

# Exploring the Limits of Single-Iteration Clarification Dialogs

Jimmy Lin<sup>1,2,3</sup>, Philip Wu<sup>1</sup>, Dina Demner-Fushman<sup>2,3</sup>, and Eileen Abels<sup>1</sup>

<sup>1</sup>College of Information Studies

<sup>2</sup>Department of Computer Science

<sup>3</sup>Institute for Advanced Computer Studies

University of Maryland

College Park, MD 20742, USA

{jimmylin,fwu,eabels}@umd.edu, demner@cs.umd.edu

## ABSTRACT

Single-iteration clarification dialogs, as implemented in the TREC HARD track, represent an attempt to introduce interaction into *ad hoc* retrieval, while preserving the many benefits of large-scale evaluations. Although previous experiments have not conclusively demonstrated performance gains resulting from such interactions, it is unclear whether these findings speak to the nature of clarification dialogs, or simply the limitations of current systems. To probe the limits of such interactions, we employed a human intermediary to formulate clarification questions and exploit user responses. In addition to establishing a plausible upper bound on performance, we were also able to induce an “ontology of clarifications” to characterize human behavior. This ontology, in turn, serves as the input to a regression model that attempts to determine which types of clarification questions are most helpful. Our work can serve to inform the design of interactive systems that initiate user dialogs.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Search Process*

## General Terms

Measurement, Experimentation

## Keywords

interactive retrieval, intermediated search, TREC HARD

## 1. INTRODUCTION

The one-shot model of information retrieval operationalized in TREC evaluations represents an oversimplification of real-world information-seeking behavior, and has often been criticized for neglecting the important role of interaction.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '06, August 6–10, 2006, Seattle, Washington, USA.

Copyright 2006 ACM 1-59593-369-7/06/0008 ...\$5.00.

Indeed, there is empirical evidence that gains in one-shot retrieval performance, as measured by metrics such as mean average precision, do not necessarily translate into higher effectiveness in interactive settings [3, 11]. However, test collections created as a result of TREC evaluations represent an affordable and repeatable way to measure *one* salient characteristic of retrieval systems. On the other hand, full-blown user studies, while potentially more insightful due to higher-degrees of realism, suffer from problems with affordability: the high-cost and time-consuming nature of such experiments limit the range of hypotheses that can be considered, the speed at which variables can be explored, and the statistical significance of results. The TREC HARD tracks grew out of the recognition that there are other interesting regions in the solution space of evaluation design which encode different tradeoffs between insightfulness, affordability, and repeatability of experiments.

The TREC HARD track [2] was originally conceived with a focus on three different ideas: richer specifications of information needs (i.e., context), finer-grained units of retrieval (i.e., passages), and limited interaction with the user (i.e., clarification dialogs). It was decided that clarification dialogs represented the most promising avenue in which to advance the state of the art, and they were retained as the sole focus in 2005. Clarification dialogs represent an attempt to reduce both the scope and duration of user–system interactions to allow practical implementation in large-scale evaluations. Instead of arbitrarily complex interface controls, interactions were limited to what can be conveyed on an HTML page—checkboxes, text input forms, and the like. Instead of repeated iterations, interactions were limited to one single cycle and bounded at three minutes in duration. Due to these factors, the HARD track had a more complex evaluation cycle: participants first submitted a set of baseline runs, along with a set of HTML forms that constituted the clarification dialogs. NIST then presented these pages to assessors, who interacted with them accordingly. The results of these interactions, captured via the CGI protocol, were sent back to the participants, who then created final runs. The goal was to compare pre- and post-clarification runs to quantify the performance gains that can be attributed to user feedback.

Explorations of clarification dialogs have proven to be challenging because most previous experiments confounded the effects of the content of clarification forms, the manner in which they are generated, and the manner in which re-

sponses are exploited. Previous results have shown little improvement in terms of standard ranked-retrieval metrics [2]. Does this finding reveal fundamental limitations about clarification dialogs, or are present systems simply unable to effectively capitalize on such interactions?

This paper describes our experiments and subsequent analysis for the TREC 2005 HARD track. To better understand the solution space of single-iteration clarification dialogs, we employed a trained intermediary to gather potentially relevant documents, construct clarification forms, and exploit user responses. 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 designing 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 headway toward achieving all of these goals. Our HARD experiments begin to quantify the added-value of having a human in the loop and the possible performance gains (Sections 3, 4, and 5). Through post-hoc analysis of clarification questions, we have discovered commonly-occurring patterns of intent, which serve as one possible ontology of clarifications (Section 6). Regression analysis reveals which types are most helpful. We also discuss how these insights can translate into system strategies for initiating and exploiting user–system dialogs (Section 7).

## 2. CLARIFICATION DIALOGS

Information need negotiation in the context of a reference interview within a library setting is a complex communication between a specialist and a user [20], 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.

Like information need negotiation in reference interviews, 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 initial 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 involve search results, whereas, in most cases, little is known about actual documents during the reference interview. Pre-search information need negotiation is essential in reference transactions because it determines *what* resources to search, which depends, for example, on answer requirements such as the format of expected results. This element of source selection has been eliminated in TREC.

As users become more accustomed to searching online electronic resources, the face-to-face reference interview is gradually being 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 depend primarily on text-based interfaces for soliciting user input. Hence, we can benefit from previous work on the email reference interview. Abels [1] 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 covering all aspects of the topic, arranged in a coherent, logical manner.

The email interview provides a real-world grounding for HARD clarification dialogs. These interactions are not merely stilted attempts at introducing elements of user studies in traditional *ad hoc* tasks, but represent realistic abstractions of information-seeking behavior. Given this understanding, we can recast the purpose of the HARD track as a means to better understand the systematic approach to asynchronous need negotiation so that such dialogs can be automatically conducted by a computer. While we acknowledge the work that has been done on analyzing real-time reference encounters (e.g., [15, 17, 18, 23]), asynchronous clarification dialogs represent a qualitatively different form of interaction.

## 3. METHODOLOGY

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

The intermediary employed the “building blocks” strategy [8, 14] with INQUERY 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 when appropriate (e.g., as a source for additional query terms). These facets were then systematically combined into complete queries, most often with INQUERY’s “soft” AND operator. Schematically, a building blocks query looks like:

$$(A_1 \vee A_2 \vee A_3 \vee \dots) \wedge (B_1 \vee B_2 \vee B_3 \vee \dots) \wedge \dots$$

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.). In addition to adding or replacing query terms based on the top hits, the negation operator was frequently used to disambiguate concepts. For most of the topics, the intermediary went through a few rounds of query reformulating until a “good query” was constructed. Following that, between 50 and 120 documents in the resulting hit list were manually examined; the actual number depended on the difficulty of the topic, the number of relevant hits, and other factors. Each examined document was assigned one of four judgments (cf. [18]):

- **Centrally relevant (CR):** based on the intermediary’s understanding of the information need, this document would be considered topically relevant.

<sup>1</sup>Philip Wu (the second author), a Ph.D. student in College of Information Studies at Maryland.

- **Peripherally relevant (PR)**: based on the intermediary’s understanding of the information need, this document would be considered relevant, but less so than documents marked centrally relevant (for example, a passing mention or a vague reference).
- **Maybe relevant (MR)**: based on the intermediary’s 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 of the topic, this judgment is assigned.
- **Not relevant (NR)**: this document would not be considered relevant.

A total of three pre-clarification runs were created automatically using the relevance judgments provided by 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 B2. Our main run, B1, consisted of CR, PR, and MR documents (in that order), followed by documents in B2 (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 (B3), we submitted an INQUERY run that used title and description terms as the query, with blind relevance feedback (top 20 *tf.idf* terms from top 10 hits).

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 or the NR pile.** Although research in information science recognizes relevance as a graded quantity [16], TREC assessors are ultimately forced to make much coarser-grained relevance judgments. 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 or the NR pile.** In searching, the intermediary makes judgments based on an interpretation of the information need; 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 adopted an inductive, bottom-up approach. Thus, the intermediary formulated 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 to ensure a consistent user interaction pattern. In addition to topic-specific questions, all clarification forms included two generic questions (located at the end of the form): “Any additional search terms?” and “Any other

comments?” Both were followed by a  $70 \times 4$  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 a more refined understanding of the information need. In addition, the intermediary performed another round of search for several difficult topics. New results were examined and relevant documents were added to CR pile. Creation of our main post-clarification run followed exactly the same procedure as the creation of our pre-clarification run, except with updated piles (run C1). In addition, we submitted two contrastive conditions: C2 used title and description terms from the topic, along with search terms supplied by the user in the clarification forms. The run C3 included additional blind relevance feedback terms to the run C2.

## 4. RESULTS

We submitted a total of 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 (out of 50 total), the B1 pre-clarification run achieved the best mean average precision across all submitted runs; for R-precision, 28 topics. For 20 topics, the C1 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 searching and assessing documents to create the pre-clarification document piles (max 170, min 35,  $\sigma = 29.7$ ). This time included 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 time statistics for the process of exploiting clarification responses, but reassessing the documents 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 the topic statements to be straightforward. 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 additional terms or phrases per topic ( $\sigma = 3.31$ ), with a maximum of fourteen.

Looking at the difference between B1 and C1, it appears that clarification had only a small impact on MAP (+3.8%)

<b>Title:</b> Three Gorges Project
<b>Description:</b> What is the status of The Three Gorges Project?
<b>Narrative:</b> 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.
<b>Clarification Questions</b>
1. <input type="checkbox"/> 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. <input type="checkbox"/> Check if yes: Is “early next century” an acceptable projected date of completion?
3. <input type="checkbox"/> Check if yes: Would articles mentioning state bank loans or foreign investment be relevant?
4. <input type="checkbox"/> 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: Sample topic and clarification questions.

	B1	B2	B3	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
	C1	C2	C3	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

**B1:** CR+PR+MR+rel feedback run

**B2:** rel feedback run

**B3:** title+desc+brf

**C1:** same as **B1**, with updated piles

**C2:** title+desc+user-supplied terms

**C3:** title+desc+brf+user-supplied terms

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

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. 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 B1 and C1, respectively. The rightmost set of bars represents an average across all topics.

The impact of user-supplied terms can be seen in the performance differences between runs B3 and C3, where title, description, and blind relevance feedback terms were expanded with user-supplied terms from the clarification forms. This resulted in a 4.4% improvement in MAP (significant at the 5% level).

## 5. ANALYSIS

A topic-by-topic analysis focused on runs B1 and C1 revealed ways in which the clarification dialogs helped or hurt. 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.<sup>2</sup> Considering this reduced

<sup>2</sup>Based on our methodology, topics with no clarification questions

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 0.405 to 0.643, 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 0.466 to 0.309). 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

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.

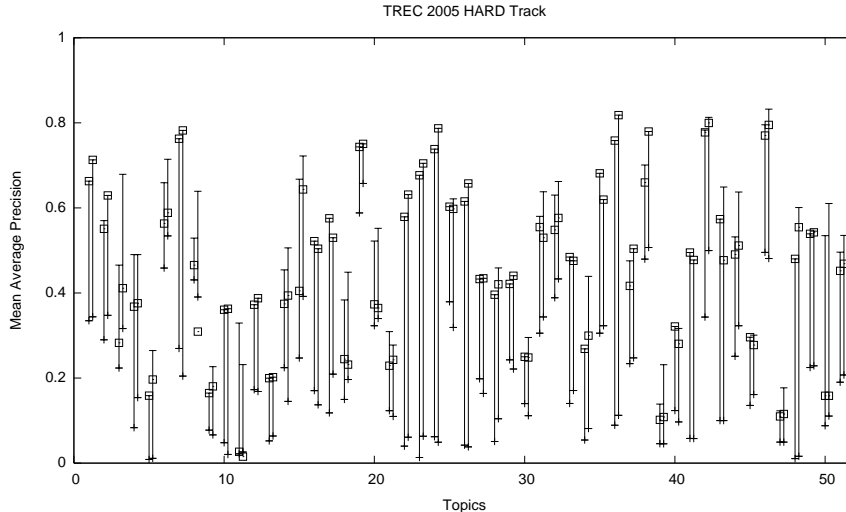


Figure 2: Comparison of mean average precision on a per-topic basis. Each pair of bars represents the median/best score range, before and after clarification. B1 and C1 scores are marked with boxes.

testset	All	A	B
# q's	50	40	32
MAP before	0.452	0.465	0.482
MAP after	0.469 (+3.8%)	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: Effects of clarification on different subsets of topics (subset A does not include topics without clarification questions; topics that exhibited the “inconsistent user” phenomenon are further excluded in subset B).

this the “inconsistent user” phenomenon.

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 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 actually clouded the intermediary’s understanding of the user’s need, mainly due to poorly-formulated clarification questions. One question, 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? Ruling out malicious intent, there are at least two possibilities: one points to a methodological flaw, and the other to the nature of information-seeking behavior itself.

Due to real-world constraints involved in coordinating the HARD track, documents were not assessed until approxi-

mately 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 [4, 20]. 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

Type	Topic	Example Clarification Question	Freq.
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: An ontology of clarification questions, with examples and frequencies of occurrence.

to the underlying need. Thus, “neutral” questioning is preferred in reference interviews so that the questions posed do not lead to biased responses [7].

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 these two issues. More carefully-constructed experiments must be conducted to better understand the shifting nature of information needs.

## 6. 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 centrally relevant or not relevant piles. Questions targeted at the peripherally relevant documents form a coherent class in terms of intent:

- **To determine the relevance threshold (RT).** We hypothesize the existence of a “relevance threshold” that guides the user in making granular judgments. Clarification questions of this type attempt to better understand this boundary in “relevance space”.

Other clarification questions fall naturally into five categories, discussed below. An example of each is found in Table 3.

- **To determine the relationship between ambiguously conjoined facets (ACF).** In most cases, information needs are composed of multiple conceptual facets, and in some cases, topic statements actually express more than one distinct information need. Often, the relationship between these facets is unclear, e.g., does a document need to contain all of the facets to be considered relevant?
- **To determine the relevance of an example concept (EC).** Is a particular concept present 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, etc. For the most part, they make poor query terms, and need to be operationalized in a particular context.

- **To determine 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.
- **To determine 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.
- **To determine the acceptability of summaries (AS).** If the topic description indicates interest in specific instances (of events, for example), would the user be interested in a general summary or overview (e.g., aggregate statistics)?

Returning to the example in Figure 1, the clarification questions would be classified as ACF, RT, RTA, CRC, respectively. The distribution of clarification types across all topics, as coded by the first author, is also shown in Table 3.

### 6.1 Regression Analysis

Given this ontology, can we determine the effectiveness of different clarification questions? We attempted to answer this question by constructing a linear regression model where the number of clarification questions in each category served as the independent variables (predictors) and the relative difference in MAP served as the dependent variable. We fixed the intercept of the regression model to zero, since asking no questions should yield no score difference. We used the 32 topics denoted as testset B in Table 2, which does not contain topics without clarification questions or topics that exhibited the “inconsistent user phenomenon”.

Overall, our regression model was statistically significant, with an  $R^2$  value of 0.66 (adjusted  $R^2$  of 0.56). Regression coefficients for each variable are shown in Table 4, along with their  $p$ -values (for convenience, the frequency of each clarification question type is also shown). Because the number

Type	Freq.	$\beta$	$p$ -value
RT	31%	0.025	0.19
ACF	10%	0.010	0.85
EC	22%	0.034	0.12
CRC	14%	0.106	< 0.01
RTA	19%	$\sim 0$	0.99
AS	3%	0.214	<< 0.01

Table 4: Regression based on clarification types.

of observations (topics) is smaller than one would normally expect for this type of analysis, these results should be taken as indicative, not conclusive. Positive values for all regression coefficients confirm our expectation that asking clarification questions correlates positively with increased MAP (to different degrees). Of all categories, AS (acceptability of summaries) and CRC (closely-related concepts) were found to be statistically significant predictors, and have the largest regression coefficients—meaning that they seemed to be the most helpful type of clarification questions to ask. However, both question types combined account for less than 20% of all clarification questions. It is interesting that RT (relevance threshold) questions are only moderately helpful, despite their prevalence. This suggests that users’ relevance profiles are rather difficult to probe, and attempts to do so are of limited effectiveness.

## 7. DIALOG STRATEGIES

Consistent with much previous work in library science, our results show that facet analysis can be a very effective tool for decomposing users’ information needs and can serve as the basis for a dialog strategy. In this section, we discuss how systems can operationalize these insights to facilitate richer user–system interactions.

Document clustering (e.g., [10, 13]) is one possible way in which facet analysis can be implemented. By noting the distribution of terms, both in the query and in the resulting hits, it may be possible to automatically categorize query terms into conceptually-related groups. This then provides an interaction opportunity for the user to either confirm or further manipulate results of the system’s analysis. Hearst [9] has shown that imposing simple constraints between sub-topics (in a two-pass retrieval scheme) can have a beneficial impact on precision, and this technique represents one simple way to exploit user responses. In further support of this analysis, Buckley and Harman [5] have discovered that missing facets are a common failure mode of many retrieval systems. Clustering also provides a concrete method for generating RTA (related topical aspects) questions—by identifying groups of documents that share terms both with the query and also with each other. In a way, such techniques as scatter–gather [10] can be viewed as a method for browsing related topic aspects.

In general, many types of clarification questions speak for the need to better model linguistic and ontological relations in queries and documents. For example, syntactic analysis can help determine if query terms should be related by conjunction or disjunction. Systems could make this decision directly if there is sufficient evidence, but there remains the opportunity to ask ACF (ambiguously conjoined facet) questions. Conjunctions and disjunctions are merely examples of

important linguistic relations that may be present in topic statements. Within the field of question answering, syntactic analysis has yielded significant improvements in performance [6], for example, differentiating between the killer of John Wilkes Booth from the man John Wilkes Booth killed (which have identical “bag of words” representations after stemming and dropping stopwords). In *ad hoc* retrieval, term relations often hold the key to high accuracy; for example, in the topic about “human smuggling”, the modification relationship between the two terms is critical, especially in a corpus that contains many articles about smuggling in general (term proximity is no substitute in this case).

Linguistic relations between query terms and terms in the result set are also important; two categories of clarification questions in our ontology (CRC and EC) specifically speak to such relations. Many of them are ontological in nature, and resources such as WordNet can be brought to bear to assist in the retrieval process. Although previous experiments with techniques such as query expansion using lexical semantic relations have yielded at best marginal improvements (e.g., [21]), they were attempted for the most part without human intervention. Placing the human in the loop has been shown to be an effective method for weeding out bad choices [12]. In a query expansion setting, linguistic and ontological relations can be employed to select candidate terms, as opposed to using only term frequencies; such interactions are, in fact, examples of CRC and EC questions.

In summary, we believe that our ontology of clarifications can be used to guide the design of interactive IR systems. In the cases where similar techniques already exist, the ontology provides a deeper explanation of *why* they work and *how* they can be improved.

## 8. LESSONS LEARNED

What have we learned about the nature of interactive retrieval and clarification dialogs through our human–in–the–loop experiments? Somewhat to our relief, we demonstrated that human involvement significantly improves IR performance. This is by no means an obvious outcome, considering that “manual runs” in previous TREC evaluations were not considerably better than fully-automated runs (e.g., [22]). The crucial difference is the element of interaction—previous manual runs, for the most part, consisted of single-shot retrieval with human-constructed queries. This setup is unrealistic in that an information-seeker does not attempt to optimize the result set from a single interaction, but rather culls relevant information from multiple iterations.

This work describes an effective strategy for leveraging human expertise and strengths of automated retrieval systems in intermediated searching. Studies of search strategies using modern ranked-retrieval systems and how to best combine human and system results represent an underexplored area in the IR literature, and our conception of need clarification as a process of creating and “shuffling” piles is novel, as far as we are aware. We have gained a better understanding of single-iteration clarification dialogs, and our experiments show that significant gains are possible, even when compared to a strong baseline. Further analysis revealed an interesting phenomenon where users’ responses to clarification questions do not appear to be consistent with their relevance judgments—more work is required to assess the significance of this finding.

Finally, this paper provides insights into three impor-

tant questions: Can clarification questions be classified into meaningful categories? Are some types more useful than others? How might a system initiate dialogs automatically? Although elicitations during face-to-face reference encounters have previously been analyzed and categorized [15, 17, 19, 23], asynchronous clarification dialogs are qualitatively different in nature. Furthermore, the setup of the HARD track allows us to isolate and focus on topical aspects of need negotiation; many of the previously cited works had to contend with categories of communications that did not relate directly to the information need, e.g., printer setup. Admittedly, some amount of realism is sacrificed, but in other ways, this work goes beyond previous studies in that we categorize the ways in which conceptual facets within an information need can be clarified. Finally, the tight coupling between the clarification dialog and system output allows us to quantitatively measure the effect of the interaction and understand the contributions of each question type.

Nevertheless, there are a number of unresolved issues relating to our clarification ontology worth mentioning. The categories presented here sprung from the mind of one single individual; given the same data, would another person come up with similar categories? Even with a fixed ontology, can humans reliably code questions? Even if the ontology is stable and questions can be reliably coded, are there other influencing factors, e.g., type of information need, domain, genre, etc.? For example, AS (acceptability of summaries) questions were found to be very useful, but the suitability of asking such questions would depend on the corpus (some collections have more summaries than others). In the same vein, there are perhaps natural associations between categories of clarification questions and types of information needs. Results from this study supply the foundation for future work on the interaction between information seekers, intermediaries, and systems.

## 9. CONCLUSION

The primary purpose of this paper was to explore the limits of single-iteration clarification dialogs, which can be grounded in asynchronous need negotiation. By involving a human intermediary in our TREC 2005 HARD experiments, we determined a plausible upper bound on retrieval effectiveness and gained a better understanding of such interactions. Furthermore, this work has yielded an “ontology of clarifications”, which can be leveraged to guide the development of future systems. We recognize the limitations of this present work, but are excited about the new frontiers in interactive information retrieval that have been identified.

## 10. ACKNOWLEDGMENTS

This work has been supported in part by DARPA cooperative agreement N66001-00-2-8910 and contract HR0011-06-2-0001 (GALE). We’d like to thank James Allan for organizing the HARD track and Doug Oard for various engaging discussions. The first author would like to thank Esther and Kiri for their loving support.

## 11. REFERENCES

- [1] E. Abels. The e-mail reference interview. *RQ*, 35(3):345–358, 1996.
- [2] J. Allan. HARD track overview in TREC 2003/2004/2005: High accuracy retrieval from documents. In *TREC 2003/2004/2005*.
- [3] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be “good enough”? User effectiveness as a function of retrieval accuracy. In *SIGIR 2005*.
- [4] M. Bates. 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, 1991.
- [5] C. Buckley and D. Harman. Reliable information access final workshop report, 2004.
- [6] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua. Question answering passage retrieval using dependency relations. In *SIGIR 2005*.
- [7] B. Dervin and P. Dewdney. Neutral questioning: A new approach to the reference interview. *RQ*, 25(4):506–513, 1986.
- [8] S. Harter. *Online Information Retrieval: Concepts, Principles, and Techniques*. Academic Press, San Diego, California, 1986.
- [9] M. Hearst. Improving full-text precision on short queries using simple constraints. In *Proceedings the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR 1996)*, 1996.
- [10] M. Hearst and J. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *SIGIR 1996*.
- [11] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR 2000*.
- [12] J. Koenemann and N. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *CHI 1996*.
- [13] D. Lawrie, W. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *SIGIR 2001*.
- [14] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, England, 1995.
- [15] R. Nordlie. “User revelation”—a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *SIGIR 1999*.
- [16] T. Saracevic. Relevance: A review of and a framework for thinking on the notion in information science. *JASIS*, 26(6):321–343, 1975.
- [17] A. Spink, A. Goodrum, D. Robins, and M. Wu. Elicitations during information retrieval: Implications for IR system design. In *SIGIR 1996*.
- [18] A. Spink and T. Saracevic. Interaction in information retrieval: Selection and effectiveness of search terms. *JASIS*, 48(8):741–761, 1997.
- [19] K. Swigger. Questions in library and information science. *Lib. and Info. Sci. Research*, 7:369–383, 1985.
- [20] R. Taylor. The process of asking questions. *American Documentation*, 13(4):391–396, 1962.
- [21] E. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR 1994*.
- [22] E. Voorhees and D. Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *TREC-8*.
- [23] M. White. Questions in reference interviews. *J. of Documentation*, 54(4):443–465, 1998.