

HARD Track Overview in TREC 2005 (Notebook)

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:

- 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. 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.

This is a notebook paper and so is short on details and is likely to contain errors that will be corrected later. Additional (and more accurate, we hope) 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 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			Participating site
Base	Final	# CFs	
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 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 essentially 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.

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 the 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.

7 Training data

The data collections from the HARD tracks of TREC 2003 and 2004 were available for training. All of that data was made available to HARD track participants courtesy of the Linguistic Data Consortium. The data was made available for use only in the HARD 2005 evaluation with the expectation that they will be destroyed at the completion of the track (i.e., after the final papers are written). The LDC is likely to release the data as collections to its members eventually.

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 totally 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 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:

- 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 Javascript, 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., "001") and the title of the topic to provide a sanity check that the annotators are, 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. Table 1 shows the rotation that was used for the submitted clarification forms (graciously generated by Diane Kelly of the University of North Carolina with very little turnaround time).

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.

A total of 92 final runs were submitted.

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. Similar, for topic 1, the assessor first did INDI1’s form, then that for CASP1, then UIUC1’s, followed by MEIJ1’s, and so on.

11 Evaluation

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

Figure 1 shows how R-precision changed between final runs and their corresponding baseline runs. More analysis of results will appear in the track presentation at TREC and in the final paper.

12 Conclusion

It appears that several sites were able to show appreciable gains using clarification forms this year.

At the time of writing, it is not clear what the fate of HARD will be in TREC 2006 and beyond.

	CASS1	DUBL1	UWAT1	MASS2	CASP2	STRA3	UWAT2	MEU1	MEU2	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. Perhaps by the final paper, these will be easier to read.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSSYSCEN-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.

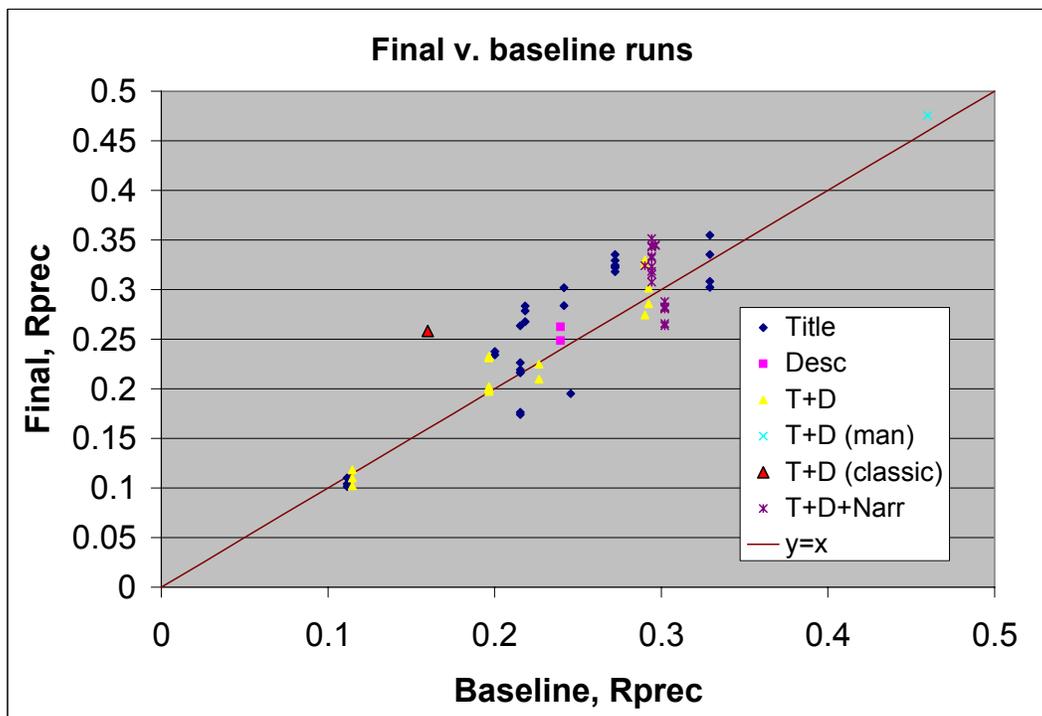


Figure 1: Comparison of R-precision values in baseline runs and runs after using a clarification form (only runs that identified a corresponding baseline run are included). Points below the $y = x$ line had final runs that were worse than their corresponding baseline runs; those above the line improved. Point colors reflect variations in the processing and may not be easily visible in a grayscale print. Note that the (excellent) outlier in the upper right is a manual run.