

Building a Reusable Test Collection for Question Answering[†]

Jimmy Lin¹ and Boris Katz²

¹College of Information Studies
Institute for Advanced Computer Studies
University of Maryland
jimmylin@umd.edu

²Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
boris@csail.mit.edu

Abstract

In contrast to traditional information retrieval systems, which return ranked lists of documents that users must manually browse through, a question answering system attempts to directly answer natural language questions posed by the user. Although such systems possess language processing capabilities, they still rely on traditional document retrieval techniques to generate an initial candidate set of documents. In this paper, we argue that document retrieval for question answering represents a different task than retrieving documents in response to more general retrospective information needs. Thus, to guide future system development, specialized question answering test collections must be constructed. We have shown that the current evaluation resources have major shortcomings, and to remedy the situation, we have manually created a small, reusable question answering test collection for research purposes. This article describes our methodology for building this test collection and discusses issues we encountered along the way regarding the notion of “answer correctness”.

1. Introduction

Question answering (QA) is an exciting and emerging field of research that lies at the intersection of computational linguistics and information retrieval. In contrast with traditional document retrieval systems, which return ranked lists of potentially relevant documents that users must then manually browse through, question answering systems attempt to directly provide users with answers to natural language questions. Although a broad range of information needs can often be stated as a question, the field of question answering is substantially narrower and focuses on a few specific question types. Much current research focuses on fact-based questions, whose answers are typically named

[†] The final version of this article appears as:

Jimmy Lin and Boris Katz. Building a Reusable Test Collection for Question Answering. *Journal of the American Society for Information Science and Technology*, 57(7):851-861, 2006.

Citations to and quotations from this work should reference that publication. If you cite this work, please check that the published form contains precisely the material to which you intend to refer.

entities such as dates, locations, proper nouns, or other short noun phrases. The following are a few examples of these so-called “factoid” questions:

When was chewing tobacco banned in baseball?
What membrane controls the amount of light entering the eye?
Who was responsible for the killing of Duncan in "Macbeth"?
How many floors are in the Empire State Building?

Other types of questions receiving attention include: “list” questions such as “What countries export oil?”, which are similar to factoid questions but have multiple answers; biographical questions such as “Who is Aaron Copland?”, which require a system to gather relevant encyclopedic facts about a person from multiple documents; and other complex questions that require reasoning and fusion of multiple “information nuggets” from different sources. Because answering these types of questions requires a core set of common capabilities, many of which are exercised by factoid questions, these relatively simple information requests have remained a staple of question answering research.

This present work focuses on factoid questions and explores the relationship between document retrieval and question answering. Because question answering seeks to retrieve finer-grained segments of information, it requires an understanding of both the natural language question and the documents within the target collection; in contrast to traditional document retrieval systems, which rely on statistics such as term frequency and inverse document frequency (Robertson, 2004), question answering systems typically employ natural language processing technology. Despite the use of more sophisticated query and document processing, question answering systems nevertheless retain document retrieval as an integral component because it significantly reduces the number of documents that must be analyzed in detail. However, we argue that the task of retrieving documents in response to a specific natural language question involves a different set of requirements and tradeoffs than retrieving documents in response to a more general retrospective information need (the *ad hoc* retrieval task).

Experimental studies with test collections represent an important paradigm in modern information research. Reusable test collections allow researchers to quickly conduct reproducible experiments that compare the effectiveness of different retrieval techniques, and are a central driving force in advancing the state of the art. In this paper we will show that currently available resources are inadequate for evaluating document retrieval in the context of question answering and should not be used to guide system development. To address this shortcoming, our group has created a reusable test collection for factoid question answering by manually gathering judgments for 110 questions on the AQUAINT corpus, a collection of approximately one million articles from the Associated Press, the New York Times, and Xinhua English News collected from 1998 to 2000. In the process, we gained many insights about the subtle factors that influence a human’s judgment of answer correctness; these valuable lessons can be applied to improve future question answering systems.

This paper is organized as follows: the next section discusses a generic question answering architecture and introduces the task of document retrieval for question answering, distinguishing it from *ad hoc* retrieval. Section 3 describes currently available resources for evaluating question answering systems, created from results of the TREC QA tracks, and demonstrates why they are inadequate. One solution is to manually create a reusable test collection—an approach we describe in Section 4. In the process of building our test collection, we contended with human variations regarding the interpretation of answer correctness; these issues are discussed in Section 5. In the penultimate section, we compare our manually-created test collection with existing resources. This article concludes with a summary of our contributions.

2. Document Retrieval for Question Answering

Functionally, most factoid question answering systems today can be decomposed into four major components (see Figure 1): question analysis, document retrieval, passage retrieval, and answer extraction (cf. Hirschman and Gaizauskas, 2001; Voorhees, 2001). The question analysis component classifies user questions into the expected semantic type of the answer. Typical approaches include the use of heuristic rules (Hovy et al. 2001) and machine learning techniques (Li and Roth, 2002), both of which may refer to a custom question type hierarchy or existing resources such as WordNet (Harabagiu et al., 2000). As an example, the expected answer type of the question “Where was Kennedy assassinated?” is *location*. The question analysis module is often also responsible for formulating one or more queries to a document retriever; these queries are used to fetch a set of potentially relevant documents from the corpus. From these documents, the passage retrieval component selects a handful of paragraph-sized fragments for subsequent analysis. Most often, passage retrieval algorithms perform a density-based weighting of query terms, i.e., they favor query terms that appear close together (see Tellex et al., 2003 for a survey). The insight is that answers to a question are likely to occur in extents containing many closely-clustered query terms while documents containing all query terms spaced far apart are less likely to contain the answer. In some systems, document and passage retrieval are performed simultaneously (e.g., Clarke et al., 2000). Finally, the answer extraction module searches the passages for the actual answers. The basic strategy is to find named entities that match the expected answer type (Srihari and Li, 1999), although Light et al. (2001) has shown this method to be insufficient. Beyond simple matching of named entities, answer extractors may also employ more advanced linguistic processing technology, such as matching syntactic relations from the questions with those from the corpus (Katz and Lin, 2003) or attempting to “justify” the answer using an abductive proof (Harabagiu et al., 2000). In general, knowledge-based approaches (e.g., Prager et al., 2000), Web-based techniques (e.g., Brill et al., 2001), and statistical methods (e.g., Echihabi and Marcu, 2003) are well represented in question answering systems.

In a typical question answering system, a document retriever is employed to produce a candidate set of documents for further linguistic processing. This is primarily done for expediency, as the speed of natural language processing techniques limits the amount of text that may be realistically processed at query time. A two-stage approach that first employs traditional document retrieval techniques to gather candidate texts, followed by

more detailed linguistic analysis, has proven to be a reasonable tradeoff. Although a document retriever is often an integral component of a question answering system, we argue that document retrieval for the purposes of answering short natural language questions represents a different task than *ad hoc* retrieval, with a distinct set of requirements.

One obvious difference between *ad hoc* retrieval and question answering is the nature of the information need. Whereas question answering deals with short, concrete requests for specific answers, *ad hoc* “topics” usually focus on more general and complex information needs:

Is it hazardous to the health of individuals to work with computer terminals on a daily basis?

Relevant documents would contain any information that expands on any physical disorder/problems that may be associated with the daily working with computer terminals. Such things as carpal tunnel, cataracts, and fatigue have been said to be associated, but how widespread are these or other problems and what is being done to alleviate any health problems.

Relevant documents in the *ad hoc* retrieval task are generally “about” the material outlined in the statement of information need. In contrast, answers to natural language questions are often “hidden” in documents that have little overall bearing to the topic, because answers are localized to small regions within a document. For example, the answer to “How many floors are in the Empire State Building?” may be found in an article about the Great Depression, which only mentions the answer incidentally. Since many models of information retrieval are fundamentally based on similarity between a query and documents in a collection, we may have to reexamine this assumption in the context of question answering.

Another immediate striking difference between *ad hoc* retrieval and document retrieval for question answering is the length of the information request: *ad hoc* topics are generally much longer than specific natural language questions. Typically, these topic statements include both a sentence-long description of the information need and a paragraph-long narrative that elaborates on background and context, and often includes a description of what types of documents should be considered relevant and irrelevant. Thus, queries generated from *ad hoc* topics are usually longer than corresponding queries for a question answering task. Accumulated experience in information retrieval research has shown that long queries behave quite differently than short queries, representing another difference between traditional *ad hoc* retrieval and document retrieval for question answering.

For the reasons discussed above, effective document retrieval techniques for the *ad hoc* task might not be effective for question answering. We cannot apply previously studied retrieval techniques and hope to produce similar improvements without first reconsidering the differences between the two tasks. A case in point is blind relevance

feedback: in traditional document retrieval, this technique has consistently proven to be beneficial, as measured by a variety of metrics such as mean average precision. Monz (2003), however, has shown that blind relevance feedback significantly decreases the number of answer-bearing documents that are retrieved in a question answering task. Blind relevance feedback is effective in raising retrieval performance because the technique augments the original query with terms drawn from potentially relevant documents, i.e., the new query with feedback terms becomes more similar to relevant documents. However, as we have discussed, document-level similarity may not necessarily correlate with presence of an answer. Another issue that Monz has studied is the effect of stemming: whereas previous studies in *ad hoc* retrieval have reported mixed results regarding its impact on precision and recall (Harman, 1991; Krovetz, 1993; Hull, 1996), he demonstrates clear precision and recall improvements that can be directly attributed to stemming document terms.

Retrieval emphasis marks another divergence between *ad hoc* retrieval and document retrieval for question answering. In the *ad hoc* task, systems place roughly equal weight on precision and recall, and metrics such as mean average precision reflect the need to balance the precision/recall tradeoff. However, recall is more important than precision in document retrieval for question answering, since obtaining a ranked list of documents is merely the first step in the question answering process. In a pipelined question answering architecture, irrelevant documents can be filtered by downstream modules, which may have access to more linguistic knowledge and better reasoning capabilities. Relevant documents that are not returned by a document retriever, however, pose serious problems. If a document containing the answer is not retrieved in the first place, then no amount of intelligent processing by subsequent modules will matter. For other types of natural language questions, e.g., list questions such as “What countries export oil?”, recall is even more important because multiple answers are desired.

We believe that document retrieval for question answering is an important task that can be fruitfully studied in isolation. Although it is only the first step in answering natural language questions, we assume that technology for retrieving a better ranked list of documents will prove beneficial to other processing modules.

3. The TREC Question Answering Tracks

Over the past few years, the question answering tracks at the Text Retrieval Conferences (TRECs) (Voorhees and Tice, 1999, 2000, 2000a; Voorhees, 2001, 2002, 2003), sponsored by the National Institute of Standards and Technology (NIST), have brought formal and rigorous evaluation methodologies to bear on the question answering task: features include blind testsets, shared corpora, comparable metrics, adjudicated human evaluations, and post-hoc stability analyses of performance metrics. The result is a benchmark that has gained community-wide acceptance—the event typically draws several dozens of teams from around the world every year. The TREC QA tracks have, in fact, become a locus of question answering research, serving not only as an annual forum for meaningful comparison of natural language processing and information retrieval techniques, but also as an efficient vehicle for the dissemination of research results. The TREC paradigm has been duplicated in similar question answering

evaluations around the world, most notably CLEF in Europe and NTCIR in Asia (both focusing on cross-language issues).

In the TREC instantiation of the question answering task, a system's response to a natural language question is a pair consisting of an answer string and a supporting document. All responses are manually judged by at least one human, who assigns one of four labels: "correct", "unsupported", "inexact", or "incorrect". In order for a response unit to be judged "correct", the answer string must provide only the relevant information and the supporting document must provide an appropriate justification for the answer string. Consider the question "What Spanish explorer discovered the Mississippi River?" A response of "Hernando de Soto" paired with a document that contains the fragment "the 16th-century Spanish explorer Hernando de Soto, who discovered the Mississippi River..." would be judged as "correct". However, the same answer string, paired with a document that contains the sentence "In 1542, Spanish explorer Hernando de Soto died while searching for gold along the Mississippi River" would be judged as "unsupported". While the answer string is correct, a human cannot conclude by reading the text that de Soto did indeed discover the Mississippi River. An answer string with extraneous words such as "Hernando de Soto discovered" would be judged "inexact". Finally, the response would be judged as "incorrect" if the answer string does not provide the information requested in the question. Thus, evaluating the response of a question answering system involves not only the answer itself, but careful consideration of the document from which the string was extracted. In this section, we will show that automatic and reliable evaluations of question answering systems are not possible outside the annual TREC cycle with currently available resources. However, we will first turn our attention to test collections for *ad hoc* retrieval as a point of comparison.

3.1 Test Collections for *ad hoc* Retrieval

The setup of the TREC question answering tracks takes many cues from the *ad hoc* tracks that were for many years the staple of document retrieval research. Although the yearly evaluations were no doubt important for the information retrieval community, their true value lies in the reusable test collections that were created from the combined effort of the participants. The notion of a reusable test collection is central to modern information retrieval research, dating back to the Cranfield experiments (Cleverdon et al., 1968). A test collection consists of a set of documents, a set of topics, and a set of relevance judgments. A topic represents a formalized information need, while the relevance judgments specify the set of documents within the collection that satisfy the information need, as assessed by the person issuing the request. A reusable test collection enables researchers to directly compare the effectiveness of different retrieval methods without involving human effort to examine the retrieved set of documents. Thus, controlled experiments in a laboratory setting can be easily conducted with rapid turnaround, outside the annual TREC cycle. In fact, the existence of these test collections obviated the need for further *ad hoc* tracks starting in TREC 9. In general, the availability of reliable and automated evaluation resources results in faster exploration of the solution space and accelerated advances in the state of the art.

For *ad hoc* test collections, which may contain hundreds of thousands of documents, exhaustively assessing the relevance of every document with respect to a particular topic is simply not practical. Instead, the pooling methodology is employed. In this setup, each team that participates in the evaluation contributes a certain number of documents to the pool (the *pool depth*) from its ranked list of results. After removing duplicates, all documents in the pool are manually assessed for relevance, and these judgments are used to score all results of all systems. Typically, each system contributes its top 100 hits (for each topic) to the pool, and is scored on all 1000 hits it returns. Zobel (1998) performed an analysis of the pooling strategy and confirmed that system rankings produced from relevance judgments gathered in this fashion are both trustworthy and fair; this means that TREC test collections can be soundly used for post-hoc evaluations, i.e., they are reusable. The performance of a new retrieval system that did not participate in the TREC evaluation, and hence did not have an opportunity to contribute to the pool, can still be accurately measured by the pooled judgments. Researchers have probed other aspects of *ad hoc* test collections, including the effect of topic size (Voorhees and Buckley 2002), the effect of incomplete judgments (Buckley and Voorhees, 2004), the effect of different evaluation metrics (Buckley and Voorhees, 2000), and different notions of relevance (Voorhees 2000; Voorhees 2001a; Sormunen 2002); in general, they have confirmed the reliability of existing test collections as a laboratory tool for experimentation and validated the general TREC methodology as an effective means for creating such test collections.

3.2 Resources for Evaluating Question Answering Systems

Given the experience with pooling in *ad hoc* retrieval, researchers have attempted to apply the same strategy to create test collections for question answering. However, as Voorhees and Tice (2000) point out, truly reusable test collections for question answering are much more difficult to build; in fact, no such collections exist today. Each year, answer patterns in the form of regular expressions and a list of relevant documents containing those answers are compiled by pooling runs submitted by participating organizations.¹ Together, these two resources have been employed by researchers to evaluate new questions answering techniques. In one type of measure, often called the strict measure, an answer is considered correct only if it matched the answer patterns *and* its supporting document was among those marked as relevant. Another type of measure, often called the lenient measure, is only concerned with matching the answer patterns. It is generally known that the strict measure underestimates answer accuracy because document-level relevance judgments are incomplete: a perfectly acceptable answer may be judged as incorrect simply because its supporting document does not appear on the list of known relevant documents. On the other hand, the lenient measure overestimates answer accuracy because documents frequently contain the answer string without actually answering the question. However, it is assumed that the combination of the two different evaluation criteria would closely approximate true question answering accuracy. We will demonstrate, however, that this assumption is not correct and that current evaluation resources (answer patterns and lists of relevant documents) cannot be reliably used for post-hoc experimentation, i.e., they cannot accurately assess the accuracy of a question

¹ Ken Litkowski usually creates the answer patterns, and NIST usually supplies the list of relevant documents.

answering system that did not participate in the original evaluation. Since our work focuses on document retrieval for question answering, we simply assume without further consideration the existence of the answer patterns; our primary concern is the quality of the document-level relevance judgments (which we discuss in the next section).

Two fundamental assumptions that contribute to the success of pooling are that participating systems in general achieve respectable performance and represent a relatively diverse set of retrieval techniques. Both of these assumptions are false in the case of question answering. The average performance of current systems is still poor, despite a few outliers (see, e.g., Voorhees, 2003). For the TREC 2004 evaluation, Voorhees reported that the median score of 92.2% of all questions is zero. As a result, the list of known relevant documents is also quite small, averaging 3.95 relevant documents per question ($\sigma = 4.07$, $\max = 23$) on the TREC 2002 testset and 3.90 document per question on the TREC 2003 testset ($\sigma = 3.84$, $\max = 25$). The histogram of the number of questions binned by the number of relevant documents is shown in Figure 2; since the TREC 2002 testset has 500 factoid questions, while the TREC 2003 testset has only 413 factoid questions, the number of questions has been normalized as a fraction. As can be seen, over a fifth of the questions in both testsets contain just one “good” relevant document. Even a casual examination of the corpus reveals the existence of many more such documents, demonstrating that the judgments are far from exhaustive. This is worrisome because a system would not be properly rewarded for answering a question correctly, unless the supporting document by chance happens to be on the list.

In addition to significant numbers of missing judgments, there is limited diversity in the types of documents that are retrieved, since most question answering teams rely on linguistically-uninformed keyword-based techniques (e.g., Srihari and Li, 1999; Clarke et al., 2000; Brill et al., 2001). Furthermore, according to Monz’s (2003) calculation, 28% of TREC 2002 participants and 21% of TREC 2003 participants simply used the results of the PRISE system provided by NIST. Although some systems do employ advanced techniques, e.g., abductive inferencing (Harabagiu et al., 2000) and feedback loops (Moldovan et al., 2002), such systems are in the minority. Besides, it is unclear what effects these advanced techniques have on the document retrieval aspect of question answering because they primarily operate on previously retrieved documents during the answer extraction stage of the QA process. Even systems that rely on advanced answer extraction techniques employ relatively standard document retrieval engines to fetch the initial list of candidate documents.

A number of existing studies on various aspects of *ad hoc* test collections should caution us about the reliability of employing pooled document-level relevance judgments to assess the performance of question answering systems. Although Zobel (1998) demonstrated that a pool depth of one hundred documents produces a fair ranking of systems, ranking stability is not invariant with respect to significantly smaller pool depths. In the TREC question answering track, the pool depth is one document, since systems are only allowed to return one response unit (an answer string and the document from which it was extracted) per question. Thus, the maximum number of relevant documents that can be discovered via the pooling strategy is bound by the number of

participant teams; in reality, however, most systems don't return the correct answer, and those that do often extract the answer from the same document. From the work of Buckley and Voorhees (2004), we know that test collections are not robust with respect to massively incomplete relevance judgments, as is potentially the case for the TREC question answering track in its current setup.

The abovementioned issues have caused us to be suspicious towards viewing existing evaluation resources as reliable question answering test collections for post-hoc experimentation. To be fair, NIST merely provides the answer patterns and relevant document lists for convenience only; they were never meant to serve as a test collection in the first place.² For lack of anything better, however, these resources have been employed by the research community in many question answering experiments, e.g., to compare the effectiveness of new techniques. Therein lies the danger: an evaluation resource that may not reliably assess system performance could potentially guide researchers towards dead-ends and prematurely close promising avenues of exploration.

To quantitatively assess the extent to which existing evaluation resources produce unreliable evaluation results, we conducted a “take one run out of the pool” experiment on the TREC 2002 results. Since we are primarily concerned with document retrieval for question answering, we only considered document-level precision, i.e., is the supporting document on the list of known relevant documents? In our experiment, we removed the contributions of a particular run to the pool, and then evaluated that run with this new reduced pool of judgments.³ This experiment simulates what would have happened if that particular run were evaluated post-hoc, i.e., it didn't participate in the evaluation. This process was repeated for every submitted run in the evaluation. We calculated the difference in document-level precision between evaluating the run on the complete set of judgments and the reduced set of judgments. Invariably, the performance dropped, i.e., the smaller pool of judgments underestimates true precision. Each of these trials produced a single data point (precision with complete judgment set and the precision drop), which we plotted in two scatter graphs: Figure 3 shows the absolute drop in performance, while Figure 4 shows the relative drop in performance (as a fraction of the original precision).

From these figures, we can see that the current set of relevance judgments for question answering does not produce trustworthy and reliable measures of document-level precision. We are most interested in rank swaps—the situation where system A performs better than system B with one set of judgments and the reverse under a different set of judgments—because they prevent us from drawing confident conclusions as to whether or not system A is “better” than system B. Consider the best-performing run in the evaluation: removing its contributions from the judgments pool and then evaluating the run would cause its precision to drop by thirty-one percentage points (40% drop in performance)! In such a scenario, the run would no longer be ranked first (by a wide margin), but would now rank third. Although the absolute drops in performance for the

² Donna Harman and Ellen Voorhees, personal communication.

³ In actuality, we removed multiple runs from the same organization together because they tended to return very similar sets of document.

other runs weren't as dramatic, the relative differences were still quite substantial. These results illustrate the unreliability of evaluating new question answering systems using the current set of relevance judgments (see Lin, 2005 for a more in-depth analysis). This is especially true for better performing systems, which are generally capable of answering more "difficult" questions. The more difficult a question is, the fewer relevant documents the pooling strategy would have gathered, thereby reducing the probability that a system would be properly rewarded for retrieving an actual relevant document. It is important to note, however, that this result applies only to post-hoc evaluation, i.e., assessing the performance of a system that did not participate in the original TREC evaluations. The reliability and stability of results for participating teams have been well established (see, for example, Voorhees, 2003 and other TREC QA track overview papers).

4. Building a Reusable QA Test Collection

In the previous section, we have discussed why presently available resources for evaluating question answering systems do not constitute a truly reusable test collection. This evaluation gap will become more pressing as the research community moves towards more difficult question types, e.g., questions that involve reasoning over and integration of facts from multiple documents. To begin addressing this problem, we have manually built a small reusable test collection for factoid question answering experiments. As previously mentioned, the collection is based on the AQUAINT corpus and consists of 110 questions (along with associated relevance judgments) selected from questions used in the 2002 TREC question answering track. Some of this work has been previously reported in Bilotti (2004) and Bilotti et al. (2004).

Our test collection was created using a much simplified one-shot variant of the search-guided relevance assessment methodology (Cormack et al., 1998; Cieri et al., 2002). Working from known answers to the questions (provided by NIST assessors), we manually crafted boolean queries with terms selected from each question and its answer—terms which we believe a "good" document is highly likely to contain. In some cases, we crafted queries that contained only keywords from the question or only keywords from the answer. Specific attempts were made to balance two competing factors: the queries must be sufficiently general to encompass the set of relevant documents, yet at the same time they had to be sufficiently restrictive to reduce the amount of manual labor involved in the assessment process. These queries were issued to an off-the-shelf IR system (Lucene), and all retrieved documents were examined manually. Although it is likely that this method will still fail to exhaustively retrieve *all* relevant documents, we assume that the resulting set of judgments is much more complete than the presently available resources.

After considering the balance between annotation detail and the amount of effort involved, we decided on a three-way judgment: supportive, unsupportive, and irrelevant. These labels correspond to the NIST guidelines of classifying response units into correct,

unsupported, and incorrect categories⁴ (recall discussions in Section 3). Supportive documents must not only contain the answer string, but a human must be able to confidently identify the answer as correct from the text. Documents that contain the answer string and discuss it in the right context, but do not actually answer the question, are marked unresponsive. Finally, irrelevant documents either do not mention the correct answer string at all, or mention it coincidentally. Examples of these judgments are shown in Table 1.

To give a concrete example, consider TREC question 1396, “What is the name of the volcano that destroyed the ancient city of Pompeii?”, whose answer is “Vesuvius” (or some variant thereof such as “Mt. Vesuvius”). We reasoned that a relevant document should contain the keywords “Pompeii” and “Vesuvius”. Therefore, we retrieved all documents from the corpus containing both these keywords and manually assessed them for relevance. For this question, we recorded fifteen supportive, three unresponsive, and ten irrelevant documents. All other documents in the collection not explicitly marked are presumed to be irrelevant, following standard TREC assumptions. An example of a clearly supportive document is AQUAINT document APW19990823.0165⁵, which states that “In A.D. 79, long-dormant Mount Vesuvius erupted, burying the Roman cities of Pompeii and Herculaneum in volcanic ash.” Document NYT20000405.0216, which states that “Pompeii was pagan in A.D. 79, when Vesuvius erupted.”, is an example of an unresponsive document. It addresses speculations that “the people of Pompeii were justly punished by the volcano eruption,” but does not explicitly mention or imply that the city was destroyed by Vesuvius. An example of an irrelevant document is NYT20000704.0049; although it contains the keywords “Pompeii” and “Vesuvius”, the text clearly does not answer the question. The article discusses winemaking in Campania, the region of Italy where both Pompeii and Vesuvius are located; it describes vineyards near the ruins of Pompeii and grape varieties that grow in the volcanic soil at the foot of Mt. Vesuvius.

One hundred and twenty questions from TREC 2002 were originally selected as the basis of our test collection. The decision to reuse old questions allowed meaningful comparisons between existing relevance judgments and our newly created resource. In the end, however, we had to discard ten questions due to a variety of issues. One common problem concerned practical constraints on the assessment effort: for example, question 1496, “What country is Berlin in?” was discarded because there were too many potentially relevant documents to go through and evaluate manually (essentially, any document that had the terms “Berlin” and “Germany” would have to be manually examined). Question 1422, “What two European countries are connected by the St. Gotthard Tunnel?” was discarded because the question itself made false assumptions about the answer. Some documents in the collection suggest that the tunnel is an essential route connecting Italy with Germany, but other documents claim that the tunnel links Italy and Switzerland. In reality, both endpoints of the tunnel and its entire extent lie within Switzerland, although its proximity to other European nations might suggest a

⁴ A fourth judgment for answers, inexact, concerns the answer string only and is not relevant for the purposes of document retrieval.

⁵ Incidentally, this document is not present in the NIST-supplied list of relevant documents.

more metaphorical interpretation. Faced with this complex situation, we decided to discard the question. Overall, malformed questions and questions with too many potential answers accounted for all instances that were ultimately excluded from our test collection.

In total, 6009 documents were manually examined using the strategy described above (on average, approximately fifty-four documents were assessed per question). Of those, 1901 were marked supportive, 298 were marked unsupportive, and 3810 were marked irrelevant. In the beginning, twenty-seven questions were doubly annotated by two different assessors for the purposes of determining inter-annotator agreement. This was accomplished in an iterative manner: first, both assessors completed assessments for ten questions independently. A discussion ensued between the assessors, in which they talked over their differences in opinion. The purpose of this discussion was not to converge on a strict set of guidelines for judging, but rather to explicate the issues involved in assessing answer correctness (experiences that we will share in the next section). In the second iteration, ten more questions were judged independently by the same two assessors; another round of discussions followed. Finally, seven more questions were annotated independently. The confusion matrix for these twenty-seven doubly-annotated questions is shown in Table 2. As can be seen, the overall agreement is approximately 84%. Disagreements in opinion occurred most often with supportive-unsupportive and unsupportive-irrelevant judgments. In comparison, there were relatively few documents which one assessor found to be supportive and the other found to be irrelevant..

5. What makes an answer “correct”?

Relevance judgments form an integral part of test collections because they define the set of documents that should be retrieved in response to a user information need. Although the notion of relevance has been much debated in the literature (e.g., Cooper, 1971; Saracevic, 1975; Harter, 1992; Barry and Schamber, 1998; Spink and Greisdorf, 2001; Mizzaro, 1999), Voorhees (2000) has shown that comparative evaluation of retrieval performance is invariant with respect to substantial difference in relevance judgments (cf. Harter, 1996). Nevertheless, it is worthwhile to explore the diverse and subtle factors that influence a person’s assessment of relevance and the notion of support in a question answering task. In the process of building our test collection, we have encountered a variety of issues regarding the human interpretation of answer correctness, which we share here. These experiences can be applied to qualitatively improve answers returned by future systems (cf. Sparck Jones, 2003).

The ambiguity of natural language means that questions often lend themselves to alternate interpretations. A case in point is whether facets of a question are restrictive or descriptive in nature, illustrated by question 1834, “Which disciple received 30 pieces of silver for betraying Jesus?” In order for a document to be judged as relevant, would it be sufficient for the document to state that Judas betrayed Jesus, or would the document explicitly need to state the amount “30 pieces of silver”? The phrasing of the question leaves open the possibility that another disciple betrayed Jesus for 40 pieces of silver, in

which case, the amount must be interpreted in a restrictive manner, i.e., a relevant document must explicitly mention this fact.

Consider another example, question 1398, “What year was Alaska purchased?”, whose answer is 1867 (as ascertained by NIST assessors). A document that contains the keywords “1867” and “Alaska” is APW19990329.0045, which says that “In 1867, U.S. Secretary of State William H. Seward reached agreement with Russia to purchase the territory of Alaska.” Is this a relevant supportive document? At first glance, probably so, but upon closer examination, one might have doubts. Bringing to bear common-sense knowledge about the world, a human might realize that the international transaction of purchasing territories is a long and complicated process that might span years. Thus, “reaching an agreement to purchase” in 1867 might not mean that Alaska was “purchased” in 1867. Consider another document, APW19991017.0082, which states that, “In 1867, the United States took formal possession of Alaska from Russia.” Would this document be considered supportive? For one, it does not mention explicitly that a purchase was involved (it could have been ceded as a result of some other treaty, for example); furthermore, the date of “taking formal possession” might be different from the purchase date. There is not only ambiguity in the interpretation of documents that might contain the answer, but also ambiguity in the interpretation of questions themselves. In the TREC question answering evaluation setting (and in our own work), the creator of the question is *not* the person performing the actual relevance judgments. Therefore, assessors must attempt to reconstruct the original intentions of the person with the original information request. These differences in interpretation, on both the question and answer end, translate into significant variations in the notion of answer correctness.

Answer granularity is another area where large differences in opinion are sometimes observed; our work confirmed experiences reported by Voorhees and Tice (2000). In a question asking for a date, how exact must the date be? Is the year an event happened sufficient, or is an exact month and day necessary? From our experiences, these judgments vary not only from assessor to assessor, but also from question to question. For a relatively recent event such as “When did World War I start?” or “When was Hurricane Hugo?”, assessors are more likely to require finer-grained dates (containing both the month and the year, for example). For other questions, such as “When did Muhammad live?”, answers as coarse-grained as “6th century” might be accepted. Similar differences in opinion occurred with other named entities such as people and place names. Is the surname of a person sufficient, or should an answer have both a given name and a surname? Is the country sufficient for a question asking about a location? We noticed differences in opinion both among different assessors and across different questions.

Finally, there are many cases where an answer is not explicitly stated in a document, but requires the assessor to bring external knowledge to bear in interpreting the text. For example, consider the question “What is Pennsylvania’s nickname?”, whose answer is “keystone state”. The vast majority of supportive documents do not explicitly relate the state with its nickname, but the connection is clear from the discourse structure of the articles (e.g., from the use of anaphoric references); our assessors did indeed find such

documents to be perfectly acceptable, but, once again, there is room for disagreement. Consider a more complicated situation concerning the question “Where did Allen Iverson go to college?” Some articles mentioned that he was a Hoya, which sports aficionados might automatically associate with Georgetown University. However, this knowledge cannot be considered “common sense”, and hence judgments of these responses varied greatly from assessor to assessor.

Our efforts in developing a reusable test collection for question answering have taught us how difficult the endeavor truly is. Rarely is there such a thing as “an obvious answer”, and there is certainly no such thing as “universal ground truth” for factoid questions; it is fruitless to try and create strict rules that govern what constitutes an answer because the notion of “correctness” varies both from person to person and from question to question (cf. Voorhees and Tice, 1999). Even if it were possible to externally impose rules that, for example, dictated the granularity of answer strings, the resulting judgments would not match real-world user needs—differences in opinion are an inescapable fact of question answering evaluation. We hope that better understanding of real-world user needs will lead to more effective question answering systems in the future.

6. Evaluating the Reusable Test Collection

In our manually created test collection, we found an average of approximately 17 supportive documents per question (median of six), over the testset of 110 questions. In comparison, the original NIST judgments averaged 3.7 relevant documents per question (median of two) over the same set of questions. Figure 4 shows a histogram of this distribution.

To verify the quality of our test collection, we evaluated all original TREC 2002 QA track submissions with our resource, and compared the system rankings with those obtained by using the original NIST judgments (measuring document-level precision only). The Kendall’s tau correlation between the two rankings was 0.87, which indicates good agreement. Kendall’s tau computes the “distance” between two rankings as the minimum number of pairwise adjacent swaps necessary to convert one ranking into the other. This value is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0; the correlation between a ranking and its perfect inverse is -1.0; and the expected correlation of two rankings chosen at random is 0.0 (cf. Voorhees and Tice, 1999). Manually examining the system rankings, we found that most rank swaps occurred when the absolute document-level precision scores were very close in the first place (within the margin of error that could be attributed to, say, choice of questions).

Since our test collection was created independently of the original NIST document-level relevance judgments (pooled from TREC 2002 participants), it does not contain system-level biases exhibited by the currently available resources. Specifically, it should reliably and fairly evaluate systems that did not participate in the original TREC 2002 evaluation, and hence can be used for post-hoc experimentation. In addition, we have shown that rankings of the TREC 2002 QA track participants produced by our test collection correlate highly with those produced by the “official” NIST judgments. In sum, we have

manually built a truly reusable test collection for question answering, and have made it available to the entire research community.⁶

7. Conclusion

In total, the creation of our test collection took approximately 230 person hours. The initial startup cost of doubly annotating nearly a quarter of the questions and discussing the differences in opinion occupied significant amounts of time. However, it was a worthwhile effort because we gained many insights into the factors that influence the interpretation of answers. In a larger effort, progress would be accelerated as this initial startup cost becomes amortized over more questions.

The use of test collections for rapid, reproducible laboratory experiments is a well-established paradigm in modern information retrieval. The creation of such test collections is time-consuming and labor-intensive, but vitally important for the advancement of the state of the art. In addition to the question answering test collection itself, the contributions of this work are fourfold: First, we explained how document retrieval for question answering is distinct from *ad hoc* retrieval, with its own set of challenges and tradeoffs. Second, we showed that currently available resources for evaluating question answering systems do not produce fair and reliable results. Third, we demonstrated that working backwards from known answers to gather relevance judgments serves as a viable strategy for building question answering test collections. Finally, we have enumerated many issues that affect the notion of answer correctness, which could be incorporated into future systems to qualitatively improve the answers they return. We hope that this work will pave the way for rapid advances in question answering technology.

Acknowledgements

We would like to recognize Matthew Bilotti and Leah Oats for their tireless annotation effort. A portion of this work was the subject of Matthew's Master's thesis. We would like to thank Donna Harman, Doug Oard, and Ellen Voorhees for helpful discussions and suggestions. This research was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program.

⁶ The test collection is available at <http://www.umiacs.umd.edu/~jimmylin/downloads/>

Figures and Tables

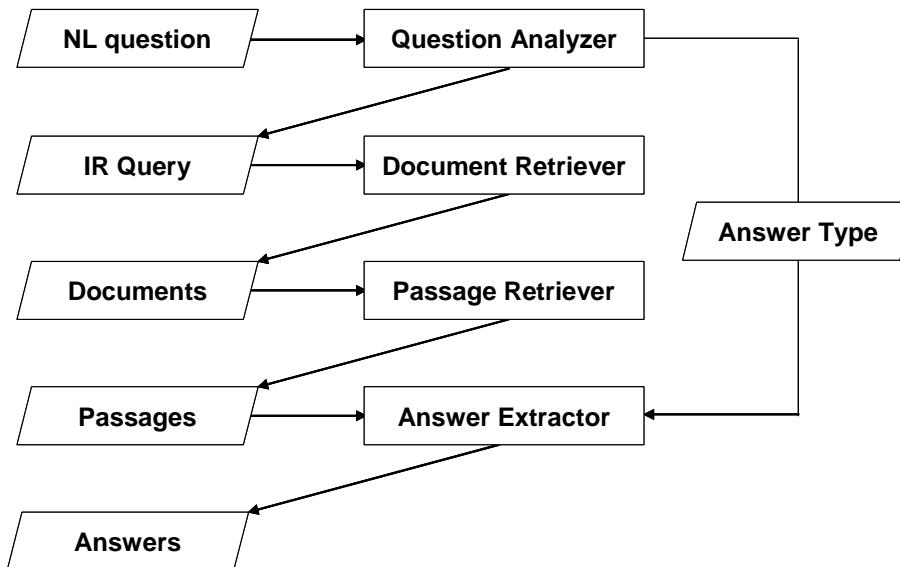


Figure 1. Typical pipelined question answering architecture

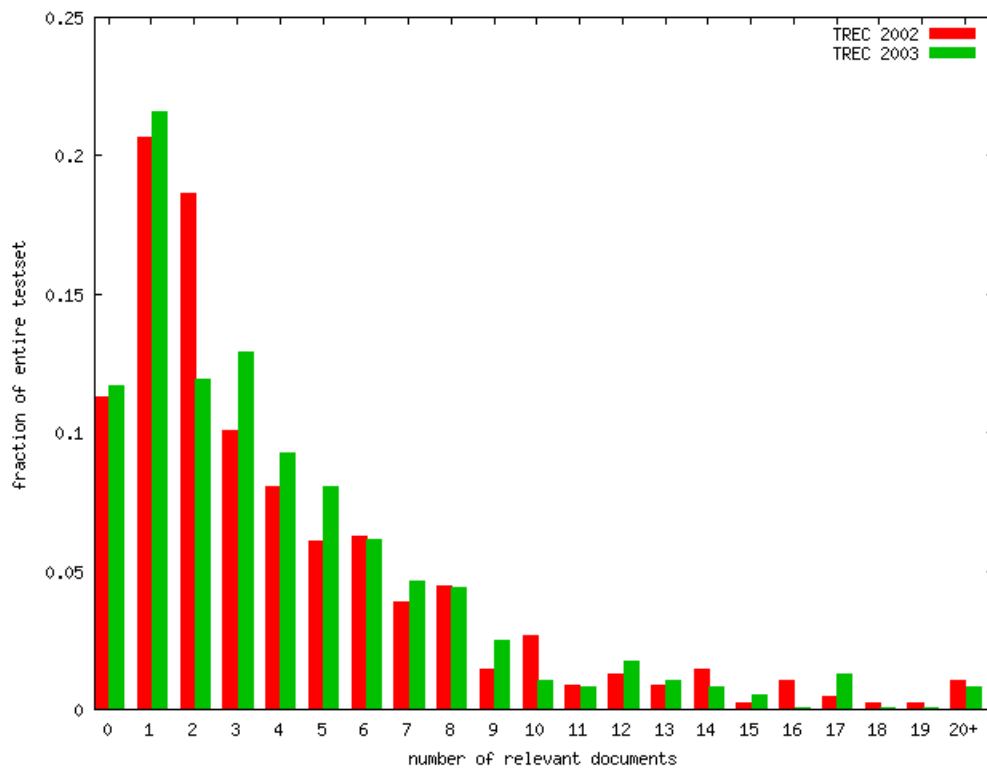


Figure 2. Histogram of questions (in terms of fraction of the entire testset) binned in terms of the number of relevant documents.

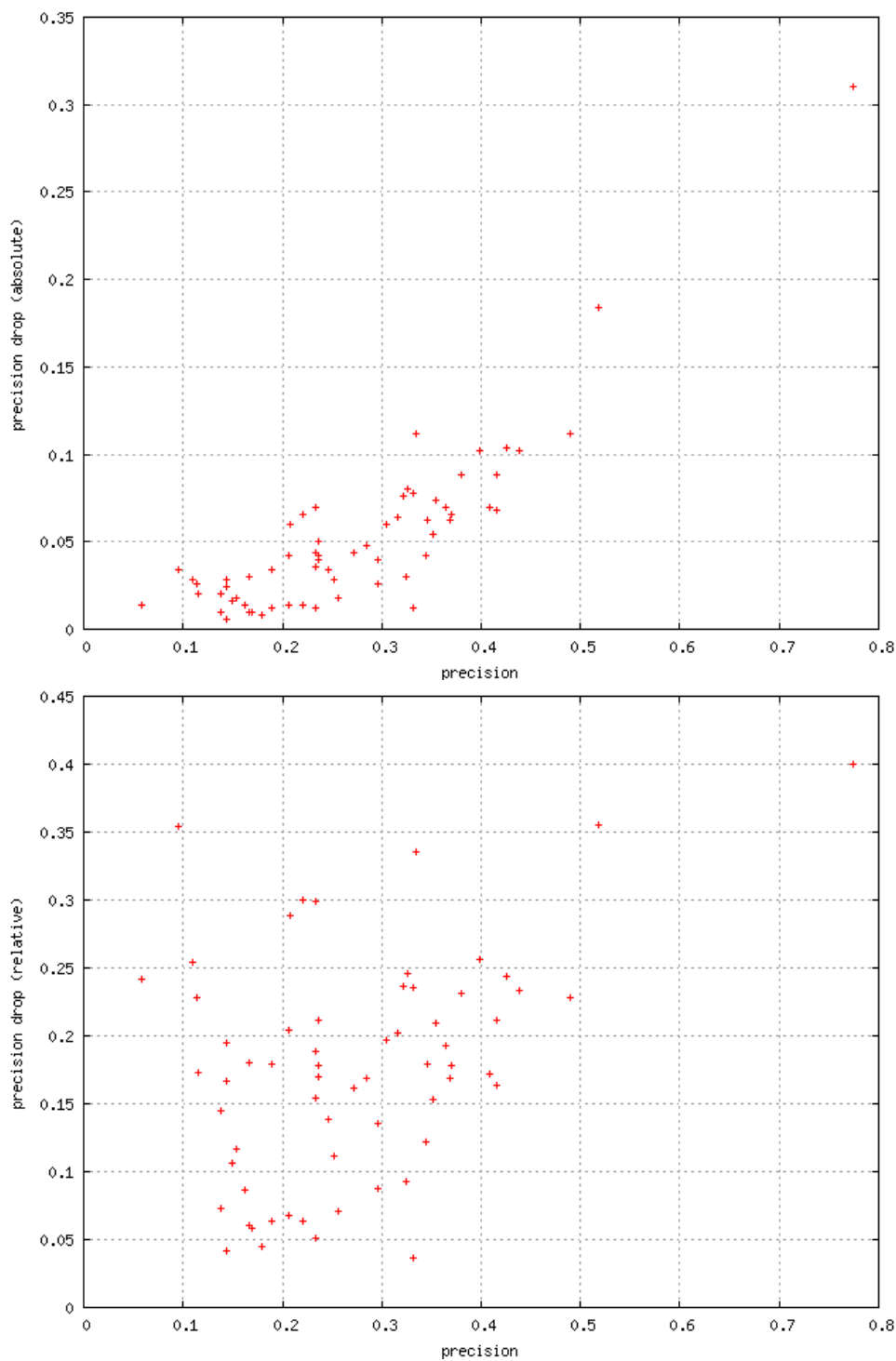


Figure 3: Results of the “take one run out of the pool” experiment. For every run, its contribution to the pool was removed, and the run was evaluated with this new reduced set of relevance judgments. The performance drop between these two conditions is plotted against the true precision of the run (absolute precision drop, top graph; relative precision drop as a fraction of the original score, bottom graph).

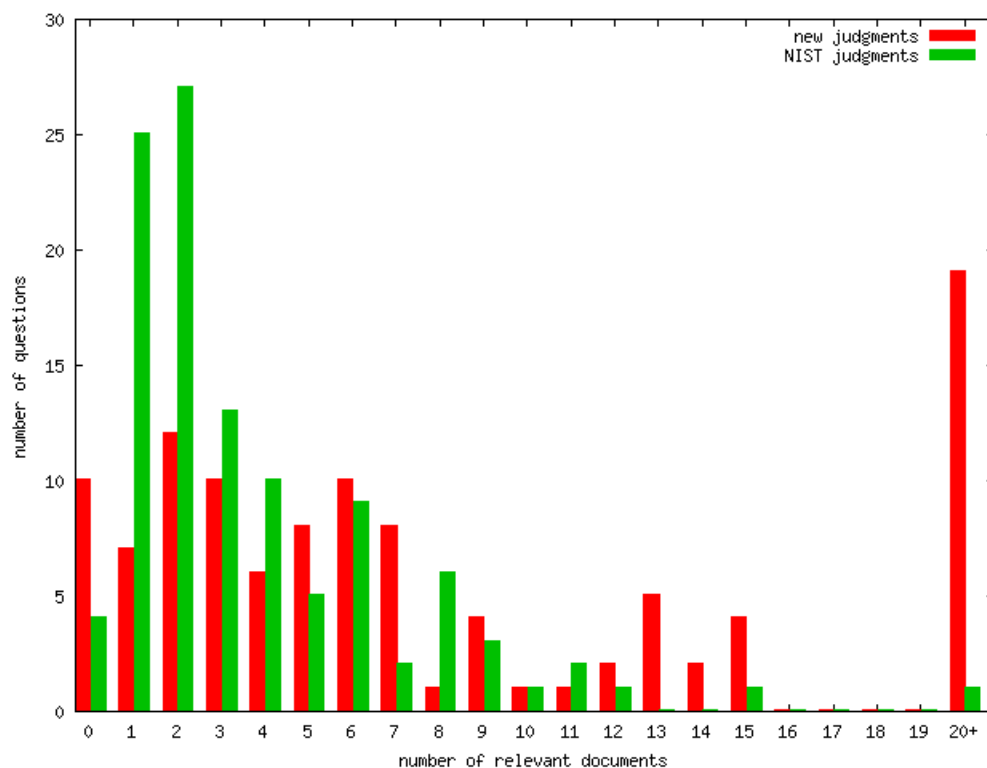


Figure 4. Histogram of number of questions binned by the number of relevant documents (old NIST judgments compared to newly created judgments).

Question:	What is the name of the volcano that destroyed the ancient city of Pompeii?
Supportive	... In A.D. 79, long-dormant Mount Vesuvius erupted, burying the Roman cities of Pompeii and Herculaneum in volcanic ash... [APW19990823.0165]
Unsupportive	... Pompeii was pagan in A.D. 79, when Vesuvius erupted... [NYT20000405.0216]
Irrelevant	... the project of replanting ancient vineyards amid the ruins of Pompeii... Coda di Volpe, a white grape from Roman times that thrives in the volcanic soils on the lower slopes of Mt. Vesuvius ... [NYT20000704.0049]

Table 1. Examples of supportive, unsupportive, and irrelevant judgments.

	supportive	unsupportive	irrelevant
supportive	306 (35.58%)	27 (3.14%)	5 (0.58%)
unsupportive	53 (6.16%)	58 (6.74%)	21 (2.44%)
irrelevant	13 (1.51%)	19 (2.21%)	358 (41.64%)

Table 2. Confusion matrix for relevance judgments by two different assessors for twenty-seven doubly annotated questions.

References

- Arampatzis, A., Tsores, T., Koster, C.H.A., & Van Der Weide, Th.P. (1998). Phrase-based Information Retrieval. *Information Processing and Management*, 34(6), 693-707.
- Barry, C., & Schamber, L. (1998). Users' Criteria for Relevance Evaluation: A Cross-Situational Comparison. *Information Processing and Management*, 34(2/3), 219-236.
- Bilotti, M. (2004). Query Expansion Techniques for Question Answering, Master's Thesis, Massachusetts Institute of Technology.
- Bilotti, M., Katz, B., & Lin, J. (2004). What Works Better for Question Answering: Stemming or Morphological Query Expansion? In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*.
- Brill, E., Lin, J., Banko, M., Dumais, S., & Ng, A. (2001). Data-Intensive Question Answering. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.
- Buckley, C., & Voorhees, E.M. (2000). Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*.
- Buckley, C., & Voorhees, E.M. (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)*.
- Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K., & Liberman, M. (2002). Corpora for Topic Detection and Tracking. In J. Allan (Ed.), *Topic Detection and Tracking: Event-Based Information Organization*. Dordrecht: Kluwer Academic Publishers.
- Clarke, C., Cormack G., Kisman, D., & Lynam, T. (2000). Question Answering by Passage Selection (MultiText Experiments for TREC-9). In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*.
- Cleverdon, C., Mills, J., & Keen, E.M. (1968). *Factors Determining the Performance of Indexing Systems*, Two volumes. ASLIB Cranfield Research Project, Cranfield, England.
- Cooper, W. (1971). A Definition of Relevance for Information Retrieval. *Information Storage and Retrieval*, 7, 19-37.
- Cormack, G., Palmer, C., & Clarke, C. (1998). Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*.

- Echihabi, A., & Marcu, D. (2003). A Noisy-Channel Approach to Question Answering. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003).
- Harabagiu, S., Pasca, M., & Maiorano, S. (2000). Experiments with Open-Domain Textual Question Answering. In Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000).
- Harman, D. (1991). How Effective is Suffixing? *Journal of the American Society for Information Science*, 42(1), 7-15.
- Harter, S. (1992). Psychological Relevance and Information Science. *Journal of the American Society for Information Science*, 43(9), 602-615.
- Harter, S. (1996). Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science*, 47(1), 37-49, 1996.
- Hirschman, L. & Gaizauskas, R. (2001). Natural Language Question Answering: The View from Here. *Natural Language Engineering*, 7(4), 275-300.
- Hovy, E., Gerber, L., Hermjakob, U., Lin, C.-Y., & Ravichandran, D. (2001). Toward Semantics-Based Answer Pinpointing. In Proceedings of the First International Conference on Human Language Technology Research (HLT 2001).
- Hull, D. (1996). Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society for Information Science*, 47(1), 70-84.
- Katz, B., & Lin, J. (2003). Selectively Using Relations to Improve Precision in Question Answering. In Proceedings of the EACL 2003 Workshop on Natural Language Processing for Question Answering.
- Krovetz, R. (1993). Viewing Morphology as an Inference Process. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993).
- Li, X., & Roth, D. (2002). Learning Question Classifiers. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002).
- Light, M., Mann, G., Riloff, E., & Breck, E. (2001). Analyses for Elucidating Current Question Answering Technology. *Natural Language Engineering*, 7(4), 325-342.
- Lin, J. (2005). Evaluation of Resources for Question Answering Evaluation. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005).

- Mizzaro, S. (1999). How Many Relevances in Information Retrieval? *Interacting With Computers*, 10(3), 305-322.
- Moldovan, D., Pasca, M., Harabagiu, S., & Surdeanu, M. (2002). Performance Issues and Error Analysis in an Open-Domain Question Answering System. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.
- Monz, C. (2003). *From Document Retrieval to Question Answering*. Ph.D. Dissertation, Institute for Logic, Language, and Computation, University of Amsterdam.
- Prager, J., Brown, E., & Coden, A. (2000). Question-Answering by Predictive Annotation. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*.
- Robertson, S. (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation*, 60(5), 503-520.
- Sormunen, E. (2002). Liberal Relevance Criteria of TREC---Counting on Negligible Documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*.
- Sparck Jones, K. (2003). Is Question Answering a Rational Task? In *Proceedings of the 2nd CoLogNET-ElsNET Symposium on Questions and Answers: Theoretical and Applied Perspectives*.
- Spink, A., & Greisdorf, H. (2001). Regions and Levels: Mapping and Measuring Users' Relevance Judgments. *Journal of the American Society for Information Science and Technology*, 52(2), 161-173.
- Srihari, R., & Li, W. (1999). Information Extraction Supported Question Answering. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*.
- Tellex, S., Katz, B., Lin, J., Marton, G., & Fernandes, A. (2003). Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*.
- Voorhees, E.M., & Buckley, C. (2002). The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*.
- Voorhees, E.M., & Tice, D.M. (1999). The TREC-8 Question Answering Track Evaluation. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*.

Voorhees, E.M., & Tice, D.M. (2000). Building a Question Answering Test Collection. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000).

Voorhees, E.M., & Tice, D.M. (2000a) Overview of the TREC-9 Question Answering Track. In Proceedings of the Ninth Text REtrieval Conference (TREC-9).

Voorhees, E.M., (2000). Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness”, *Information Processing and Management*, 36(5), 697-716, 2000.

Voorhees, E.M., (2001). Overview of the TREC 2001 Question Answering Track. In Proceedings of the Tenth Text REtrieval Conference (TREC 2001).

Voorhees, E.M. (2001a). Evaluation by Highly Relevant Documents. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001).

Voorhees, E.M. (2002). Overview of the TREC 2002 Question Answering Track. In Proceedings of the Eleventh Text REtrieval Conference (TREC 2002).

Voorhees, E.M. (2003). Overview of the TREC 2003 Question Answering Track. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2003).

Zobel, J. (1998). How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998).