Evaluating Summaries and Answers: Two Sides of the Same Coin?

Jimmy Lin[1,3] **and Dina Demner-Fushman**[2,3]
[1]College of Information Studies
[2]Department of Computer Science
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
jimmylin@umd.edu, demner@cs.umd.edu

Abstract

This paper discusses the convergence between question answering and multi-document summarization, pointing out implications and opportunities for knowledge transfer in both directions. As a case study in one direction, we discuss the recent development of an automatic method for evaluating definition questions based on *n*-gram overlap, a commonly-used technique in summarization evaluation. In the other direction, the move towards topic-oriented summaries requires an understanding of relevance and topicality, issues which have received attention in the question answering literature. It is our opinion that question answering and multi-document summarization represent two complementary approaches to the same problem of satisfying complex user information needs. Although this points to many exciting opportunities for system-building, here we primarily focus on implications for system evaluation.

1 Introduction

Recent developments in question answering (QA) and multi-document summarization point to many interesting convergences that present exciting opportunities for collaboration and cross-fertilization between these largely independent communities. This position paper attempts to draw connections between the task of answering complex natural language questions and the task of summarizing multiple documents, the boundaries between which are beginning to blur, as anticipated half a decade ago (Carbonell et al., 2000).

Although the complementary co-evolution of question answering and document summarization presents new directions for system-building, this paper primarily focuses on implications for evaluation. Although assessment of answer and summary quality employs different methodologies, there are many lessons that each community can learn from the other. The summarization community has extensive experience in intrinsic metrics based on *n*-gram overlap for automatically scoring system outputs against human-generated reference texts—these techniques would help streamline aspects of question answering evaluation. In the other direction, because question answering has its roots in information retrieval, much work has focused on extrinsic metrics based on relevance and topicality, which may be valuable to summarization researchers.

This paper is organized as follows: In Section 2, we discuss the evolution of question answering research and how recent trends point to the convergence of question answering and multi-document summarization. In Section 3, we present a case study of automatically evaluating definition questions by employing metrics based on *n*-gram overlap, a general technique widely used in summarization and machine translation evaluations. Section 4 highlights some opportunities for knowledge transfer in the other direction: how the notions of relevance and topicality, well-studied in the information retrieval literature, can guide the evaluation of topic-oriented summaries. We conclude with thoughts about the future in Section 5.

2 Convergence of QA and Summarization

Question answering was initially conceived as essentially a fine-grained information retrieval task. Much research has focused on so-called factoid questions, which can typically be answered by named entities such as people, organizations, locations, etc. As an example, a system might return "Bee Gees" as the answer to the question "What band did the music for the 1970's film 'Saturday Night Fever'?". For such well-specified information needs, question answering systems represent an improvement over traditional document retrieval systems because they do not require a user to manually browse through a ranked list of "hits". Since 1999, the NIST-organized question answering tracks at TREC (see, for example, Voorhees 2003a) have served as a focal point of research in the field, providing an annual forum for evaluating systems developed by teams from all over the world. The model has been duplicated and elaborated on by CLEF in Europe and NTCIR in Asia, both of which have also introduced cross-lingual elements.

Recently, research in question answering has shifted away from factoid questions to more complex information needs. This new direction can be characterized as a move towards answers that can only be arrived at through some form of reasoning and answers that require drawing information from multiple sources. Indeed, there are many types of questions that would require integration of both capabilities: extracting raw information "nuggets" from potentially relevant documents, reasoning over these basic facts to draw additional inferences, and synthesizing an appropriate answer based on this knowledge. "What is the role of the Libyan government in the Lockerbie bombing?" is an example of such a complex question.

Commonalities between the task of answering complex questions and summarizing multiple documents are evident when one considers broader research trends. Both tasks require the ability to draw together elements from multiple sources and cope with redundant, inconsistent, and contradictory information. Both tasks require extracting finer-grained (i.e., sub-document) segments, albeit based on different criteria. These observations point to the convergence of question answering and multi-document summarization.

Complementary developments in the summarization community mirror the aforementioned shifts in question answering research. Most notably, the DUC 2005 task requires systems to generate answers to natural language questions based on a collection of known relevant documents: "The system task in 2005 will be to synthesize from a set of 25–50 documents a brief, well-organized, fluent answer to a need for information that cannot be met by just stating a name, date, quantity, etc." (DUC 2005 guidelines). These guidelines were modeled after the *information synthesis* task suggested by Amigó et al. (2004), which they characterize as "the process of (given a complex information need) extracting, organizing, and inter-relating the pieces of information contained in a set of relevant documents, in order to obtain a comprehensive, non-redundant report that satisfies the information need". One of the examples they provide, "I'm looking for information concerning the history of text compression both before and with computers", looks remarkably like a user information need current question answering systems aspire to satisfy. The idea of topic-oriented multi-document summarization isn't new (Goldstein et al., 2000), but only recently have the connections to question answering become explicit. Incidentally, it appears that the current vision of question answering is more ambitious than the information synthesis task because in the former, the set of relevant documents is not known in advance, but must first be discovered within a larger corpus.

There is, however, an important difference between question answering and topic-focused multi-document summarization: whereas summaries are compressible in length, the same cannot be said of answers.[1] For question answering, it is difficult to fix the length of a response *a priori*: there may be cases where it is impossible to fit a coherent, complete answer into an allotted space. On the other

[1] We would like to thank an anonymous reviewer for pointing this out.

1	*vital*	american composer
2	*vital*	musical achievements ballets symphonies
3	*vital*	born brooklyn ny 1900
4	*okay*	son jewish immigrant
5	*okay*	american communist
6	*okay*	civil rights advocate
7	*okay*	had senile dementia
8	*vital*	established home for composers
9	*okay*	won oscar for "the Heiress"
10	*okay*	homosexual
11	*okay*	teacher tanglewood music center boston symphony

Table 1: The "answer key" to the question "Who is Aaron Copland?"

hand, summaries are condensed representations of content, and should theoretically be expandable and compressible based on the level of detail desired.

What are the implications, for system evaluations, of this convergence between question answering and multi-document summarization? We believe that the two fields have much to benefit from each other. In one direction, the question answering community currently lacks experience in automatically evaluating unstructured answers, which has been the focus of much research in document summarization. In the other direction, the question answering community, due to its roots in information retrieval, has a good grasp on the notions of relevance and topicality, which are critical to the assessment of topic-oriented summaries. In the next section, we present a case study in leveraging summarization evaluation techniques to automatically evaluate definition questions. Following that, we discuss how lessons from question answering (and more broadly, information retrieval) can be applied to assist in evaluating summarization systems.

3 Definition Questions: A Case Study

Definition questions represent complex information needs that involve integrating facts from multiple documents. A typical definition question is "What is the Cassini space probe?", to which a system might respond with answers that include "interplanetary probe to Saturn", "carries the Huygens probe to study the atmosphere of Titan, Saturn's largest moon", and "a joint project between NASA, ESA, and ASI". The goal of the task is to return as many interesting "nuggets" of information as possible about the target entity being defined (the Cassini space probe, in this case) while minimizing the amount of irrelevant information retrieved. In the two formal evaluations of definition questions that have been conducted at TREC (in 2003 and 2004), an information nugget is operationalized as a fact for which an assessor could make a binary decision as to whether a response contained that nugget (Voorhees, 2003b). Additionally, information nuggets are classified as either *vital* or *okay*. Vital nuggets represent facts central to the target entity, and should be present in a "good" definition. Okay nuggets contribute worthwhile information about the target, but are not essential. As an example, assessors' nuggets for the question "Who is Aaron Copland?" are shown in Table 1. The distinction between vital and okay nuggets is consequential for the score calculation, which we will discuss below.

In the TREC setup, a system response to a definition question is comprised of an unordered set of answer strings paired with the identifier of the document from which it was extracted. Each of these answer strings is presumed to have one or more information nuggets contained within it. Although there is no explicit limit on the length of each answer string and the number of answer strings a system is allowed to return, verbosity is penalized against, as we shall see below.

To evaluate system output, NIST gathers answer strings from all participants, hides their association

[NYT19990708.0196] Once past a rather routine apprenticeship, which included three years of study with Nadia Boulanger in Paris, Copland became one of the few American composers to make a living from composition.
Nugget present: 1

[NYT20000107.0305] A passionate advocate of civil rights, Copland conducted a performance of the "Lincoln Portrait" with Coretta Scott King as narrator.
Nuggets present: 6

[NYT19991117.0369] after four prior nominations, he won an Oscar in 1949 for his music for "The Heiress"
Nugget present: 9

Figure 1: Examples of judging actual system responses.

with the runs that produced them, and presents all answer strings to a human assessor. Using these responses and research performed during the original development of the question (with an off-the-shelf document retrieval system), the assessor creates an "answer key"; Table 1 shows the official answer key for the question "Who is Aaron Copland?".

After this answer key has been created, NIST assessors then go back over each run and manually judge whether or not each nugget is present in a particular system's response. Figure 1 shows a few examples of real system output and the nuggets that were found in them.

The final score of a particular answer is computed as an F-measure, the harmonic mean between nugget precision and recall. The β parameter controls the relative importance of precision and recall, and is heavily biased towards the latter to model the nature of the task. Nugget recall is calculated solely as a function of the vital nuggets, which means that a system receives no "credit" (in terms of recall) for returning okay nuggets. Nugget precision is approximated by a length allowance based on the number of vital and okay nuggets returned; a response longer than the allowed length is subjected to a verbosity penalty. Using answer length as a proxy to precision appears to be a reasonable compromise because a pilot study demonstrated that it was impossible for humans to consistently enumerate the total number of nuggets in a response, a necessary step in calculating nugget precision (Voorhees, 2003b).

The current TREC setup for evaluating definition questions necessitates having a human "in the loop". Even though answer keys are available for questions from previous years, determining if a nugget was actually retrieved by a system currently requires human judgment. Without a fully-automated evaluation method, it is difficult to consistently and reproducibly assess the performance of a system outside the annual TREC cycle. Thus, researchers cannot carry out controlled laboratory experiments to rapidly explore the solution space. In many other fields in computational linguistics, the ability to conduct evaluations with quick turnaround has lead to rapid progress in the state of the art. Question an-

Let

r # of *vital* nuggets returned in a response
a # of *okay* nuggets returned in a response
R # of *vital* nuggets in the answer key
l # of non-whitespace characters in the entire answer string

Then

$$\text{recall } (\mathcal{R}) = r/R$$
$$\text{allowance } (\alpha) = 100 \times (r+a)$$
$$\text{precision } (\mathcal{P}) = \begin{cases} 1 & \text{if } l < \alpha \\ 1 - \frac{l-\alpha}{l} & \text{otherwise} \end{cases}$$

Finally, $F(\beta) = \dfrac{(\beta^2+1) \times \mathcal{P} \times \mathcal{R}}{\beta^2 \times \mathcal{P} + \mathcal{R}}$

$\beta = 5$ in TREC 2003, $\beta = 3$ in TREC 2004.

Figure 2: Official definition of F-measure.

swering for definition questions appears to be missing this critical ingredient.

To address this evaluation gap, we have recently developed POURPRE, a method for automatically evaluating definition questions based on *idf*-weighted unigram co-occurrences (Lin and Demner-Fushman, 2005). This idea of employing *n*-gram co-occurrence statistics to score the output of a computer system against one or more desired reference outputs has its roots in the BLEU metric for machine translation (Papineni et al., 2002) and the ROUGE (Lin and Hovy, 2003) metric for summarization. Note that metrics for automatically evaluating definitions should be, like metrics for evaluating summaries, biased towards recall. Fluency (i.e., precision) is not usually of concern because most systems employ extractive techniques to produce answers. Our study reports good correlation between the automatically computed POURPRE metric and official TREC system ranks. This measure will hopefully spur progress in definition question answering systems.

The development of automatic evaluation metrics based on *n*-gram co-occurrence for question answering is an example of successful knowledge transfer from summarization to question answering evaluation. We believe that there exist many more opportunities for future exploration; as an example, there are remarkable similarities between information nuggets in definition question answering and recently-proposed methods for assessing summaries based on fine-grained semantic units (Teufel and van Halteren, 2004; Nenkova and Passonneau, 2004).

Another promising direction of research in definition question answering involves applying the Pyramid Method (Nenkova and Passonneau, 2004) to better model the vital/okay nuggets distinction. As it currently stands, the vital/okay dichotomy is troublesome because there is no way to operationalize such a classification scheme within a system; see Hildebrandt et al. (2004) for more discussion. Yet, the effects on score are significant: a system that returns, for example, all the okay nuggets but none of the vital nuggets would receive a score of zero. In truth, the vital/okay distinction is a poor attempt at modeling the fact that some nuggets about a target are more important than others—this is exactly what the Pyramid Method is designed to capture. "Building pyramids" for definition questions is an avenue of research that we are currently pursuing.

In the next section, we discuss opportunities for knowledge transfer in the other direction; i.e., how summarization evaluation can benefit from work in question answering evaluation.

4 Putting the Relevance in Summarization

The definition of a meaningful extrinsic evaluation metric (e.g., a task-based measure) is an issue that the summarization community has long grappled with (Mani et al., 2002). This issue has been one of the driving factors towards summaries that are specifically responsive to complex information needs. The evaluation of such summaries hinges on the notions of relevance and topicality, two themes that have received much research attention in the information retrieval community, from which question answering evolved.

Debates about the nature of relevance are almost as old as the field of information retrieval itself (Cooper, 1971; Saracevic, 1975; Harter, 1992; Barry and Schamber, 1998; Mizzaro, 1998; Spink and Greisdorf, 2001). Theoretical discussions aside, there is evidence suggesting that there exist substantial inter-assessor differences in document-level relevance judgments (Voorhees, 2000; Voorhees, 2002); in the TREC *ad hoc* tracks, for example, overlap between two humans can be less than 50%. For factoid question answering, it has also been shown that the notion of answer correctness is less well-defined than one would expect (Voorhees and Tice, 2000; Lin and Katz, 2005 in press). This inescapable fact about the nature of information needs represents a fundamental philosophical difference between research in information retrieval and computational linguistics. Information retrieval researchers accept the fact that the notion of "ground truth" is not particularly meaningful, and any prescriptive attempt to dictate otherwise would result in brittle and overtrained systems of limited value. A retrieval system must be sensitive to the inevitable variations in relevance exhibited by different users.

This philosophy represents a contrast from computational linguistics research, where ground truth does in fact exist. For example, there is a single correct parse of a natural language sentence (modulo truly ambiguous sentences), there is the notion of a correct word sense (modulo granularity issues), etc. This view also pervades evaluation in machine translation and document summarization, and is implicitly codified in intrinsic metrics, except that there is now the notion of multiple correct answers (i.e., the reference texts).

Faced with the inevitability of variations in humans' notion of relevance, how can information retrieval researchers confidently draw conclusions about system performance and the effectiveness of various techniques? Meta-evaluations have shown that while some measures such as recall are relatively meaningless in absolute terms (e.g., the total number of relevant documents cannot be known without exhaustive assessment of the entire corpus, which is impractical for current document collections), relative comparisons between systems are remarkably stable. That is, if system A performs better than system B (by a metric such as mean average precision, for example), system A is highly likely to out-perform system B with any alternative sets of relevance judgments that represent different notions of relevance (Voorhees, 2000; Voorhees, 2002). Thus, it remains possible to determine the relative effectiveness of different retrieval techniques, and use evaluation results to guide system development.

We believe that this philosophical starting point for conducting evaluations is an important point that summarization researchers should take to heart, considering that notions such as relevance and topicality are central to the evaluation of the information synthesis task. What concrete implications of this view are there? We outline some thoughts below:

First, we believe that summarization metrics should embrace variations in human judgment as an inescapable part of the evaluation process. Measures for automatically assessing the quality of a system's output such as ROUGE implicitly assume that the "best summary" is a statistical agglomeration of the reference summaries, which is not likely to be true. Until recently, ROUGE "hard-coded" the so-called "jackknifing" procedure to estimate average human performance. Fortunately, it appears researchers have realized that "model averaging" may not be the best way to capture the existence of many "equally good" summaries. As an example, the Pyramid Method (Nenkova and Passonneau, 2004), represents a good first attempt at a realistic model of human variations.

Second, the view that variations in judgment are an inescapable part of extrinsic evaluations would lead one to conclude that low inter-annotator agreement isn't necessarily bad. Computational linguistics research generally attaches great value to high kappa measures (Carletta, 1996), which indicate high human agreement on a particular task. Low agreement is seen as a barrier to conducting reproducible research and to drawing generalizable conclusions. However, this is not necessarily true—low agreement in information retrieval has not been a handicap for advancing the state of the art. When dealing with notions such as relevance, low kappa values can most likely be attributed to the nature of the task itself. Attempting to raise agreement by, for example, developing rigid assessment guidelines, may do more harm than good. Prescriptive attempts to define what a good answer or summary should be will lead to systems that are not useful in real-world settings. Instead, we should focus research on adaptable, flexible systems.

Third, meta-evaluations are important. The information retrieval literature has an established tradition of evaluating evaluations post hoc to insure the reliability and fairness of the results. The aforementioned studies examining the impact of different relevance judgments are examples of such work. Due to the variability in human judgments, systems are essentially aiming at a moving target, which necessitates continual examination as to whether evaluations are accurately answering the research questions and producing trustworthy results.

Fourth, a measure for assessing the quality of automatic scoring metrics should reflect the philosophical starting points that we have been discussing. As a specific example, the correlation between an automatically-calculated metric and actual human preferences is better quantified by Kendall's τ than by the coefficient of determination R^2. Since relative system comparisons are more meaningful than absolute scores, we are generally less interested in correlations among the scores than in the rankings of systems produced by those scores. Kendall's τ 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. Typically, a value of greater than 0.8 is considered "good", although 0.9 represents a threshold researchers generally aim for.

5 Conclusion

What's in store for the ongoing co-evolution of summarization and question answering? Currently, definition questions exercise a system's ability to integrate information from multiple documents. In the process, it needs to automatically recognize similar information units to avoid redundant information, much like in multi-document summarization. The other research direction in advanced question answering, integration of reasoning capabilities to generate answers that cannot be directly extracted from text, remains more elusive for a variety of reasons. Finer-grained linguistic analysis at a large scale and sufficiently-rich domain ontologies to support potentially long inference chains are necessary prerequisites—both of which represent open research problems. Furthermore, it is unclear how exactly one would operationalize the evaluation of such capabilities.

Nevertheless, we believe that advanced reasoning capabilities based on detailed semantic analyses of text will receive much attention in the future. The recent flurry of work on semantic analysis, based on resources such as FrameNet (Baker et al., 1998) and PropBank (Kingsbury et al., 2002), provide the substrate for reasoning engines. Developments in the automatic construction, adaptation, and merging of ontologies will supply the knowledge necessary to draw inferences. In order to jump-start the knowledge acquisition process, we envision the development of domain-specific question answering systems, the lessons from which will be applied to systems that operate on broader domains. In terms of operationalizing evaluations for these advanced capabilities, the field has already made important first steps, e.g., the Pascal Recognising Textual Entailment Challenge.

What effect will these developments have on summarization research? We believe that future systems will employ more detailed linguistic analysis. As a simple example, the ability to reason about people's age based on their birthdates would undoubtedly be useful for answering particular types of questions, but may also play a role in redundancy detection, for example. In general, we anticipate a move towards more abstractive techniques in multi-document summarization. Fluent, cohesive, and topical summaries cannot be generated solely using an extractive approach—sentences are at the wrong level of granularity, a source of problems ranging from dangling anaphoric references to verbose subordinate clauses. Only through more detailed linguistic analysis can information from multiple documents be truly synthesized. Already, there are hybrid approaches to multi-document summarization that employ natural language generation techniques (McKeown et al., 1999; Elson, 2004), and researchers have experimented with sentential operations to improve the discourse structure of summaries (Otterbacher et al., 2002).

The primary purpose of this paper was to identify similarities between multi-document summarization and complex question answering, pointing out potential synergistic opportunities in the area of system evaluation. We hope that this is merely a small part of a sustained dialogue between researchers from these two largely independent communities. Answering complex questions and summarizing multiple documents are essentially opposite sides of the same coin, as they represent different approaches to the common problem of addressing complex user information needs.

6 Acknowledgements

We would like to thank Donna Harman and Ellen Voorhees for many insights about the intricacies of IR evaluation, Bonnie Dorr for introducing us to DUC and bringing us into the summarization community, and Kiri for her kind support.

References

Enrique Amigó, Julio Gonzalo, Victor Peinado, Anselmo Peñas, and Felisa Verdejo. 2004. An empirical study of information synthesis task. In *Proceedings of ACL 2004*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING/ACL 1998*.

Carol Barry and Linda Schamber. 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34(2/3):219–236.

Jaime Carbonell, Donna Harman, Eduard Hovy, Steve Maiorano, John Prange, and Karen Sparck-Jones. 2000. Vision statement to guide research in Question & Answering (Q&A) and Text Summarization.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

William S. Cooper. 1971. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37.

David K. Elson. 2004. Categorization of narrative semantics for use in generative multidocument summarization. In *Proceedings of INLG 2004*, pages 192–197.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Jamie Callan. 2000. Creating and evaluating multi-document sentence extract summaries. In *Proceedings of CIKM 2000*.

Stephen P. Harter. 1992. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9):602–615.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of HLT/NAACL 2004*.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceeding of HLT 2002*.

Jimmy Lin and Dina Demner-Fushman. 2005. Automatically evaluating answers to definition questions. Technical Report LAMP-TR-119/CS-TR-4695/UMIACS-TR-2005-04, University of Maryland, College Park.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT/NAACL 2003*.

Jimmy Lin and Boris Katz. 2005, in press. Building a reusable test collection for question answering. *Journal of the American Society for Information Science and Technology*.

Inderjeet Mani, Therese Firmin, David House, Gary Klein, Beth Sundheim, and Lynette Hirschman. 2002. The TIPSTER SUMMAC text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Kathleen R. McKeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of AAAI-1999*.

Stefano Mizzaro. 1998. How many relevances in information retrieval? *Interacting With Computers*, 10(3):305–322.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The Pyramid Method. In *Proceedings of HLT/NAACL 2004*.

Jahna C. Otterbacher, Dragomir R. Radev, and Airong Luo. 2002. Revisions that improve cohesion in multi-document summaries: A preliminary study. In *Proceedings of the ACL 2002 Workshop on Automatic Summarization*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Tefko Saracevic. 1975. Relevance: A review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343.

Amanda H. Spink and Howard Greisdorf. 2001. Regions and levels: Mapping and measuring users relevance judgments. *Journal of the American Society for Information Science and Technology*, 52(2):161–173.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of EMNLP 2004*.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of SIGIR 2000*.

Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716.

Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems, Springer-Verlag LNCS 2406*.

Ellen M. Voorhees. 2003a. Evaluating the evaluation: A case study using the TREC 2002 question answering track. In *Proceedings of HLT/NAACL 2003*.

Ellen M. Voorhees. 2003b. Overview of the TREC 2003 question answering track. In *Proceedings of TREC 2003*.