Inf Retrieval (2006) 9:723–755
DOI 10.1007/s10791-006-9006-4

Precision prediction based on ranked list coherence

Steve Cronen-Townsend* · **Yun Zhou** · **W. Bruce Croft**

Received: 7 July 2004 / Accepted: 29 August 2005 / Published online: 9 September 2006

Abstract We introduce a statistical measure of the coherence of a list of documents called the *clarity score*. Starting with a document list ranked by the query-likelihood retrieval model, we demonstrate the score's relationship to query ambiguity with respect to the collection. We also show that the clarity score is correlated with the average precision of a query and lay the groundwork for useful predictions by discussing a method of setting decision thresholds automatically. We then show that passage-based clarity scores correlate with average-precision measures of ranked lists of passages, where a passage is judged relevant if it contains correct answer text, which extends the basic method to passage-based systems. Next, we introduce variants of document-based clarity scores to improve the robustness, applicability, and predictive ability of clarity scores. In particular, we introduce the ranked list clarity score that can be computed with only a ranked list of documents, and the weighted clarity score where query terms contribute more than other terms. Finally, we show an approach to predicting queries that perform poorly on query expansion that uses techniques expanding on the ideas presented earlier.

Keywords Performance prediction · Language models · Query expansion

S. Cronen-Townsend (✉) · Y. Zhou · W. B. Croft
Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, MA 01003
e-mail: steve.cronen-townsend@esko.com

Present address: Esko-Graphics, 40 Westover Rd., Ludlow, MA 01056
e-mail: stct@esko-graphics.com

Y. Zhou
e-mail: yzhou@cs.umass.edu

W. B. Croft
e-mail: croft@cs.umass.edu

1. Introduction

With the increasing use of Web search engines to search the enormous, diverse, and dynamic collection of text on the Web, it has become obvious to more people that some queries are not as effective as others. The difficult part of finding a desired piece of information is often the process of interactively guessing effective queries and repeatedly refining those guesses using the knowledge gained of what confusable documents are present in the index at search time. The process typically ends when a query is found that ranks a desired document in top 20, or so, documents.

The clarity score measures the coherence of a ranked list of documents, that is, the extent to which they use similar language. A query yielding a ranked list containing a huge variety of documents in word-usage at the top ranks can be distinguished from a query yielding a list of top documents that use language similarly, since it will have a lower clarity score. Thus clarity scores are related to the degree of *query ambiguity* in the collection, where query ambiguity is defined as the degree to which the query retrieves documents in the given collection with similar word usage.

Moreover, there is a relation between the coherence of a returned ranked list of documents and the chances of that list containing many relevant documents. In particular, in a ranked list containing documents of greatly differing word usage generally *at most one of the documents is relevant* since they are all very different. However, in a coherent ranked list containing documents with similar word usage, the likely options are that many of the documents are relevant or none of them are relevant. If the document list was ranked using the query, it is much more likely that many are relevant. Thus the more coherent a returned ranked list is, in general, the more likely it is to contain many relevant documents. So high clarity score (low ambiguity) queries are likely to perform better than low clarity score (high ambiguity) queries. Though they can be computed without any reference to what is being searched for, clarity scores are correlated with the effectiveness of the query in the given collection.

The use of a coherence-based score to predict retrieval effectiveness is particularly appropriate for content-based queries, since even in cases where there is only one relevant document, the system putting documents with similar content at the top of the list may still help predict a satisfactory result. Other types of queries used in web searching, such as "home-page" queries (Craswell and Hawking, 2003) often have only a single relevant document and prediction using coherence scores will consequently be less effective.

Information retrieval systems may now be imagined that treat individual queries differently depending on their predicted effectiveness. The correlation with average precision that may make this possible is explored along with simple ideas for automatically setting decision thresholds for decisions based on clarity score. In addition, the basic correlation is shown to hold up in a weaker form in a passage-based system, and important variants of the clarity scores are introduced.

Clarity scores are defined in Section 2 where they are based on documents as the unit of text. We go on to explore the facets of clarity scores that establish their usefulness as a research tool, their relation to ambiguity in Section 3 and their ability to predict retrieval performance in Section 4. We then introduce passage-based clarity scores in Section 5. The connection to ambiguity is the same for passage-based scores, so we move on to predicting passage question answering performance in Section 6 where the predictive ability is less than in the document context, but still intriguing. Next, we introduce important variants of clarity score in Section 7 to increase the robustness, applicability, and predictive power of document-based clarity scores. In particular, we introduce ranked list clarity scores, which are independent of document retrieval scores, in Section 7.1 and a version that weights query

terms in Section 7.2. In Section 8 we approach the task of predicting how the performance of a query is effected by query expansion. This study is important since it provides a first step towards a system trying to choose the best performing retrieval method for each query, thereby improving system performance. Clarity score-related ideas are built on creatively in this study, and Section 9 on other work bears out that clarity scores have not mainly been used in research for straightforward performance prediction, but have been used as a source of ideas to apply in information retrieval research.

2. Document-based clarity

The clarity score method was first introduced in Croft et al. (2001) and then studied more thoroughly in Cronen-Townsend et al. (2002). Here we introduce the basic technique and add improved estimation.

To calculate a clarity score for a query in a given collection, a *relevance model* is first formed (Lavrenko and Croft, 2001, 2003). This model represents the language usage in documents closely related to the query, and can be thought of as a collection-dependent query model. The relevance model is then compared to the *collection model*, representing the average language usage in the collection. The degree of difference between the two models is the clarity score for the query. Under this scheme, a query matching documents using very generic language (on average) receives a score near zero, and a query matching documents using a certain specialized vocabulary (on average) receives a relatively high score. Since not strongly preferencing documents of any common word usage is a good way to average documents to get something like the collection model, incoherent ranked lists lead to low clarity scores. Similarly, averaging with a bias for documents high in a coherent ranked list leads to higher clarity scores.

In our case, the query and collection models are examples of statistical language models (Croft and Lafferty, 2003). Our models approximate words as occurring independently and therefore are simply probability distributions over all terms in the collection vocabulary.

2.1. Document processing

In order to present meaningful clarity scores while explaining our methodology, we must first mention the processing of documents in this study to allow estimation of language models. All the calculations and experiments presented in this paper use the Lemur Toolkit[1] for language modeling-based information retrieval (Ogilvie and Callan, 2002). We index collections with all characters converted to lower case and punctuation characters replaced with spaces. We then remove single characters and digits, and terms on the InQuery stop list (Broglio et al., 1994).

We use Krovetz stemming (Krovetz, 1993) to attempt to put words with different suffixes, for example, into the same class. As was noted in Cronen-Townsend et al. (2003), stemming improves clarity scores by enabling them to better compute coherence. In particular, having the system treat occurrences of "computer" as distinct from "computers" can lead to meaningless variation in coherence measures due to variation in the relative usage of "computer" and "computers" between documents.

[1] Freely available from `http://www.cs.cmu.edu/~lemur/`.

2.2. Calculating clarity

We first explain the simplest model, the collection model. The probability for each term in the collection vocabulary is simply estimated as the relative frequency of the term in the collection (the number of times the term occurs in the collection divided by the the total number of term occurrences in the collection).

We discuss document models next, since documents form the large units of text in this section. The relevance models used to compute clarity scores are formed by weighted mixtures of document models.

We estimate the model for document, D, by relative frequencies of terms in the document linearly combined with collection relative frequencies. In particular,

$$P(w|D) = \lambda P_{ml}(w|D) + (1 - \lambda)P_{coll}(w), \tag{1}$$

where $P_{ml}(w|D)$ is the relative frequency of term w in document D, $P_{coll}(w)$ is the relative frequency of the term in the collection as a whole. A constant value of λ gives Jelinek-Mercer smoothing (Zhai and Lafferty, 2001), as was used in the original clarity score work (Cronen-Townsend et al., 2002). Using a document-dependent value of λ, given by

$$\lambda = \frac{\|D\|}{\|D\| + \mu}, \tag{2}$$

gives Dirichlet smoothing (Zhai and Lafferty, 2001) with prior μ, as can verified by substitution. Both types of smoothing are used in this work.

The relevance model is given by

$$P(w|Q) = \sum_{D} P(w|D)P(D|Q), \tag{3}$$

where w is any term, Q the query, D is a document or the model estimated from the corresponding single document, and summation is done over all documents. This can be interpreted as a weighted average of document models, $P(w|D)$, with weights given by $P(D|Q)$.

For the weights, $P(D|Q)$, in Eq. (3) we perform a query likelihood retrieval step (Song and Croft, 1999). We estimate the likelihood of an individual document model generating the query as

$$P(Q|D) = \prod_{q \in Q} P(q|D), \tag{4}$$

and obtain $P(D|Q)$ by Bayesian Inversion with uniform prior probabilities for documents. The document-independent term $P(Q)$ is incorporated by requiring that $P(D|Q)$ sum to one over all documents in the collection.

The clarity score for the query is simply the relative entropy, or Kullback-Leibler (KL) divergence (Cover and Thomas, 1991), between the query and collection language models (probability distributions),

$$\text{clarity score} = D(P(w|Q)\|P_{coll}(w)). \tag{5}$$

This is given by

$$D(P(w|Q)\|P_{coll}(w)) = \sum_{w \in V} P(w|Q) \log_2 \frac{P(w|Q)}{P_{coll}(w)}, \quad (6)$$

where V is the entire vocabulary of the collection.

The efficiency of clarity score computation with Eq. (6) is dominated by the estimation of the relevance model by Eq. (3), since the collection model is precomputed at index-time. Throughout this paper we estimate relevance models for document-retrieval by truncating the summation in Eq. (3) at the top 500 documents. Since $P(D|Q)$ generally falls off sharply well before this cutoff, this cutoff has very little effect on the clarity scores.

2.3. Smoothing

Clarity score computation uses document models in two separate steps. The first step is estimating the likelihood of query generation by a model of each document using Eq. (4). The second step uses mixing weights derived from those scores to combine document models via Eq. (3).

The original clarity work used the same Jelinek-Mercer smoothing for both of these steps with $\lambda_1 = \lambda_2 \equiv 0.6$ in Eq. (1) (Cronen-Townsend et al., 2002). We find it important for performance to smooth the document models with Dirichlet smoothing for the first step (query likelihood, subscript 1) and Jelinek-Mercer smoothing for the second step (mixing, subscript 2). This is consistent with what is found in relevance model retrieval, where the relevance model forms the expanded query and is used for retrieval (Lavrenko, 2004). Our finding is also consistent with Zhai and Lafferty (2001), who tested Jelinek-Mercer smoothing, Dirichlet smoothing and absolute discounting for information retrieval and found that Dirichlet smoothing outperforms Jelinek-Mercer smoothing on all of the 9 test collections they tested with short queries. Since the first step to forming a relevance model is a query likelihood retrieval step, it is not surprising that a better retrieval result (i.e. with Dirichlet smoothing) leads to a better model for clarity score computation.

In this paper we use two smoothing conditions for clarity calculation. Both use $\mu_1 = 1000$ for the query likelihood retrieval step. *Light smoothing* uses $\lambda_2 = 0.9$ for the document models in the mixing step, and *heavy smoothing* uses $\lambda_2 = 0.1$ in the mixing step. Thus, in heavy smoothing, the probability estimate for each term is made up of only 10% of its document relative frequency and 90% of its collection relative frequency.

The two smoothing conditions were chosen by exploring the variation of a performance measure (rank correlation with average precision, see Section 4.1 for full details) as a function of λ_2 for all collections studied when $\mu_1 \equiv 1000$. As the λ_2 is decreased from $\lambda_2 = 0.9$, the behavior of performance on all collections is similar. The performance increases slightly (nearly monotonically) as λ_2 is decreased to values below 0.1. At some small value of λ_2 (typically around 0.001) the performance starts falling off. Since the point where the performance falls off varies somewhat by collection, we chose $\lambda_2 = 0.1$ as a safe level of smoothing that was far from the falling off point for all collections. Thus our heavy smoothing condition of $\lambda = 0.1$ seems unlikely to be too much smoothing for untested collections and should offer reliably good performance for untested collections.

We use heavy smoothing and refer to it as *standard* smoothing for the basic document-based clarity scores described so far. Light smoothing clarity scores differ mainly in scale (they are larger) and are used for some examples where ease reading score values, is desired.

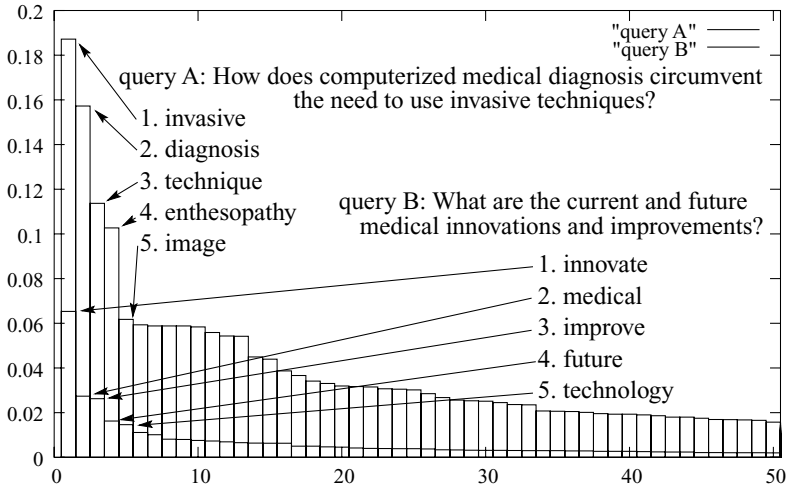

Fig. 1 Term clarity score contributions for the top terms for two same-topic queries from the TREC Query Track

The slight reordering that occurs in a query set ranked by clarity score when one changes from light to heavy smoothing benefits the correlation with average precision slightly but makes some displays of values harder to understand.

2.4. Two examples

Figure 1 shows the top 50 individual term contributions to the summation in Eq. (6) (clarity score) for two related queries[2] using our light smoothing settings (see Section 2.3) to make the values larger and more readable. The queries are query A: "How does computerized medical diagnosis circumvent the need to use invasive techniques?" with a clarity score of 3.53 and query B: "What are the current and future medical innovations and improvements?" with a score of 0.73. Top contributing terms are those whose probabilities most stand out in comparison to the collection model, such as "invasive" and "diagnosis" for query A. The total clarity score is the total of all bars for the appropriate query if one imagines extending the plot to include all vocabulary terms. The figure shows that the relevance model for query A is much more unusual than the collection model, and shows the contributions for these spikes. This is due to its highly-scoring documents using the same certain terms, what we call coherence. Query B has much lower maximum contributions and a lower total (clarity score). The fact that the medical term "enthesopathy" was one of the top contributing terms in query A's clarity score while the top terms of query B's language model are all fairly general is a good indicator that query A is a better performing query than query B. For Query A, "Enthesopathy" occurred in documents that had high query likelihood scores, leading to an estimate of its probability in the relevance model well above its collection model probability. For Query B, by contrast, the only terms that stand out are fairly general terms, indicating that $P(D|Q)$ is less focussed (peaked) on a coherent set of documents. Without reference to the information need (i.e. without having a system understand what the user wants) definitive

[2] Both from the TREC 9 Query Track and designed for TREC topic 96.

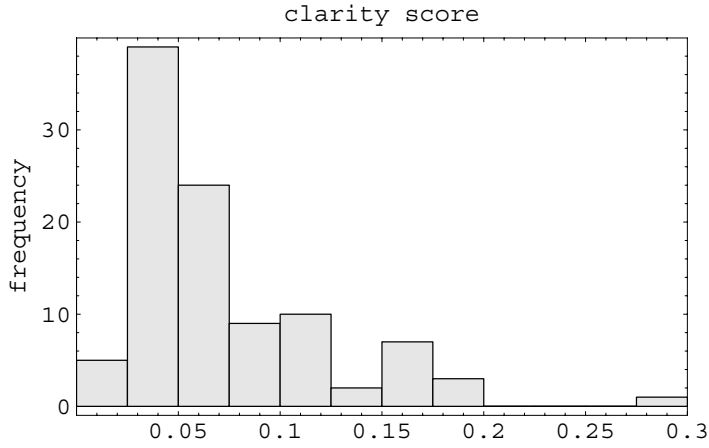

Fig. 2 Histogram of the standard clarity score for the 100 queries of TREC 7 + 8

prediction is impossible. However, the coherence of the ranked list for query A is a good indicator, since the documents that use language similarly could mostly be relevant to the user.

In Fig. 2 the clarity scores computed using standard (heavy smoothing) clarity scores on TREC 7 + 8 are shown. TREC 7 + 8 refers to the documents from TREC disks 4 and 5 excluding Congressional Record and queries that are the titles of topics #351 − #450. This test set was chosen since the standard clarity scores of its queries form the basis of many later figures in this paper. The histogram serves to suggest a typical density distribution for the clarity scores of a set of well-designed queries. The rightmost bar is for the single query "anorexia nervosa bulimia" with clarity score 0.28. The query "postmenopausal estrogen Britain" with clarity score 0.19 contributes to the next highest bar and "supercritical fluids," (clarity score 0.16) is one of the contributors to the third highest scoring bar. The query "illegal technology transfer" with clarity score 0.019 is one of the lowest scoring queries in the set and contributes to the leftmost bar. Note that the first three queries listed, with relatively high clarity scores, all plausibly match coherent text in the collection. The query "illegal technology transfer" consists of terms co-occurring in documents in varying contexts and gets a relatively low clarity score.

3. Query ambiguity

Every query to an information retrieval system is ambiguous to some degree, in the sense that it could specify the word usage in a relevant document to a greater or lesser extent. We use the term *query ambiguity* to represent this nonspecificity. This usage is related strongly to the common English usage of the term ambiguity, but is notably different from the term's meaning in linguistics, for example, where it relates to how difficult it is to decide the truth of a statement. Since queries are not statements, our usage need not cause confusion.

One way of quantifying query ambiguity with respect to a collection of documents, is with a measure of the coherence of the top retrieved documents. A query that has top-scoring documents all using the same rare terms in similar proportions is taken to be a low ambiguity query. Whereas a query returning a mixture of documents all using different rare terms is

taken to be a high ambiguity query. Clarity scores can be thought of as a quantification of query ambiguity in this way; high clarity score queries are low in ambiguity, and low clarity score queries are high in ambiguity. Since coherence of the top documents has nothing to do with relevance, the ambiguity of a query, thought of in this way, does not depend on the notion of relevance or even on the existence of *any* relevant documents in the collection. This interpretation also highlights the role of the collection in clarity scores, since a query can only be scored using the documents present in a certain collection. The same query could have a relatively high clarity score in one collection and a relatively low clarity score in a different collection.

For example, imagine a user interested in news mentions of World Cup soccer issuing the query "World Cup" to an information retrieval system accessing the TREC AP88 collection of news articles from the Associated Press in 1988. Although there are articles mentioning World Cup soccer in the collection, articles about World Cup chess tournaments are predominant among the articles that use the terms "world" and "cup" most frequently. If those two query terms are the only evidence the system has about what the user means, it is impossible for the system to return the soccer articles consistently higher in the ranked list than the articles about World Cup chess tournaments. In this example, the query's ambiguity is greatly increased by the particular documents that happen to be present; if the documents about World Cup chess were not in the collection, the query would be less ambiguous.

Despite the fact that the user might not have known that there was a World Cup in anything other than the sport of soccer, he or she would, typically, get a ranked list with chess articles predominating near the top and with some soccer articles mixed in. Since the mixed list is less coherent (and therefore closer to the collection model) than a list of documents all about world cup soccer, the clarity score measure can detect such query ambiguity. Most real-world query ambiguity is not between things as close as World Cup soccer and World Cup chess.

The relationship between clarity score and query ambiguity is further demonstrated in Fig. 3. This figure shows the construction of an artificial highly ambiguous query ("fashion model railroad", bottom center) by adding single terms, and gives the (light smoothing) clarity score for each related query. Light smoothing was chosen to make the figure more clear by making the scores of order 1 and the score differences larger than if heavy smoothing were used. To the left and right sides of the intentionally highly ambiguous query are related queries with much less ambiguity. Arrows between boxes indicate the addition of a single query term.

The lowest clarity score query in Fig. 3 is the single term query "model", but the score of the most ambiguous three term query "fashion model railroad" is only negligibly higher. The related three term queries "model railroad gauge" and "fashion model photographer" have about double the clarity score. Moreover, the two two-term queries (sharing two of its three

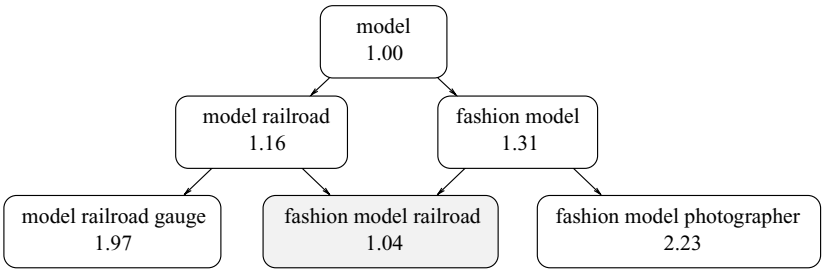

Fig. 3 Standard Clarity scores some queries to TREC 1+2+3 related by addition of a single term (arrows)

terms) are both significantly higher in clarity score. This example shows that combining the two senses (left side and right side in Fig. 3) created a low clarity score query, relative to other three term queries. Moreover, this example suggests that comparisons between clarity scores of differing length queries are sensible, and not dominated by query length. This is consistent with clarity scores being based on KL-divergence between probability distributions over the same events, and hence on the same scale for queries of different lengths.

Further discussion of the relationship between clarity scores and ambiguity can be found in Cronen-Townsend and Croft (2002).

4. Document retrieval prediction

A query with a relatively high clarity score is likely to be effective in retrieving relevant documents in a simple information retrieval system. The aim of this section is substantiating this claim.

4.1. Document retrieval

For document retrieval we use the query likelihood method. Scoring documents in this way has already been described as the first step of constructing a relevance model in order to compute a clarity score (Section 2.2). We use the identical query likelihood scores, again with Dirchlet prior $\mu = 1000$, for document retrieval.

4.2. Measuring performance

We measure the performance of each query by its average precision score with query likelihood retrieval, computed by the `trec_eval` package (Buckley, n.d.). The precision of a set of documents is defined as the fraction of relevant documents in the set and the average precision of a ranked list is the mean of the precision scores for ordered sets up to and including each relevant document. Average precision is an overall measure of the quality of a ranked list.

Because the probability distributions of the scores are unknown, an appropriate, distribution-free, test of correlation is the Spearman rank correlation test (Gibbons and Chakraborty, 1992). A rank correlation of 1 indicates perfect agreement between the ranking by average precision and the ranking by clarity score, and a rank correlation of -1 indicates opposite rankings. The null distribution (the distribution of the score if the rankings are unassociated) is well-approximated by a normal distribution for sample sizes as large as 50 (our smallest sample size in this study), making estimation of p-values feasible.

4.3. Document-based clarity and average precision

The rank correlation between standard clarity scores and retrieval average precision (uninterpolated) is shown in column 5 of Table 1 for a variety of TREC collections. The seventh line of the table, for example, means that using a combination of TREC 7 and 8 Ad Hoc Tracks, with the documents of TREC disks 4 and 5 minus the Congressional Record, and title queries from TREC topics 351-450, the rank correlation coefficient of the standard clarity score with average precision is 0.62. Retrieval is done by ranking each document by its model's likelihood of generating the query, using Eq. (4). The document model, Eq. (1), is estimated for each document using Dirichlet smoothing with $\mu = \mu_1 = 1000$ in Eq. (2). Relevance models are mixed from Jelinek-Mercer smoothed document models with $\lambda = \lambda_2 = 0.1$

Table 1 Correlation between ranking queries by average precision and by standard clarity score

TREC	Disks	Topics	Number	Spearman R
$1+2+3$	1&2	51–150	100	0.50
4	2&3	201–250	50	0.49
5	2&4	251–300	50	0.46
6	4&5	301–350	50	0.61
7	4&5-CR	351–400	50	0.64
8	4&5-CR	401–450	50	0.61
$7+8$	4&5-CR	351–450	100	0.62
agg QT	1	51–100	1804	0.55

(heavy smoothing condition) in Eq. (1). Heavy smoothing is chosen for reliably high correlation with average precision, as explained in Section 2.3. All queries are titles of TREC topics (average length 2.9, sample standard deviation 1.1) except for the query track where there are an average of about 36 unique queries (after processing) of varying length (average 4.8, sample standard deviation 2.3) for each of the topics 51–100 and TREC 4 where the description fields are used instead (average length 8.4 terms). The highest p-value for a set of standard standard clarity scores in this table is 0.0006 (TREC 5) and the lowest is 8×10^{-120} (aggregate Query Track), indicating that all correlations are statistically significant.

The strength of the correlations (shown in column 5 of Table 1) is their real importance. The small p-values indicate that the correlation is extremely unlikely to occur by chance in unrelated rankings. The extent of the correlation is also visible in Fig. 4 as a linear trend for average precision of queries to increase as their score increases. The rightmost circled point represents the query "anorexia nervosa bulimia" with clarity score 0.28. The circled point with clarity score 0.19 and average precision near zero is "postmenopausal estrogen Britain", and the circled point at clarity score 0.16 and average precision nearly 0.9 represents

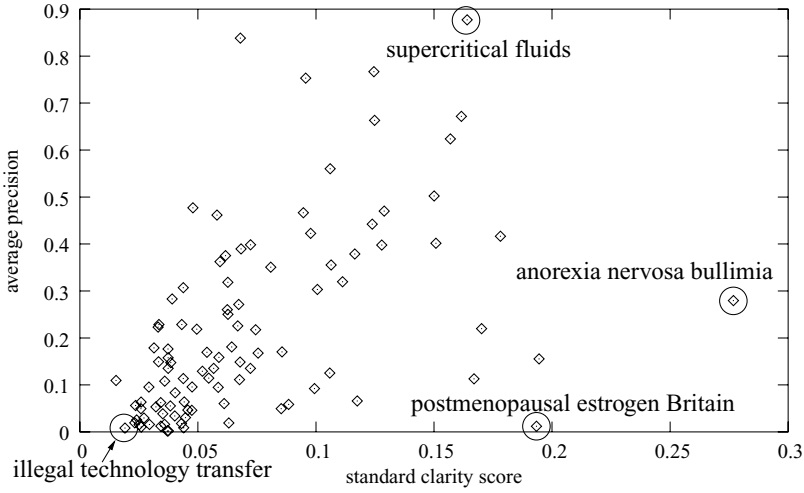

Fig. 4 Scatter plot of average precision versus standard clarity score for the 100 queries of TREC $7+8$

"supercritical fluids." The query "illegal technology transfer" is represented by the circled point with clarity score 0.019 and average precision near zero.

4.4. Decisions using document-based clarity

An information retrieval system can now be imagined that treats queries predicted to be ineffective differently than other queries. For example, such a system might suggest that the user try improving a suspicious query, rather than simply showing search results for the query. To make such a prediction based on clarity score, the system would need a decision threshold to indicate how low a clarity score was sufficient to predict that the query was going to be ineffective. Queries scoring below the threshold would be handled differently from those scoring above the threshold. Computing such a threshold without relevance information is addressed in this section. Such a system would make some errors, and the error rate would be crucial in determining whether the prediction would be helpful or not. This further supports the strength of the correlations (i.e., $|R|$) being the important statistic, since a higher correlation leads to fewer predictive errors.

For straightforward use of clarity scores, a system must decide how high (or low) a clarity score is sufficient to justify some decision about how to treat that query. Our basic idea is to estimate the probability distribution over clarity scores using single term queries randomly sampled from the collection vocabulary. Estimating the clarity score distribution in this way can be done at index-time for any collection.

Continuing as in Cronen-Townsend et al. (2002), we set a threshold simply by requiring that a query have a higher score than a certain set percentage of one-term queries. The exact percentage can be set to roughly match estimates of an optimal threshold for the task at hand determined using test collections.

To demonstrate the approach, we chose the "task" of deciding if a given query is in the upper or lower half of queries in a given test set in average precision performance. We compute estimated Bayes-optimal thresholds for classifying queries as in the upper half or lower half of average precision for each test set as in Cronen-Townsend et al. (2002). Then, for each of the test collections, we estimate a one-term query distribution of standard clarity scores by randomly sampling the vocabulary to form queries. This sampling is independent of any human-generated queries, and gives some sense of the distribution of clarity scores for the collection at hand. We found that 80%-thresholds for standard clarity best matched the Bayes-optimal thresholds for predicting whether a test query was in the upper or lower half of the set in performance. The results are shown in Table 2. The "relative" column gives the relative discrepancy between the automatic threshold value and the optimum threshold. We are pleased to note that the use of an 80% threshold for document-based clarity scores

Table 2 Automatic and Bayes optimal thresholds for standard clarity scores

TREC	auto:80%	Optimal	Relative
1+2+3	0.066	0.044	+50.%
4	0.047	0.082	−43%
5	0.054	0.050	+8.0%
6	0.053	0.074	−28%
7, 8, 7+8	0.049	0.060	−18%
agg QT	0.053	0.046	+15%

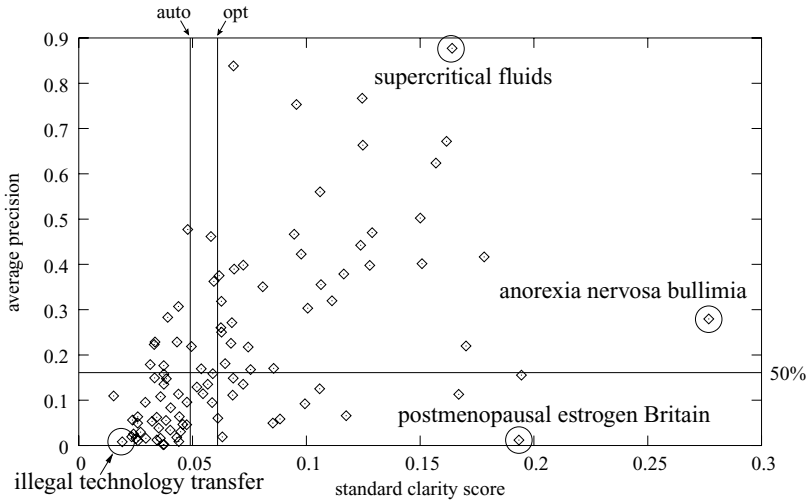

Fig. 5 Scatter plot of average precision versus standard clarity score for the 100 queries of TREC 7 + 8 with thresholds indicated

originally proposed for $\lambda_1 = \lambda 2 \equiv 0.6$ clarity still holds up for the new clarity scores with our heavy smoothing condition ($\mu_1 = 1000$, $\lambda_2 = 0.1$). We refer to this as the automatic method for setting thresholds.

Figure 5 shows the 50% threshold in average precision (labeled "50%") as well as the Bayes-optimal threshold in clarity score (labeled "opt"). These lines divide the scatter plot into four quadrants. Queries in the upper right quadrant and the lower left quadrant have correct predictions while queries in the upper left and lower right quadrants have mistaken performance predictions based on their clarity scores.

In Fig. 5 the automatic clarity score threshold is also indicated and labeled "auto." It is interesting to note that switching from the Bayes optimal threshold to the automatic threshold for this data results in about the same number of queries switching from incorrect to correct classification as switch from correct to incorrect classification. That is to say, there are about as many high- as low-performing queries in the region between the optimal and automatic thresholds, so the Bayes error is close to uneffected by changing from the optimal to the automatic threshold in this case. Thus the relative discrepancy of -18% from the optimum threshold value to the automatic threshold value is not very important.

5. Passage-based clarity scores

We seek to demonstrate that clarity scores function similarly with passages as the multiple-word units of text as they do for document-based retrieval systems. To do this we demonstrate the correlation of passage based clarity scores with overall measures of the quality of a ranked list of passages.

In our case, passage retrieval is provided by a straightforward passage-based question answering system using query likelihood retrieval (Corrada and Croft, 2004).[3] Given a question,

[3] With answer modeling turned off.

the passage question-answering system first performs a document retrieval step and then creates highly-overlapping passages from the top documents. Finally, it scores the best passage from each document to form a ranked list of passages. This list, with no processing, is the output of the system.

5.1. Passage processing

As a first step in obtaining passages, an initial document retrieval is done with a Lemur tf.idf retrieval with the question forming the query, in order to get a wide range of question-related documents from which to extract passages. Passages are then created by sliding a 250 character window through each retrieved document with a one word increment and recording a passage for each position of the window. For each document, all overlapping passages are scored by query likelihood and the best scoring passage from each document is retained for the ranked list. The ranked lists are truncated at the 90 top-scoring passages, since that value attains near-optimum MRR performance over each of TREC 2000, TREC 2001, and TREC 2002 question answering test beds with this system. These 90 passages are used for estimating a relevance model for clarity score calculation.

5.2. Calculation

We calculate passage-based clarity scores as in the document-based case in Section 2.2, with the 90 scored passages in the role of the documents, D. The collection model, C, still refers to a model of all the collection's documents aggregated together.

For the passage-based case we adjusted several parameters that are left fixed in the document-based case. In particular, we adjusted the passage cutoff (restricted to be less than or equal to 90) and the number of top terms compared to the collection model to form clarity scores (a restriction on the number of terms summed over in Eq. (6)). We use a variety of Dirichlet prior values significantly lower than the 1000 that is our fixed standard setting for documents. Adjusting these parameters was necessary due to the smaller size of passages (hence smaller Dirichlet priors) and the different patterns of word occurrence that occur in answers to questions of a certain type (e.g. questions whose answer is a location). The necessity of adjusting these parameters based on the test data, at times, to achieve meaningful clarity scores indicates the passage-based scores are less robust than the document-based scores. This is thought to be due to the smaller samples of text used to estimate the underlying models.

6. Passage question answering prediction

It is desirable for an operational question answering system to have the ability to identify questions that the system will not be able to answer well. Current systems are designed for questions with short factual answers. Even within this restricted domain, there are many questions that may cause a system to match diverse text from the collection, such as "What time of year do most people fly?". For the simple passage retrieval system we have described, we now demonstrate that the coherence of the language use in a ranked list of passages is correlated, in many cases, with the degree to which passages containing correct answer text appear high in the ranked list.

In a full question answering system that goes on to extract brief answer text from the ranked list of passages, this could be used to decide when not to answer a question, rather

than giving an unreliable answer. This option would be taken in cases when passage retrieval performance was predicted to be low enough that answer extraction is unlikely to succeed. A version of this started being incorporated in the TREC QA track in 2002 (Voorhees, 2002). In TREC evaluations a system is now allowed to answer "NIL" to a question indicating its belief that no answer exists in the document collection.

In this study we focus on validating clarity scores as tool in passage retrieval question answering, where the system returns a ranked list of passages. Passages are deemed relevant if they contain correct answer text. We show that there exists an association between the clarity score of a ranked list of passages and average-precision based measures of the ranked list. This association is analogous to the association between between document-based clarity scores of queries and average precision.

6.1. Passage retrieval

Questions have single characters and digits removed and are Krovetz stemmed (Krovetz, 1993) as is the collection, the same as was done in the document-based case. Passages are scored for retrieval by query likelihood (Eq. (4)) with Dirichlet smoothing with prior μ adjusted for each question type. As was already described in Section 5.1, the best scoring passage from each document is retained and the ranked list truncated to the 90 top-scoring passages. The redundancy in explanation is due to the fact that for passage-based systems, as well as for document-based systems, a query likelihood retrieval is a necessary first step in clarity score calculation, leading to its already having been described in that context.

6.2. Measuring performance

For evaluation purposes, passages that contain correct answer text within them (that is, they match a NIST-supplied pattern) are judged relevant to the question. So passages take the role of documents in a document-based system and matching a supplied answer text pattern takes the role of relevance.

Mean reciprocal rank (MRR) is a standard metric for evaluating the performance of question answering systems (Voorhees, 2000) but measures the list only down to the first relevant document, and not overall. Since clarity scores measure the overall coherence of a ranked list, they correlate best with an overall measure of ranked list quality.

As noted in Cronen-Townsend et al. (2003), average precision-based measures should relate better to the ease of extracting one brief answer from the passages. This is because answer extraction will be easier, in general, if the answer text is repeated often in the ranked list. This repetition is something that does not effect MRR at all, but does increase average precision based measures.

We measure performance by the average precision of the ranked list of answer passages. The definition is the same as in document-based retrieval with the passages treated as small documents. In our case, the set of relevant passages is all the highly overlapping passages from the top ranked documents that contain the correct answer text. Consider, for example, a document containing one instance of the answer. The system starts a window at the beginning of the document and makes one passage for each position of the window as it advances a word at a time. This process begins generating relevant passages as soon as all the correct answer text is inside the window and continues to generate one relevant passage per window position until some piece of the correct answer text exits the other side of the window. Since the system only ranks one passage per document, it is penalized in average precision for not returning all the other overlapping relevant passages from any document where the system

Table 3 Correlations for TREC 2000, 2001, and 2002 QA

Collection	Num.	R
TREC 2000	692	0.225
TREC 2001	500	0.196
TREC 2002	499	0.148

finds a relevant passage. We call this measure the *overlap-penalized average precision.* It forms an overall measure of the quality of a ranked list of passages in our system, but with values on a small scale.

Since there are about 50 relevant passages, on average, per document containing relevant passages in our collections, the maximum overlap-penalized average precision is reduced by about a factor of 50 from a typical document retrieval experiment. Having the system score every overlapping passage would create a new source of variability in the clarity scores, since the nearly identical passages in the ranked list with nearly identical scores would amount to coherence in the ranked list. What is important for this study is that overlap-penalized average precision measures the overall quality of the ranked list returned by the system, and this is the same ranked and scored list that is used to compute clarity scores.

The other extreme for measuring performance is to only judge the ranked list based on the relevance of passages it contains. This limits the degree to which the system is penalized for a poor document retrieval step; as long as at least one relevant passage was extracted, a perfect score of one is possible. We call this measure the *reduced average precision*, since the number of passages counted as relevant is reduced.

6.3. Passage-based clarity scores and average precision

We begin examining the relationship between passage clarity scores and passage question answering performance by considering the three entire test sets of questions for the TREC 2000, TREC 2001, and TREC 2002 Question Answering tracks. We use overlap-penalized average precision (as defined in Section 6.2) as our primary performance measure.

Table 3 shows the correlation between question clarity score and overlap-penalized average precision in our system in the three TREC test beds studied. The correlations are all significant with the highest P-value (for TREC 2002) being 5×10^{-4}.

Table 4 shows the system parameters used to obtain the Table 3 results. The "passages" column refers to the number of passages mixed to form query models (cutoff in Eq. (3)). It is analogous to the cutoff of 500 document models used in document retrieval and has a maximum of 90 because the passaging system only returns 90 scored passages. The "terms" column indicates the number of (truncated) terms in the relevance models, which are compared with the collection model to form the clarity score in Eq. (6). "MRR" indicates mean

Table 4 Tuned system parameters and two measures of system performance on TREC 2000, TREC 2001, TREC 2002 aggregate QA data

Question type	μ_1	λ_2	*Passages*	*Terms*	MRR	No correct
TREC 2000	440	0.2	90	10	0.402	33%
TREC 2001	230	0.2	90	10	0.319	45%
TREC 2002	140	0.2	20	10	0.288	52%

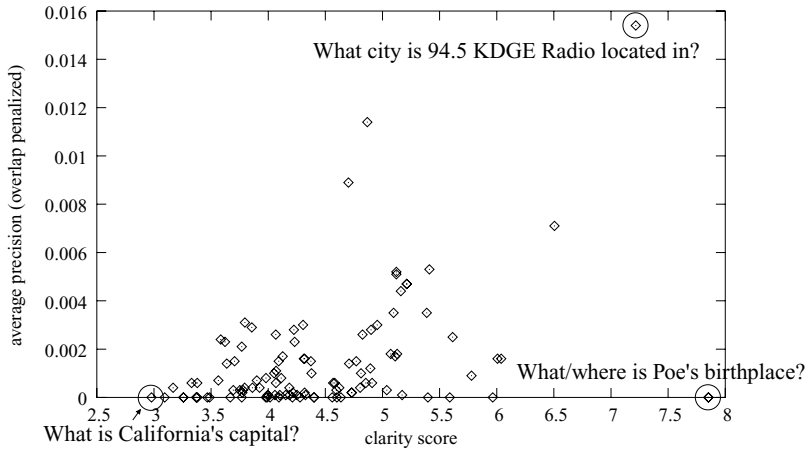

Fig. 6 Scatter plot of the overlap-penalized average precision versus clarity score for the 109 "Location" questions of the TREC 2000 QA Track

reciprocal rank with credit for first correct passages within the top 20, and "no correct" is the number of queries with no correct answers among the passages extracted. We speculate that small number of top terms (e.g. 10) helps correlations by exaggerating the difference between the most focussed relevance models and those formed, for example, for questions where none of the passages in the system are actually relevant (the "no correct" case).

The strength of the correlations in the aggregate question sets (Table 3) is significantly lower than the correlations found with clarity scores in document retrieval, making it very difficult to imagine using these correlations for predictive purposes. To explore this situation and to uncover meaningful cases where clarity scores correlate more strongly with performance, we study the TREC 2000 (TREC 9) data broken down by question answer type. We use the University of Pennsylvania classification scheme for the TREC 2000 questions (Morton, 2001) focusing on answer type. We use the questions in the categories "Amount," "Famous, "Location," "Person," "Time," and "Miscellaneous" which each contain enough questions for collecting meaningful statistics.

Figure 6 shows the scatter plot of overlap-penalized average precision versus clarity score for each of the 109 "Location" questions in the TREC 2000 Question Answering data. "Location" questions are the class where performance is most correlated with average precision, with Spearman $R = 0.358$. Four extreme scoring queries are circled in the figure, though two of the circles coincide. The highly scoring and highly performing question is "What city is 94.5 KDGE Radio located in?" This question matches coherent passages that usually contain the correct answer. The high-scoring and zero average precision circle encloses two identically performing questions "What is Poe's birthplace?" and "Where is Poe's birthplace?" and the low-scoring and zero average precision circled query is "What is California's capital?". In these cases, the passaging system fails to create any passages containing correct answer text, assuring that the average precision of the ranked list will be zero. The clarity score calculation takes the same ranked list and gives it a coherence-based score that is some positive value. These are what we refer to as system-unanswerable questions.

Table 5 shows the correlations over the six question classes containing sizable numbers of questions. By testing our system on each class separately we found interesting variations

Table 5 Correlations broken down by question type for TREC 2000 QA Track data

Question type	Num.	R	P-value
A (amount)	53	uncorr	–
F (famous)	83	0.133	0.11
L (location)	109	0.358	0.00010
P (person)	113	0.256	0.0033
T (time)	73	0.309	0.0044
X (misc)	168	0.254	0.00053
AFLPTX	599	0.233	6×10^{-9}

Table 6 Tuned system parameters and two measures of system performance on TREC 2000 QA data

Question type	μ_1	λ_2	*Passages*	*Terms*	MRR	No correct
A (amount)	450	0.1	90	10	0.185	53%
F (famous)	140	0.4	90	10	0.560	13%
L (location)	740	0.1	25	35	0.430	25%
P (person)	430	0.4	90	100	0.515	32%
T (time)	380	0.2	90	10	0.210	59%
X (misc)	670	0.6	30	10	0.386	31%
AFLPTX	440	0.4	90	10	0.402	33%

over the question types in clarity scores' correlation with performance in our system. We found no correlation for "Amount" questions, and the strongest correlation for "Location" questions. By separating the questions in to classes we also gained the ability to tailor the system slightly for each class of question. Table 6 shows the parameter settings used in the calculation leading to Table 5.

We believe much of the variation over question type seen in Table 5 is due to intrinsic statistical differences to correct answer passages in our systems for the different classes. These rank correlations are comparable numerically to the results found with the original clarity scores and a similar, but sentence-based passaging system with non-overlapping passages (Cronen-Townsend et al., 2003). However, these results are significantly stronger since we obtain similar rank correlation without removing system-unanswerable questions.

For example, in Fig. 6, there are 27 system-unanswerable questions, with no relevant passages ranked by the system. These queries get zero average precision, and lie along the x-axis, but they have non-zero clarity scores depending on the coherence of the ranked lists they return. Such cases hurt the correlation of clarity with average precision by forming many points tying for the lowest average precision rank, spread out over the x-axis in the scatter plot. Despite the mid-ranking correction we apply, many ties in the rankings diminishes the meaning of the Spearman R statistic (Gibbons and Chakraborty, 1992). Heuristically, they are also consistent with a linear relationship of zero slope on the scatter plot, and being mixed with answerable questions leads to lower correlation values and less ability to predict retrieval performance.

The system having no correct passages for a given question is usually due to a shortcoming of the system taking 1 best passage per document and selecting a (higher scoring) nonrelevant passage rather than a relevant one, but it is sometimes a deficiency of the preliminary document

retrieval step and occasionally a question with no correct answer in the entire collection. Since a searcher does not generally know if a given collection or system contains an answer to a factual question at the time the question is posed, the robustness of our technique (in the test beds studied) to a small proportion of unanswerable questions is promising.

We next examine an alternative measure of our system's performance, where the system is still evaluated with average precision, but it is the *reduced* average precision where the relevant passages are considered to be the passages that the system actually retains (best-scoring passage in their document and one of the best 90 overall) that also contain correct answer text. A system-unanswerable question (no relevant passages in the ranked list) still receives a score of zero on this measure, but as long as there is at least one relevant passage for the question in our system, a perfect score of 1 is possible. Here the system is not explicitly penalized for any passages that are not returned (since they are all considered nonrelevant). Figure 7 is a scatter plot of reduced average precision versus clarity score for the "Location" questions in the TREC 2000 data. The same four queries are labeled as in Fig. 6. One can see that a perfect score (reduced average precision equals 1) is sometimes obtained. The degree of correlation is visually similar to that in Fig. 6 and the rank correlation is $R = 0.320$ which is slightly less than for the overlap-penalized case. This is probably due to the increased number of ties for the top average precision rank, hurting the meaningfulness of the statistic. For this reason, we rely on overlap-penalized average precision for evaluating our system and use it for all other tables and figures.

Studying the reduced average precision performance of our system offers one surprise. The correlation for "Miscellaneous" questions in the TREC 2000 data that exists for overlap-penalized average precision does not exist for the alternative measure. In all other question classes the correlation is merely reduced from that observed with overlap-penalized average precision. This seems to be an artifact of a large number of (tying) system-unanswerable questions, a good proportion of them with relatively high clarity scores (due to relatively coherent retrieval results) and some (tying) perfect-scoring questions with relatively low clarity scores. The questions not receiving a reduced average precision score of zero or one do exhibit correlation.

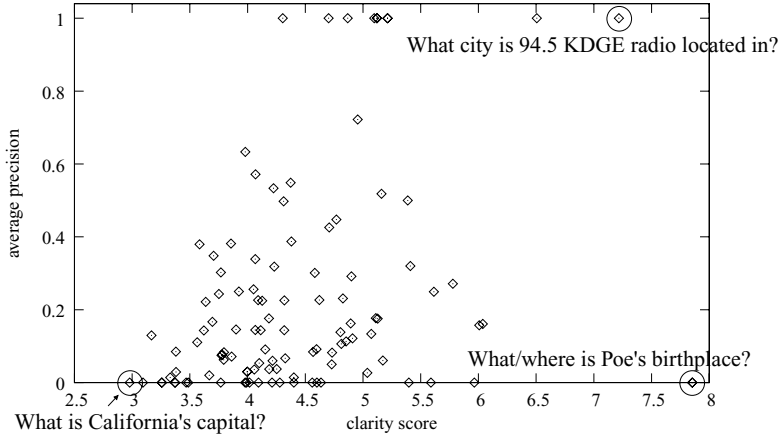

Fig. 7 Scatter plot of the average precision judged only on returned passages versus clarity score for the 109 "Location" questions of the TREC 2000 QA Track

6.4. Decisions using passage-based clarity

Even the highest correlation we observed in our question answering data is not comparable to the best correlation we achieve with standard clarity scores in document retrieval. In particular, the highest correlation in question answering is $R = 0.36$ for "Location" questions in TREC 2000 (see Table 5) while the highest correlation observed for standard clarity score in document retrieval is $R = 0.64$ for the TREC 7 test set (see Table 1, a superset of TREC $7 + 8$ data is shown in Fig. 5). Achieving the reported level of correlation for "Location" questions also required adjusting several parameters for the system, the questions being classified, and "location" being a good class for exhibiting the correlation. For all of these reasons, it is hard to imagine a straightforward use of precision prediction based on clarity scores in question answering. The fact that the basic correlation has been demonstrated, however, paves the way for creative use of clarity-related techniques in studying the question answering task.

7. Variants of clarity score

In this section we introduce two variants of document-based clarity scores called ranked list clarity scores and weighted clarity scores. These variants have been developed for several reasons.

Ranked list clarity was originally developed in a successful attempt to improve the robustness of clarity score correlations with average precision. Additionally, The ranked list relevance models it uses only require document ranks (and not probability scores) and thus the technique is applicable to any retrieval situation that results in a ranked list without accompanying probability scores, such as in cases of relevance feedback. Weighted clarity scores were originally introduced in an attempt to increase clarity score correlations with average precision. Though they succeed in that, their greatest effect may be in introducing the notion of weighted divergences to information retrieval. Through that idea, they influenced our own application to predicting query expansions, presented in Section 8, which also makes use of ranked list relevance models.

7.1. Ranked list clarity scores

Ranked list clarity scores are an alternative technique for computing clarity scores that allow them to be used with any retrieval model, extending their range of applicability considerably. They do not use the documents' probability scores from retrieval, merely the ranking (which may or may not be based on such scores) to form a relevance model. The relevance model in this approach is literally a model of the ranked list of documents returned by the query. This ranked list can even be modified by user feedback or any other technique that results in a ranked list.

For example, a system using relevance model retrieval (Lavrenko and Croft, 2001, 2003) scores documents with Kullback-Leibler divergence between two distributions. The scores are not probabilities and hence cannot be used as the mixing weights in Eq. (3). Additionally, implementations typically use the rank-equivalent cross-entropy for scoring, and in this case the scores are not even positive. But such a system does create a ranked list, which can be used to form ranked list clarity scores.

7.1.1. Calculation and smoothing

In our implementation of ranked list clarity scores, we simply replace $P(D|Q)$ by a simple function of the document's rank. We have tried two schemes for this function. We refer to the relevance models produced in this way as *ranked list relevance models*. The first scheme is *flat cutoff*, where $P(D|Q)$ is a constant if the rank is less than or equal to the cutoff rank and zero otherwise. For the second scheme, *linear cutoff*, we use a linearly decreasing function of the rank that goes to zero at the cutoff rank plus one. Specifically,

$$P(D \mid Q) = \begin{cases} \dfrac{2(c+1-r)}{c(c+1)} & \text{for } r \leq c, \\ 0 \quad \text{for } r > c \end{cases} \tag{7}$$

where r is the rank of document D and c is the cutoff rank.

Ranked list clarity behaves similarly to standard clarity with respect to choice of smoothing methods and works well for the same heavy and light smoothing conditions on all tested collections. So we generally use the standard, heavy smoothing, condition.

7.1.2. Ranked list clarity and average precision

The correlation results between ranked list clarity scores and average precision are shown in Table 7 for eight TREC test collections. The clarity scores were computed using our heavy smoothing condition ($\mu_1 = 1000$, $\lambda_2 = 0.1$) and a linear cutoff with $c = 60$ in Eq. (7). Heavy smoothing was used since, again, it offers high and reliable performance. A small linear cutoff, like $c = 60$, is necessary in computing ranked list clarity in order to focus estimation of the model on top-ranked documents. In case of the standard clarity, the focus of estimation is determined by document likelihood scores, $P(D|Q)$, which are gotten by Bayesian inversion of query likelihood scores, and the high cutoff at 500 documents has little effect on standard clarity scores.

The cases for which ranked list clarity outperforms standard clarity (compare Tables 1 and 7) are interesting. The ranked list clarity technique is significantly better on TREC $1 + 2 + 3$, TREC 4, and the TREC 9 aggregate Query Track. These are all cases where there are significant differences in the query sets, as compared to the title queries of the remaining TREC Ad Hoc collections we tested. In the case of TREC $1 + 2 + 3$ the difference is less honed and consistent title queries, in TREC 4 it is long description queries, and in the aggregate Query Track it is large a large variety of query lengths and styles. So there is some evidence that ranked list clarity scores are more robust in the face of query variability.

Table 7 Correlation between ranking queries by average precision and by ranked list clarity score

TREC	Disks	Topics	Number	Spearman R
1+2+3	1&2	51–150	100	0.60
4	2&3	201–250	50	0.67
5	2&4	251–300	50	0.49
6	4&5	301–350	50	0.53
7	4&5-CR	351–400	50	0.61
8	4&5-CR	401–450	50	0.51
7 + 8	4&5-CR	351–450	100	0.57
agg QT	1	51–100	1804	0.61

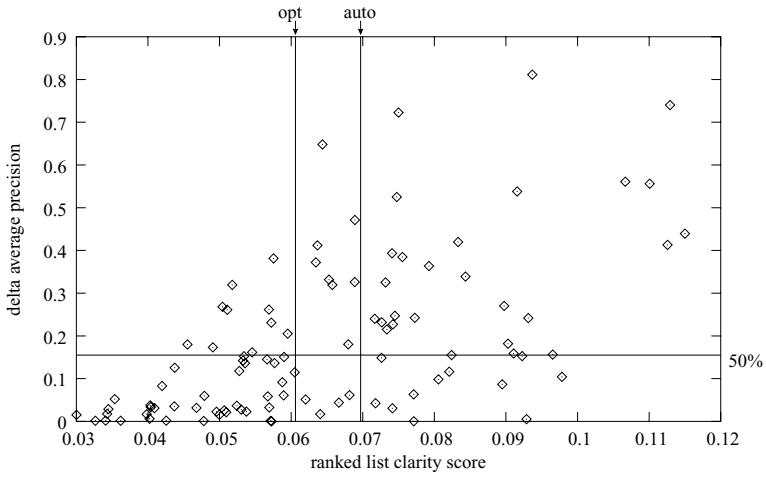

Fig. 8 Scatter plot of average precision versus ranked list clarity score for the 100 queries of TREC $1 + 2 + 3$

In fact, the reason we developed ranked list clarity scores is that clarity scores, as originally defined (Cronen-Townsend et al., 2002), have very little correlation with average precision in TREC $1 + 2 + 3$. With $\lambda_1 = \lambda_2 \equiv 0.6$ we find $R = 0.18$ (P-value 0.039) for the rank correlation with $\lambda = 0.6$ query likelihood retrieval. With Dirichlet smoothing of the query likelihood retrieval with $\mu_1 = 1000$ and using $\lambda_2 = 0.1$ this comes up to 0.50 (as listed in Table 1). The ranked list clarity score technique brings the correlation (with $\mu = 1000$ query likelihood retrieval) up to 0.60 (as listed in Table 7). The scatter plot of this data is shown in Fig. 8. Carefully smoothing the document models used in clarity calculation makes the correlation in TREC $1 + 2 + 3$ comparable to that in all other collections we have tested, but ranked list clarity produces an even stronger correlation.

7.1.3. Decisions using ranked list clarity

Making a decision based on ranked list clarity score relies on having a suitable decision threshold. We precede again as in Section 4.4. For the "task" of deciding whether a query is in the upper- or lower-performing half of queries in a test set, this method leads to a uniform heuristic to set a decision threshold at a clarity score higher than 40% of one term queries, tuned on all the data. Table 8 shows the automatic threshold and the relative discrepancy from the Bayes-optimal to the automatic threshold for each collection (in the "relative" column).

Table 8 Automatic and Bayes optimal thresholds for ranked list clarity scores

TREC	auto:40%	Optimal	Relative
$1 + 2 + 3$	0.069	0.060	+15%
4	0.058	0.050	+16%
5	0.064	0.059	+8.5%
6	0.060	0.070	−14%
7, 8, 7 + 8	0.059	0.062	−4.8%
agg QT	0.065	0.053	+23%

Figure 8 shows the ranked list clarity scores for the TREC $1+2+3$ collection using heavy smoothing. The 50% threshold in average precision (labeled "50%") as well as the Bayes-optimal threshold in clarity score (labeled "opt") and the automatic threshold (labeled "auto"), are shown. The average precision threshold line and either of the clarity score threshold lines divide the scatter plot into four quadrants. Queries in the upper right quadrant and the lower left quadrant have correct predictions while queries in the upper left and lower right quadrants have mistaken performance predictions based on their clarity scores.

The ranked list clarity score automatic threshold shown in Fig. 8 seems a significant distance from the Bayes optimal threshold. However, switching from the Bayes optimal threshold to the automatic threshold for this data, again results in about the same number of queries switching from incorrect to correct classification as switch from correct to incorrect classification. That is to say, there are about as many high- as low-performing queries in the region between the optimal and automatic thresholds, so the Bayes error is close to uneffected by changing from the optimal to the automatic threshold in this case.

7.2. Weighted clarity scores

Weighted clarity scores are based on the idea that differences in term usage between the query model and collection model are not equally significant for all terms. Differences in the usage of query terms are considerably more important than differences in the usage of other terms. Taking this into account increases the correlation with average precision while introducing theoretical machinery into information retrieval that opens up new research possibilities and makes the prediction of query expansion possible (presented in Section 8).

7.2.1. Calculation and smoothing

The clarity score was defined in Eq. (5) as the relative entropy (or Kullback-Leibler divergence) between a query's relevance model in the collection, $P(w|Q)$, and a model of the entire collection $P(w \mid \text{coll})$. We extend clarity scores by using the weighted relative entropy (Taneja and Tuteja, 1984; Arndt, 2001)

$$D(A||B;U) = \frac{1}{E(A;U)} \sum_{events,i} u_i a_i \log_2 \frac{a_i}{b_i}, \tag{8}$$

where A and B represent probability distributions and U represents a vector of weights over events. The normalization factor $E(A;U)$ is given by $E(A;U) = \sum_j a_j u_j$, where a_i and b_i represents the probability of event i according to the A and B distributions, respectively. The weighted relative entropy is the expectation value of the quantity $\text{Log}_2 \frac{A}{B}$ using a weighted version of the A distribution instead of the unmodified A distribution as in standard relative entropy.

A generalized clarity measure is then defined by the weighted relative entropy from the query model, $P(w|Q)$, to the collection model $P(w|\text{coll})$

$$\text{clarity} = \sum_{w \in V} \frac{u(w)P(w|Q)}{\sum_{w' \in V} u(w')P(w'|Q)} \log_2 \frac{P(w|Q)}{P(w|\text{coll})}, \tag{9}$$

where $u(w)$ are the term weights. When $u(w) \equiv 1$ for all terms in the vocabulary this expression is just the usual KL divergence and leads to the clarity scores used previously

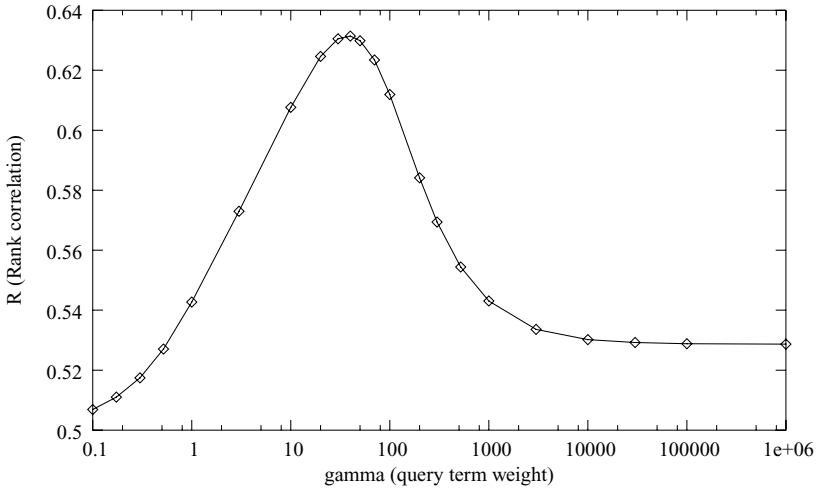

Fig. 9 Rank correlation between clarity score and query likelihood document retrieval average precision as a function of the query term weight for the TREC 9 Query Track aggregate

(Cronen-Townsend et al., 2002). The ordinary KL divergence is the expectation value in the first distribution of the difference in the log probabilities of an event in the two distributions. The weighted relative entropy is an expectation value of the same quantity with a *weighted* version of the first distribution being used to calculate the expectation value.

Query terms are the most important terms in information retrieval systems. To reflect this, we tried giving each query term weight $u(q) = \gamma$ and all other terms weight $u(w) = 1$. Here γ represents how many times more significant the occurrence of a query term is than the occurrence of another term in a document. We also tried giving each query term q the weight $n(q)\gamma$ where $n(q)$ is the number of times the given term appears in the query. Even in test collections such as the TREC 9 Query Track aggregate where query terms do repeat on occasions, the results of the two schemes are nearly identical. With either implementation, a nearly identical improvement is seen in the Spearman rank correlation between the clarity scores and the average precisions of the queries for all values of γ.

As shown in Fig. 9 for the aggregate TREC 9 Query Track (with light smoothing), the weighted clarity score predicts average precision better as the weight γ of query terms in measuring the degree of difference from the collection model is increased, until a peak is reached (at $\gamma = 40$) and the correlation returns slowly to a value comparable the unweighted value as γ is increased further. The value $R = 0.54$ at $\gamma = 1$ (unweighted) is significantly higher than the value of $R = 0.39$ reported for the original clarity score method for the Query Track aggregate (Cronen-Townsend et al., 2002). This difference is due to our use of Dirichlet smoothing, rather than linear smoothing, for the scoring of documents by query likelihood. As the relative importance of query terms to the comparison is increased, the correlation rises to a maximum of $R = 0.63$ at $\gamma = 40$.

At $\gamma = \infty$, the clarity score is computed over just the terms that appear in the query. This setting gives only slightly less correlation with retrieval performance ($R = 0.53$) than using the entire relevance model for comparison and no weighting ($R = 0.54$). Thus, in applications requiring retrieval precision prediction where a full relevance model is not needed, a *reduced relevance model* may be computed solely for the purpose of clarity score computation. For these reduced relevance models the probabilities only need be estimated for the query terms

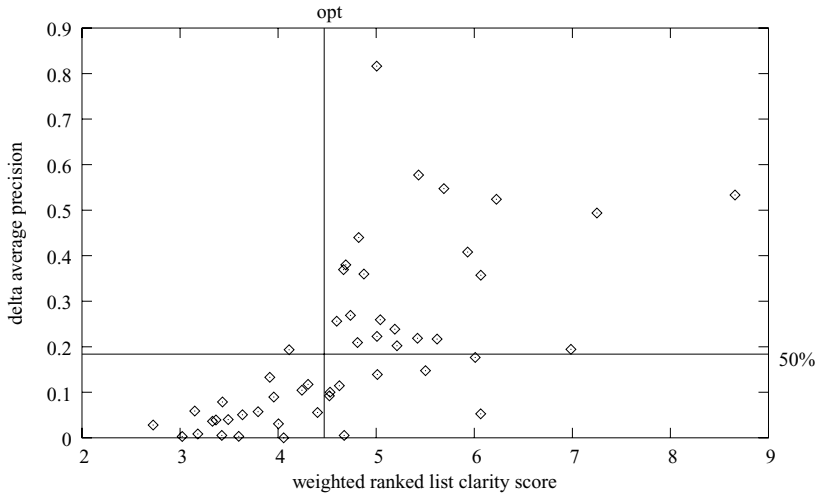

Fig. 10 Scatter plot of average precision versus weighted ranked list clarity score for the 50 queries of TREC 4

themselves. This calculation can be done with just the lists of documents containing each query term and the number of term occurrences in each. Since this information is typically stored in indices for information retrieval, this computation can be made extremely efficient. The closeness of $\gamma = \infty$ and $\gamma = 1$ performance holds for all tested collections and usually the $\gamma = \infty$ correlation is actually stronger than the $\gamma = 1$ correlation.

Weighted clarity scores behave in a complicated manner when mixed with heavy smoothing. In plots like Fig. 9 but with heavy smoothing, a shoulder forms to the left of the main peak which accounts for the increase in heavy smoothing correlation for $\gamma = 1$. But the dip between the shoulder occurs at varying values of γ, making it impossible to set a uniform policy for choosing γ to achieve good performance across collections. For this reason weighted clarity scores are best computed with light smoothing probability estimates; this policy is followed throughout this work.

A scatter plot of weighted ranked list clarity scores for TREC 4 is shown in Fig. 10 for light smoothing. Note the higher range of clarity score values than in previous scatter plots of document-based clarity scores (see Figs. 5 and 8.). This is due to a combination of the light smoothing and the query term weighting, which both increase clarity scores, on average.

7.2.2. Weighted clarity and average precision

Both the basic clarity score and the ranked list variant can be computed with weighting. For either type of clarity score, The optimal value of γ for correlation with retrieval performance does not vary much over the collections we tested. The peak of the curve is at slightly higher values of γ for most collections other than the Query Track, leading to good performance at $\gamma = 100$ for all collections. For ranked list clarity scores using $\gamma = 70$ leads to slightly better performance. The results of applying these uniform settings across collections are shown in Table 9. P-values for the rank correlation listed range from 0.0004 (TREC 6 ranked list clarity) to 1×10^{-152} (aggregate Query Track ranked list clarity). Again all values are statistically significant indicating clarity scores and average precision are related.

Table 9 Correlation between ranking queries by average precision and by weighted version of standard and ranked list clarity scores

				Spearman R	
TREC	Disks	Topics	Number	Standard $\gamma = 100$	Ranked list $\gamma = 70$
$1 + 2 + 3$	1&2	51–150	100	0.56	0.62
4	2&3	201–250	50	0.65	0.75
5	2&4	251–300	50	0.61	0.56
6	4&5	301–350	50	0.58	0.48
7	4&5-CR	351–400	50	0.67	0.54
8	4&5-CR	401–450	50	0.70	0.63
$7 + 8$	4&5-CR	351–450	100	0.68	0.57
agg QT	1	51–100	1804	0.61	0.62

Figure 10 is a scatter plot of the weighted ranked list clarity for most correlated case, TREC 4 with ranked list clarity. It is interesting that this highest correlation occurs with long, description field, queries. The 50% threshold in average precision (labeled "50%") as well as the Bayes-optimal threshold in clarity score (labeled "opt") are shown, and, as explained before, divide the scatter plot into four quadrants. Queries in the upper right quadrant and the lower left quadrant have correct predictions while queries in the upper left and lower right quadrants have mistaken performance predictions based on their clarity scores. In this case, very few queries fall into the two mistaken prediction quadrants. In fact, only one query below the Bayes-optimal threshold score in Fig. 10 has an average precision that puts it in the top half of the test queries, and only just barely.

7.2.3. Decisions using weighted clarity

To make simple decisions based on weighted clarity score, a decision threshold must be set. The automatic thresholding method does not work well for weighted clarity scores, unfortunately. This is why no automatic threshold is shown in Fig. 10. Sampled single term queries no longer lead to a reasonable estimate of the score distribution on real queries, since the weighting of query terms makes clarity scores extra sensitive to differences in query length and composition. The weighting process itself makes the scores depend much more highly on the query terms themselves. Thus unweighted clarity scores are currently a better choice for applications in real systems because they can be automatically thresholded, despite the slightly higher correlation with average precision found for weighted clarity scores.

8. Application: Query expansion

Query expansion is a well-known technique that has been shown to improve *average* retrieval performance. This section describes an approach to the task of improving retrieval through deciding automatically whether to use unexpanded retrieval results or expanded retrieval results. Simple comparisons of clarity scores are not sufficient for the purpose, so we compare ranked list relevance models directly. We are aided by calculating the top terms in clarity score contribution (again the whole clarity score is overly averaged for our purposes). As we will see in Section 9 on other work, the creative use of clarity-related ideas has proved useful to other researchers, as well.

Query expansion has not been used in many operational systems because of the fact that it can greatly degrade the performance of a system for certain individual queries. Our implementation is best suited to the task of predicting when a query will perform significantly worse after query expansion than before expansion, thereby improving the consistency of retrieval while having a small effect on the average performance, due to the small proportion of significantly bad-to-expand queries.

We develop a method for deciding whether or not to apply query expansion to a particular query. Our method requires a system to perform both the unexpanded and expanded retrieval steps internally, and then compares the results of the two retrieval steps to decide which results to present to the user. We make ranked list models of the unexpanded and expanded ranked lists and compare the models directly, to sense poor expansions. Since our method models the two ranked lists, it can be used to choose between any two retrieval techniques whether they are based on language modeling, or not.

8.1. Document retrieval: unexpanded and expanded queries

The first step in sensing poor expansion results is performing both the unexpanded and expanded retrieval steps. The change in average precision (expanded minus unexpanded) becomes our figure-of-merit, which we term the *improvement* of a query. We seek to predict when the improvement of a query will be significantly negative. In these cases a system should use the unexpanded query results. We refer to a query with negative improvement as bad-to-expand and a query with positive improvement as good-to-expand.

In our approach to the query expansion prediction task, we use query likelihood retrieval (Song and Croft, 1999) for unexpanded retrieval and relevance models (Lavrenko and Croft, 2003) for expanded retrieval. Relevance model retrieval is a conceptually simple, principled, and effective expansion technique. It simply uses the relevance model as the expanded query, ranking documents by their similarity to it.

Query likelihood retrieval is a one step process. We use Dirichlet smoothing with $\mu_1 = 1000$ throughout this study. This uniform setting gives reasonable performance across all TREC collections tested.

Relevance model retrieval is a three step process and each step requires individualized smoothing for good performance (Lavrenko, 2004). The first two steps, query likelihood scoring of documents and mixing of document models, construct the relevance model (query model). The third step performs retrieval with this relevance model. For the first two steps we use our light smoothing condition ($\mu_1 = 1000$, $\lambda_2 = 0.9$). For the retrieval step we use Jelinek-Mercer smoothing with $\lambda = 0.2$. We keep the parameters constant across all collections to keep the emphasis on applying our methods to other collections where relevance information is not available.

Since we did not tune the parameters for each collection our results are not strictly comparable to those in previous papers on retrieval where the parameters are tuned for each test collection to obtain best performance.

8.2. Model comparison scores

The precise task we are interested in is predicting queries that should not be expanded (highly negative improvement) with a score that does not depend on relevance information. To do this, we compare a ranked list model of the unexpanded retrieval ranked list (model A) with a ranked list model of the ranked list produced with the expanded query (model B). With this comparison, our goal is to sense when the expanded retrieval has strayed from the

sense of the original query. Comparison scores focus on important terms in the unexpanded query and are high when the documents in the expanded results use the terms much less frequently than do the documents in the unexpanded results. This often indicates a poor expansion outcome (highly negative improvement). In this case the system would show the user the unexpanded retrieval results instead of the expanded retrieval results. We call this strategy *selective query expansion*.

For models of the two retrieval results (unexpanded and expanded) we use the ranked list models described previously with a flat cutoff 100 (i.e. top 100 documents all equally weighted in the mixture). We use our light smoothing condition and no query term weighting, since these settings provide marginally better prediction performance than other choices. Similarly, more complicated weighting schemes for the documents, other than the simple flat cutoff, provided no net benefit.

We compare the models with the weighted relative entropy, Eq. (8), as $D(A||B;U)$. We have tried several different schemes for the weights U. Using equal weights over all vocabulary terms (standard KL divergence) considers differences between models for each vocabulary term as equally important.

An important insight toward achieving better prediction performance is that differences in how two models use each vocabulary term are *not* all equally important. One wants to weight "important" terms in the unexpanded model highly. But weighting terms based on their relative frequency leads to measures with no perceptible relationship to expansion performance since it compares the two models primarily on generic and commonly occurring terms.

To choose important terms, we use the top T terms in contribution to the clarity score of Model A. We compute score of each term as

$$\text{contrib}(w) = P(w) * \text{Log}_2 \frac{P(w)}{P(w|\text{coll})}, \tag{10}$$

and take the T highest terms. These contributions measure how unusual a terms' usage is in the ranked list relative to generic language of the collections and are a suitable measure of importance.

The top T terms are all given weight 1 and all other terms are given weight zero. Relative insensitivity to the value of T as long as it is in the range 5 to 50 is observed in testing over TREC $1+2+3$, 5, 6, 7, 8, and aggregate Query Track. Query terms (after stopping) generally are chosen as important terms, but are given the same weight in the comparison as other important terms in the model of the unexpanded ranked list.

The comparison is sensitive to the choice of weights and we tried several other alternatives. Methods basing importance on term probability, weight generic common terms too highly, and methods basing a term's clarity contribution seem to make poor choices of relative weights. Choosing a fixed number of top terms by their clarity contribution and weighting them all evenly is our best performing method.

Combining these consideration gives a model comparison score as

$$\text{comparison score} = \sum_{w \in \tau} P(w|A) \log_2 \frac{P(w|A)}{P(w|B)}, \tag{11}$$

where w represents a term, A represents the ranked list relevance model of the unexpanded results, and B represents the same for the expanded results, and τ represents the set of T top terms.

Interpreting Eq. (11), the model comparison score is an expectation of the difference in log probabilities for the important unexpanded terms in models of the two ranked lists. It is an

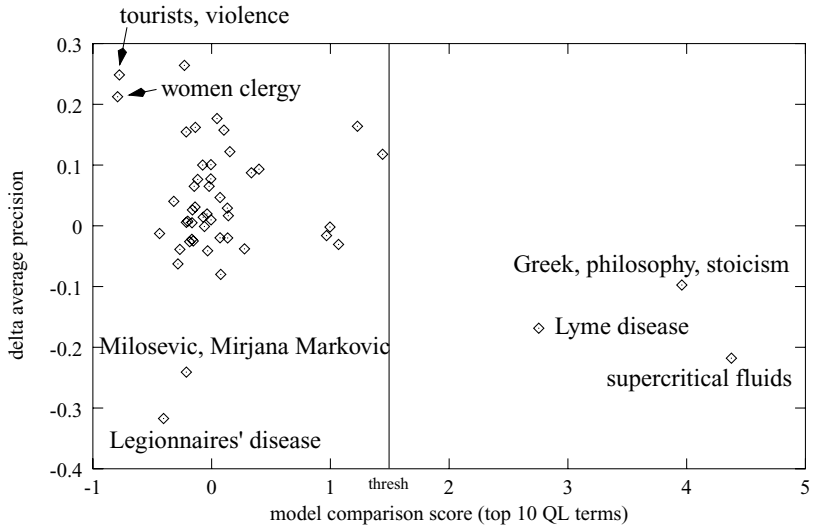

Fig. 11 A scatter plot of Δ average precision and model comparison scores for TREC 8

average (under the A distribution) of $\mathrm{Log}_2 P(w|A) - \mathrm{Log}_2 P(w|B)$ when w is each of these important terms. Thus a high and positive value indicates that the important terms are used much less frequently in the in the expanded model. This often indicates an expansion that has strayed from the original sense of the query. A negative score indicates that the expanded retrieval uses the important terms more frequently, which often indicates a good result, and a score of zero means the two ranked lists use the terms evenly.

8.3. Model comparison scores and delta average precision

The average precision change (improvement) versus model comparison score is shown in Fig. 11 for the 50 queries of TREC 8 Ad Hoc Track, with extreme examples labeled. One can see that extremely high scores are an indicator that the expansion may be performing poorly and high magnitude negative scores are indeed an indication that expansion may be performing well. The very high scores are well separated from the other queries' scores.

An automatic threshold set to be above 95% of randomly sampled one term queries from the vocabulary is shown.[4] Different thresholds were tried an our system performed best when the threshold was set high, so that queries with a higher comparison score were as likely as possible to have negative delta average precision. This threshold also suits the task of improving retrieval consistency by avoiding a small proportion of bad expansion.

Examining extreme cases sheds some light on these results. Of three highest scoring queries in TREC 8, all three perform poorly on expansion. There are no good-to-expand queries with positive scores close to these scores, suggesting that automatic decisions are a possibility (the lowest scoring of the three is nearly 2 times as highly scoring as the next nearest query). Also, the two lowest scoring queries ("tourists, violence" and "women clergy") do perform very well on expansion.

[4] Computed the analogous way to clarity score thresholds.

Table 10 Selective mean average precision with estimated Bayes optimum thresholds

TREC	Queries	Rel model	Model comp	Perfect choice
1 + 2 + 3	51–150 title	0.2490	ident	0.2589
5	251–300 title	0.1609	0.1644(∗)	0.1837
6	301–350 title	0.2013	0.2115(∗)	0.2468
7	351–400 title	0.2524	0.2212	0.2711
8	401–450 title	0.2715	0.2756(∗)	0.3011
Agg QT	51–100: 1804 var	0.2219	0.2188	0.2387

There are bad-to-expand queries, however, that the method fails to detect. One is "Milosevic, Mirjana Markovitch" where the TREC topic indicates that a document must refer to some variant of Mirjana Markovitch's name to be relevant. The name, however, is drowned out by other important terms that occur more frequently in the expanded results, producing a low comparison score. This is despite the fact that the expanded results do not use the name as much and are hence irrelevant, leading to a highly negative improvement. The other missed prediction is "Legionaires' disease" where documents can contain the terms "legionaire" (meaning soldier) and "disease" (and related words) yet not be about Legionaires' disease, leading to a low comparison score despite its bad-to-expand status.

8.4. Decisions using model comparison scores

The decision we wish to make for each query is whether or not to use the expanded results or drop them and use the unexpanded results. In this section we are concerned with implementing this selective query expansion method.

Table 10 shows the comparison of the mean average precision for three retrieval methods. The column "Relevance Model" shows results using relevance model retrieval for all queries. The column "Perfect Choice" uses the relevance information and chooses expansion only when it performs better. "Model Comp" is our selective query expansion procedure with a 95% threshold to decide not to expand a given query. The mean average precision marked "(*)" are higher than using relevance model retrieval for every query, indicating our method helps more than it hurts, on average, for these test sets.

Table 11 shows the breakdown of how queries higher than the threshold perform in tests of our selective query expansion method. The threshold is set to a model comparison score that exceeds 95% of one-term queries. Above-threshold queries are divided into three

Table 11 Breakdown of above-95%-threshold query performance for selective query expansion

Collection	Rel model	Model comp	Above threshold		
			Good	Neut	Bad
TREC 5	0.1609	0.1621	0	2	1
TREC 6	0.2013	0.2197	0	3	3
TREC 8	0.2715	0.2812	0	0	3
TREC 1 + 2 + 3	0.2490	0.2451	1	0	0
TREC 7	0.2524	0.2394	2	1	1
QT agg	0.2219	0.2217	13	32	11

performance classes: "good" where Δ, the improvement, is greater than 0.05, "neut" where $-0.05 < \Delta < 0.05$, and "bad" where $\Delta < -0.05$. The method is successful in collections above the line and unsuccessful in collection below the line. We do not expect to see large *mean* average precision improvements since the method is tuned to detect a small percentage of queries that perform very poorly on expansion. For certain collections (e.g. TREC 5, 6, 8), the effect of the method with this high automatic thresholding is quite good: all the queries above threshold are indeed bad to expand or neutral, hence some poor expansions are avoided. For these collections, more consistency in retrieval results is obtained.

Two of the failures of the model comparison method with automatic thresholding (below the line in 11) can be analyzed. In TREC 7, there are only 3 very bad-to-expand queries ($\Delta < 0.1$) using relevance model retrieval for expansion. Here the method is given very little room to improve the retrieval. In the case of the aggregate Query Track, the queries exhibit a high degree of variability. Queries range from titles to long natural language queries (Buckley, 2000). We speculate that this variability makes the application of our method more difficult.

9. Relationship to other work

9.1. Query performance prediction

Prediction of query performance has long been of interest in information retrieval, though early attempts met with little success. The work of Cronen-Townsend et al. (2002) demonstrated some of the first success at addressing this challenge.

A variety of work explores similar issues with a different focus. In work on automatic query expansion, Carpineto et al. (2001) use a weight similar to individual term contributions to the clarity score of a query (as shown in Fig. 1, for example) to rank and weight terms within Rocchio query expansion. Pirkola and Jarvelin (2001) also focus on automatic query expansion and examine individual term contributions to the retrieval effectiveness of queries. They have success in using collection statistics to identify the most important query term when there is no information as to the actual relevance of the documents to the query. Sullivan (2001), in seeking to model the difficulty of questions, models long question text directly and compares questions with an existing set of questions whose effectiveness at retrieving relevant documents (when viewed as information retrieval-style queries) has been measured. Rorvig (2000) speculates about using the dispersion of the top documents as a measure of query difficulty. The idea is only tested averaged over all queries in a test set and averaged over systems, a large difference from our measure on individual queries with respect to a collection.

Recently, the Robust Track was proposed in TREC to investigate poorly-performing queries Voorhees (2003). Amati et al. (2003) proposed a related measure to the clarity score of a query to capture query difficulty in this context.

9.2. Clarity scores in other research

The initial studies of clarity scores in predicting document retrieval performance (Cronen-Townsend and Croft, 2002) and quantifying query ambiguity (Cronen-Townsend and Croft, 2002) have led to various interesting applications. Clarity scores have been used to improve performance in the link detection task in topic detection and tracking (Lavrenko et al., 2002) by modifying the measure of similarity of two documents. The method judges two

document models as more similar if they are also far from general English usage, in addition to using words comparably to each other. Clarity scores have also been used to evaluate an ambiguity reduction technique by Allan and Raghavan (2002). Turpin and Hersh (2004) failed to find a correlation between clarity scores and user performance in the TREC Interactive Track and Diaz and Jones (2004) extended clarity scores to include time features. Brants et al. (2002) uses clarity scores in document segmentation. Shah and Croft (2004) use the clarity score contributions of query terms to select terms to expanded with WordNet, boosting a query's ability to get a the first relevant document in the ranked list as high as possible.

9.3. Predicting query expansion

Automatic query expansion techniques have been researched extensively, for example (Robertson, 1984; Buckley et al., 1994; Xu and Croft, 2000; Lavrenko and Croft, 2003). The degree to which the techniques can lead to very poor performances for some queries is a recognized issue in these studies. There also has been much work (Buckley and Salton (1995) for example) on using user feedback to improve retrieval ranked lists. Such systems can suffer from the same sort of straying from the original sense of the query as automatic expansion systems, and are candidates for the use of techniques, such as ours, designed to detect such straying. We know of no published works on predicting or sensing automatically when such techniques fail and see our work as a small step in that direction.

10. Conclusions

We have extended the original clarity score method in important ways, including more careful attention to smoothing which gives significant improvements in the correlation between clarity scores and document retrieval performance. Further extensions of the original method include ranked list clarity scores and query term weighting. The first provides additional correlation in difficult test sets, and the second provides further increases in correlation, though automatically thresholding weighted scores remains an unsolved problem. The basic correlation has been shown to apply in passage question answering as well for a variety of data sets. Clarity scores are a potentially useful tool for information retrieval researchers and system designers.

As an additional application of these techniques, a novel framework is introduced for predicting query expansion failures. Since this method is based on ranked list language models it may be applied to any choice between retrieval methods. Though it provides only the first steps of a solution to a very difficult problem in information retrieval, we believe this framework can be built upon by other researchers. Through the use of clarity score techniques, retrieval methods can begin to be chosen individually and automatically for each query. We believe that this sort of approach provides a promising avenue for enhancing information retrieval systems.

Acknowledgments We would like to thank Andrés Corrada-Emmanuel for providing the necessary question answering data. We are also indebted to Victor Lavrenko for his expert advice in the area of smoothing language models and to several anonymous reviewers who helped us improve this paper greatly. This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

Allan, J., & Raghavan, H. (2002). Using part of speech patterns to reduce query ambiguity. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 307–314).

Amati, G., Carpineto, C., & Romano, G. (2003). Fondazione ugo bordoni at TREC 2003: Robust and web track. In *Proceedings of the Twelth Text REtrieval Conference(TREC 2003)*. NIST Special Publication 500-255, in press.

Arndt, C. (2001). *Information measures: information and its description in science and engineering*. Berlin, New York: Springer.

Brants, T., Chen, F., & Tsochantaridis, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (pp. 211–218). ACM Press.

Broglio, J., Callan, J. P., & Croft, W. B. (1994). INQUERY system overview. In *Proc. TIPSTER Text Program (Phase I)* (pp. 47–67). Morgan Kaufmann.

Buckley, C. (2000). The TREC-9 query track. In E. Voorhees & D. Harman (Eds.) *Proceedings of the Ninth Text REtrieval Conference(TREC-9)* (pp. 500–249). NIST Special Publication.

Buckley, C. (n.d.). `trec_eval` information retrieval evaluation package. Available from `ftp://ftp.cs.cornell.edu/pub/smart`.

Buckley, C., & Salton, G. (1995). Optimization of relevance feedback weights. In *Proc. of the 18th Annual ACM SIGIR Conference* (pp. 351–357).

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1994). Automatic query expansion using SMART(TREC 3). In *Text REtrieval Conference* (pp. 69–80). NIST Special Publication 500-225.

Carpineto, C., de Mori, R., Romano, G., & Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems, 19*(1), 1–27.

Corrada, A., & Croft, W. B. (2004). Answer models for question answering passage retrieval. *To Appear in Proceedings of the 27th Annual International ACM SIGIR*.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley-Interscience.

Craswell, N., & Hawking, D. (2003). Overview of the TREC 2003 web track. In *Proceedings of the Twelth Text REtrieval Conference(TREC 2003)*. NIST Special Publication 500-255, in press.

Croft, W. B., Cronen-Townsend, S., & Lavrenko, V. (2001). Relevance feedback and personalization (A language modeling perspecvtive. *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*.

Croft, W. B., & Lafferty, J. (Eds.) (2003). *Language modeling for information retrieval*. Dordrecht: Kluwer Academic.

Cronen-Townsend, S., Corrada-Emmanuel, A., & Croft, W. B. (2003). Predicting question effectiveness, Technical Report IR-282, Center for Intelligent Information Retrieval, University of Massachusetts.

Cronen-Townsend, S., & Croft, W. B. (2002). Quantifying query ambiguity. In *Proc. of Human Language Technology 2002* (pp. 94–98).

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 299–306).

Diaz, F., & Jones, R. (2004). Using temporal profiles of queries for precision prediction. *To Appear in Proceedings of the 27th Annual International ACM SIGIR*.

Gibbons, J. D., & Chakraborty, S. (1992). *Nonparametric statistical inference*, 3rd ed. New York, New York: Marcel Dekker.

Krovetz, R. (1993). Viewing morpholgy as an inference process. In *Proc. of the 16th Annual ACM SIGIR Conference* (pp. 191–202).

Lavrenko, V. (2004). Personal communication.

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., & Thomas, S. (2002). Relevance models for topic detection and tracking. In *Proc. of Human Language Technology 2002* (pp. 104–110).

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. *Research and Development in Information Retrieval* (pp. 120–127).

Lavrenko, V., & Croft, W. B. (2003). *Relevance models in information retrieval* (pp. 11–56). Kluwer Academic.

Morton, T. (2001). Personal Communication.

Ogilvie, P., & Callan, J. (2002). Experiments using the Lemur toolkit. In *Proc. of the Tenth Text Retrieval Conference, (TREC-10)* (pp. 103–108).

Pirkola, A., & Jarvelin, K. (2001). Employing the resolution power of search keys. *Journal of the American Society for Information Science and Technology, 52*(7), 575–583.

Robertson, S. (1984). On term selection for query expansion, *Journal of Documentation, 46*, 359–364.
Rorvig, M. (2000). A new method of measurement for question difficulty. In *Proceedings of the 2000 Annual Meeting of the American Society for Information Science, Knowledge Innovations, 37*, 372–378.
Shah, C., & Croft, W. B. (2004). Evaluating high accuracy retrieval techniques. In *SIGIR '04: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval* (pp. 2–9). ACM Press.
Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference* (pp. 279–280).
Sullivan, T. (2001). Locating question difficulty through explorations in question space. In *Proceedings of the 1st ACM/IEEE Joint Conference on Digital Libraries* (pp. 251–252).
Taneja, H. C., & Tuteja, R. K. (1984). Characterization of a quantitative-qualitative measure of relative information. *Information Sciences, 33*, 217–222.
Turpin, A., & Hersh, W. (2004). Do clarity scores for queries correlate with user performance? In *Proc, of the Fifteenth Australian Database Conference (ADC2004)* (pp. 85–91).
Voorhees, E. (2003). Overview of the TREC 2003 robust retrieval track. In *Proceedings of the Twelth Text REtrieval Conference(TREC-2003)* (pp. 195–201). NIST Special Publication 500-255, in press.
Voorhees, E. M. (2000). Overview of the TREC-9 question answering track, In E. Voorhees & D. Harman (Eds.) *Proceedings of the Ninth Text REtrieval Conference(TREC-9)*. NIST Special Publication 500-249.
Voorhees, E. M. (2002). Overview of the TREC 2002 question answering track. In E. Voorhees (Ed.) *Proceedings of the Eleventh Text REtrieval Conference(TREC-9)*. NIST Special Publication 500-251.
Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems, 18*(1), 79–112.
Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *Research and Development in Information Retrieval* (pp. 334–342).