HARD Track Overview in TREC 2004
High Accuracy Retrieval from Documents

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

Abstract

The HARD track of TREC 2004 aims to improve the accuracy of information retrieval through the use of three techniques: (1) query metadata that better describes the information need, (2) focused and time-limited interaction with the searcher through "clarification forms", and (3) incorporation of passage-level relevance judgments and retrieval. Participation in all three aspects of the track was excellent this year with about 10 groups trying something in each area. No group was able to achieve huge gains in effectiveness using these techniques, but some improvements were found and enthusiasm for the clarification forms (in particular) remains high. The track will run again in TREC 2005.

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, short-duration, 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—that 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.

The HARD track ran for the second time in TREC 2004. It used a new corpus and a new set of 50 topics for evaluation. All topics included metadata information and clarification forms were considered for each of them. Because of the expense of sub-document relevance judging, only half of the topics were used in the passage-level evaluation.

A total of 16 sites participated in HARD, up from 14 the year before. Interest remains strong, so the HARD track will run again in TREC 2005, but because of funding uncertainties will only address a subset of the issues. Exactly what is included and how it takes place will be determined by interested participants. Information about the track will be available at the track's Web page, `http://ciir.cs.umass.edu/research/hard` (the contents of the site are not predictable after 2005).

Topic creation, clarification form entry, and relevance judging were all carried out by the Linguistic Data Consortium (LDC) at the University of Pennsylvania (http://www.ldc.upenn.edu). The annotation work was supported in part by the DARPA TIDES project.

Evaluation of runs using the judgments from the LDC was carried out by NIST.

The remainder of this document discusses the HARD 2004 track and provides an overview of some of its results. Additional details on results are available in the TREC papers from the participating sites.

2 HARD Corpus

The HARD 2004 evaluation corpus itself consisted 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

This information was made available to participating sites with a research license. The data was provided free of charge, though sites interested in retaining the data after the HARD track ended were required to make arrangements with the LDC to do so.

3 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.

The following is a sample topic from the evaluation corpus (topic HARD-428). Some particularly long sections of the topic have been elided.

```
<topic>

<number>
HARD-428
</number>

<title>
International organ traffickers
</title>

<description>
Who creates the demands in the international ring of organ trafficking?
</description>

<topic-narrative>
Many countries are institutionalizing legal measures to prevent the
selling and buying of human organs. Who, in the ring of international
organ trafficking, are the "buyers" of human organs? Any information
that identifies 'where' they are or 'who' they may be will be
considered on topic; the specificity of info does not matter. Also,
the story must be about international trafficking. Stories that only
contain information about the "sellers" of organs or those that focus
on national trafficking will be off topic.
</topic-narrative>

<metadata-narrative>
Subject (CURRENT EVENTS) is chosen as it is expected that such
articles will have more information about the identities of the
parties involved. Genre (NEWS) is expected to exclude stories that
tends to focus on ethical matters.
</metadata-narrative>

<retrieval-element>
passage
</retrieval-element>

<metadata>
   <familiarity>
   little
   </familiarity>

   <genre>
   news-report
   </genre>
```

```
   <geography>
   any
   </geography>

   <related-text>
     <on-topic>
     Every day, 17 Americans die of organ failure. In Israel, the average
     wait for a kidney transplant is four years. In response, a global gray
     market has bloomed. In India, for example, poor sellers are quickly...
     </on-topic>

     <relevant>
     At least 30 Brazilians have sold their kidneys to an international
     human organ trafficking ring for transplants performed in South
     Africa, with Israel providing most of the funding, says a legislative...
     </relevant>
   </related-text>

   <subject>
   CURRENT EVENTS
   </subject>
</metadata>
</topic>
```

4 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 for evaluation purposes.)
- 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 comes 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 [Voorhees, 2005]. 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.

5 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.)

6 Clarification forms

A unique aspect of the HARD track is that it provides access to the person who formulated the query and will be doing the annotation. It allows sites to get a small amount of additional information from that person by providing a small Web page as a form with clarification questions, check boxes, etc. for the searcher to fill in.

The assessor spent no more than three (3) minutes filling out the form for a particular topic. If some portions of a form were not filled out when the time expired, those portions were left blank. Sites were aware of the time limit and were encouraged to keep their forms small—however, several (perhaps most) sites built longer forms intending to get whatever they could within three minutes rather than building forms designed to be filled in quickly.

In order to avoid implementation issues, systems were required to restrict the forms to simple HTML without Javascript, images, and so on. They were also told what would be the hardware configuration used by annotators, so they could tailor the presentation appropriately if desired.

The LDC reported that the annotators enjoyed filling out clarification forms immensely—if only because it was an entirely new type of annotation task for them.

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 had 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 five topics were dropped from both the HARD and SOFT evaluations. (They could have been kept for the SOFT evaluation, but then the scores of the two evaluations 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. (That meant that it was possible for a site to submit 1000 items in a ranked list, but have fewer than 1000 documents ranked.)

8.2 Passage-level evaluation

Two passage measures were explored for HARD 2004. The first was the same one used in HARD 2003, passage R-precision. Some research at UMass Amherst demonstrated an extremely strong bias in favor of short passages, so a second measure was also explored.

8.2.1 Passage R-Precision

In a nutshell, this evaluation measure considers the "true" relevant R passages as found by annotators. It considers the top R passage returned by a system and counts the proportion of characters that overlap relevant passages. It incorporates a penalty for repeating text in multiple passages. More details are provided below.

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 an evaluation measure reported for the track because it is a cutoff-based measure that tracks mean average precision extremely closely in document evaluations.

The following is an operational definition of passage recall and precision as used in the evaluation. 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 to recall and precision: F = (2*prec*recall)/(prec+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, 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 -1) 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.

8.2.2 Passage-level bpref

Some explorations at UMass Amherst showed that passage R-precision could be improved dramatically by splitting existing passages into smaller pieces. For example, by splitting the top-ranked passages into 32 pieces and then using the top R of those (rather than the top R original passages), the value of passage R-precision increased by 128%.

Although numerous measures were considered, a variation of bpref [Buckley and Voorhees, 2004] was finally selected. In this measure, the top 12,000 characters of the system's ranked list of passages was considered (intended to correspond roughly to 10 normal sized passages).

As a document evaluation measure, bpref considers two sets of documents: a relevant set and a non-relevant set. The assumption is that if a document A is taken from the first set and B is taken from the second, then the user has a binary preference that A be ranked higher than B. The measure counts the proportion of times that the user's implied set of preferences is satisfied. A perfect system would rank all known relevant documents above all known non-relevant documents, would thereby satisfy all of the user's preferences, and receive a score of 1.0. The worst possible score is zero, and systems will normally score somewhere in the middle.

To extend this measure to passages, we consider character-level preferences. We assert that all relevant characters should be presented before any non-relevant characters and count the proportion of preferences that are satisfied. Note that the choice of character as the base unit is arbitrary and made for reasons of simplicity. It could have been word, phrase, or even sentence, but each of those would require algorithmic decisions about boundaries between units that are not necessary for character-level decisions. We believe (though have not investigated) that different units will merely change the scale of results.

9 Protocol

The HARD 2004 track ran from May through August of 2004. On June 25th, sites received the 50 evaluation topics, but without any of the metadata fields provided. That is, they received just the title, description, and narrative information, a format consistent with past "ad hoc" TREC tracks.

Using that base information, sites were asked to do their best to rank documents for relevance and return the ranked list of documents (not passages). These were the "baseline runs" and were due to NIST on July 9th.

In addition, sites could optionally generate up to two clarification forms that the LDC annotators would fill out. These forms were due to the LDC on July 16th

On July 29th, the filled-out forms were returned to sites and the metadata fields of the topics were released to all sites, regardless of whether they used clarification forms. Sites could use any of that information to produce improved ranked lists. The final runs, incorporating everything they could, were due to NIST on August 5th.

As described above, one baseline run and one final run were used from each site. The top 85 documents from each of those runs were pooled together and used by the LDC for judging. For topics that required passage-level judgment, the annotator marked passages as relevant as soon as a relevant document was found.

10 Participation

The following 16 sites participated in the HARD track of TREC 2004. The first three columns indicate whether the site used metadata values, clarification forms, or passage retrieval in any of their submitted runs.

Meta	CF	Psgs	Site
Y	Y	Y	Chinese Academy of Science, Institute of Software [Sun et al., 2005]
N	Y	N	Clairvoyance Corporation [Evans et al., 2005]
N	Y	Y	Indiana University [Yang et al., 2005]
N	Y	N	Microsoft Research Cambridge [Zaragoza et al., 2005]
Y	N	N	The Robert Gordon University [Harper et al., 2005]
Y	N	N	Rutgers University [Belkin et al., 2005]
?	?	?	Tsinghua University
Y	N	Y	University of Chicago [Levow, 2005]
?	?	?	University of Cincinnati
N	Y	Y	University of Illinois at Urbana-Champaign [Jiang and Zhai, 2005]
N	Y	Y	University of Maryland & Johns Hopkins University [He et al., 2005]
Y	Y	Y	University of Massachusetts Amherst [Abdul-Jaleel et al., 2005]
N	Y	N	University of North Carolina at Chapel Hill [Kelly et al., 2005]
Y	N	N	University of Twente[Rode and Hiemstra, 2005]
N	Y	Y	University of Waterloo & Bilkent University [Vechtomova and Karamuftuoglu, 2005]
Y	Y	Y	York University [Huang et al., 2005]
7	10	8	COUNTS

(No information was reported for Tsinghua University or the University of Cincinnati, and they did not provide a paper on this track to TREC for publication.)

It is interesting to note the wide range of ways that the different purposes of the track were exploited. Only three sites used all three possible components of the track. The clarification forms were the most popular, but not by a wide margin.

11 Results

This section provides a sketch of some of the results found by participating sites. Further and more detailed information is available in the sites individual papers.

11.1 Use of metadata

For the most part, sites built models for the geography, genre, and subject metadata categories. They typically used text classification techniques to decide whether a document matched the category. Some sites used the Web to collect more data relevant to the category. And some built manual term lists for classification (mostly for geography information).

In general, sites were unable to demonstrate substantial gains in effectiveness using metadata. Since metadata differentiated between relevant and merely on-topic documents, a run using metadata should score much better on "hard" measures (where only relevant documents are counted as relevant) and "soft" measures (where on-topic documents are also counted as relevant). Several runs were able to improve in that direction, though not by huge margins.

Some of these results are because topics tended not to require the metadata to improve performance. For example, *AIDS in Africa* is obviously a non-US topic, and being told that it is not US is of little value.

The University of North Carolina asked (in clarification forms) the user how many times they had searched before for each topic. They then showed that users who had claimed low familiarity in metadata also had not previously searched often for this topic. They did not use the metadata to aid retrieval, but cleverly used the clarification form to show how familiarity metadata could be collected [Kelly et al., 2005].

The University of Waterloo also did not use metadata for retrieval, but did a very nice analysis using the familiarity metadata. Users with low familiarity selected fewer phrases in Waterloo's clarification forms. User's with low familiarity were helped by the clarification forms but users with much familiarity were hurt [Vechtomova and Karamuftuoglu, 2005].

11.2 Use of clarification forms

Clarification forms allowed sites to ask the user anything about the topic that could be expressed in simple HTML. Most requested information asked for judgments on keywords, documents, or passages. One site asked whether presented passages were of about the right length, presumably to get a handle on the right amount of information that should be returned. Several sites included free-form entry of phrases, other keywords, or related text at the end of their clarification forms.

When sites asked for keywords, they had usually found words or phrases that their system suspected were related to the topic. These might be words or phrases appearing in top-ranked documents, synonyms of query words found using Wordnet (for example), extracted noun phrases or named entities, or ranges of time that where relevant material would appear.

Document-style requests generally asked for a judgment of relevant for the passage. That was often the title and a few keywords from a document, the passage most likely to be relevant ("best passage"), or a cluster of documents represented by titles and/or key words. The set of documents, passages, or clusters chosen for presentation were either the top-ranked set or a set modified to incorporate some notion of novelty—i.e., do not present two highly similar documents for judgment.

Clarification forms were very popular, very fun, provided an open ended framework for experimentation, and were by those counts very successful. On the other hand, most sites limited themselves to keyword and text relevance feedback rather than trying more novel techniques, so the "open ended" nature has not (yet) encouraged new ideas.

The value of clarification forms remains elusive to determine. Many sites saw some gains from their clarification forms, but there were several sites that achieved their best performance—or nearly their best—on the baseline runs. Unquestionably work should consider on clarification forms because they are popular, though until more impressive gains are seen, their value will debatable.

11.3 Use of passages

As described in Section 8, two measures for passage retrieval were considered, but others were compared. Two get a sense of how similar they were, we investigated the correlation between bpref at 12,000 characters. (That measure was declared "primary" in the track guidelines, but sufficiently late in the process that some sites fit to the passage R-precision measure.)

- Precision at 12,000 characters measured the proportion of characters that were relevant in the top 12,000 characters. It showed a 99% correlation.
- Character R-precision (similar to passage R-precision, but a character-oriented evaluation where R is the total number of relevant *characters* not passages). It showed an 88% correlation.
- Passage F1 at 30 passages retrieved showed a 90% correlation.
- Passage precision at 10 passages showed an 80% correlation.
- Passage R-precision (last year's official measure) showed a 45% correlation.

If nothing else, these results should suggest that sites training their systems to optimize passage R-precision should not be expected to do well on the character bpref measure.

Passage retrieval systems often use fixed-length passages of some number of words or characters, treating those passages as if they were documents. Some sites tried to generate appropriately sized passages using HMMs, retrieving and then merging highly ranked adjacent sentences, or looking for runs of text where the query terms are highly dense. Most sites scored passages and then combined the passage score with the document score in one way or another.

There was substantially more activity in passage retrieval for HARD 2004 than in 2003. However, the issue of how best to resolve variable-length passage retrieval with variable-length passage "truth" judgments remains open and begs for substantially more exploration. There are clear problems with the passage R-precision measure, but the character bpref is also not without issues. Unfortunately, the HARD 2005 track will be dropping passage retrieval because of funding issues.

11.4 Overall results

When measured by topicality (i.e., when on-topic and/or relevant documents are the target), the top runs were all automatic and used both the title and description. Some top runs used clarification forms, passage retrieval, and the (hard) related text information. A few top runs used the geography and genre metadata fields and a couple used the topic narrative and (soft) related text.

When measured by relevance (i.e., only relevant documents were the target), the top runs used similar information, though all top runs used the (hard) related text.

For passage retrieval evaluation, the best runs were usually automatic (though the second ranked run was manual), used the title and scription, incorporated a clarification form, and did passage retrieval. Interestingly, the fifth ranked run was a document run with no passages marked. Some sites were able to find advantage to the geography and genre metadata, and some used related text and narrative. Note that related text (of both kinds) was more often used in top performing document retrieval systems than in top performing passage retrieval systems.

No top run by any of the measures used the familiarity field.

12 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.

The HARD track will continue in TREC 2005. Funding considerations have forced the removal of passage retrieval from the evaluation. Topics deemed by the Robust track to be difficult will be used rather than developing new topics, though they will be judged against a new corpus. Familiarity metadata will be collected, but not used in any particular way by the annotators.

Acknowledgments

The coordination of the HARD track at the CIIR would not have been possible without the help of Fernando Diaz, Mark Smucker, and Courtney Wade. The way the track was organized was by consensus of participating researchers and they all deserve credit for the shape the track eventually took.

The work 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

[Abdul-Jaleel et al., 2005] Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D., and Wade, C. (2005). UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC 2004*. Appears in this volume.

[Belkin et al., 2005] Belkin, N., Chaleva, I., Cole, M., Li, Y.-L., Liu, L., Liu, Y.-H., Muresan, G., Smith, C., Sun, Y., Yuan, X.-J., and Zhang, X.-M. (2005). Rutgers' HARD track experiences at TREC 2004. In *Proceedings of TREC 2004*. Appears in this volume.

[Buckley and Voorhees, 2004] Buckley, C. and Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of SIGIR*, pages 25–32.

[Evans et al., 2005] Evans, D. A., Bennett, J., Montgomery, J., Sheftel, V., Hull, D. A., and Shanahan, J. G. (2005). TREC-2004 HARD-track experiments in clustering. In *Proceedings of TREC 2004*. Appears in this volume.

[Harper et al., 2005] Harper, D. J., Muresan, G., Liu, B., Koychev, I., Wettschereck, D., and Wiratanga, N. (2005). The Robert Gordon University's HARD track experiments at TREC 2004. In *Proceedings of TREC 2004*. Appears in this volume.

[He et al., 2005] He, D., Demner-Fushman, D., Oard, D. W., Karakos, D., and Dhudanpur, S. (2005). Improving passage retrieval using interactive elicitation and statistical modeling. In *Proceedings of TREC 2004*. Appears in this volume.

[Huang et al., 2005] Huang, X., Huang, Y. R., Wen, M., and Zhong, M. (2005). York University at TREC 2004: HARD and genomics tracks. In *Proceedings of TREC 2004*. Appears in this volume.

[Jiang and Zhai, 2005] Jiang, J. and Zhai, C. (2005). UIUC in HARD 2004 – passage retrieval using HMMs. In *Proceedings of TREC 2004*. Appears in this volume.

[Kelly et al., 2005] Kelly, D., Dollu, V. D., and Fu, X. (2005). University of North Carolina's HARD track experiments at TREC 2004. In *Proceedings of TREC 2004*. Appears in this volume.

[Levow, 2005] Levow, G.-A. (2005). University of Chicago at TREC 2004: HARD track. In *Proceedings of TREC 2004*. Appears in this volume.

[Rode and Hiemstra, 2005] Rode, H. and Hiemstra, D. (2005). Conceptual language models for context-aware text retrieval. In *Proceedings of TREC 2004*. Appears in this volume.

[Sun et al., 2005] Sun, L., Zhang, J., and Sun, Y. (2005). ISCAS at TREC-2004: HARD track. In *Proceedings of TREC 2004*. Appears in this volume.

[Vechtomova and Karamuftuoglu, 2005] Vechtomova, O. and Karamuftuoglu, M. (2005). Approaches to high accuracy retrieval: Phrase-based search experiments in the HARD track. In *Proceedings of TREC 2004*. Appears in this volume.

[Voorhees, 2005] Voorhees, E. M. (2005). Overview of the TREC 2004 question answering track. In *Proceedings of TREC 2004*. Appears in this volume.

[Yang et al., 2005] Yang, K., Yu, N., Wead, A., La Rowe, G., Li, Y.-H., Friend, C., and Lee, Y. (2005). WIDIT in TREC-2004 genomics, HARD, robust and web tracks. In *Proceedings of TREC 2004*. Appears in this volume.

[Zaragoza et al., 2005] Zaragoza, H., Craswell, N., Taylor, M., Saria, S., and Robertson, S. (2005). Microsoft Cambridge at TREC-13: Web and HARD tracks. In *Proceedings of TREC 2004*. Appears in this volume.