HARD Track Overview in TREC 2005
High Accuracy Retrieval from Documents

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

1 Introduction

TREC 2005 saw the third year of the High Accuracy Retrieval from Documents (HARD) track. The HARD track explores methods for improving the accuracy of document retrieval systems, with particular attention paid to the start of the ranked list. Although it has done so in a few different ways in the past, budget realities limited the track to "clarification forms" this year. The question investigated was whether 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—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.

The following summarizes the changes from the HARD track in TREC 2004 [Allan, 2005]:

- There was no passage retrieval evaluation as part of the track this year.
- There was no use of metadata this year.
- The evaluation corpus was the full AQUAINT collection. In HARD 2003 the track used part of AQUAINT plus additional documents. In HARD 2004 it was a collection of news from 2003 collated especially for HARD.
- The topics were selected from existing TREC topics. The same topics were used by the Robust track [Voorhees, 2006]. The topics had not been judged against the AQUAINT collection, though had been judged against a different collection.
- There was no notion of "hard relevance" and "soft relevance", though documents were judged on a trinary scale of not relevant, relevant, or highly relevant.
- Clarification forms were allowed to be much more complex this year.
- Corpus and topic development, clarification form processing, and relevance assessments took place at NIST rather than at the Linguistic Data Consortium (LDC).
- The official evaluation measure of the track was R-precision.

The HARD track's Web page may also contain useful pointers, though is not guaranteed to be in place indefinitely. As of early 2006, it was available at http://ciir.cs.umass.edu/research/hard.

For TREC 2006, the HARD track is being "rolled into" the Question Answering track. The new aspect of the QA track is called "ciQA" for "complex, interactive Question Answering." The goal of ciQA is to investigate interactive approaches to cope with complex information needs specified by a templated query.

2 The Process

The HARD track proceeded as follows. This process follows roughly that of past years' tracks, though it simpler because passage retrieval was not an issue.

At the end of May, the track guidelines were finalized. Sites knew then that the evaluation corpus would be the AQUAINT collection (see Section 4), so could begin indexing the data and/or training their systems (see Section 7).

On June 15, 2005, participating sites received the set of 50 test topics from NIST (see Section 5).

Three weeks later, on July 7, sites had to submit the "baseline" ranked lists produced by their system (see Section 8). These runs ideally represented the best that the sites could do with only "classic" TREC topic information.

On the same day, sites were permitted to submit sets of clarification forms, where each set contained a form for each topic in the test set. The clarification form could contain almost anything that the site felt an answer would be useful for improving the accuracy of the query (e.g., possibly relevant passages, keywords that might reflect relevance). See Section 9 for more details.

For the next two weeks, assessors at NIST filled out clarification forms for the topics. On July 25, the clarification form responses were shipped to the sites.

On August 8, the sites submitted new "final" ranked lists that utilized information from the clarification forms (see Section 10).

Between then and early September, the assessors judged documents for relevance (see Section 6). Relevance assessments ("qrels") were made available to the researchers on September 9, 2005.

3 Participation

A total of 16 sites submitted 122 runs for the track. The following breakdown shows how many runs each site submitted, broken down by baseline and final runs, as well as the number of clarification forms submitted.

# runs			
Base	Final	# CFs	Participating site
0	10	2	Chinese Academy of Sciences
1	8	2	Chinese Academy of Sciences NLPR
4	6	2	Indiana University
2	7	2	Meiji University
1	11	2	Rutgers University
2	6	2	SAIC/U. of Virginia
1	1	1	University College Dublin
1	6	3	University of Illinois at Urbana-Champaign
3	3	1	University of Maryland, College Park
4	4	2	University of Massachusetts
1	3	3	University of North Carolina
2	4	2	University of Pittsburgh
1	7	2	University of Strathclyde
2	6	2	University of Twente
2	4	3	University of Waterloo
3	5	3	York University

4 HARD Corpus

For TREC 2005, the HARD track used the AQUAINT corpus. That corpus is available from the Linguistic Data Consortium for a modest fee, and was made available to HARD participants who were not a member of the LDC for no charge. The LDC's description of the corpus[1] is:

> The AQUAINT Corpus, Linguistic Data Consortium (LDC) catalog number LDC2002T31 and isbn1-58563-240-6 consists of newswire text data in English, drawn from three sources: the Xinhua News Service (People's Republic of China), the New York Times News Service, and the Associated Press Worldstream News Service. It was prepared by the LDC for the AQUAINT Project, and will be used in official benchmark evaluations conducted by National Institute of Standards and Technology (NIST).

The corpus is roughly 3Gb of text and includes 1,033,461 documents (about 375 million words of text, according to the LDC's web page). All documents in the collection were used for the HARD evaluation.

5 Topics

Topics were selected from among existing TREC topics that almost no system was able to handle well in previous years. Because those old topics were to be judged on a new corpus (AQUAINT), they were manually vetted to ensure that at least three relevant documents existed in the AQUAINT corpus. These topics were also used by the TREC 2005 Robust track [Voorhees, 2006].

The topic numbers used were: 303, 307, 310, 314, 322, 325, 330, 336, 341, 344, 345, 347, 353, 354, 362, 363, 367, 372, 374, 375, 378, 383, 389, 393, 394, 397, 399, 401, 404, 408, 409, 416, 419, 426, 427, 433, 435, 436, 439, 443, 448, 622, 625, 638, 639, 648, 650, 651, 658, and 689.

6 Relevance judgments

Topics were judged for relevance by the same assessor who answered the clarification forms for the topic (see Section 9 for more information on clarification forms). In the first two years of HARD, that same person also created the original topic statement; however, because topics were re-used, it was not possible to use the same person for the original step. No attempt was made to ensure that this year's assessor's notion of relevance would match that of the original assessor.

Six assessors worked on the fifty topics, as follows:

Assessor A: 347 399 401 404 408 409 419 426
Assessor B: 625 638 639 648 650 651 658 689
Assessor C: 427 433 435 436 439 443 448 622
Assessor D: 303 322 345 354 362 363 367 383 393
Assessor E: 336 341 353 372 375 378 394 397
Assessor F: 307 310 314 325 330 344 374 389 416

Documents were judged as either not relevant, relevant, or highly relevant. For purposes of this track, judgments of *relevant* and *highly relevant* were treated as the same.

[1] At http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T31 as of May 2006.

7 Training data

The data collections from the HARD tracks of TREC 2003 [Allan, 2004] and 2004 [Allan, 2005] were available for training. All of that data was made available to HARD track participants courtesy of the Linguistic Data Consortium. The data was provided for use only in the HARD 2005 evaluation with the expectation that it will be destroyed at the completion of the track (i.e., after the final papers are written). The HARD 2004 corpus and topics are now available for purchase from the LDC as catalogue numbers LDC2005T28 and LDC2005T29[2].

The TREC 2004 HARD track used a corpus of news from 2003, had 49 topics with several metadata fields. Topics, relevance judgments, and clarification forms were provided.

The TREC 2003 HARD track corpus was a set of 372,219 documents totaling 1.7Gb from the 1999 portion of the AQUAINT corpus, along with some US government documents from the same year (Congressional Record and Federal Register). The topics were somewhat like standard TREC topics, but included lots of searcher and query metadata. Topics, relevance judgments, and clarification forms were provided.

8 Baseline submissions

Submissions of baseline runs were in the standard TREC submission format used for ad-hoc queries. Up to 1000 documents were provided in rank order for each of the 50 topics. The details were in a file with lines containing a topic number, a document ID, the document's rank against that topic, and its score (along with some other bits of bookkeeping information). Every topic was required to have at least one document retrieved, and it could have anywhere from one to 1,000 documents.

Sites were asked to provide the following information:

1. *Was this an entirely automatic run or a manual run?* Two baseline runs were manual, all others were automatic.
2. *Did you use the title, description, and/or narrative fields for this run?* The runs included 9 using just the title field, 3 using just description, 8 combining title and description, and 10 also adding in the narrative.
3. *To what extent did you use earlier relevance judgments on the topics?* One run claimed to have used the judgments of these topics against prior TREC corpora.
4. *A short description of the run.*
5. *Preference in terms of judging of this run?* Only one baseline run per site was included in the judging pool.

9 Clarification forms

All 16 participating sites submitted at least one clarification forms: two submitted one form, ten submitted two forms, and four sites submitted three. All submitted forms were filled out, even though the track guidelines only guaranteed that two would be.

Clarification forms were filled out by the NIST assessors using the following platform:

[2] Described at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T28 and . . . LDC2005T29, respectively, as of May 2006.

- Redhat Enterprise Linux, version “3 workstation”
- 20-inch LCD monitor with 1600x1200 resolution, true color (millions of colors)
- Firefox Web browser, v1.0.3
- No assumption that the machine is connected to any network at all. (The goal was to have it disconnected from all networks of any sort, but that proved infeasible in the NIST environment.)

In past years, the contents of the clarification forms were strictly controlled to allow only a limited subset of HTML. This year, virtually all restrictions were lifted, meaning that sites could include Javacript, Java, images, or the like. The following restrictions were made:

- The forms had to assume they were running on a computer that is disconnected from all networks, so all necessary information had to be included as part of the form. If it required multiple files, they all had to be within the same directory structure. Sites could not assume that all of its clarification forms would be on the same computer.
- It was not possible to invoke any cgi-bin scripts
- It was not possible to write to disk

Clarification forms could be presented in almost any layout, but had to include the following items:

- **<form action=“/cgi-bin/clarification_submit.pl” method=“post”>**
  This indicates the script where the output was generated (all it did was output the selected information).
- **<input type=“hidden” name=“site” value=“XXXXn”>**
  Here, “XXXX” is a 4-letter code designating the site (provided in the lead-up to the baseline submission) and “n” is a run number. The run numbers reflected the priority order of the form. That is, XXXX1 will be processed then XXXX2 and so on.
- **<input type=“hidden” name=“topicid” value=“000”>**
  Indicates the topic number, a 3-digit code with zeros padding as needed (001 rather than 01 or 1).
- **<input type=“submit” name=“send” value=“submit”>**
  This is the submit button that had to appear somewhere on the page.

In addition, sites were strongly encouraged to include somewhere on the page the topic number (e.g., “303”) and the title of the topic to provide a sanity check that the annotators were, indeed, answering the correct questions.

For each submission, all clarification forms were put in a single directory (folder) with the name indicated (e.g., NIST1). Each clarification form inside that directory was also a directory with the name of the submission and the topic number (e.g., NIST1_043 for topic 43 of the NIST1 submission).

Inside *that* directory, the main clarification form was called index.html. It could access any files from within the directory hierarchy, using relative pathnames. For example, “logo.gif” would refer to the file NIST1/NIST1_043/logo.gif within the directory structure, and “../logo.gif” would refer to NIST1/logo.gif”.

Sites were asked the following information about each submitted form:

1. Did you use clustering to generate this form?
2. Did you use text summarization, either extractive or generative?

3. Did you use document-level feedback? That is, did you ask the user to judge an entire document for relevance, even if you did so using a title, passage, or keywords from the document?

4. Did you ask the user to judge selected passages of text, independent of the documents they came from?

5. Did you ask the user to judge keywords for relevance, independent of the documents they came from?

6. If you used any techniques not listed above, briefly list them at the bullet-list level of detail.

7. Did you use any sources of information beyond the query and AQUAINT corpus and, if so, what were they?

The assessors spent no more than three minutes per form no matter how complex the form was. The three minutes included time needed to load the form, initialize it, and do any rendering, so unusually complex or large forms were implicitly penalized. At the end of three minutes, if the assessor had not pressed the "submit" button, the form was timed out and forcibly submitted (anything entered up to that point was saved).

NIST recorded the time spent on the form returned for each form. That information was returned in a separate file along with all of the clarification form responses. Assessors were never permitted more than 180 seconds per form, but some of the reported times were greater than 180 because of the time it took for the system to "shut down" a form if the time limit expired.

Clarification forms were presented to annotators in an order to minimize the chance that one form would adversely (or positively) impact the use of another form. Tables 1 and 2 shows the rotation that was used for the submitted clarification forms.

10 Final submissions

Final submissions incorporated information gleaned from clarification forms and combined that with any other retrieval techniques to achieve the best run possible.

The following questions were asked for each submission:

1. *Which of your baseline runs is an appropriate baseline?* There were 26 submissions that indicated that the final run did not have a corresponding baseline run. This often reflected a site's providing a new "baseline" or trying out a technique that was developed after the baseline runs and so had no corresponding baseline.

2. *Which of your clarification forms was used to generated this final run?* There were 33 final runs that indicated they did *not* use a clarification form.

3. *Other than the clarification form's being answered, was this an entirely automatic run or a manual run?* Only four of the final runs were marked as being manual runs; the remaining 88 were automatic.

4. *Did you use the title, description, and/or narrative fields for this run?* Here, 28 runs used just the title, 2 used just the description, 39 combined the title and description, and 23 also included the narrative.

5. *To what extent did you use earlier relevance judgments on the topics?* A total of 13 runs indicated that they used the earlier relevance judgments.

6. *A short description of the run.*

7. *What is the preference in terms of judging of this run?* Only one final run from each site was included in the judging pool.

	NCAR1	MARY1	INDI2	STRA2	UIUC3	UIUC1	NCAR3	TWEN2	PITT1	YORK2	CASP1	CASS2	NCAR2	PITT2	MASS1	SAIC1	YORK1
T1	28	30	23	5	19	3	20	12	15	16	2	17	32	22	29	7	21
T2	29	31	24	6	20	4	21	13	16	17	3	18	33	23	30	8	22
T3	30	32	25	7	21	5	22	14	17	18	4	19	34	24	31	9	23
T4	31	33	26	8	22	6	23	15	18	19	5	20	1	25	32	10	24
T5	32	34	27	9	23	7	24	16	19	20	6	21	2	26	33	11	25
T6	33	1	28	10	24	8	25	17	20	21	7	22	3	27	34	12	26
T7	34	2	29	11	25	9	26	18	21	22	8	23	4	28	1	13	27
T8	1	3	30	12	26	10	27	19	22	23	9	24	5	29	2	14	28
T9	2	4	31	13	27	11	28	20	23	24	10	25	6	30	3	15	29
T10	3	5	32	14	28	12	29	21	24	25	11	26	7	31	4	16	30
T11	4	6	33	15	29	13	30	22	25	26	12	27	8	32	5	17	31
T12	5	7	34	16	30	14	31	23	26	27	13	28	9	33	6	18	32
T13	6	8	1	17	31	15	32	24	27	28	14	29	10	34	7	19	33
T14	7	9	2	18	32	16	33	25	28	29	15	30	11	1	8	20	34
T15	8	10	3	19	33	17	34	26	29	30	16	31	12	2	9	21	1
T16	9	11	4	20	34	18	1	27	30	31	17	32	13	3	10	22	2
T17	10	12	5	21	1	19	2	28	31	32	18	33	14	4	11	23	3
T18	11	13	6	22	2	20	3	29	32	33	19	34	15	5	12	24	4
T19	12	14	7	23	3	21	4	30	33	34	20	1	16	6	13	25	5
T20	13	15	8	24	4	22	5	31	34	1	21	2	17	7	14	26	6
T21	14	16	9	25	5	23	6	32	1	2	22	3	18	8	15	27	7
T22	15	17	10	26	6	24	7	33	2	3	23	4	19	9	16	28	8
T23	16	18	11	27	7	25	8	34	3	4	24	5	20	10	17	29	9
T24	17	19	12	28	8	26	9	1	4	5	25	6	21	11	18	30	10
T25	18	20	13	29	9	27	10	2	5	6	26	7	22	12	19	31	11
T26	19	21	14	30	10	28	11	3	6	7	27	8	23	13	20	32	12
T27	20	22	15	31	11	29	12	4	7	8	28	9	24	14	21	33	13
T28	21	23	16	32	12	30	13	5	8	9	29	10	25	15	22	34	14
T29	22	24	17	33	13	31	14	6	9	10	30	11	26	16	23	1	15
T30	23	25	18	34	14	32	15	7	10	11	31	12	27	17	24	2	16
T31	24	26	19	1	15	33	16	8	11	12	32	13	28	18	25	3	17
T32	25	27	20	2	16	34	17	9	12	13	33	14	29	19	26	4	18
T33	26	28	21	3	17	1	18	10	13	14	34	15	30	20	27	5	19
T34	27	29	22	4	18	2	19	11	14	15	1	16	31	21	28	6	20
T35	28	30	23	5	19	3	20	12	15	16	2	17	32	22	29	7	21
T36	29	31	24	6	20	4	21	13	16	17	3	18	33	23	30	8	22
T37	30	32	25	7	21	5	22	14	17	18	4	19	34	24	31	9	23
T38	31	33	26	8	22	6	23	15	18	19	5	20	1	25	32	10	24
T39	32	34	27	9	23	7	24	16	19	20	6	21	2	26	33	11	25
T40	33	1	28	10	24	8	25	17	20	21	7	22	3	27	34	12	26
T41	34	2	29	11	25	9	26	18	21	22	8	23	4	28	1	13	27
T42	1	3	30	12	26	10	27	19	22	23	9	24	5	29	2	14	28
T43	2	4	31	13	27	11	28	20	23	24	10	25	6	30	3	15	29
T44	3	5	32	14	28	12	29	21	24	25	11	26	7	31	4	16	30
T45	4	6	33	15	29	13	30	22	25	26	12	27	8	32	5	17	31
T46	5	7	34	16	30	14	31	23	26	27	13	28	9	33	6	18	32
T47	6	8	1	17	31	15	32	24	27	28	14	29	10	34	7	19	33
T48	7	9	2	18	32	16	33	25	28	29	15	30	11	1	8	20	34
T49	8	10	3	19	33	17	34	26	29	30	16	31	12	2	9	21	1
T50	9	11	4	20	34	18	1	27	30	31	17	32	13	3	10	22	2

Table 1: Rotation used to fill out clarification forms (the right edge of the table continues in Table 2). The rows of the table correspond to topics and the columns to clarification forms from sites. For example, the form indicates that NCAR's primary clarification form (NCAR1) will be the 28th considered for topic 1, the 29th for topic 2, ..., the 1st for topic 8, and so on. Similarly, for topic 1, the assessor first did INDI1's form (see Table 1), then that for CASP1, then UIUC1's, followed by MEIJ1's, and so on.

11 Overview of submissions

As mentioned above, 16 sites participated. The following statistics provide some details of the submissions. Note that the information is largely self-reported and has not been rigorously verified, so it is possible that it may be somewhat inaccurate.

A total of 30 baseline runs were submitted from 15 sites. One of those 15 sites made use of the earlier judgments for the topics (on a different corpus and using a different assessor).

A total of 35 sets of clarification forms were submitted. The average time per form on a single question was 116.5 seconds, with a minimum of five seconds and a maximum of 180 seconds. (In fact, one query's form reported taking 676 seconds, but the more than 8 additional minutes were presumably consumed by the system trying to force the form to close after the three minutes had expired.)

Every site had at least one form that took the full three minutes, and many had a dozen or two that took that long. The University of Massachusetts had the distinction of being the only site that used the full

	CASS1	DUBL1	UWAT1	MASS2	CASP2	STRA3	UWAT2	MEIJ1	MEIJ2	RUTG2	YORK3	RUTG1	SAIC2	INDI1	TWEN1	UIUC2	UWAT3
T1	26	11	25	13	6	27	14	4	33	18	31	24	8	1	9	10	34
T2	27	12	26	14	7	28	15	5	34	19	32	25	9	2	10	11	1
T3	28	13	27	15	8	29	16	6	1	20	33	26	10	3	11	12	2
T4	29	14	28	16	9	30	17	7	2	21	34	27	11	4	12	13	3
T5	30	15	29	17	10	31	18	8	3	22	1	28	12	5	13	14	4
T6	31	16	30	18	11	32	19	9	4	23	2	29	13	6	14	15	5
T7	32	17	31	19	12	33	20	10	5	24	3	30	14	7	15	16	6
T8	33	18	32	20	13	34	21	11	6	25	4	31	15	8	16	17	7
T9	34	19	33	21	14	1	22	12	7	26	5	32	16	9	17	18	8
T10	1	20	34	22	15	2	23	13	8	27	6	33	17	10	18	19	9
T11	2	21	1	23	16	3	24	14	9	28	7	34	18	11	19	20	10
T12	3	22	2	24	17	4	25	15	10	29	8	1	19	12	20	21	11
T13	4	23	3	25	18	5	26	16	11	30	9	2	20	13	21	22	12
T14	5	24	4	26	19	6	27	17	12	31	10	3	21	14	22	23	13
T15	6	25	5	27	20	7	28	18	13	32	11	4	22	15	23	24	14
T16	7	26	6	28	21	8	29	19	14	33	12	5	23	16	24	25	15
T17	8	27	7	29	22	9	30	20	15	34	13	6	24	17	25	26	16
T18	9	28	8	30	23	10	31	21	16	1	14	7	25	18	26	27	17
T19	10	29	9	31	24	11	32	22	17	2	15	8	26	19	27	28	18
T20	11	30	10	32	25	12	33	23	18	3	16	9	27	20	28	29	19
T21	12	31	11	33	26	13	34	24	19	4	17	10	28	21	29	30	20
T22	13	32	12	34	27	14	1	25	20	5	18	11	29	22	30	31	21
T23	14	33	13	1	28	15	2	26	21	6	19	12	30	23	31	32	22
T24	15	34	14	2	29	16	3	27	22	7	20	13	31	24	32	33	23
T25	16	1	15	3	30	17	4	28	23	8	21	14	32	25	33	34	24
T26	17	2	16	4	31	18	5	29	24	9	22	15	33	26	34	1	25
T27	18	3	17	5	32	19	6	30	25	10	23	16	34	27	1	2	26
T28	19	4	18	6	33	20	7	31	26	11	24	17	1	28	2	3	27
T29	20	5	19	7	34	21	8	32	27	12	25	18	2	29	3	4	28
T30	21	6	20	8	1	22	9	33	28	13	26	19	3	30	4	5	29
T31	22	7	21	9	2	23	10	34	29	14	27	20	4	31	5	6	30
T32	23	8	22	10	3	24	11	1	30	15	28	21	5	32	6	7	31
T33	24	9	23	11	4	25	12	2	31	16	29	22	6	33	7	8	32
T34	25	10	24	12	5	26	13	3	32	17	30	23	7	34	8	9	33
T35	26	11	25	13	6	27	14	4	33	18	31	24	8	1	9	10	34
T36	27	12	26	14	7	28	15	5	34	19	32	25	9	2	10	11	1
T37	28	13	27	15	8	29	16	6	1	20	33	26	10	3	11	12	2
T38	29	14	28	16	9	30	17	7	2	21	34	27	11	4	12	13	3
T39	30	15	29	17	10	31	18	8	3	22	1	28	12	5	13	14	4
T40	31	16	30	18	11	32	19	9	4	23	2	29	13	6	14	15	5
T41	32	17	31	19	12	33	20	10	5	24	3	30	14	7	15	16	6
T42	33	18	32	20	13	34	21	11	6	25	4	31	15	8	16	17	7
T43	34	19	33	21	14	1	22	12	7	26	5	32	16	9	17	18	8
T44	1	20	34	22	15	2	23	13	8	27	6	33	17	10	18	19	9
T45	2	21	1	23	16	3	24	14	9	28	7	34	18	11	19	20	10
T46	3	22	2	24	17	4	25	15	10	29	8	1	19	12	20	21	11
T47	4	23	3	25	18	5	26	16	11	30	9	2	20	13	21	22	12
T48	5	24	4	26	19	6	27	17	12	31	10	3	21	14	22	23	13
T49	6	25	5	27	20	7	28	18	13	32	11	4	22	15	23	24	14
T50	7	26	6	28	21	8	29	19	14	33	12	5	23	16	24	25	15

Table 2: Continuation of Table 1; this table appears to the right of that table.

three minutes of annotator time for *every* form. Those forms were apparently designed to collect as much information as possible during clarification time for later processing to determine which questions were most useful [Diaz and Allan, 2006].

A total of 92 final runs were submitted across the 16 sites. Of those, three runs made use of the past judgments. Different sites used different parts of the topics for their runs:

- 28 runs were title-only queries
- 2 runs were description-only queries
- 38 runs combined the title and description
- 24 runs included the narrative along with the title and description

All runs were automatic (not counting clarification form interaction) except for those submitted by the University of Maryland, where experiments used a trained intermediary to collect potentially useful information for a clarification form [Lin et al., 2006].

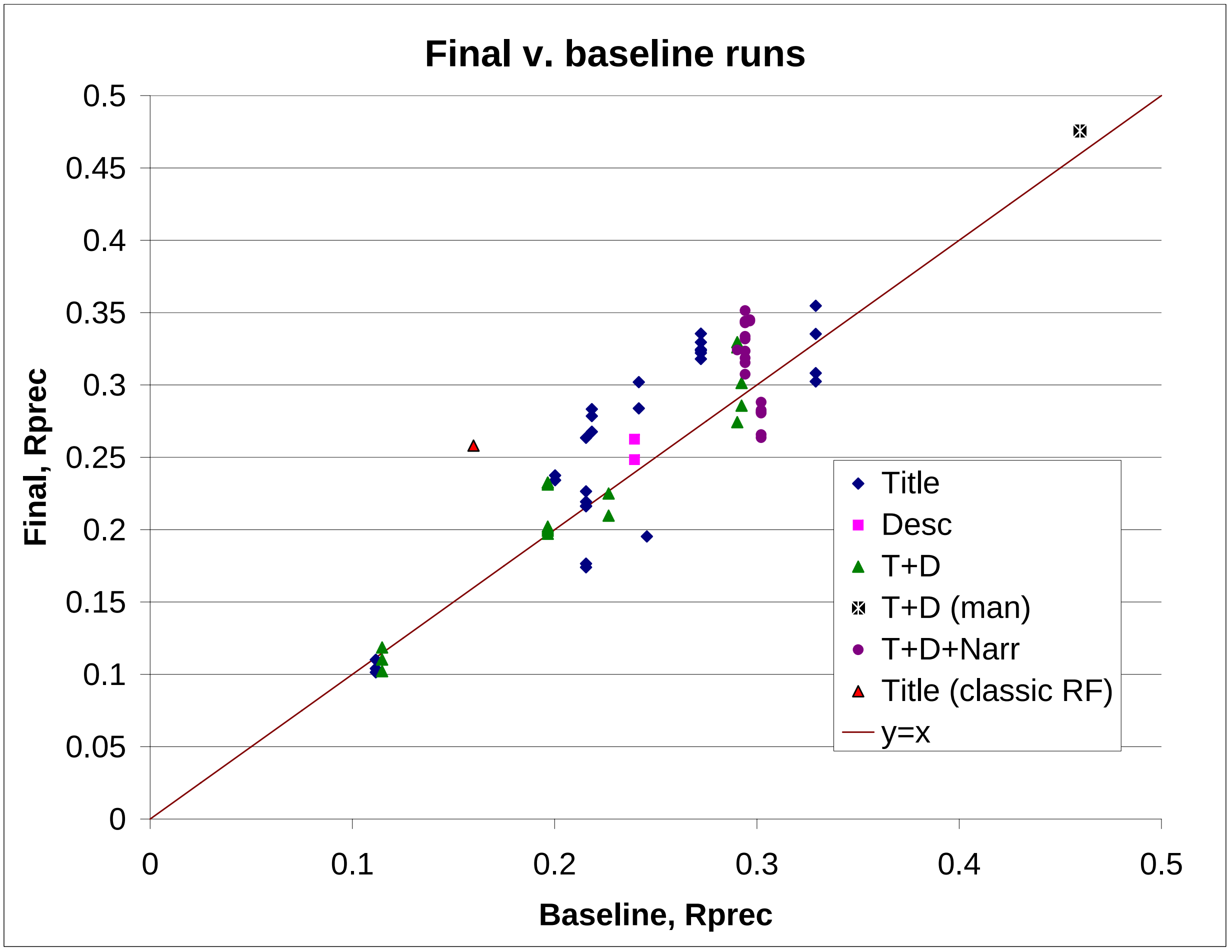

Figure 1: Comparison of R-precision values in baseline runs and runs after using a clarification form (only runs that identified a corresponding baseline run are included).

12 Discussion

System output was evaluated by R-precision, defined as precision at R documents retrieved, where R is the number of known relevant documents in the collection.

Figure 1 shows overall performance as impacted by clarification forms. Recall that when a final run was submitted, sites were asked to indicate which of their baseline runs was used as a starting point. The graph includes a point for each such baseline-final pair. Because (by chance) different baseline runs never had the same score, points that make up vertical lines represent multiple final runs that used the same baseline run. For example, the run at baseline R-precision of 0.3291 was used for four final runs that had R-precision ranging from 0.3024 to 0.3547.

Point colors and shape reflect which portions of the topic were used for the query, though the differences may not be easily visible in a grayscale print. The (excellent) outlier labeled "T+D (man)" in the upper right is the manual run from the University of Maryland. The red triangle at baseline 0.1599 and final 0.2581, labeled "Title (classic RF)", is a special run created by NIST, and is discussed further below.

12.1 General observations

Just considering baseline runs, the automatic runs had R-precision scores ranging from 0.1116 to 0.3291. Using the title, description, and narrative seemed to be helpful, since four different sites achieved comparable scores. However, some baselines without the narrative performed just as well, and the best automatic baseline used only the title.

Ultimately, the goal of the HARD track was to explore the value added by clarification forms. That means that it is the improvement from baseline to final that is more interesting. In the graph, points below the $y = x$ line had final runs that were worse than their corresponding baseline runs; those above the line improved. Most of the sites were able to improve on their baseline performance.

12.2 Classic relevance feedback

In previous years of the HARD track, there was a concern that simple relevance feedback of documents might be a simpler and more effective type of clarification form. To explore that issue this year, NIST volunteered to provide a form that was purely relevance feedback. To do that, NIST ran a baseline system and then created a clarification form that included the top-ranked documents, asking that they be judged as relevant or not. The baseline system was Prise3, a system based on the Lucene open source IR engine, so it used a tf-idf style of retrieval. Prise3 was the same system used to create new topics for other tracks this year. The title field to retrieve the top 50 documents.

The clarification form listed, along with the query's title and description, the title of the top 50 documents. The assessor could click on a link to see the full text of a document if needed. The assessor used his or her three minutes to judge as many documents as possible, and then a new query was created using that information. Because Prise3 did not support relevance feedback at that time, the final run version 11.0 of the well-known SMART system. The system was tuned using the Robust 2004 topics on the past corpus, without paying any special attention ot the topics that were re-used for HARD this year. The tuning used the top-ranked five relevant documents on that corpus, as an estimate of what might come back from the clarification forms. The tuned parameters were the weighting scheme (ltc.lnc), the number of feedback terms to select (50), and the Rocchio parameters ($\alpha = 4, \beta = 2$).

The red triangle in Figure 1 shows the performance of the NIST feedback runs, where the baseline performance is the Prise3 system and the final performance is for the SMART system. The point is dramatically above the $y = x$ line, showing the dramatic improvement (more than 60%) this approach can cause. Unfortunately, the baseline was substantially below the better-performing systems, making it difficult to know whether simple relevance feedback would be equally effective at different qualities of baseline system. The results suggest that having a "pure" relevance feedback clarification form from every system might be a useful point for comparison.

12.3 Results by query

To a limited degree, it appears that better performing baselines result in larger gains from the clarification form. Figure 2 shows a breakdown of the same runs with each query represented. The graph shows a clear suggestion that it is easier to improve better-performing queries, but also demonstrates that poor-performing queries can be improved and have more *room* for improvement.

Another way of looking at the same question is to explore the absolute gain as a function of baseline R-precision. Figure 3 shows the same queries as Figure 2, but the y-axis shows the gain rather than value of R-precision. There is a very slight trend toward lower gain given higher baseline R-precision, but the fit is poor and the slope is almost horizontal. The graph suggests a strong negative correlation to the eye, but it

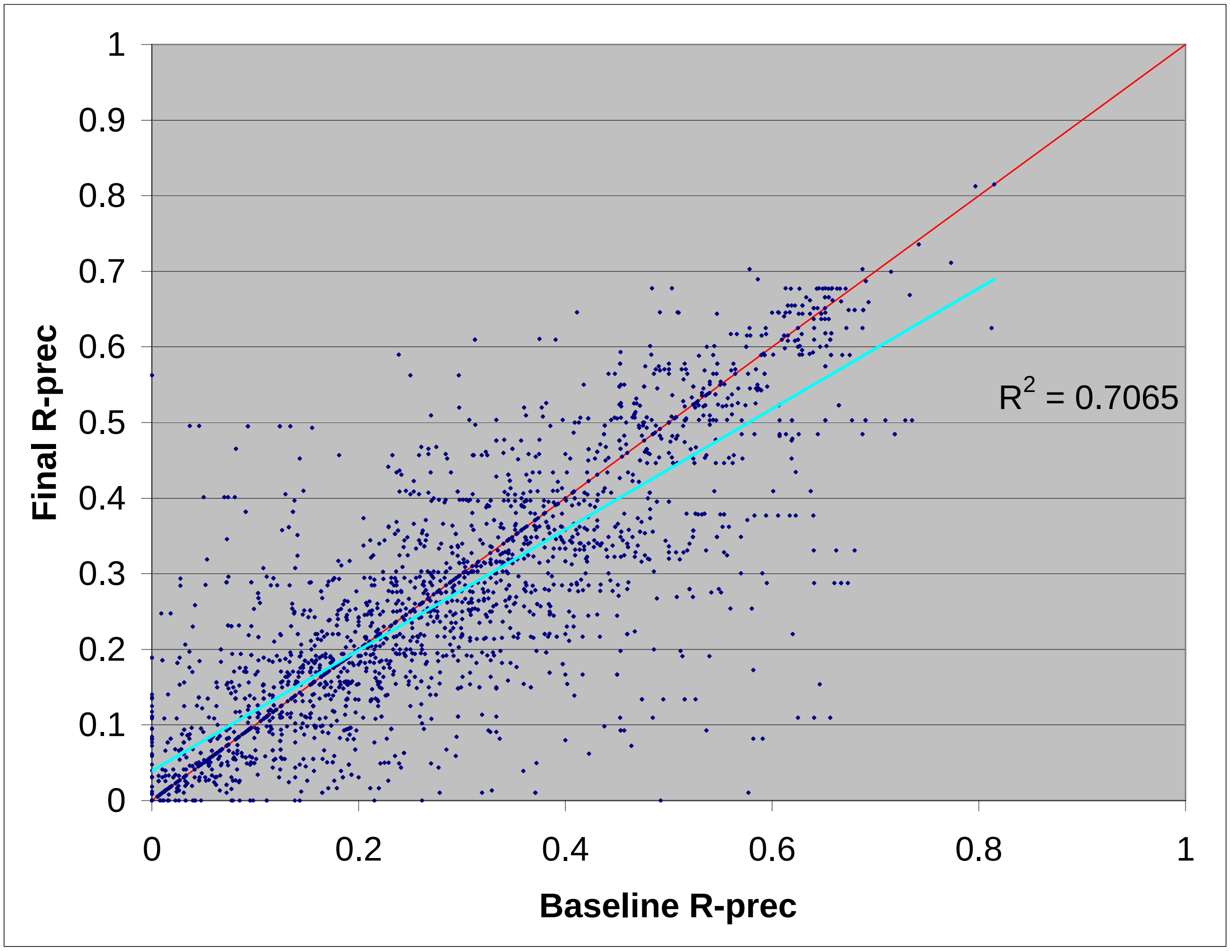

Figure 2: Comparison of R-precision values in baseline and final runs on a query-by-query basis. Each query from each run pair represented in Figure 1 is represented by a point. The $y = x$ line is shown as well as a linearly fit trend line.

is an artifact of the absolute loss being capped by the value of baseline R-precision—that is, if the baseline R-precision is 0.02, it is not possible to lose more than 0.02, but the gain can be quite large.

Figure 4 shows the absolute gain as a function of the number of relevant documents in the query. Again, there is a very weak trend toward more gain given more relevant documents in the pool. But the graph very clearly shows that the variance of the gain is large across all queries, regardless of the number of relevant documents they have.

Finally we consider the possibility that gain is correlated with the amount of time spent in clarification forms. Figure 5 shows that having annotators spend more time providing clarification information did not in and of itself increase realized gain. (Any effect may be obscured because a third of the interactions with annotators were truncated at 180 seconds, meaning we do not know how much time they actually might have spent.)

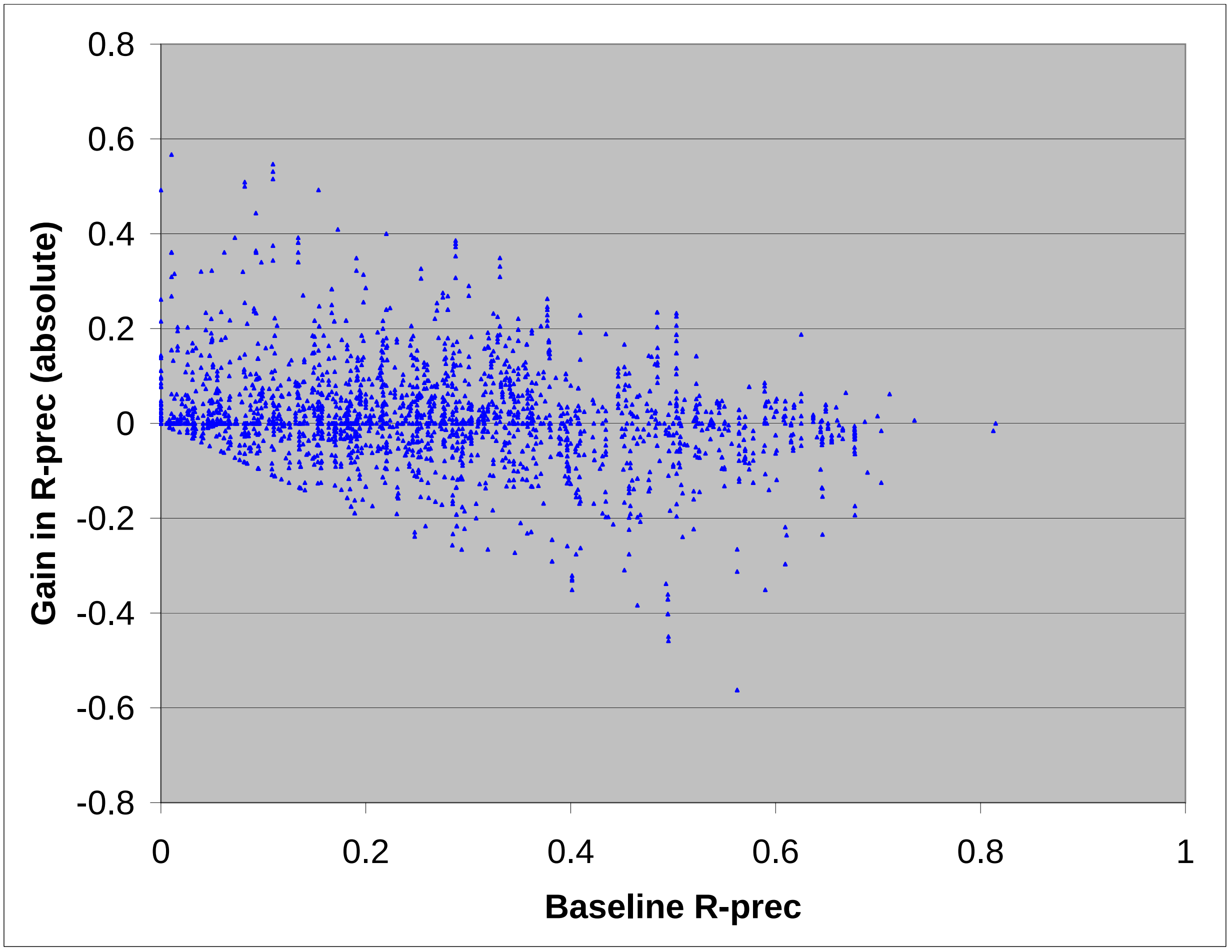

Figure 3: Comparing absolute gain in R-precision to baseline R-precision value.

12.4 Comparing two individual runs

It is illuminating to compare two runs that did well in the overall evaluation. We will consider the top performing title-only queries from two different groups.

1. Run MASStrmS is the automatic run with highest final R-precision. It started with baseline run MASSbaseTEE3 that had R-precision of 0.3291. It incorporated information from clarification form MASS1 and then achieved a final R-precision of 0.3547. That represents a 0.0256 gain in R-precision, an 8% relative improvement.
2. Run UIUChCFB3 is the automatic run with second highest final R-precision. It started with baseline run UIUC05Hardb0 that had R-precision of 0.2723. It incorporated information from clarification form UIUC3 and then achieved a final R-precision of 0.3355. That represents a 0.0623 gain in R-precision, a 23% relative improvement.

Figure 6 shows scatterplots of baseline and final R-precision values for the two runs, with UMass' run on the left and UIUC's on the right. For most queries in the UMass results, the final runs are almost identical to the baseline runs. However, a handful of queries with very low baseline scores show remarkable improvement, accounting for most of the gain in that system. This run appears to represent a very conservative query modification strategy, a reasonable choice given the high quality baseline.

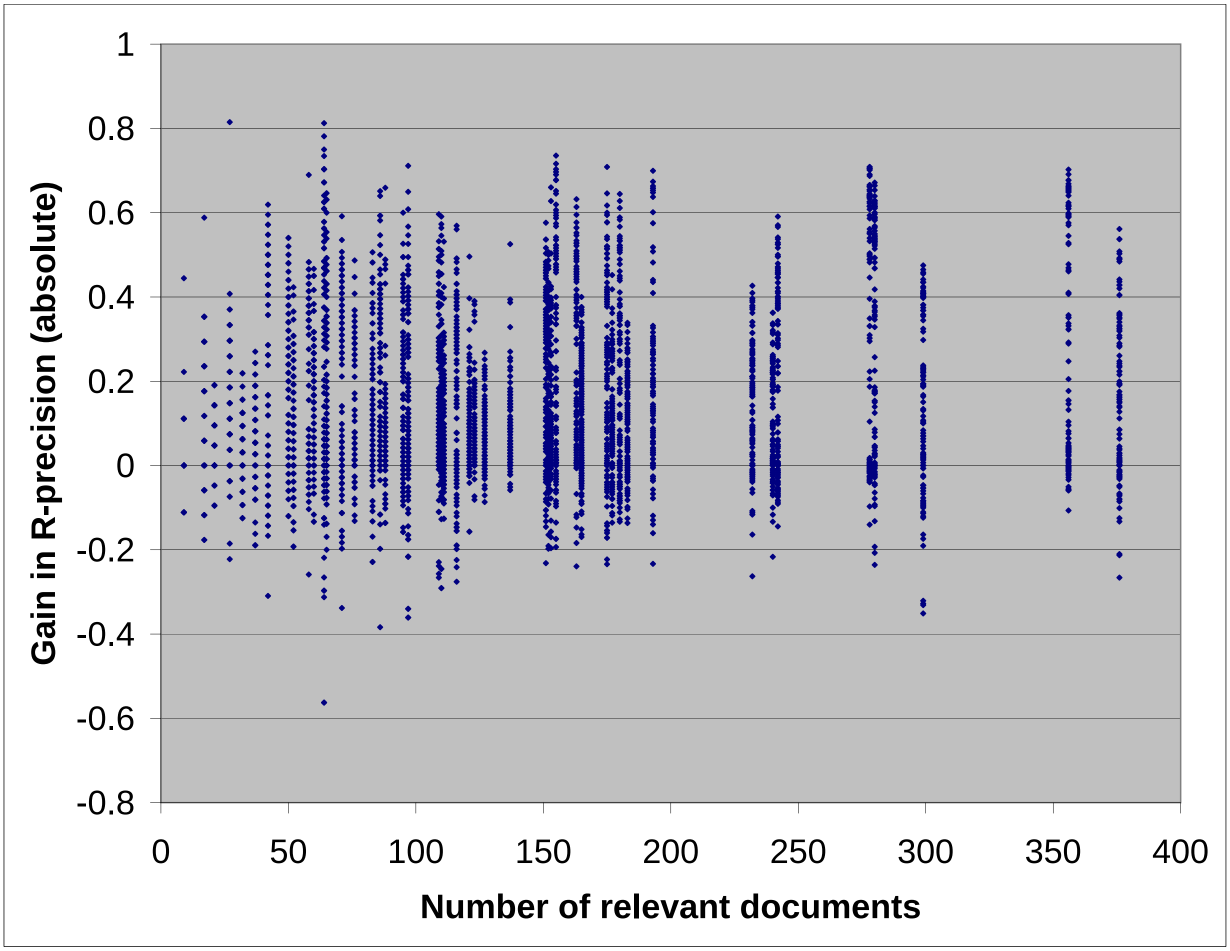

Figure 4: Comparing absolute gain in R-precision to the number of relevant documents for a query.

The UIUC run, in contrast, shows dramatic changes between the baseline and final runs. A large number of queries improve and a handful are significantly harmed. The strategy here is clearly much riskier and often pays off handsomely, trimming much of the baseline performance difference between UMass and UIUC.

Finally we do a direct comparison of how queries performed in the two systems. Figure 7 has an entry on the x-axis for every query. The queries are sorted by the final R-precision value of the query in the UIUChCFB3 run, the solid (blue) line that degrades smoothly from the upper left to the lower right. The corresponding baseline performance for that query is represented by (blue) diamonds.

The UMASStrmS final R-precision values are represented by the jagged (brown) line that roughly follows the trend of the UIUChCFB3 line, with the (red) triangles indicating baseline effectiveness.

Query effectiveness at the two sites follows a similar trend, but huge differences are common, with each site out-performing another by large margins in some cases. For example, query 651 show comparable baseline performance for the two sites, but successful clarification only by UIUC. Query 409 shows roughly the opposite result. Query 389 shows a case where UMass had substantially higher baseline performance, but UIUC's final run topped the UMass effectiveness by a good bit.

Comparing two systems provides only a glimpse of what is happening during clarification and final runs. It does suggest that different approaches work better for different queries, leading to the obvious question of whether it is possible to combine the clarification forms or to predict when one style is likely to be more

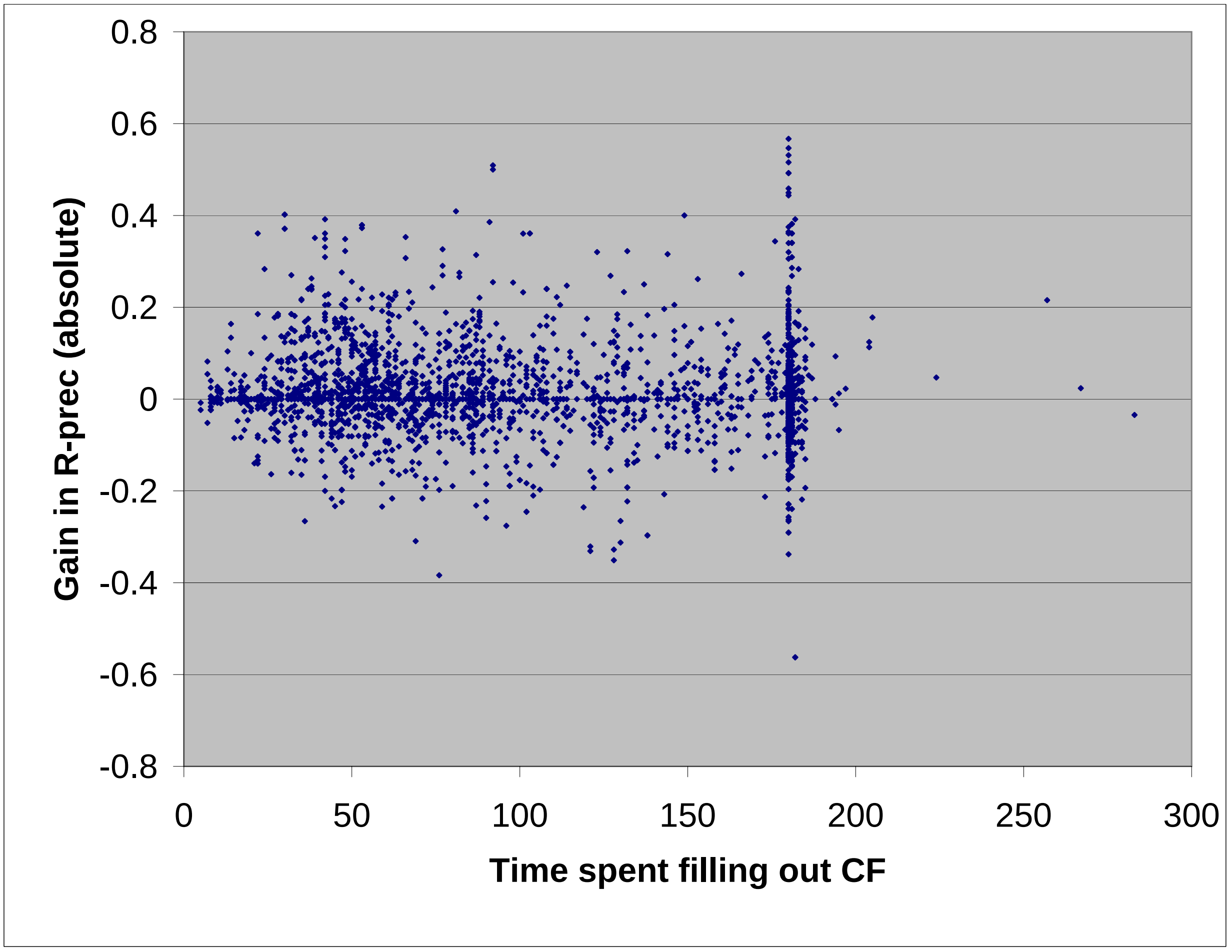

Figure 5: Comparing absolute gain in R-precision to time spent in the clarification form. Note that the density of scores at 180 seconds corresponds to the maximum time allowed in a form. The handful of scores beyond 180 seconds represent clarification forms that were difficult to "shut down" (see Section 9).

useful.

13 Conclusion

Several sites were able to show appreciable average gains from using clarification forms. None of the gains was consistently dramatic, however, begging the question of whether the time spent clarifying a query was a worthwhile investment. Further amplifying that question, it is worth pointing out that the best best automatic Robust track run beat all of the automatic baseline and final HARD track runs. (Of course, it is unknown whether a clarification form based on that run would improve the results further.)

This year there was an interesting variety of clarification forms tried. Forms of user-assisted query expansion were very popular, but sites also considered relationships between terms [Diaz and Allan, 2006, Yang et al., 2006], passage feedback [Diaz and Allan, 2006], incorporated summarization [Jin et al., 2006], and even used elaborate visualizations based on self organizing maps [He and Ahn, 2006]. The track itself did not provide clear support for any of these approaches.

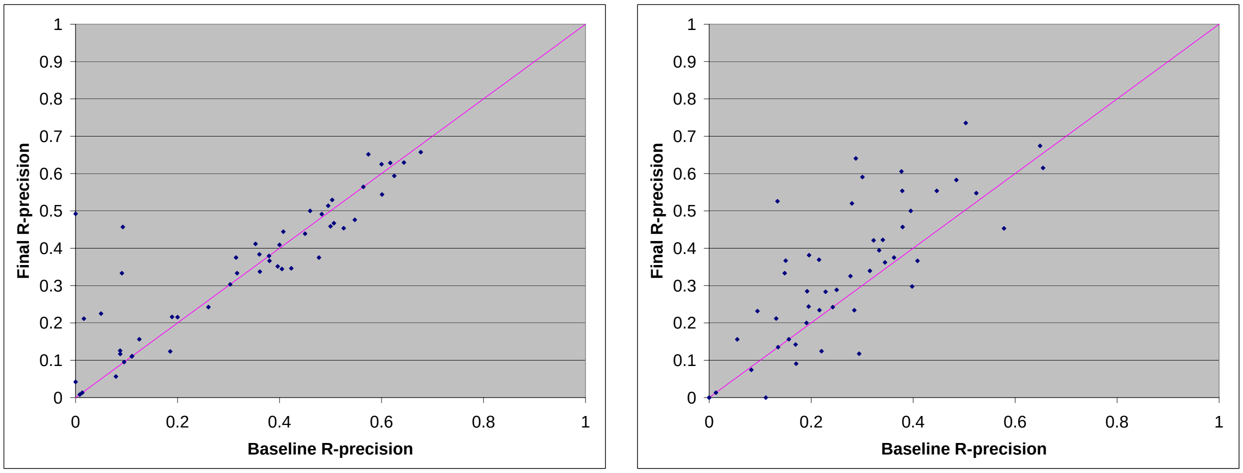

Figure 6: Comparison of baseline and final R-precision values for the MASStrmS run (left) and for the UIUChCFB3 run (right), broken down by query.

It is important to note that the clarification forms do not represent "interactive information retrieval" experiments. They provide a highly focused and very limited type of interaction that can (potentially) improve the effectiveness of document retrieval. Whether these clarification forms can be deployed in a way that pleases a user or that will actually be used is an entirely different question, one that would have to be tested in a more realistic environment.

After a three-year run, TREC 2005 was the end of the HARD track. For TREC 2006, it is being made part of the Question Answering track as "ciQA", or "complex, interactive Question Answering." The goal of ciQA is to investigate interactive approaches to cope with complex information needs specified by a templated query.

Acknowledgments

As do most tracks in TREC, the HARD track owes a debt of gratitude to Ellen Voorhees at NIST, who not only provided the typical TREC infrastructure, but who was an important contributor in the shaping of the track's goals and evaluation approaches. The Linguistic Data Consortium kindly provided training data to participants and words of wisdom to NIST organizers. Stephanie Strassel and Meghan Glenn of the LDC were particularly generous with their time and support.

Finally, the track would not have been successful without the involvement of its participants—not only by participating, but by designing the track during discussions in person and on-line. Diane Kelly at the University of North Carolina should be singled out for rapid production of the rotation tables used to assign clarification forms to annotators.

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

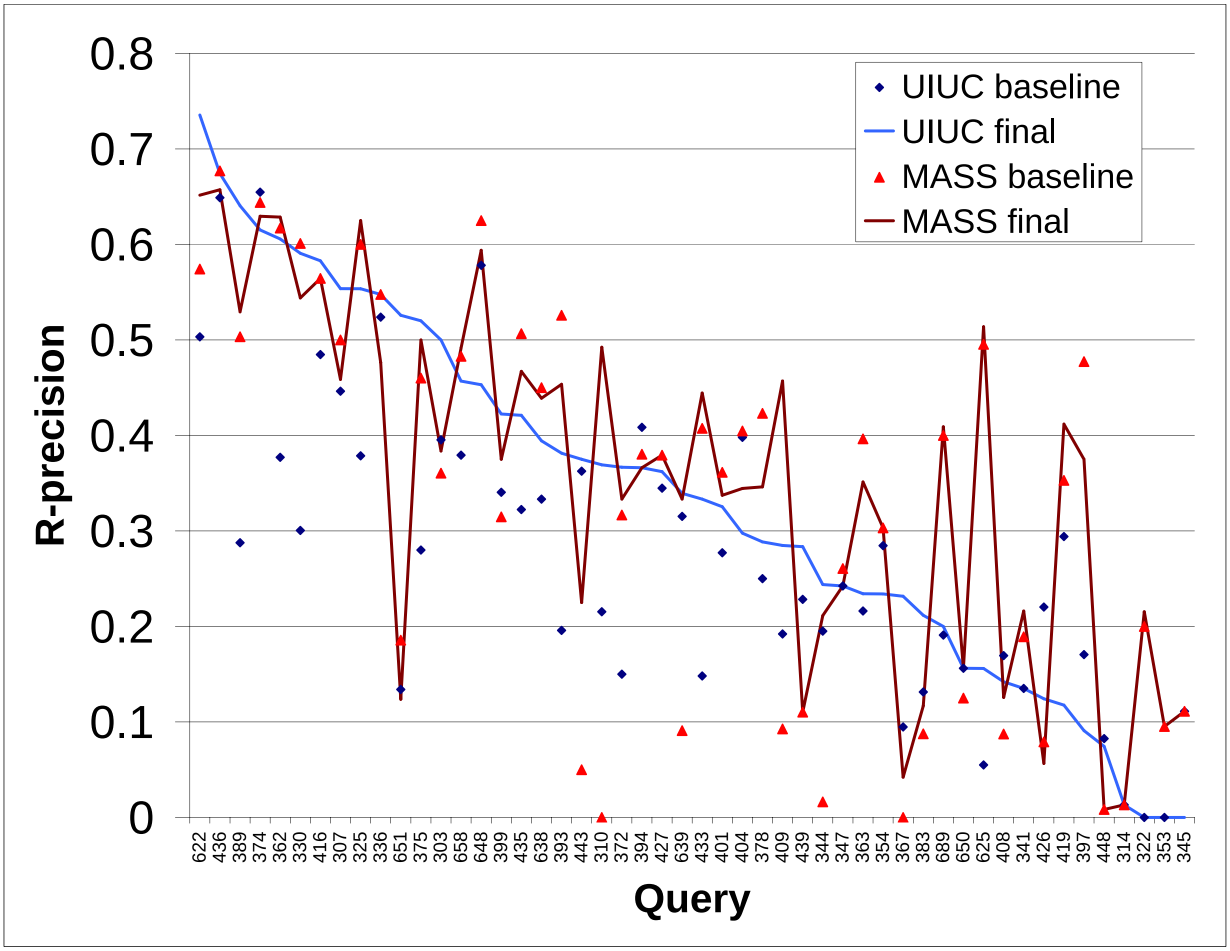

Figure 7: A query-by-query comparison of baseline and final R-precision values for the MASStrmS and UIUChCFB3 runs in Figure 6. Each query is a point on the x-axis; the queries are ordered by the final R-precision score of the UIUChCFB3 run.

References

[Allan, 2004] Allan, J. (2004). HARD track overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of TREC 2003*, pages 24–37. NIST special publication 500-255. Also on-line at http://trec.nist.gov.

[Allan, 2005] Allan, J. (2005). HARD track overview in TREC 2004: High accuracy retrieval from documents. In *Proceedings of TREC 2004*, pages 25–35. NIST special publication 500-261. Also on-line at http://trec.nist.gov.

[Baillie et al., 2006] Baillie, M., Elsweiler, D., Nicol, E., Ruthven, I., Sweeney, S., Yakici, M., Crestani, F., and Landoni, M. (2006). University of Strathclyde at TREC HARD. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Belkin et al., 2006] Belkin, N., Cole, M., Gwizdka, J., Li, Y.-L., Liu, J.-J., Muresan, G., Roussinov, D., Smith, C., Taylor, A., and Yuan, X.-J. (2006). Rutgers information interaction lab at TREC 2005: Trying HARD. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Diaz and Allan, 2006] Diaz, F. and Allan, J. (2006). When less is more: Relevance feedback falls short and term expansion succeeds at HARD 2005. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[He and Ahn, 2006] He, D. and Ahn, J. (2006). Pitt at TREC 2005: HARD and Enterprise. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Jin et al., 2006] Jin, X., French, J. C., and Michel, J. (2006). SAIC & University of Virginia at TREC 2005: HARD track. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Kelly and Fu, 2006] Kelly, D. and Fu, X. (2006). University of North Carolina's HARD track experiment at TREC 2005. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Kudo et al., 2006] Kudo, K., Imai, K., Hashimoto, M., and Takagi, T. (2006). Meiji University HARD and Robust track experiments. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Lin et al., 2006] Lin, J., Abels, E., Demner-Fushman, D., Oard, D. W., Wu, P., and Wu, Y. (2006). A menagerie of tracks at maryland: HARD, enterprise, QA, and genomics, oh my! In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Lv and Zhao, 2006] Lv, B. and Zhao, J. (2006). NLPR at TREC 2005: HARD experiments. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Rode et al., 2006] Rode, H., Ramírez, G., Westerveld, T., Hiemstra, D., and de Vries, A. P. (2006). The Lowlands' TREC experiments 2005. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Tan et al., 2006] Tan, B., Velivelli, A., Fang, H., and Zhai, C. (2006). Interactive construction of query language models – UIUC TREC 2005 HARD track experiments. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Vechtomova et al., 2006] Vechtomova, O., Kolla, M., and Karamuftuoglu, M. (2006). Experiments for HARD and Enterprise tracks. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Voorhees, 2006] Voorhees, E. M. (2006). Overview of the TREC 2005 robust retrieval track. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Wen et al., 2006] Wen, M., Huang, X., An, A., and Huang, Y. (2006). York University at TREC 2005: HARD track. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Yang et al., 2006] Yang, K., Yu, N., George, N., Loehrlen, A., McCaulay, D., Zhang, H., Akram, S., Mei, J., and Record, I. (2006). WIDIT in TREC 2005 HARD, Robust, and SPAM tracks. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.

[Zhang et al., 2006] Zhang, J., Sun, L., Lv, Y., and Zhang, W. (2006). Relevance feedback by exploring the different feedback sources and collection structure. In *Proceedings of TREC 2005*. On-line at http://trec.nist.gov.