HARD Track Overview in TREC 2004 (Notebook)
High Accuracy Retrieval from Documents

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

1 Introduction

The High Accuracy Retrieval from Documents (HARD) track explores methods for improving the accuracy of document retrieval systems. It does so by considering three questions:

1. Can additional metadata about the query, the searcher, or the context of the search provide more focused and therefore accurate results? These metadata items generally do not directly affect whether or not a document is on topic, but they do affect whether it is relevant. For example, a person looking for introductory material will not find an on-topic but highly technical document relevant.

2. Can highly focused interaction with the searcher be used to improve the accuracy of a system? Participants created "clarification forms" generated in response to a query—and leveraging any information available in the corpus—and were filled out by the searcher. Typical clarification questions might ask whether some titles seem relevant, whether some words or names are on topic, or whether a short passage of text is related.

3. Can passage retrieval be used to effectively focus attention on relevant material, increasing accuracy by eliminating unwanted text in an otherwise useful document? For this aspect of the problem, there are challenges in finding relevant passages, but also in determining how best to evaluate the results.

This is a notebook paper and so is short on details. Additional information will be provided in the final paper. The HARD track's Web page may also contain useful pointers:

`http://ciir.cs.umass.edu/research/hard`

2 Participation

The following 16 sites participated in the HARD track. A summary of each group's activity is provided below. The summaries were written by the site and are listed in alphabetical order.

Chinese Academy of Science, Institute of Software

The ISCAS team participated in all the three aspects of the HARD task. They focus on studying the problem of the combination of the user- and query-information from clarification forms and metadata. All the submitted results are constructed automatically though some of them are time consuming. They provided two kinds of clarification form. One is a list of keywords that might appear in relevant documents, the other is a list of the title and keywords of the top 10 relevant documents. For passage retrieval, they first cut these documents to small pieces and then do the same run as document retrieval. They used almost all kinds of the metadata provided in topics. The details are as follows:

Subject & Familiarity: Google is used as a resource for query expansion. The metadata subject and familiarity are used together to choose the related texts from the searched result of Google. First input the title of a query to Google. Then choose the related web site from the top 3 sites. If the web site at Google directory is classified the same as the subject at metadata, then the web site is a related site. How many texts within the related site should be used as related texts is decided by metadata familiarity based on an assumption, that is, the less the user is familiar with a topic, the more he wants to know.

Genre: A term list related to different Genres is constructed manually. Then re-rank the document list by the number of these terms in the documents.

Geography: Collect the terms related to US from Internet. The terms include the name and abbreviation of the main cities and all states of US. Re-rank the document list by counting the number of these terms in the documents.

Related Text: The relevant texts are used as the basis for automatic query expansion.

Finally, they combined all the information above by different ways and wish could get more accurate ranked lists.

Clairvoyance Corporation

The Clairvoyance team participated in the HARD Track, submitting three runs. The experiments focused primarily on exploiting user feedback through clarification forms for query expansion and largely ignored other features of the documents or metadata. Two non-baseline runs contrasted alternative strategies for clustering documents in the response set of a topic-one based on simple re-grouping of responding documents and the other based on reprocessing of the response set into small sub-documents (passages) and then clustering. In both cases, the grouped documents were presented to judges in evaluation forms and the groups of documents selected as being on topic were used as the source of query expansion terms. A more detailed description follows.

All Clairvoyance processing was managed under the CLARIT system, thus, terms (features) were based on morphologically-normalized linguistic phrases for document indexing and query processing. Furthermore, all documents were processed as sets of sub-documents and scoring of documents was based on best-scoring sub-documents.

The baseline submission (CL102TDN) reflected an automatic pseudo-relevance feedback run. Queries were formed from the Title, Description, and Narrative [TDN] portions of each topic then submitted to the full test corpus. Two hundred terms were selected from the top-responding ten documents for each query, weight-normalized (N(t)), and merged with the terms in the initial query (which were weight-boosted: 2+N(t)). The expanded query was resubmitted to obtain the principal results set. There were no feedback forms associated with this run.

As a "standard" experimental run (CLAI2), the Clairvoyance team used a version of its "quintad" approach of TREC 2003. Given the ranked results of the baseline run, the top-200 ranked documents for each query were segregated as a response set. For each topic, beginning with the first ranked document, a query was created from the document by parsing its top-scoring sub-document and then submitted to the topic's response set. The source document and the four other closest documents were combined into a group (or pseudo-cluster) and removed from the response set. This process was repeated using the next remaining highest ranked document in the response set for the next query. This process continued until a total of ten five-document groups was formed. Note that, using this method, the response set is reorganized with a bias toward the initial top-scoring documents. User-feedback forms were generated by presenting each of the ten groups in the order in which they were produced. Each group was headed by a "summary" set of 35 terms, taken from the five top-scoring sub-documents in the group, and showed each document's title. (If the source document had no title, a pseudo-title was created by taking the first 100 characters from the beginning of the best-hit sub-document of the document, respecting word boundaries.) Users were asked to judge which groups seemed relevant to the topic and indicated their judgments by selecting the appropriate category: "on topic"; "unsure"; "not on topic"; or "unjudged". All best-scoring sub-documents in any quintad that received a user-feedback score of "on-topic" were reserved for further processing. If there were fewer than five "on-topic" quintads, all best-scoring sub-documents from up to five "on-topic" and "unsure" marked groups were used. Two hundred terms were extracted from the selected documents, their weights were normalized, and they were merged back into the TDN version of the query, with boosting. This expanded query was submitted to the full test database and the top-1000 ranked results were provided to TREC as the output of this run. In those cases when there were insufficient "on-topic" or "unsure" marked groups, the original response set from the baseline run was used as the new result, but modified by excluding any documents from groups that may have been marked "not on topic" and promoting to the top of the ranked list any documents that may have been marked "on-topic".

As a new experimental run (CLAI1), the Clairvoyance team used a new approach to response set clustering. As with the quintad run (CLAI2), the baseline run's top-200 ranked documents for each query were segregated as a response set. This set was reprocessed as a database of short sub-documents. These short sub-documents were re-ranked against the original query and up to the top-500 responding sub-documents were reserved as response set-2. The sub-documents in response set-2 were grouped using asymmetric clustering. (There was no limit on the number of sub-documents that could form a group or the number of groups, but, practically speaking the number of clusters that resulted from this process was 5-25.) Each cluster was score-ranked against the original query, using the cluster centroid vector as a representation of the cluster content. User-feedback forms were generated by presenting the groups in rank order; each group was preceded by a set of summary terms based on the centroid vector and showed the top-five document titles corresponding to the best sub-documents in the cluster. Any summary terms that were also query terms were front-ranked and highlighted. Judges marked groups using the same categories as for the quintad run. Feedback was processed in the same fashion as for quintad runs, but was based on the short sub-documents from response set-2.

Indiana University

No information provided.

Microsoft Research Cambridge

Microsoft Research Cambridge invited feedback on selected snippets via the clarification forms, and also invited users to enter free form (positive or negative) phrases. Main area of investigation was methods of feature selection for feedback from the chosen snippets. Various selection formulae were tried on training data (last year's HARD data and also the training data supplied for this year's track).

Rutgers University

No information provided.

The Robert Gordon University

RGU investigated the effect of exploiting the topic metadata to re-rank their initial baseline run. They used the Lemur toolkit (LTK) to obtain a baseline ranking, using title and description for each topic, and OKAPI BM25 weighting. Then, they focussed on re-ranking this baseline for each topic, based on queries generated specifically to rank separately by genre, geography, and familiarity. The baseline and metadata-derived rankings were then combined using an evidence combination approach.

RGU were interested in applying machine learning techniques to build classifiers for ranking documents, based on the data provided by the training topics. In fact, the training data proved either too sparse, or too biased given the small number of topics, and this approach was abandoned. Instead, for the genre specification, they manually generated a single query for re-ranking by "opinion/editorial", and they did not attempt to generate re-rankings for genre "news" or genre "other". For the geography specification, they conjectured that US (resp. non-US) geography could be approximated using a single query for re-ranking comprising US (resp. non-US) places names. Familiarity rankings were generated based on topic-specific queries generated by a novel approach based on the clarity-measure (see below).

RGU wanted to explore a principled way of combining the evidence provided by the individual rankings, and specifically using Dempster-Shafer evidence combination. This proved problematic due to the high variability of scores within each ranking, and the difficulty of consistently normalising scores from different sources of evidence. Instead, they generated rank-scores for each ranking, by effectively assigning a score of (1001-rank) for each document (assuming 1000 documents in the baseline and un-retrieved documents assigned a score of zero). The individual rankings were combined by averaging the document scores, where each source of evidence (baseline, genre, geography, and familiarity) was weighted in the range zero to one, according to the estimated accuracy and perceived importance of that source. A number of settings for the source weightings were tried, and submitted in the final runs.

RGU devised a specific hypothesis concerning ranking documents by familiarity, namely: Users unfamiliar with a topic will prefer documents in which representative terms occur, and users familiar with a topic will prefer documents in which highly discriminating terms occur. If these sets of representative and discriminating terms could be identified for each topic, then the term sets could be used as a query to re-rank the baseline according to familiarity. RGU operationalised this hypothesis on the basis of two variants of the clarity measure, referred to as clarity-representation and clarity-discrimination. Using these measures, and for each topic, they generated specific queries for ranking by 'little' familiarity and 'much' familiarity respectively.

Tsinghua University

No information provided.

Illinois Urbana-Champaign

University of Illinois at Urbana-Champaign focused on evaluating a new method that employs hidden Markov models for passage retrieval. Most of the current passage retrieval methods extract only fixed-length passages. The optimal length of a passage, however, depends on both the query and the individual document. UIUC aimed at accurately identifying variable-length passages. They used a simple 5-state hidden Markov model to retrieve the most relevant passage from a document. A pseudo-feedback mechanism was naturally incorporated into the system. UIUC also exploited some user feedback through clarification forms.

University of North Carolina

In this year's HARD track, UNC took advantage of the one-shot interaction provided by the clarification form to investigate the effectiveness of various techniques for eliciting additional information from searchers about their information problems. UNC's experiments were motivated by several interests. First, the UNC team was interested in creating a feedback technique for use in situations where a searcher's initial query is unclear or ambiguous. Although previous research has successfully developed and evaluated the clarity measure [Cronen-Townsend et al., 2002] for predicting query ambiguity, it remains unclear what steps should be taken to clarify ambiguous queries once they are identified. Thus, UNC sought to extend the work on the clarity measure by investigating techniques that could potentially be used to follow-up ambiguous queries. Second, the UNC team was interested in developing and evaluating a generic, document-independent feedback technique that could be used in multiple information-seeking situations. This interest was motivated by the supposition that traditional relevance feedback techniques, which typically present top-ranked documents or keywords to searchers for feedback, are unlikely to work well in situations where ambiguous queries are posed because there is a large chance that documents retrieved in response to such queries will be irrelevant.

In designing their clarification forms, UNC considered several things: techniques used by sites in last year's HARD track and the results of these techniques, feedback provided by last year's searchers about the clarification forms, and previous research investigating elicitation techniques in real-world and electronic environments. Rather than present searchers with a list of keywords or snippets from retrieved documents, the team chose to pose three open-ended questions to searchers. Two of these questions required searchers to respond using natural language, that is, in complete sentences or phrases, while one of these questions asked searchers to respond using key terms. Preliminary analysis of searchers' responses to the clarification form indicate that searchers provided, on average, seven new terms per topic in response to one of the natural language questions, and only 1.5 new terms per topic in response to the key term question. UNC used the Lemur IR toolkit (TFIDF retrieval model) to conduct all of the experiments; the information obtained from the clarification form was used for query expansion. Because of their research interest, UNC did not use any meta-data in their experiments, nor did they perform passage retrieval.

University of Chicago

The University of Chicago's participation in the HARD track focused on passage retrieval and on the exploitation of metadata information to improve HARD-relevance. Passage retrieval employed query-specific merger of relevant 2-3 sentence pseudo-documents to extract passages. The use of metadata fields, primarily GEOGRAPHY and GENRE, contrasted the use of evidence from explicit lexical cues and from language model perplexity scores to identify documents meeting the desired criteria. Documents were reranked based on their match to the specified metadata criteria.

University of Cincinnati

The University of Cincinnati used the WordNet dictionary (actually, the nouns division of it) and topics metadata information (more precisely, the information contained in topic/related-text/relevant node) in order to expand initial search query, that was generated from the terms contained in topic/title node.

Baseline run was performed by generating the initial search query from the terms contained in topic/title inner text, after removing stopwords. Document ranking is evaluated using the Okapi BM25 function [Robertson et al., 1998], with Robertson/Sparck Jones weight [Robertson and Sparck Jones, 1976].

For our final run, we used two different sources for a query expansion – clarification forms and the topics metadata information mentioned above. For each topic, the clarification form was generated by taking the terms contained in the topic/title node for the current topic. For each term in the topic, the clarification form listed all possible term synsets and synsets that are either in hyponym or meronym relation with the term synsets. A user was able to choose everything that was synonymous or semantically related to the topic terms.

A query for the final run is formed by expanding a baseline query with top ten weighted terms taken from the clarification forms and topics metadata information. The weight is evaluated using the co-occurrence of an expansion term and the baseline query terms.

University of Maryland

The University of Maryland/Johns Hopkins University team used maximal marginal relevance to remove redundancy when generating clarification forms and applied language modeling to estimate the desired passage extent. Two baseline runs used keywords automatically selected from the full topic description (title, description, and narrative). One run was submitted with no query expansion, the other included terms extracted from top 10 returned documents using blind relevance feedback techniques.

The clarification forms were designed to seek evidence about the type of information that was sought, additional cues that could help locate the desired information, and desired passage extent. A list of named entities (i.e., person names, organization names, and locations) was extracted from the top 10 documents from the unexpanded baseline run and displayed to the assessor in the clarification form. The assessor was asked to select any appropriate named entities, and a text box was provided in which additional entity names could be manually entered. For cases in which passages rather than full documents were requested, as many as five highly ranked passages from highly ranked documents were also presented to the assessor on the clarification form. Those passages were selected in a greedy manner, favoring passages similar to the topic description, but dissimilar to passages that had already been selected for display. This "marginal relevance" factor was included to maximize the diversity of the resulting set of five passages. The assessor was asked to designate passages that could be recognized as relevant and to indicate whether the passage was too long, about the right length, or too short. When full documents were requested, passages from the same document were concatenated and presented as a document surrogate for judgment. An assessment of passage length was not requested in that case. Entities and passages selected by the assessor using the clarification from and relevant text identified in the provided metadata were added to the query using heuristically established weights to perform an improved search (which was then used as a basis for passage retrieval if passages were requested).

The passage retrieval module relied on the density of query terms to identify possible passage locations, combining evidence from document scores with term density to compute a single rank order on passages (i.e., passages from the same document could appear at different locations in the list). Passage extent was estimated based on three types of features: (i) similarity of each paragraph to the query, (ii) similarity of each paragraph to its preceding paragraph, (iii) differences between sequential inter-paragraph similarity values. Linear discriminant analysis was applied to the resulting feature vectors to identify the projection that most differentiated relevant paragraphs from irrelevant ones. This scalar-valued projection was then modeled as the output of a Hidden Markov Model (HMM) with two states, relevant and irrelevant. This HMM was trained using HARD-03 passages (for which ground truth is known). Cross-validation showed a substantial improvement in R-precision over HARD-03 results.

University of Massachusetts Amherst

UMass Amherst investigated the four areas in HARD 2004: clarification forms, fixed length passage retrieval, variable length passage retrieval, and metadata. For clarification forms, they studied various traditional methods of eliciting passage-level relevance. In addition to passage-level judgments, UMass Amherst also explored the use of named entities for interactive query expansion and temporal feedback for document re-ranking. They returned fixed-length passages for the twenty five passage topics. These passages were ranked using a support vector machine that used term-based statistics as features. The margin from the boundary was used to score the passages. In addition to retrieving fixed-length passages, UMass Amherst experimented with extracting passages from documents at retrieval time based to produce passages of different lengths. On various runs, they utilized the related text, genre, and geography metadata. To use related text, they created a language model for all related text and mixed this with a model for the topic's title and description and then retrieved documents using this final mixture model. For genre and geography metadata types, UMass Amherst built a SVM classifier for each. They reranked documents based on their classifier scores.

University of Twente

UTwente's HARD track research focuses on the usage of the meta-data coming along with the search topics. They have neither used clarification forms for relevance feedback nor have they performed retrieval on passage-level.

Looking back at last years' HARD track experiments from other sites, every context category was handled in a different way, if the provided meta-data was used at all. Instead of introducing another set of new techniques, UTwente's basic research hypothesis was that statistical language models are a sufficient mean to be applied as a universal representation for all context categories. Obviously, language models can be utilized effectively as subject classifiers, but they tried to apply them on other user meta-data as well, in order to come up with a uniform framework for contextual text retrieval.

Due to time limitations, UTwente restricted their experiments this year to three exemplary types of meta-data only, namely the subject, geography and related text category. In order to construct the language models for subject and geography classification, they used three different sources of data: manual annotation, keywords in APE section of the corpus, and the training data. The later ranking according to query and "meta-query" was calculated by standard language modeling score functions and combined using an evidence combination approach.

Besides looking for the system performance in total, the given context categories were also compared with respect to their ability to improve the retrieval.

University of Waterloo

The University of Waterloo/Bilkent University team focused this year on the use of noun phrases in relevance feedback and searching. Several syntactico- statistical methods for extracting noun phrases and single terms from text were developed and comparatively evaluated in the interactive query expansion task using clarification forms. A phrase search algorithm was also developed and tested. Another area of investigation was document classification by genre into opinion-editorial and news articles.

York University

The York team uses Okapi BSS (Basic Search System) as the basis to participate in the HARD Track. Our experiments mainly focus on exploiting various feedback methods for query expansion and term weighting formulae. Both document level index and passage level index are built for the retrieval purpose.

Two sets of clarification forms (CFs) were submitted. The first one concentrated on short paragraph feedback and the second one concentrated on keyword feedback within the context. The CF results were then used in relevance feedback algorithms. From the metadata, only the granularity, geography and relevant document information were used. The geography information was used to filter out non-relevant documents and the relevant document information was used to automatically expand the query terms. The following two algorithms were applied in the final submissions. Both of them are based on the same topics and CF results.

Algorithm 1: For each topic, York did both document level search and passage level search. Then these two searches were combined into one. The basic assumption for this combination is: if an article is hit by both searches, it should be assigned more weight than others that are hit by only one search. After initial results are generated, blind feedback was used to do the query expansion. Then the final results were generated by using the same algorithm.

Algorithm 2: For each topic, York did only document level search or passage level search according to the value of "retrieval-element" (document or passage). The search terms are automatically extracted from topics and CF results. The terms extracted from CF results were classified into two sets: positive and negative terms. Positive weights were assigned to positive terms and negative weights were assigned to negative terms respectively.

In summary, York used Okapi BM25 for passage retrieval. But for document level retrieval, York applied a modified version of BM25, named as BM50, by adding a correction factor which is based on the length of document. This correction factor is added at the end of the usual BM25 function, and is served as an adjusting factor that gives relative low weight to those documents with considerable short or long lengths. For the passage level evaluation, the automatic run 'yorku04ha1' achieves the best result (0.358) in terms of Bpref measure at 12K characters.

3 HARD Corpus

The evaluation corpus consists entirely of English text from 2003, most of which is newswire. The specific sources and approximate amounts of material are:

Source	Abbrev	Num docs	Size (Mbs)
Agence France Press	AFP	226,777	497
Associated Press	APW	236,735	644
Central News Agency	CNA	4,011	6
LA Times/Wash Post	LAT	34,145	107
New York Times	NYT	27,835	105
Salon.com	SLN	3,134	28
Ummah Press	UMM	2,557	5
Xinhua (English)	XIN	117,516	183
Totals		652,710	1,575

4 Topics

Topics were an extension of typical TREC topics: they included (1) a statement of the topic and (2) a description of metadata that a document must satisfy to be relevant, even if it is on topic. The topics were represented in XML and included the following components:

- *number* is the topic's number–e.g., HARD-003.
- *title* is a short, few word description of the topic.

- *description* is a sentence-length description of the topic.
- *topic-narrative* is a paragraph-length description of the topic. This component did not contain any mention of metadata restrictions. It is intended purely to define what is “on topic.”
- *metadata-narrative* is a topic author’s description of how metadata is intended to be used. This description helps make it clear how the topic and metadata were intended to interact.
- *retrieval-element* indicates whether the judgments (hence retrieval) should be at the *document* or *passage* level. For HARD 2004, half of the topics were annotated at the passage level.
- The following metadata fields were provided:
  - *familiarity* had a value of *little* or *much*. It affected whether a document was relevant, but not whether it was on topic.
  - *genre* had values of *news-report*, *opinion-editorial*, *other*, or *any*. It affected whether a document was relevant, but not whether it was on topic.
  - *geography* had values of *US*, *non-US*, or *any*. It affected whether a document was relevant, but not whether it was on topic.
  - *subject* describes the subject domain of the topic. It is a free-text field, though the LDC attempted to be consistent in the descriptions it used. It affected whether or not a document was on-topic.
  - *related-text.on-topic* provided an example of text that the topic’s author considered to be on-topic but not relevant.
  - *related-text.relevant* provided an example of text that the topic’s author considered to be relevant (and therefore also on-topic).

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

5 Relevance judgments

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

- OFF-TOPIC means that the document does not match the topic. (As is common in TREC, a document without any judgment is assumed to be off topic.)
- ON-TOPIC means that the document does match the topic but that it does not satisfy the provided metadata restrictions. Given the metadata items listed above, that means it either does not satisfy the FAMILIARITY, GENRE, or GEOGRAPHY items (note that SUBJECT affects whether a story is on topic).
- RELEVANT means that the document is on topic *and* it satisfies the appropriate metadata.

In addition, if the *retrieval element* field is *passage* then each judgment will come with information that specifies which portions of the documents are 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 will be from the “<” in the “<DOC>” tag of the original document (an offset of zero would mean include the “<” character). The length will indicate the number of bytes that are included. If a document contains multiple relevant passages, the document will be listed multiple times.

The HARD track used the standard TREC pooling approach to find possible relevant documents. The top 85 documents from one baseline and one final run from each submitted system were pooled (i.e., 85 times 16 times 2 documents). The LDC considered each of those documents as possibly relevant to the topic.

Across all topics, the LDC annotated 36,938 documents, finding 3,026 that were on topic and relevant and another 744 that were on topic but not relevant. Topics ranged from one on topic and relevant document to 519; from 1 on topic but not relevant document to 70.

6 Training data

The LDC provided 20 training topics and 100 judged documents per topic. The topics incorporated a selection of metadata values and came with relevance judgments.

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. (No one took advantage of such a capability.)

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 (HARD) and ignoring (SOFT) the query metadata. Ranked lists were also evaluated incorporating passage-level judgments. We discuss each evaluation in this section.

Five of the 50 HARD topics (401, 403, 433, 435, and 450) had no relevant (and on topic) 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 judgments and one for SOFT.

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.

8.2 Passage-level evaluation

A variety of measures were considered and will be reported on for the final paper.

9 Conclusion

The second year of the HARD track appears to have been much more productive for most sites. With better training data and a clearer task definition earlier, groups were able to carry out more careful and interesting research.

At the time of the notebook writing, it is not known what has been learned. The final paper will contain details on that.

Acknowledgments

This work 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

[Cronen-Townsend et al., 2002] Cronen-Townsend, S., Zhou, Y., and Croft, W. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 299–306.

[Robertson and Sparck Jones, 1976] Robertson, S. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(May-June):129–146.

[Robertson et al., 1998] Robertson, S., Walker, S., and Beaulieu, M. (1998). Okapi at TREC-7: automatic ad-hoc, filtering, VLC and interactive. In *Proceedings of the Seventh Text Retrieval Conference*, pages 253–264.