tracks, many decisions were made quite late in the process. Next year they need to happen more quickly.

Acknowledgments

The HARD corpus creation and annotation was funded by DARPA under the TIDES program. NIST provided coordination and evaluation support for the work in the track as part of its support for TREC.

The coordination at the CIIR was supported in part by the Center for Intelligent Information Retrieval, in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903, and in part by NSF grant number IIS-9907018. Any opinions, findings and conclusions or recommendations expressed in this material are the author’s and do not necessarily reflect those of the sponsor.

References

[AbdulJaleel et al., 2004] AbdulJaleel, N., Corrada-Emmanuel, A., Li, Q., Liu, X., Wade, C., and Allan, J. (2004). UMass at TREC 2003: HARD and QA. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Belkin et al., 2004] Belkin, N., Kelly, D., Lee, H.-J., Li, Y.-L., Muresan, G., Tang, M.-C., Yuan, X.-J., and Zhang, X.-M. (2004). Rutgers' HARD and web interactive track experiments at TREC 2003. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Grunfeld et al., 2004] Grunfeld, L., Kwok, K., Dinstl, N., and Deng, P. (2004). TREC 2003 robust, HARD and QA track experiments using PIRCS. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[He and Demner-Fushman, 2004] He, D. and Demner-Fushman, D. (2004). HARD experiment at Maryland: From need egotiation to automated HARD process. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Hearst et al., 1995] Hearst, M. A., Karger, D. R., and Pedersen, J. O. (1995). Scatter/gather as a tool for the navigation of retrieval results. In *The proceedings of the 1995 AAAI Fall Symposium on Knowledge Navigation.*

[Ma et al., 2004] Ma, L., Tan, W., Chen, Q., Ma, S., , Shi, S., Xiao, S., Wang, H., and Wang, H. (2004). THUIR at TREC 2003: HARD experiments. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Ramakrishnan et al., 2004] Ramakrishnan, G., Bellare, K., Shah, C., and Paranjpe, D. (2004). Generic text summarization using wordnet for novelty and hard. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Robertson et al., 2004] Robertson, S., Zaragoza, H., and Taylor, M. (2004). Microsoft cambridge at TREC-12: HARD track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Shanahan et al., 2004] Shanahan, J., Bennett, J., Evans, D. A., Hull, D. A., and Montgomery, J. (2004). Clairvoyance Corporation experiments in the TREC 2003 high accuracy retrieval from documents (HARD) track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Shen and Zhai, 2004] Shen, X. and Zhai, C. (2004). Active feedback - UIUC TREC-2003 HARD experiments. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Srikanth et al., 2004] Srikanth, M., Ruiz, M., and Srihari, R. (2004). UB at TREC 12: HARD and genomics tracks. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Vechtomova et al., 2004] Vechtomova, O., Lam, E., and Karamuftuoglu, M. (2004). Interactive search refinement techniques for HARD tasks. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

[Wu et al., 2004] Wu, Z., Du, L., Sun, L., and Ye, S. (2004). TREC12 HARD track at ISCAS. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.