A Menagerie of Tracks at Maryland: HARD, Enterprise, QA, and Genomics, Oh My!*

Jimmy Lin,[a,c] Eileen Abels,[a] Dina Demner-Fushman,[b,c]
Douglas W. Oard,[a,c] Philip Wu,[a] and Yejun Wu,[a,c]

[a]College of Information Studies
[b]Department of Computer Science
[c]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
{jimmylin,eabels,oard,fwu}@umd.edu,
demner@cs.umd.edu, wuyj@glue.umd.edu

Abstract

This year, the University of Maryland participated in four separate tracks: HARD, enterprise, question answering, and genomics. Our HARD experiments involved a trained intermediary who searched for documents on behalf of the user, created clarification forms manually, and exploited user responses accordingly. The aim was to better understand the nature of single-iteration clarification dialogs and to develop an "ontology of clarifications" that can be leveraged to guide system development. For the enterprise track, we submitted official runs to the Known Item Search and the Discussion Search tasks. Document transformation to normalize dates and version numbers was found to be helpful, but suppression of text quoted from earlier messages and expansion of the indexed terms for a message based on subject line threading proved to not be. For the QA track, we submitted a manual run of "other" questions in an effort to quantify human performance on the task. Our genomics track participation was in collaboration with the National Library of Medicine, and is primarily reported in NLM's overview paper.

1 Introduction

Information retrieval is a complex, interdisciplinary field that lies at the intersection between computer science, library and information science, linguistics, and other related fields. In addition, specialized expertise is necessary for working in certain domains. For TREC, we have assembled a team whose members bring a broad range of experiences to tackle this complex challenge. This year, we participated in four tracks: HARD (Section 2), enterprise (Section 3), QA (Section 4), and genomics (Section 5).

2 HARD Track

The one-shot model of information retrieval operationalized in TREC is an oversimplification of real-world information-seeking behavior, and has often been criticized for neglecting the important role of interaction. The development of single-iteration clarification dialogs in the HARD track begins to address some of these concerns by introducing limited elements of interaction, but at the same time preserving the benefits provided by a large-scale, shared evaluation.

Explorations of clarification dialogs have proven to be challenging because most previous experiments confound the effects of clarification form content and the manner in which responses are

*Authors listed alphabetically with the exception of the first author.

exploited. Results from previous years have shown little improvement in terms of standard ranked-retrieval metrics. Does this finding reveal fundamental limitations about such interactions, or are present systems simply unable to effectively capitalize on such interactions?

In an effort to better understand the solution space of single-iteration clarification dialogs, we employed a trained intermediary to manually gather potentially relevant documents, construct clarification forms, and exploit user responses accordingly. We had three major goals:

- to establish a plausible upper bound on the effectiveness of single-iteration clarification dialogs;
- to develop an "ontology of clarifications" that can be used as a basis for the design of more nuanced automated systems that initiate and exploit clarification dialogs;
- to generate insights that will guide future work on the design of interfaces and strategies for maximizing the utility of user-system interaction.

We have made significant headway in achieving the first goal as a direct result of our HARD experience. Through post-hoc analysis of clarification questions, we have also discovered commonly-occurring patterns of intent, thus serving as the beginnings of an ontology of clarifications. This paves the way to accomplishing our third goal, which we leave for future work.

The description of our HARD experience is organized as follows: Section 2.1 relates HARD clarification dialogs to information need negotiation in a reference interview setting. Section 2.2 describes our methodology for producing the initial ranked list and exploiting user feedback to generate a final set of results. Section 2.3 presents results from our pre- and post-clarification runs. In Section 2.4, we examine how clarification dialogs improved performance in many cases, but also hurt in others. Section 2.5 describes an inductive attempt to organize commonly-observed patterns into a preliminary ontology of clarifications.

2.1 The Nature of Clarification Dialogs

Information need negotiation in the context of a reference interview within a library setting is a complex communication between an information specialist and a user (Taylor, 1962), which begins with the user describing his or her requirements. Through a series of interactions, *both* parties arrive at a better understanding of the information need and a mutually agreeable search strategy for acquiring the desired information.

Like information need negotiation in a reference interview, clarification dialogs aim at gaining a better understanding of the user's requirements. However, the critical difference is that the reference interview usually *precedes* the actual search process, whereas clarification dialogs in HARD occur after an initial search is performed (this setup is a methodological necessity in order to evaluate pre- and post-clarification performance). Thus, HARD clarification dialogs can involve search results, whereas little is known about actual documents during the reference interview. Pre-search information need negotiation is absolutely essential in the real world because it largely determines *what resources* to search, which depends, for example, on the format of expected results. This element of source selection has been eliminated in TREC.

As electronic resources move increasingly online, the face-to-face reference interview is being gradually replaced by other media: initially, the telephone, and now, email and online chat. HARD clarification dialogs share many formal properties with email reference interviews: both are asynchronous and primarily depend on text-based interfaces for soliciting user input. Hence, we can benefit from previous work on the email reference interview. Abels (1996) identified five approaches often used in need negotiation over email: (1) piecemeal, (2) feedback, (3) bombardment, (4) assumption, and (5) systematic. Her analysis showed that the systematic approach was most successful and most efficient in terms of the number of messages exchanged. In the systematic approach, the information specialist responds to a request with a list of open- and closed-ended questions that covered all aspects of the topic, arranged in a coherent, logical manner. From this point of view, the purpose of the HARD track is to better understand the systematic approach to asynchronous need negotiation so that such dialogs can be automatically conducted by a computer.

2.2 Methodology

To establish an upper bound on the effectiveness of single-iteration clarification dialogs, we employed a trained intermediary[1] in all phases of our HARD experiments. This section describes our methodology for creating pre- and post-clarification runs.

The intermediary primarily employed the "building blocks" strategy (Harter, 1986; Marchionini, 1995) with INQUERY (Version 3.1p1 for Solaris) to gather relevant documents on behalf of the user (who is also the assessor; we use these two terms interchangeably). First, conceptual facets were identified from the topic statement and captured with a disjunction of synonymous or related terms using INQUERY's "soft" OR operator. External tools such as Google and Wikipedia were used as appropriate. These facets were then systematically combined into complete queries, most often with INQUERY's "soft" AND operator. The first ten or so hits were examined to determine if the query was "good"; if not, the intermediary reformulated the query, taking advantage of additional terms that may have appeared in the top hits and INQUERY's full range of query operators (hard boolean operators, proximity operators, etc.). Once a "good query" was constructed, the intermediary manually examined each document in the resulting hit list. The aim was to assess the top 100 hits, but the actual number of documents examined varied, depending on the difficulty of the topic, the number of relevant hits, and other factors. Each examined document was assigned one of four judgments:

- **Centrally relevant (CR)**: based on the intermediary's understanding of the information need, this document would be considered topically relevant.
- **Peripherally relevant (PR)**: based on the intermediary's understanding of the information need, this document would be considered relevant, but less so than than documents marked centrally relevant (for example, a passing mention or a vague reference).
- **Maybe relevant (MR)**: based on the intermediary's incomplete understanding of the information need, this document may be relevant. Ambiguity in TREC topic statements often force the intermediary to make assumptions, draw inferences, etc. If a document would be considered relevant based on a particular interpretation, this judgment is assigned.
- **Not relevant (NR)**: this document would not be considered relevant.

Creation of our official pre-clarification run was accomplished automatically using the relevance judgments of the intermediary. Based on *tf.idf* scores, 20 terms were selected from the documents marked centrally relevant. These terms were combined with terms from the topic title and topic description using INQUERY's weighted sum operator (weight of 3.0 for title terms, 1.0 for all others). This ranked list was submitted as run **MARYB2**. Our main run, **MARYB1**, consisted of CR, PR, and MR documents (in that order), followed by documents in **MARYB2** (with duplicates removed). Documents in each of the three piles were simply arranged in the order they were examined in the search process. As an automatic baseline, we submitted an INQUERY run that used title and description terms as the query, with blind relevance feedback (top 20 terms from top 10 hits in terms of *tf.idf* scores); this run was called **MARYB3**.

We conceived of the clarification process as a reshuffling of documents between the four piles created by the intermediary. Clarification questions were explicitly created with one of two goals:

- To move documents in the PR pile into either the CR pile or the NR pile. We hypothesize that the user's mental model includes a boundary for making hard decisions about document-level relevance; these questions are aimed at a better understanding of this threshold.
- To move documents in the MR pile into either the CR pile or the NR pile. In the search process, our intermediary makes relevance judgments based on a particular interpretation of the information need as captured in the topic statement; this often involves drawing inferences, making assumptions, etc. The purpose of these questions is to verify the correctness of the interpretation.

Although a major goal of our research is the development of an "ontology of clarifications", we consciously decided to adopt a inductive, bottom-up approach. Thus, the intermediary formulated

[1] Philip Wu, a Ph.D. student in College of Information Studies at Maryland.

Title: Three Gorges Project
Description: What is the status of The Three Gorges Project?
Narrative: A relevant document will provide the projected date of completion of the project, its estimated total cost, or the estimated electrical output of the finished project. Discussions of the social, political, or ecological impact of the project are not relevant.

Clarification Questions

1. ☐ Check if yes: Must a relevant article mention the date of completion and total cost and estimated electrical output? Leave unchecked if it is sufficient to discuss any one of these facets.
2. ☐ Check if yes: Is "early next century" an acceptable projected date of completion?
3. ☐ Check if yes: Would articles mentioning state bank loans or foreign investment be relevant?
4. ☐ Check if yes: Would articles discussing the cost (or completion date) of a subcomponent of the project be relevant? For example, "power transmission project" or "the first construction phrase".

Figure 1: [HARD] Sample topic and clarification questions.

questions as appropriate, without reference to any pre-existing ontologies, questions types, or stylized phrasing of questions (e.g., question templates). However, all questions were constructed so that responses could be captured via checkboxes. This ensured a consistent user interaction pattern.

In addition to topic-specific questions, all clarification forms included two generic questions (located at the end): "Any additional search terms?" and "Any other comments?" Both were followed by a 70×4 text box for free-formed input.

As a complete example, Figure 1 shows the topic statement for topic 416 "Three Gorges Project", along with the four clarification questions generated by our intermediary. In this example, the second question was targeted at PR documents, while the other questions were targeted at MR documents.

After receiving clarification responses from the user, our intermediary shuffled the document piles based on (hopefully) a more refined understanding of the information need. In our conception of the ideal interaction, the clarification forms would supply sufficient evidence for the intermediary to confidently eliminate the PR and MR piles completely. Realistically however, documents with uncertain relevance still remained.

Creation of the official ranked list for our post-clarification run followed exactly the same procedure as the creation of our pre-clarification run, except with different piles (run **MARY05C1**). In addition, we submitted two contrastive conditions: **MARY05C2** used title and description terms from the topic, along with additional search terms supplied by the user in the clarification forms. **MARY05C3** added additional blind relevance feedback terms to **MARY05C2**. For both runs, terms were combined using INQUERY's weighted sum operator, with a weight of 3.0 given to title terms, and 1.0 to all other terms.

2.3 Results

In total, we submitted three pre-clarification and three post-clarification runs for the HARD track. Official results are shown in Table 1: **median** is the mean of the per-topic median score of all submitted runs, **best** is the mean of the best per-topic score of all submitted runs, and **best auto** is the highest-scoring automatic run. In total, 30 pre-clarification and 92 post-clarification runs were submitted by all participants. For 29 topics, the **MARYB1** pre-clarification run achieved the best mean average precision across all submitted runs; for R-precision, 28 topics. For 20 topics, the **MARY05C1** post-clarification run achieved the best mean average precision across all submitted runs; for R-precision, 17 topics.

Our intermediary spent an average of 109 minutes per topic preparing the document piles for the

	MARYB1	**MARYB2**	**MARYB3**	**median**	**best**	**best auto**
MAP	0.452	0.368	0.252	0.190	0.496	0.304
R-Prec	0.460	0.386	0.292	0.252	0.513	0.329
	MARY05C1	**MARY05C2**	**MARY05C3**	**median**	**best**	**best auto**
MAP	0.469	0.233	0.263	0.207	0.535	0.322
R-Prec	0.476	0.286	0.301	0.264	0.545	0.355

MARYB1: CR+PR+MR+rel feedback run
MARYB2: rel feedback run
MARYB3: title+desc+brf
MARY05C1: same as **MARYB1**, with updated piles
MARY05C2: title+desc+user-supplied terms
MARY05C3: title+desc+brf+user-supplied terms

Table 1: [HARD] Official results (pre-clarification on top and post-clarification on bottom).

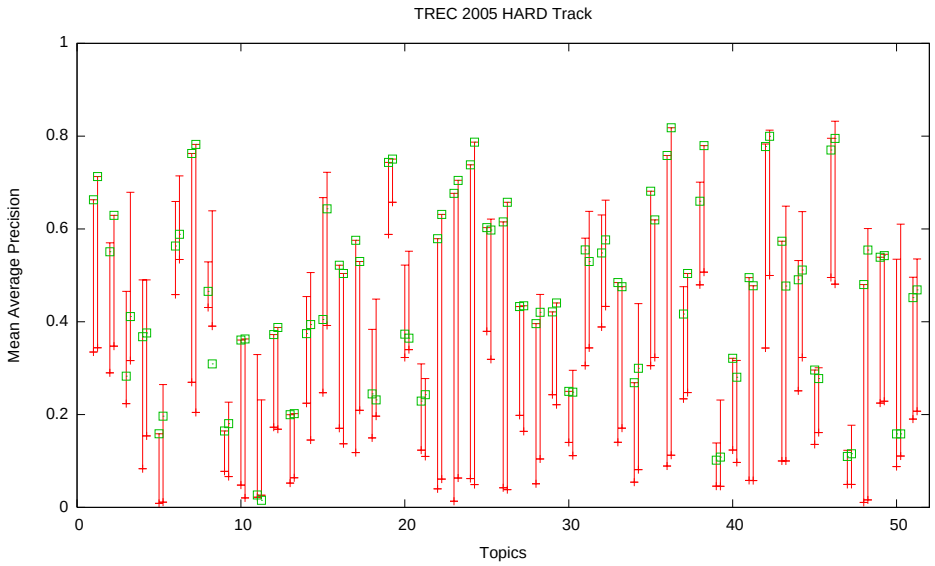

Figure 2: [HARD] Comparison of mean average precision on a per-topic basis. Each pair of bars represents the median/best score range, before and after clarification. **MARYB1** and **MARY05C1** scores are marked with boxes. The rightmost bars represent the average across all topics.

pre-clarification runs (max 170, min 35, $\sigma = 29.7$). This time includes analyzing the topic statement, formulating a "good" query, and performing the relevance assessment. For about half a dozen topics, the intermediary had difficulty generating a good query and finding relevant documents; the advice of other team members was sought, but this time is not included in the figures mentioned above. We did not keep detailed timing statistics for the process of exploiting clarification responses, but reassessing document relevance based on user feedback took approximately ten to thirty minutes per topic.

A total of 89 clarification questions was posed across all 50 topics (discounting the two generic questions present for every topic), for an average of 1.8 questions per topic ($\sigma = 1.47$). Topic 341 "airport security" had the most clarification questions, with seven. Ten topics had no specific clarification questions: the intermediary found them to be straightforwardly defined. Disregarding the ten topics without specific clarification questions, the average number of questions per topic jumps to 2.2 ($\sigma = 1.31$).

For thirty-five of the topics, clarification responses included additional search terms supplied by the user. In fifteen of the forms, clearly demarked phrases were entered. There was an average of 3.66 terms/phrases per topic ($\sigma = 3.31$), with a maximum of fourteen.

testset	All	A	B
# q's	50	40	32
MAP before	0.452	0.465	0.482
MAP after	0.469 (+3.7%)	0.485 (+4.3%)	0.519 (+7.7%)
$\delta[0.10, \infty]$	8	8	8
$\delta[0.05, 0.10)$	12	9	9
$\delta[-0.05, 0.05)$	22	16	14
$\delta[-0.10, -0.05)$	4	3	1
$\delta[-\infty, -0.10)$	4	4	0

Table 2: [HARD] Effect of clarification dialog on different subsets of the topics.

Looking at the difference between **MARYB1** and **MARY05C1**, it appears that clarification had only a small impact on mean average precision (+3.7%) and R-precision (+3.5%). A Wilcoxon signed-rank test reveals that the difference in MAP is significant at the 1% level, while the difference in R-precision is significant at the 5% level. For additional significance results reported in this section, the Wilcoxon signed-rank test is used. Figure 2 shows the effects of the clarification dialog on mean average precision on a per-topic basis. Each pair of closely-spaced bars represents a single topic: the left bar represents the range of the median to best score before clarification; the right bar, after clarification. Boxes indicate the performance of **MARYB1** and **MARY05C1**, respectively. The rightmost set of bars represents an average across all topics.

2.4 Analysis

A topic-by-topic analysis focused on runs **MARYB1** and **MARY05C1** revealed ways in which the clarification dialogs were improving or hurting retrieval effectiveness. We arbitrarily divided topics into five bins, according to the relative differences between pre- and post-clarification MAP: $\delta \geq 0.10$, $0.05 \leq \delta < 0.10$, $-0.05 \leq \delta < 0.05$, $-0.10 \leq \delta < -0.05$, and $\delta < -0.10$. As can be seen in Table 2, eight topics fell in the last two bins, where clarification decreased MAP by at least 5%.

First, we narrowed our examination to topics for which there were clarification questions (forty topics). This is shown as testset **A** in Table 2.[2] Considering this reduced set of topics, we observe a gain of 4.3% in terms of MAP (significant at the 1% level). We then manually examined each topic in order to better understand ways in which the clarification dialog helped or hurt.

For many topics, it is easy to see why clarification dialogs improved performance. A better understanding of the user's information need brings the intermediary's relevance judgments more in sync with those of the user. The most dramatic example of this is with topic 362 "human smuggling", where MAP jumped from .405 to .643, a gain of 59%. The topic called for reports about incidents of human smuggling for monetary gain. The clarification questions confirmed that the element of monetary gain must be present, and that summaries of smuggling rings and smuggling statistics were not relevant.

Somewhat distressing are seven topics in which the clarification dialog resulted in a decrease in MAP of at least 5%. For example, the mean average precision of topic 336 "black bear attacks" dropped 34% (from .466 to .309). To the clarification question "Does a document need to mention frequency of attacks **and** cause of attacks **and** method of control to be considered relevant?", the assessor answered "yes", indicating that documents with missing facets were not considered relevant. However, analysis of the final qrels show that many documents missing the abovementioned facets were nevertheless marked relevant. In other words, the assessor's answer to the clarification question did not match the actual criteria used in the assessment! We have dubbed this the "inconsistent user" phenomenon.

[2]Based on our methodology, topics with no clarification questions should have exactly the same pre- and post-clarification MAP. However, due to differences in the source of term statistics for query expansion terms, there were slight differences in MAP.

Type	Topic	Clarification Question	%
RT	(404) Ireland, peace talks	Would a general reference to violence without specifying particular acts be relevant?	28 (31%)
ACF	(344) Abuses of E-Mail	Does an article need to discuss both cases of email abuse and steps taken to prevent abuse to be relevant?	9 (10%)
EC	(344) Abuses of E-Mail	Would email hoaxes be considered "abuse"?	20 (22%)
CRC	(336) Black Bear attacks	Would other species of bears (brown bear, grizzly bear...) be of interest?	12 (13%)
RTA	(341) Airport Security	Would articles about tightened security policy on airport employees be relevant?	17 (19%)
AS	(362) human smuggling	Would a summary of a smuggling ring be relevant?	3 (3%)

Table 3: [HARD] Examples of clarification questions and their prevalence.

In fact, examining all eleven topics where clarification dialogs caused a drop in MAP revealed eight cases of the "inconsistent user" phenomenon. For these topics, the feedback received in the clarification dialog was misleading and contradicted the users' relevance criteria as reflected in the final judgments. Results of removing these topics from testset **A** are shown in Table 2 as testset **B**. On these topics, clarification dialogs yielded an increase of 7.7% in mean average precision (significant at the 1% level). The table shows that topics in the worst-performing bin (MAP decrease greater than 10%) can all be attributed to this cause.

In three of the eleven topics under consideration, the clarification dialog did actually cloud the intermediary's understanding of the user's need, mainly due to poorly-formulated clarification questions. One case, for example, asked whether or not "details" were necessary. This being a vague term, the intermediary and user ultimately had different notions of what "details" meant.

What is the cause for this "inconsistent user" phenomenon? After ruling out malicious intent, there are at least two possibilities: one points to a methodological flaw, while the other stems from the nature of information-seeking behavior itself.

Due to real-world constraints involved in coordinating the HARD track, documents were not assessed until approximately a month after the clarification questions had been answered (in order to allow ample time for participants to prepare their final runs). During this time, the assessors may have already forgotten their original answers: instability in relevance criteria over long periods of time could be the source of observed user inconsistencies. This is exacerbated by the fact that this year's topics did not represent "real" information needs, since the topic statements were not constructed by the assessors themselves.

Research in information science, however, suggests that inconsistencies in users' notions of relevance may be an inescapable fact of real-world information-seeking behavior. The TREC evaluation methodology assumes a static information need against which documents are evaluated for relevance, when, in truth, information needs are themselves constantly shifting and evolving as users learn more about the subject (Taylor, 1962; Bates, 1991). Therefore, it is entirely conceivable that the mere act of participating in the clarification dialog altered the users' needs. Since our clarification questions were created based on documents assumed to be relevant by the intermediary, we are already circumscribing the bounds of the user's relevance space. Most of our clarification questions can be considered "leading", which may influence the assessor to respond in a calculated manner that runs counter to the true underlying need. Thus, "neutral" questioning is preferred in reference interviews because it focuses on what users really want to know, as opposed to the best available resources and documents (Dervin and Dewdney, 1986).

In truth, inconsistencies in users' notions of relevance are most likely caused by a combination of both factors described above. Unfortunately, the current HARD methodology conflates the two issues. More carefully-constructed experiments must be conducted to better understand the shifting nature of information needs.

2.5 Towards an Ontology of Clarifications

In the process of generating clarification questions, we noticed that a number of common patterns began to emerge, even though we did not impose any sort of pre-existing theory or ontology in a top-down manner. Analysis of our HARD results included an attempt to induce an "ontology of clarifications" in a bottom-up manner by observing similarities in the intent of clarification questions. This task was undertaken by the first author, who then manually coded all clarification questions according to the induced ontology.

As previously described, we view the clarification dialog as an opportunity to better understand the user's information need so that peripherally relevant and maybe relevant documents can be sorted into either the relevant or not relevant piles. Questions targeted at the peripherally relevant documents form a coherent class in terms of question intent:

- **Determining the relevance threshold (RT).** We hypothesize the existence of a "relevance threshold" that guides the user in making hard judgments about document relevance. Clarification questions of this type attempt to better understand this boundary in "relevance space".

Other clarification questions fall naturally into five categories based on question intent, shown below. Examples of each type can be found in Table 3.

- **Determining the relationship between ambiguously conjoined facets (ACF).** Many topic statements actually express multiple, related information needs; one can view this as multiple facets of a larger topic. Often, the relationship between these facets is unclear, e.g., does a document need to contain all of the facets simultaneously to be considered relevant?
- **Determining the relevance of an example concept (EC).** Is a particular concept found in one or more documents an example of a concept mentioned in the topic statement? For example, topic 347 concerns wildlife extinction: it is unclear whether documents about plants are relevant. The intermediary therefore formulated a clarification question to better understand the user's definition of "wildlife". A specific subclass of this type concerns so-called "meta-terms", such as pros/cons, advantages/disadvantages, impact, etc. For the most part, they make poor query terms, and need to be operationalized in a particular context.
- **Determining the relevance of a closely-related concept (CRC).** Does the user's interest in a particular concept A extend to a closely-related concept A'? A and A' may be ontologically related via hypernymy, hyponymy, antonymy, etc.
- **Determining the relevance of related topical aspects (RTA).** Is the user interested in topics that are conceptually related, but not directly requested? Topics often focus on a particular aspect of a larger concept; these questions ascertain whether users might consider other aspects of the larger concept relevant.
- **Determining the acceptability of summaries (AS).** If the topic statement indicates interest in specific instances (of events, for example), would the user be interested in a general summary?

Going back to the complete example in Figure 1, the clarification questions would be classified as ACF, RT, RTA, CRC, respectively.

The distribution of question types across all topics, as coded by the first author, is also shown in Table 3. It can be seen that RT questions are the most prevalent, followed by EC questions. On the other end of the spectrum, only three AS questions were observed.

Naturally, this is very preliminary work, but it demonstrates that although clarification questions are tailored to a specific topic, broader generalizations can be captured. These results lead to many interesting questions for future work: Given the same data, will another researcher come up with the same or similar categories? Even fixing the ontology, can humans reliably code questions? Is there any correlation between question types and retrieval effectiveness, i.e., are certain types of clarification questions "better" than others? Is the relative distribution of question types specific to this set of topics? Our HARD experiments raise more questions than they answer. Ultimately, we hope that such a clarification ontology can guide the design of systems that automatically initiate clarification dialogs when necessary and appropriately exploit user responses without human

intervention—in short, we want to develop computer systems capable of conducting reference interviews. This, indeed, is a lofty goal, but we have taken a small first step in enriching user-system interactions.

3 Enterprise Track

Informal text genre such as electronic mail (email) have become ubiquitous in recent years, but much of what we know about email search is based on small-scale experiments with private collections or on resource-intensive qualitative study designs that would be costly to replicate or extend. The TREC 2005 enterprise track offered the first opportunity to create a large public test collection for evaluation of content-based email search using World Wide Web Consortium (W3C) mailing lists. The University of Maryland participated in the Known Item Search (KI) and the Discussion Search (DS) tasks. The goal for KI was to find a specific email that the user knows to exist. DS was a more conventional *ad hoc* retrieval task in which the user is searching for arguments in favor or against some point in an email archive. This might be used, for example, to assemble design rationale when considering a change to a previously published W3C standards document. A partially relevant document in the DS task is an email that addresses the specified topic in new (not quoted) text; a fully relevant document in the DS task is an email that contributes at least one pro or con related to the specified topic in new (not quoted) text.[3] Because we performed no specific processing to detect pro or con points, all results reported in this section are scored based on (at least) partial relevance.

In this first year of the track, participating teams performed topic development and relevance judgment. We participated in both activities. A total of 150 KI topics were created, 25 of which were used for training. A total of 60 DS topics were created, one of which was subsequently removed from the collection because no relevant documents were found. No DS training topics were available, so all results in this paper were produced by systems tuned using only the 25 KI training topics. Results from this first year of the enterprise track should therefore be considered preliminary—the main goal in the first year of a new track is to explore the design space for tasks and metrics and to create an initial test collection for use in formative evaluations. As best we can tell, the enterprise track excelled at both tasks in 2005.

3.1 Methods

In our enterprise track experiments, we tested the following ideas: (1) document expansion using adjacent messages in time-ordered threads, (2) inclusion or exclusion of quoted text from earlier messages, and (3) normalization of date and version expressions. Queries were constructed from the "query" field, or from both the "query" and "narrative" fields. The expanded, unexpanded, and normalized collections were indexed with INQUERY (version 3.1p1 for Solaris).

3.1.1 Collection Parsing

The W3C mailing lists were crawled from w3c.org in June 2004. Each message in the collection was embedded in a Web page with extensive HTML/XML markup (usually generated by the hypermail utility program) to format the most important fields from the message for display to end users. We used a Java SAX parser to recover the original RFC-822 header structure and to extract the body text. For messages that the SAX parser failed to parse (such as X-Mail messages), additional processing using Perl scripts was performed. This yielded 174,311 messages with a total size of 515 MB. We distributed our parsed collection to interested participants in the track in order to minimize duplicated effort.

3.1.2 Detection of Quoted Text

Since the goal of the DS task was to find emails that contributed a pro or a con to the topic in new (not quoted) text, quoted text was identified before indexing. Heuristics for tagging text that

[3]TREC 2005 Enterprise Track Guidelines, at http://www.ins.cwi.nl/projects/trec-ent/wiki/index.php/Main_Page

had been automatically copied from an earlier message were developed based on inspection of the collection. Two types of quoted text were tagged: (1) lines that start with > (or | >, " >, "), and (2) lines below text such as "Forwarded message," "Original message," "Mensagem original," "Mensaje original," or "In/On/At (time) (somebody) wrote/writes/said." These heuristics are imperfect, and after we submitted our runs we found one more quotation pattern (lines below "Reply Separator"); additional as-yet undetected patterns probably remain to be found. Regardless, the patterns that we did use provide a reasonable basis for exploring the effect of suppression of (probably duplicated) quoted text within the threads (Wu and Oard, 2005).

3.1.3 Document Expansion with Threads

Email threads have the potential to organize groups of messages around a topic (Lam et al., 2002). We constructed threads automatically based on subject line repetition patterns that are indicative of the use of the reply function in widely-used email clients. This was done by using Lucene[4] to index the text from the subject field with a short stopword list consisting of "re:," "fw:," and "fwd:"), searching the collection with every subject line as a query, removing duplicate results, and then sorting the resulting threads in chronological order (Wu and Oard, 2005). Note that this simple process creates a single sequence rather than the richer tree representation normally associated with threading. The process yielded 22,252 multi-message threads, plus an additional 68,899 "threads" that each consisted of a single message with a unique (or empty) subject line. The mean thread length for multi-message threads was 4.7 messages, with a median thread length of 3 messages. Although threading using subject line reply detection is not perfect, it serves as a useful basis for beginning to explore the effect of thread-based document expansion.

Messages in multi-message threads were then expanded by adding one copy of its chronologically preceding message (if any), one copy of its chronologically successive message (if any), and one additional copy of its own text (to achieve the effect of downweighting the expansion text). We chose this approach because document expansion seeks to enhance the chance of an appropriate lexical match between the query and the indexed terms, while temporal locality would be expected to limit the adverse effect of topic drift that might be more severe over longer periods. Automatically included text that was quoted from earlier messages might be helpful (further upweighting central concepts) or harmful (introducing an additional potential source for topic drift). We therefore tried both variants.

3.1.4 Date and Version Normalization

Various forms of date and software/standard version expressions are present in the collection, and some of the 25 KI training topics included these types of expressions. The tokenization stage of information retrieval systems designed for other tasks will, however, mishandle some such expressions. We therefore automatically normalized the representations of those expressions to enhance the chance of appropriate partial matches between query and document terms. For example, "12-Jan-2000" would be transformed to "day12 January 2000" and "HTML 4.0" would be transformed to "HTML four pointx zero." If a date expression was invalid, such as 14/14/2002, no transformation would be performed. When ambiguous cases arose, such as "1/6/2001" which could be either January day6 2001 in American format or June day1 2001 in European format, the American style was used. Table 4 shows examples of these transformation patterns.

3.2 Relevance Assessment

The participants created 60 topics for the discussion search task. Each topic was assigned to two participating groups to judge relevance in order to support computation of interannotator agreement. The top 50 retrieved documents from the four highest priority runs from each group were pooled. We judged the five topics that we had created (DS22, DS29, DS38, DS42, and DS49) and four others (DS8, DS21, DS32, and DS45). The final author of this paper judged topics DS8, DS21,

[4]Lucene (http://lucene.apache.org/) was used only for threading; all enterprise track experiments were conducted using INQUERY.

Before transformation	After transformation
12-Jan-2000	day12 January 2000
Jan 6-12 or Jan 6 - 12	January day6 - day12
6∼12 January or 6-12 January	day6 - day12 January
6 January 2001	day6 January 2001
Jan 6 2001 or Jan. 6 2001	January day6 2001
2001-01-06*	January day6 2001
2001/01/06 or 2001/1/6	January day6 2001
1/6/2001	January day6 2001
01/06/12	January day6 2001
6/Jan or 06/Jan or 6/January	January day6
jan 2001	January 2001
HTML-4.1	HTML four pointx one
HTML5.1c1	HTML five pointx one c1
J2SDK1.4.1	J2SDK one pointx four pointx one
gcc-make 3.7	gcc-make three pointx seven
gcc-12.7.2.1p1	gcc twelve pointx seven pointx two pointx one p1

*note: no transformation is performed if this pattern appears in a URL.

Table 4: [Enterprise] Examples of transformation patterns.

DS22, DS29, DS45, and DS49. Another information studies graduate student judged topics DS38, DS42, and DS49. According to NIST statistics, there were an average of 529 messages per topic across the 59 pools (ranging from 249–865). Our assessment process required between 3 and 10 hours per topic. The official rule for relevance judgment was that if a message is: (1) on topic and provides pro/con, it is relevant; (2) on topic but provides no pro/con, it is partially relevant. In our practice, if a message was felt to be marginally on topic or if the judge was unsure whether or not it was on topic, it was typically judged as non-relevant.

3.3 Results and Discussion

We submitted five runs for the DS task and four runs for the KI task, and we scored one additional pair of runs locally. For DS, the top 1,000 documents retrieved for each topic were submitted. For KI, the top 100 documents retrieved for each topic were submitted.

TitleDefault (DS) / KIDefault (KI). This is our baseline run, with the messages indexed in their original form (i.e., no suppression of quoted text, no expansion, and no transformation). Queries were constructed from all words in the "query" field, which was intended to model the query that a searcher might have posed. The DS run contributed to the judgment pools.

ThrQuot (DS) / KIThrQuot (KI). This run explored the effect of document expansion. Quoted text was retained, document expansion was performed, and no transformations were applied. Queries were again constructed from all words in the "query" field. The DS run contributed to the judgment pools.

ThrNoQuot (DS) / KIThrNoQuot (KI). This run explored the effect of quoted text on document expansion. Quoted text was removed, document expansion was performed, and no transformations were applied. Queries were again constructed from all words in the "query" field. The DS run contributed to the judgment pools.

ThrNoQNarr (DS). This run explored the sensitivity to richer specifications of information needs. Quoted text was removed, document expansion was performed, and no transformations were applied. Queries were constructed from all words in both the "query" and "narrative" fields.

Run	MAP	BPref	P10	R-Prec
TitleDefault	0.3754	0.3752	**0.5102**	0.4034
ThrQuot	0.2023	0.2411	0.3475	0.2666
ThrNoQuot	0.1972	0.2439	0.3763	0.2663
ThrNoQNarr	0.1742	0.2219	0.2983	0.2516
TitleTrans	**0.3828**	**0.3846**	0.5051	**0.4120**
TitleNewText	0.3474	0.3539	0.4949	0.3793

Table 5: [Enterprise] Comparison with the Baseline (TitleDefault).

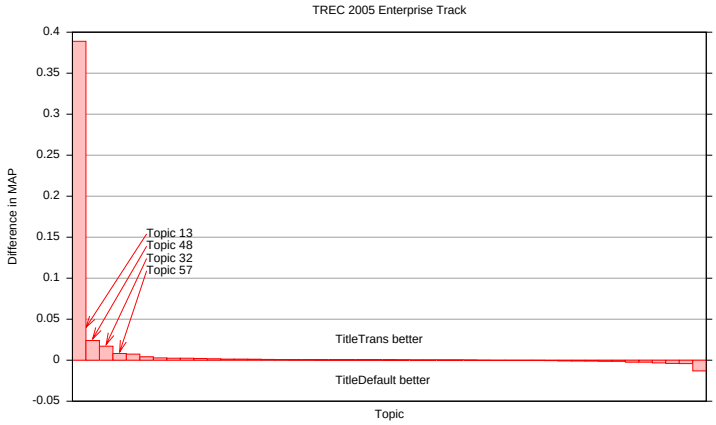

Figure 3: [Enterprise] Difference in MAP between TitleTrans and TitleDefault, partial relevance. Topics are sorted by the difference in MAP (12 identical cases not shown).

No narrative field was available for the KI task, so this run was submitted only for the DS task. The DS run contributed to the judgment pools.

TitleTrans (DS) / KITrans (KI). This run explored the effect of document transformation. Quoted text was retained, no document expansion was performed, and document transformation was applied to normalize date and version references. Queries were constructed from all words in the "query" field and transformed in the same manner. These runs were submitted for official scoring, but the DS run did not contribute to the judgment pools.

TitleNewText (DS) and KITitleNewText (KI). This run explored the effect of quoted text without document expansion. Quoted text was removed, no document expansion was performed, and no transformations were applied. Queries were constructed from all words in the "query" field. These runs was not submitted; they were scored locally.

After receiving the official results and relevance judgments, we found two errors in our submitted runs. Query 51 in the DS task was incorrectly formatted. That error affected all of our DS runs. More seriously, a document formatting error invalidated the results for our baseline system (TitleDefault / KIDefault). We have subsequently corrected these two errors, and the results reported below are based on rescoring those runs locally and are therefore unofficial.

Table 5 shows retrieval effectiveness measures for the DS task. Comparing TitleTrans with TitleDefault, we see substantial improvement on one topic from transforming dates and version numbers, with no substantial impact on any other topics (see Figure 3). Examination of the results on a topic-by-topic basis indicates that when queries included date or version expressions (e.g., topics 13, 32, and 57), transformation was indeed somewhat helpful.

A Wilcoxon signed-rank test for paired samples indicated that the reduction in Mean uninterpolated Average Precision (MAP) between TitleDefault and the four other untransformed conditions

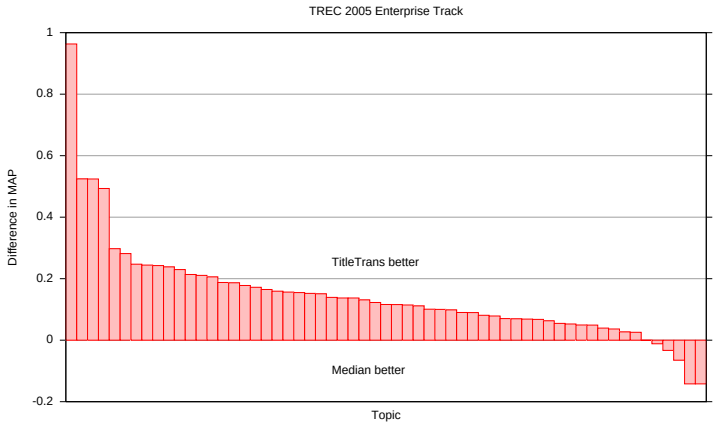

Figure 4: [Enterprise] Difference in MAP between TitleTrans and Median, partial relevance. Topics are sorted by the difference in MAP.

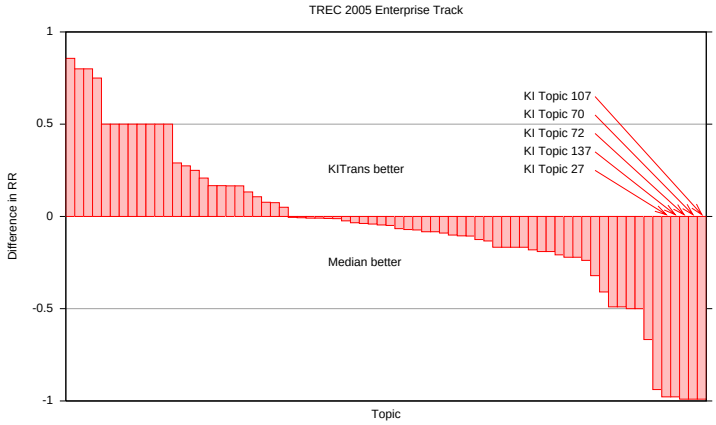

Figure 5: [Enterprise] Difference in Reciprocal Rank between KITrans and Median. Topics are sorted by difference in RR (53 identical cases not shown).

was statistically significant at $p < 0.05$. From this we conclude that expanding documents using chronologically sorted threads in the way that we tried was harmful, regardless of whether quoted text was retained or removed, and that removing quoted text was harmful when searching unexpanded documents. Additional analysis will be needed to characterize the reasons for the reduction of effectiveness that resulted from our narrowly focused way of exploiting automatically detected reply chains.

No statistically significant difference was found between ThrQuot and ThrNoQuot, suggesting that the beneficial effect of reinforcing central terms by including quoted text and the harmful effects of topic drift are fairly closely balanced when threading information is available. Finally, the MAP for ThrNoQNarr is statistically significantly below ThrNoQuot, indicating that the additional information in the narrative field of the topic description did not generally prove to be helpful.

TitleTrans yielded our best results when evaluated using MAP, BPref, and R-precision. Compared with the median across all runs submitted by any team, our TitleTrans run outperformed the median AP for 53 out of 59 topics (see Figure 4). Since that run did not contribute to the judgment pools, BPref may be the more appropriate measure for comparing systems over test collections with incomplete relevance judgments (Buckley and Voorhees, 2004). Our TitleTrans run outperformed the median BPref for 50 of 59 topics, tied with the median for 3 topics, and yielded the best score

7 times. A topic-by-topic analysis for BPref yields a plot very similar to Figure 4. Median results are likely to be biased somewhat low in the first year of a new track, but this analysis suggests that conventional term weighting and document scoring techniques such as those implemented in INQUERY perform reasonably well for this task.

For the Known Item task, KITrans was the best of our 4 runs. As Figure 5 shows, when compared with the median Reciprocal Rank (RR) across all 67 submitted runs on a topic-by-topic basis, there are 47 topics for which KITrans yielded a RR below the median and just 25 topics for which the RR of our KITrans run was above the median. The two cases (i.e., topic 70 and 107) in which we completely missed emails that the median system placed at rank 1 seem particularly notable. Examination of topic KI70 reveals that the target has a small amount of new text, along with a fairly long list of words under "Forwarded message." Our KITitleNewtext run (which removes the quoted text) placed the target message at rank 1, so removing quoted text is clearly helpful in some cases. KITrans missed the target message for topic KI107 entirely, apparently because the KI107 query "access tables on the web" consists of very common terms. In that case, we still had high term frequencies for those terms down to the last submitted position (rank 100). A phrase query might have been helpful in that case. KITrans also had a similar problem with KI137 and KI72. The KI72 query is "workflow management in bioinformatics." Our top ranked messages in that case have high term frequencies for workflow, management, or bioinformatics, which misses the importance of having the three terms together in that case. Finally, the KI27 query is "XML DSig 99." Our top ranked documents have "XML-DSig" in either the "To" or the "Cc" field; restricting matches to the subject and body fields might generally be a good idea (except when a person's name or email address appears in the query).

4 Question Answering Track

Our participation in the question answering track this year focused on evaluation methodology for complex questions, as exemplified by the "other" questions—shorthand for "tell me anything interesting about the target which I haven't already asked." Unlike factoid or list questions, which can be straightforwardly answered by one or more short phrases, it is difficult to even define what the purpose of such questions are, never mind getting different humans to agree on what a "good answer" is.

For the past few years, NIST has employed an evaluation methodology based on the notion of "nuggets", or relevant facts. A nugget answer key (consisting of "vital" and "okay" nuggets) is generated manually after the assessor reads through all system responses. The assessor then uses this answer key to identify relevant nuggets in system responses.

The obvious downside of this approach is that the process requires human intervention. Given the success of automatic evaluation metrics based on n-gram co-occurrences such as BLEU (Papineni et al., 2002) for machine translation and ROUGE (Lin and Hovy, 2003) for summarization, we have recently introduced POURPRE, an automatic metric for scoring answers to complex questions based on many of the same ideas (Lin and Demner-Fushman, 2005). Rankings generated by POURPRE correlate well (Kendall's τ) with official rankings on previous TREC datasets, thus validating its usefulness for automatic system evaluation. We extended the same experiment to this year's "other" questions. Results on all available "other" and "definition" test sets are shown in Table 6 (compared to two ROUGE conditions, default and with stopword removal).

Although POURPRE still outperforms ROUGE on this year's testset, the observed correlations are much lower. We hypothesized that this was due to a large number of questions with very few vital nuggets, causing much sharper quantization in nugget recall, which is weighted much more heavily than precision. To test this hypothesis, we created a variant answer key in which all nuggets were considered vital, and reran our experiments. These results are shown in Table 7; for brevity, only the best POURPRE and ROUGE configurations are shown. In addition, we observed high correlation between the "all vital" score and the official scores.

The vital/okay distinction attempts to capture the intuition that some nuggets are more important than others, but such a binary distinction may be too coarse-grained as reflected in the current calculation of F-score. To better understand if the current evaluation methodology is actually measuring "the right thing", we submitted a run of "other" questions that was manually created by a

Run	POURPRE				ROUGE	
	micro, cnt	macro, cnt	micro, *idf*	macro, *idf*	default	stop
2005 "other" ($\beta = 3$)	0.598	**0.709**	0.679	0.698	0.662	0.670
2004 "other" ($\beta = 3$)	0.785	**0.833**	0.806	0.812	0.780	0.786
2003 "definition" ($\beta = 3$)	0.846	**0.886**	0.848	0.876	0.780	0.816
2003 "definition" ($\beta = 5$)	0.890	**0.878**	0.859	0.875	0.807	0.843

Table 6: [QA] Kendall's τ correlation between rankings generated by POURPRE/ROUGE and official scores.

Run	Official	POURPRE	ROUGE
2005 "other" ($\beta = 3$)	0.853	0.807	0.789
2004 "other" ($\beta = 3$)	0.919	0.868	0.840
2003 "definition" ($\beta = 3$)	0.920	0.915	0.872
2003 "definition" ($\beta = 5$)	0.927	0.910	0.886

Table 7: [QA] Kendall's τ correlation between rankings generated by POURPRE/ROUGE and scores based on variant answer key where all nuggets are considered vital.

human. These results are shown in Figure 6. Surprisingly, our F-score of 0.299 does not appear to be much better than the best-performing automatic run (F-score of 0.248). For 44 questions (out of a total of 75), the median F-score of the automatic systems was zero. For 27 questions, our manual run received an F-score of zero, i.e., no vital nuggets were retrieved. Our manual run beat the median score of automatic systems in only 47 questions, and beat the best automatic system in only 15 questions.

Results from our manual run hold implications for the current evaluation methodology. Assuming that our manual run represents a "good answer", there are at least two possibilities for its low score: the task is not sufficiently well-defined, and that the current evaluation metric does not recognize variations in what constitutes a "good answer". We recommend that the question answering community take a step back to more closely examine these issues, in order to ensure that the TREC evaluations are actually measuring something meaningful.

5 Genomics Track

We directed our main effort in the genomics track into joint work with several teams at the National Library of Medicine (NLM) for the *ad hoc* and text categorization tasks; see (Aronson et al., 2005). Nevertheless, we submitted a Maryland-only *ad hoc* run, primarily to diversify the judgment pool since retrieved documents obtained during system development differed significantly from those of the other NLM base systems.

Topics for this year's *ad hoc* task were divided into five categories, called templates, that covered a broad range of information needs, from specific questions about experimental procedures (e.g., "How to open up a cell through a process called electroporation?"), to very complex questions about the biological impact or role of mutations of a given gene (e.g, "The role of 185delAG mutation in BRCA1 gene in ovarian cancer.").

Our experiments focused on query expansion techniques using INQUERY. The TREC-MEDLINE 2004 collection was indexed without stemming; additional controlled vocabulary terms (MeSH) were also indexed in a separate field. During system development, we tested several query expansion techniques on a small training set of ten topics provided by track organizers (two per template): expanding gene names, e.g., GST fusion protein to Glutathione Transferase; expanding disease names, e.g., ovarian cancer to malignant neoplasm of ovary, ovarian carcinoma, and neoplastic process; adding template-specific terms, e.g., mutation and mutagenesis for template five; and searching for

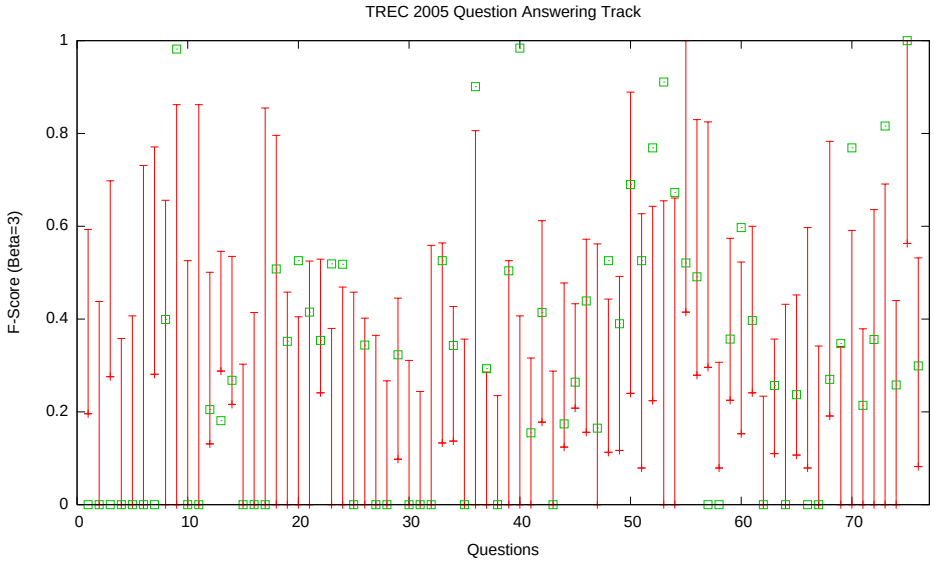

Figure 6: [QA] F-score of "other" questions (bars indicate median/best range). Our manual run is marked with boxes.

template-specific MeSH terms. All expansions were performed automatically using the available domain ontologies and tools provided by the other NLM team members.

Ultimately, we submitted only the run based on expansion of disease names from the University of Maryland. This run, **MARYGEN1**, achieved a MAP of 0.173 and R-precision of 0.195; this puts us below the mean of median per-topic performance across all submissions (0.222 MAP and 0.248 R-precision). For a more detailed description and analysis, please refer to the paper submitted by NLM (Aronson et al., 2005).

6 Conclusion

Overall, we are satisfied with our TREC experience this year, not only in terms of the quantitative results we achieved, but also in terms of the insights we gained about the nature of different types of information-seeking behaviors. Our collaborations across disciplinary boundaries have yielded insights that could not have been obtained with a more homogeneous group, thus affirming the value of an integrated approach to information retrieval.

7 Acknowledgments

This work has been supported in part by DARPA cooperative agreement N660010028910, the Joint Institute for Knowledge Discovery at the University of Maryland, and the National Library of Medicine. The authors wish to thank Erin Greenwell for helping with the enterprise track relevance judgments. The first author would like to thank Kiri for her kind support.

References

Eileen G. Abels. 1996. The e-mail reference interview. *RQ*, 35(3):345–358.

Alan R. Aronson, Dina Demner-Fushman, Susanne M. Humphrey, Jimmy Lin, Hongfang Liu, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, Lorraine K. Tanabe, and W. John Wilbur. 2005. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*.

Marcia J. Bates. 1991. The Berry-Picking search: User interface design. In M. Dillon, editor, *Interfaces for Information Retrieval and Online Systems: The State of the Art*, pages 51–61. Greenwood Press, New Jersey.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*.

Brenda Dervin and Patricia Dewdney. 1986. Neutral questioning: A new approach to the reference interview. *RQ*, 25(4):506–513.

Stephen P. Harter. 1986. *Online Information Retrieval: Concepts, Principles, and Techniques*. Academic Press, San Diego, California.

Derek Lam, Steven L. Rohall, Chris Schmandt, and Mia K. Stern. 2002. Exploiting e-mail structure to improve summarization. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW 2002)*.

Jimmy Lin and Dina Demner-Fushman. 2005. Automatically evaluating answers to definition questions. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2003)*.

Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, England.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.

Robert S. Taylor. 1962. The process of asking questions. *American Documentation*, 13(4):391–396.

Yejun Wu and Douglas Oard. 2005. Indexing emails and email threads for retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 665–666.