Predicting Question Effectiveness

Steve Cronen-Townsend Andrés Corrada-Emmanuel W. Bruce Croft

{crotown, corrada, croft}@cs.umass.edu
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003

ABSTRACT

We develop a method for predicting the quality of the passage retrieval component in question answering systems. Since high-quality passages form the basis for accurate answer extraction, our method naturally extends to prediction of an entire system's effectiveness at extracting a correct answer for a given question. Such prediction of question performance may lead to ways of guiding users in improving questions unlikely to succeed. Our metric is also a necessary research step towards systems that automatically tailor their methods to suit each individual question.

Building on previous work on predicting the performance of queries in retrieving documents, we show how to compute the *clarity score* for questions using passage-based collections. We show that this score is correlated with the average precision in a TREC-9 based system, breakdown the correlation by question type, and discuss example questions. We also study a more general set of questions extracted from a Web log to help make the case for the general usefulness of performance prediction based on question *clarity score*.

Keywords
language models, clarity scores, question answering

1. INTRODUCTION

Recognizing questions that are likely to perform poorly is crucial in real question answering(QA) systems. Any such system will get questions that the system will not be able to answer well. Such questions may include those that are personal and inappropriate to the information the system has, such as "Where am I?," as well as questions that are outside of the domain of the document collection being queried. There are also questions which are not the simple fact questions for which current systems are designed, such as "What caused the fall of the Roman Empire?" In addition there are questions that are appropriate but are vague in the sense that they cause the system to match diverse text from the collection. For example, "What time of year do most people fly?" Identifying such questions is crucial, since it is the first step in developing systems that go on to advise a user when the probability of a system obtaining a useful answer is low and guide the user in reformulating the question. It is also a necessary research step towards developing systems that adapt to each individual question to achieve superior results for the user.

This work establishes a technique that should allow question answering systems to automatically identify when they are asked a question unlikely to yield a brief, factual answer. In this paper we do not address out of domain questions and only consider questions with at least one answer within the QA system. Our technique is based on the clarity score concept from the work of Cronen-Townsend, Zhou, and Croft[7]. In that work, the clarity score of a query in a given document collection is shown to be a useful predictor of the query's effectiveness in retrieving documents from the collection.

In the present work, we adapt the clarity score concept for the passage retrieval step in question answering systems. Candidate answer passages are treated as small documents and improvements in estimation techniques are introduced. We show that the clarity score of a question with respect to passages in a passage retrieval system, appropriately calculated, correlates with the average precision of the passage retrieval. Here passages are deemed correct if they contain correct answer text anywhere within them. Since good passage rankings are required to extract correct answers, clarity scores should also be useful in predicting whether a full question answering system will be able to extract a correct answer from the retrieved passages.

2. PASSAGE RETRIEVAL

2.1 System Overview

Our passage-based question-answering system uses the TREC-9 QA track data. Given a question, it first creates candidate answer passages and then it ranks them, seeking to put passages more likely to contain answers high in the list. No processing of the ranked passages is performed, and the ranked and scored list itself is the output of the system.

The creation of candidate passages for a question is done by starting with a *document* retrieval step using the question as a query. The retrieved documents are parsed into over-

lapping passages a maximum of 250 characters long, which are then ranked according to the likelihood of a language model of the passage text generating the question[1].

The passaging in our system is done by sliding a window that contains an integral number of sentences forward by one sentence repeatedly. Each position of the window creates one passage. Thus neighboring passages generally overlap by at least one whole sentence. In order to maintain a 250 character limit for each new passage, however, overlap with the preceding passage is sometimes sacrificed by leaving out overlapping sentences to keep the new passage within the limit. So in a region of document text with short sentences, each sentence is repeated a few times, while a very long sentence is usually in a passage by itself that does not overlap with neighboring passages.

Candidate passages are created with a 250 character limit to allow comparison to existing systems. Additionally, overlapping passages of this kind are found to lead to higher mean average precision retrieval with our language modeling approach. Although pure windowed systems that do not respect sentence boundaries are found to perform even better, the experiments become quite costly in our current implementation. With our sentence-based windowed system, we can efficiently study issues related to the overlapping of passages. Full sentences are also required for further processing in complete question answering systems.

The retrieval system uses no stop word list (though punctuation and single characters are removed) and no stemming. It was found that leaving stop words in increased the performance of our passage retrieval slightly.

2.2 Measuring Performance

To measure the performance of a question in our question-answering system we use the *average precision* of the ranked list of answer passages. The *precision* of a set of passages is defined as the fraction of relevant passages in the set. The *average precision* for a ranked list of passages is simply the mean of the precision scores for ordered sets up to and including each relevant document. We calculate average precisions using the `trec_eval` package[3]. In the question answering case, where there are often few correct passages in the system, this means that perfect average precision scores of 1 are attainable if the, say, two correct passages are ranked first and second. In probabilistic systems, such as ours, each candidate passage for a question is scored by the system since the system believes, in some sense, that there is some non-zero probability that each is relevant. An important aspect of judging the quality of a retrieval by the average precision of the ranked list of passages is that average precision does not penalize the system for small scores on irrelevant passages that are below the scores of all relevant passages.

Average precision is an overall measure of the quality of the ranked list of passages, in the sense that it measures the degree to which the list has passages that contain correct answers ranked highly. In this case (TREC-9) the correctness is judged simply by seeing if the passages contain textual patterns supplied by TREC for evaluations. These patterns were designed by humans to be concise and to match the passages that were judged relevant when the track ran.

Clarity scores correlate much more highly with performance measured with average precision than with MRR (mean reciprocal rank) which is the older standard for evaluating the performance of question answering systems[13]. This is due, in part, to the fact that clarity scores produce an overall measure of the coherence of the language use of top passages, whereas MRR is more susceptible to random fluctuations since it only reflects the position of the first answer-containing passage in the list and not the list's overall quality. Considering that question answering systems often use passage retrieval as only a first step in producing an answer list to show to users, measuring performance in a way that accounts for overall quality of the answer passage list makes sense.

For example, a candidate answer passage list where the only correct answer appears in the second position, would not be nearly as useful to a system as a candidate answer passage list where the only correct answers appear in passages at ranks 3 through 10. The repetition in the second retrieval would aid a system in extracting a short answer correctly. The MRR score for the first case is 0.5 (the reciprocal of the first correct answer's rank of 2) and the MRR score for the second retrieval is 0.33. So MRR clearly scores the first retrieval more highly. However, average precision scores the first as 0.5 and the second as $(\frac{1}{3}+\frac{2}{4}+\frac{3}{5}+\frac{4}{6}+\frac{5}{7}+\frac{6}{8}+\frac{7}{9}+\frac{8}{10})/8 \approx 0.64$. Since average precision, in this case, rates the retrieval with repeats farther down the list more highly, it clearly captures some aspect of the quality of a ranked list of passages that is important to a full question answering system that extracts short answers from the retrieved passages, making use of repetitions. As was shown by Dumais et al[8], using this redundancy leads to more effective question answering, particularly in the web environment.

3. PASSAGE-LEVEL CLARITY SCORES

The most important aspect of this method is the computation of passage-level clarity scores. Essentially, we treat the passages as tiny documents and compute clarity scores as by Cronen-Townsend et al[7], with some improvements.

The intuition behind clarity scores is that the language model for a jumbled ranked list where top-ranked documents use language very differently is measurably different than the language model for a ranked list with a relatively focussed set of top documents that use language similarly. In particular, the model for the first case (a poor retrieval) should be more like the overall collection model. In the second case of a focussed ranked list with top documents using language similarly (usually a good retrieval), the model has very large probabilities for topical words that make its model unlike a model of the collection as a whole. The clarity score is a measure of the difference between the query/question model and the collection model and should be low in the case of a jumbled retrieval and high in the case where top documents use language similarly. In the case of a jumbled retrieval, at most one of the top documents generally can be relevant and, in the case of a focussed retrieval, many of the documents are often relevant. Thus there is a connection between the clarity score value for a retrieval step and the measured performance of that step. Since a ranked list is formed in response to a question, the clarity score can also be thought

[1]Efforts are also underway to create better rankings of candidate passages with more sophisticated language modeling techniques. In this study, however, we focus on the extent to which we can predict the quality of the answer passage retrieval in the simplest version of our passage retrieval system.

of as a measure of the effectiveness (in terms of ranked list coherence) of that question given the collection and retrieval system.

To formally define clarity scores we adopt a language modeling perspective and begin by constructing a unigram language model for the question. This model gives probability estimates for finding individual terms in a passage related to the question. Individual passage models contribute to this overall question model proportionally to their likelihood of generating the question via random sampling. That is to say, passages that use many terms in the question frequently contribute most to the weighted average. This makes the final model look most like the model for documents that use many question terms frequently. Lavrenko and Croft have dubbed this a "method one" relevance model[11].

Mathematically, the question language model is given by

$$P(w|Q) = \sum_{S \in C_Q} P(w|S)P(S|Q), \qquad (1)$$

where w is any term, Q the question, S is a passage model, and C_Q is the set of passages for the question Q. In principle we think of C_Q as containing all possible candidate answer passages, while currently we restrict this to a finite set of passages generated by a separate passage generation step for each question.

To estimate a question model via (1) we first must estimate $P(S|Q)$, the probability of each answer passage, given the question. These probabilities are the mixing weights for a weighted average of the document models $P(w|S)$. Since these weights are probabilities, dividing by their sum of one is not explicit.

We make the estimate of $P(S|Q)$ by first estimating $P(Q|S)$ for all answer passages S as

$$P(Q|S) = \prod_{q \in Q} P(q|S), \qquad (2)$$

with the product running over all question terms q. We obtain $P(S|Q)$ by Bayesian inversion with uniform prior probabilities for all passages.

A crucial improvement in the method for QA is smoothing the probability estimates of words made from the small sample of text (the passage S) with the entire collection to make a passage language model. In our case, we smooth using the counts from the entire TREC-9 collection, according to

$$P(w|S) = \lambda P_{ml}(w|S) + (1 - \lambda)P_{coll}(w), \qquad (3)$$

where $P_{ml}(w|S)$ is just the relative frequency of term w in S and $P_{coll}(w)$ is the relative frequency of w in the whole collection. The parameter λ is set to 0.6 throughout this paper. When scoring documents according to their likelihood of generating the question (with equation 2) we use smoothed language models from (3), as well, where the term w is set to each question term q successively.

With this question language model, we compute the clarity score by calculating how different the model is from the language of the collection as a whole. We measure the difference with the Kullback-Liebler divergence[5], $D(P(w|Q)||P_{coll}(w))$, from the question language model to the collection language model. In particular,

$$\text{clarity score} = \sum_{w \in V} P(w|Q) \log_2 \frac{P(w|Q)}{P_{coll}(w)}, \qquad (4)$$

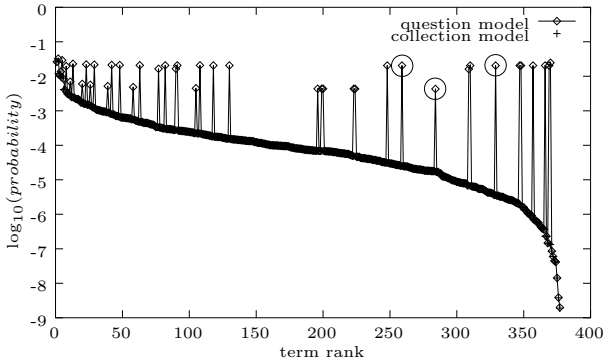

Figure 1: The question model for the question "Where was Tesla born?" shown with the collection model(TREC-9), over the terms that occur in the passages used to estimate the question model. The terms are presented in the order in which they occur in the collection. The collection model "+" symbols blur together to form the thick black line that agrees closely with the question model for most terms. The three circled points in the question model that stand out well above the collection model represent terms "unit," "inventor," and "yugoslavia," from left to right.

where V is the entire vocabulary of the collection.

Clarity score estimation was done with Krovetz stemming[10], despite the fact the retrieval was done without stemming, because we found experimentally that stemming significantly improves the correlation between clarity scores and unstemmed retrieval performance. Krovetz stemming was also used in the original clarity score studies[6, 7]. We now believe that treating trivially different word forms (e.g. singular and plural) as the same is important for the accuracy of clarity scores' assessment of ranked list coherence.

We originally guessed that the overlapping nature of the passages would pose a problem for the predictive nature of clarity scores, since the models created overestimate the relative frequency of the words which overlap. Detailed experiments showed, however, that performance was slightly enhanced by the overlap. Thus no special accommodation was made for the overlapping nature of the passages.

3.1 Examples

3.1.1 High clarity score question

Figure 1 shows the question and collection language models for the question 'Where was Tesla born?" from TREC-9. The terms that occur in the documents used to estimate the question model are plotted in the order they occur in the collection model. Thus the leftmost term is "the" and terms become increasingly rare as one moves towards the right side of the graph.

The bulk of the question model and the collection model are nearly identical, indicating that many commonly used words occur with similar relative frequency in text related to the question as in the collection as a whole. There are certain terms, however, that are vastly more probable in the question model than in the collection as a whole. These jargon terms for the question lead to jumps in the question model between adjacent terms. A jump is basically showing

a transition between a general term and a question model jargon term. Three large spikes are circled. They occur for terms "unit", "inventor", and "yugoslavia" in order from left to right. The term "unit" occurs with high probability because a Tesla is a unit of magnetic field strength. The final two points, that are both high for the question model represent "nikola" and "tesla". These are on the far right side of the graph because they are the most rare terms in all of TREC-9 that occur in the passages used to estimate the clarity score of this question.

3.1.2 Passages from a single document

Now we consider the TREC-9 question "What do penguins eat?" The extreme nature of this example, for which only one document is retrieved in the whole TREC-9 collection, allows us to illuminate some fine points of the whole system.

For this example question, our system extracts 24 passages from the single retrieved document ("LA120389-0149" from the LA Times). Of these 24 passages, only three actually contain question terms. Those three passages are:

1 Krill are a staple for a wide range of antarctic creatures, including whales, seals, penguins and a host of other sea birds and fish.

2 Besides studying krill, Surveyor scientists will study the feeding habits of fur seals and penguins, species that consume large quantities of krill. They also will study the microscopic algae and plants that krill feed on.

3 But scientists do not know what will happen to the antarctic food chain as krill are increasingly harvested for human and livestock consumption.

Of these passages, all three are judged correct by the TREC-9 patterns, since all three contain the word "krill" which is a correct answer. Interestingly, passages 1 and 2 contain only the question term "penguin" and the third passage only contains the terms "do" and "what". In this case our system got lucky and got credit for passage 3 in rank 3 even though it was ranked highly in the second stage for reasons that would not normally lead to a correct answer passage. The quality of the initial document retrieval from which candidate passages were made led to this success.

In this extreme case, the question language model is primarily a mixture of the models of the first and second passages with a little bit of the third and tiny amounts of the models for the other 21 passages from this document. Although the third passage contains two question terms, they are very common terms and the smoothing estimates used in scoring the other documents are not vastly lower for these terms, so its probability of generating the question does not stand out like the passages that use the rare term "penguin".

We also tested clarity score computations using only documents containing at least one question term (as in [7]) as well as those mixing in documents containing no question terms with a pure smoothing probability of generating the question (21 passages in this case). This question had the maximal difference in the two clarity scores, but it was still only 0.08%. This indicates that the effect of this approximation would likely have been slight, if we had used it.

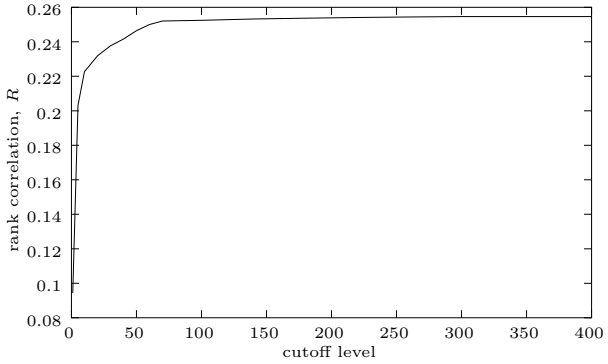

Figure 2: The Spearmann rank correlation coefficient R betweeen the average precision of all answerable $AFLPTX$ questions and their clarity scores as a function of the number of top passages used to estimate the clarity scores.

4. CLARITY AND AVERAGE PRECISION

4.1 Overall

Clarity scores of questions, computed with the methods described in Section 3, correlate significantly with the average precision of questions, determined as described in Section 2. The overall Spearmann rank correlation[9] for the 488 answerable questions is $R = 0.255$ with a P-value of 9.6×10^{-9} as compared to unassociated rankings. The rank correlations, though still strong enough to be useful, are weaker than those reported for document retrieval, where values of R hover around 0.5[7]. This difference is possibly due to the smaller samples of text (less than 250 characters) used to estimate passage models, as compared to the entire documents used to estimate document models in document retrieval work.

Figure 2 shows the way the correlation reaches the maximum value for this system as the number of top passages used to estimate the clarity scores is increased. There is essentially no correlation when the single top passage is used to estimate the clarity scores, since clarity then becomes just a weak measure of the specificity in the language use in that passage. As passages are added, however, the correlation to performance rises rapidly to 99% of the maximum value when 100 top documents are used and reaches the maximum value attainable for our system when about 340 top documents are used.

As more top passages are used to estimate the clarity scores of the questions, the correlation stops improving. This is related, in our system, to the number of candidate answer passages in our system that actually do contain correct answers (according to the patterns). There are an average of 25.6 correct answer passages per answerable question in our system, with a sample standard deviation of 42.7. This distribution is skewed with some passages having many correct answers (a maximum of 602 with the next highest number being 242). When the cutoff is large enough all of the correct answer passages are used in estimating the clarity scores of the questions. Almost all of the time adding more top documents does not help the correlations.

Question Type	Num.	R	P-value
A (amount)	35	0.171	0.16
F (famous)	76	0.148	0.10
L (location)	100	0.308	0.0011
P (person)	90	0.245	0.010
T (time)	48	0.350	0.0082
X (misc)	139	0.266	0.00090

Table 1: The rank correlation of clarity scores with average precision in TREC-9 QA track, for the 488 questions (of 600) that are answerable in our system. The average precisions of the questions were computed with a QA system using question likelihood on non-overlapping 250-character passages copied from retrieved documents, with models smoothed with counts from all of TREC-9. Clarity scores were computed using a cutoff at 340 documents and linear smoothing with 0.6 times the passage term relative frequencies plus 0.4 times the term relative frequencies in the entire collection. The P-values are the estimated probabilities of seeing as high or higher an R value with two unassociated rankings.

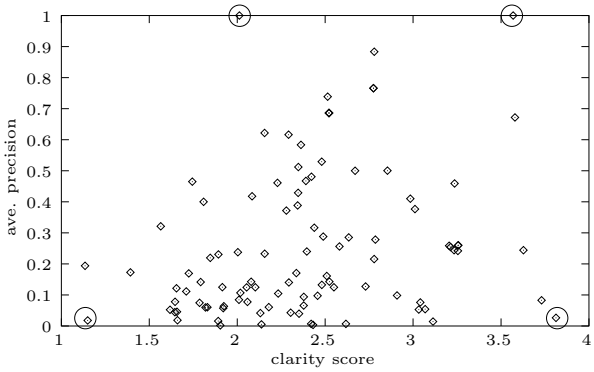

Figure 3: The average precision of the 100 answerable location-type questions in our system plotted against the clarity score of the question. Each point represents a single question. One of the two best-performing questions is "Where was Tesla born?" at average precision 1 and clarity score 3.57. This and the other 3 circled questions are discussed in Section 4.3.2.

4.2 Breakdown by Question Type

The breakdown of the correlation by question type is shown in Table 1. The questions are accurately classified by a Univerity of Pennsylvania classification scheme [12]. Only six question types with significant number of answerable questions are used in this study. These are A (amount), F (famous), L (location), P (person), T (time), and X (miscellaneous).

We note that there appears to be some statistically meaningful variation among the correlations for different question types. In particular F-type questions, questions asking what someone or something is best known for, or for a definition, pose difficulties for prediction in our system.

4.3 Examples

4.3.1 Challenging question types

Amount questions (A-type) have the highest p-value in Table 1. The highest clarity score question in this group is "How many types of lemurs are there?" This question has a rather specific question model, since fairly focused language co-occurs with "lemur" in the collection. However, since "type" does not co-occur in passages with "lemur" (the types are usually called "species") our ranking method leads to a very similar model to the one generated by the single term query "lemur" using the same candidate passages. Since there are many passages that mention lemurs, the chance of one mentioning the number of types of lemurs is small, leading to a low average precision, rather than a high value as one would guess from the high average precision of the question. Questions, like this, that lead to fairly specific language models without being very good at matching text near answers seem to be somewhat typical in this class (though they often have several jargon terms). There are also correlation problems due to questions that have essentially one correct answer passage in our system. A few such question have a fairly large influence on the correlation coefficient, since there are only 33 questions total.

The F questions have a lower Spearmann rank correlation between clarity score and average precision than the A questions, and there are more than twice as many questions. Some examples of F questions are "Who is Zebulon Pike?" and "Define thalassemia.". The first is relatively high in precision and about the median in clarity score for this class. Here we suspect that the complicated patterns used to select many different answers allow relatively high precision retrieval to be lower than expected in clarity, since top passages may be judged correct for containing very different answer text, leading to less coherence. In cases like the second, uses of the key term or terms have a fairly consistent use of language near them in the highly ranked passages, but without often having the required definition or reason for fame. Separating definition questions into their own class and assessing the degree to which these difficulties are characteristic of these classes is future work which will benefit from a much larger test collection of questions and documents.

4.3.2 Location Class Examples

The average precision of location (L-type) questions versus their clarity scores is shown in Figure 3. Four extreme case questions are circled and will be discussed.

The top-right question (high clarity score and high average precision) is "Where was Tesla born?", as discussed in section 3.1.1, with a clarity score of 3.57 and an average precision of 1 because the two correct passages in our system are ranked numbers one and two. The language in top-ranked passages is also quite coherent, leading to the high clarity score.

The lower-left question (low clarity score and low average precision) is "Where is Venezuela?" This is also a good prediction case. Here, a low clarity score would lead one to guess, correctly, that question is low performing. In fact, questions whose coordinates in graphs such as figure 3 lie nearly along a line from the lower-left of the figure to the upper right are good cases for prediction. In this case, the common words have little effect in our question-likelihood

system, making the question nearly identical to the one term query "Venezuela." That query ranks passages with diverse language highly, leading to the low clarity score. It turns out that few of them high in the list contain the correct answer of "South America," leading to the question's low average precision.

The upper-left question (lower half clarity score and high average precision) is "What is the location of Rider College?" In this case, there is only a single correct passage in our system, which happens to be ranked number one. Since there is only one correct passage the coherence with the other highly-ranked passages is limited, leading to a modest clarity score. Since clarity scores measure coherence of language use, our system clearly needs repetitions of answers in the collection to predict well, a factor that should aid our method in larger, web-searching systems.

Such cases of low clarity score and high average precision are extremely rare. Generally speaking, if the language use among the very top passages is diverse, only one of them will contain a correct answer, leading to a lowered average precision. Even if several topmost documents contain the only answers, the language would be more coherent making the clarity score higher, even though the average precision would stay one. Hence the prediction in this case would be enhanced, since a high clarity score question would be guessed to be high in average precision.

The final circled query, lower right (high clarity score and low average precision), is "What was Poe's birthplace?" In this case, "Poe" and "birthplace" were never used near each other in the TREC-9 data so our system's candidate passage creation was thrown off and did not include many possible passages from TREC-9 that do contain the answer. In this case, this led to a low average precision. Compounding the situation is a limitation of the judgements by pattern, which in this case marks two passages as correct because they contain "Boston" even though there is no mention of birth. As it turns out, our system receives a much lower average precision because of these two passages it was "supposed" to rank near the top, though, as seen by a human, they are incorrect passages. When the question is rephrased as "Where was Poe born?" (as was done within TREC-9) it receives a middle clarity score and middle average precision, meaning our prediction method is accurate for the new question. So some part of the lack of correlation is a limitation of the metrics used. This is another problem that would have been ameliorated by more data, making it likely that "Poe" and "birthplace" would be used together.

The explanations behind different cases of prediction performance are varied, as one would expect when looking at instances of something based on human language, where the choice of particular wording in a certain passage, say, might have a large impact on statistical measures for a certain question. It is worth noting that small number of relevant passages can cause problems and successes for the system, but, in general, performance seems to benefit somewhat from the trend that rare passsages that match question terms well often contain correct answers.

5. WEB QUESTIONS

To demonstrate the usefulness of these techniques with more general questions and with a Web-like collection, we study questions extracted from the Excite query log searching the Web test collection WT10g[1]. Although the TREC-9 questions were, in part, taken from the same log of questions issued to the Excite search engine on December, 20th, 1999[13], the questions we automatically extract have much greater diversity, and we test them on diverse Web-derived documents. Our goal is to show how clarity-score based techniques can be used to help detect vague or unfocussed questions. Since there are no relevance judgements on these queries for these documents, we devised a simple evaluation to demonstrate that clarity scores correlate with an expert searcher's judgement of whether a question is focussed and specific enough to retrieve consistent answers on the Web[2].

The log contains about 2.5 million queries. A subset of 500 questions was sampled from 147,726 questions extracted from the log by virtue of beginning with a common question word variant and having a question mark. We further removed questions which contained non-stop words that were not present in the WT10g collection. We also filtered out sexual questions, compound questions, and questions with obvious misspellings[3], arriving at a final list of 445 questions.

The final list of questions was given to an experienced

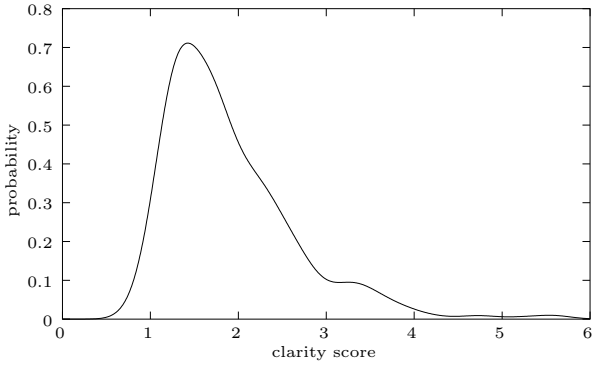

Figure 4: The clarity scores for our final set of 445 well formed questions randomly sampled from the Excite query log and processed as described in the body of the paper. The scores were computed with 500 top documents used to estimate the question model for each question, and smoothed as described in the main text. The distribution of clarity scores was estimated with kernel smoothing of bandwidth 0.2.

Class	Number	Ave. clarity	P-value
Predicted effective	223	2.03	0.00026
Predicted ineffective	222	1.81	0.00020

Table 2: The results of a human expert predicting if a question was focussed and specific enough to retrieve consistent results on the web. The P-values are compared to random tagging with a probability of 0.5 a question going into each class and were estimated via sampling experiments with 1 million samples.

[2]We do not believe that all measures of queries should be learned from expert's judgements. An important strength of clarity scores is that the way they are computed is both simple in terms of the statistics of word usage and theoretically meaningful, aiding research progress.

[3]Misspellings seem to vastly raise a question's clarity score

information retrieval researcher to tag, according to whether the expert believed each question was specific enough to retrieve a consistent answer on the Web.

In order to calculate clarity scores of the questions in this case, the WT10g web test collection and the question both had words on the InQuery stop list removed[4] and were passed through the Krovetz stemmer[10]. Clarity scores of the questions were calculated with the top 500 documents for each according to question likelihood and document models were smoothed via linear smoothing with 0.6 times the relative frequency of the term in the document plus 0.4 times the relative frequency of the term in the whole collection. The clarity scores, kernel smoothed with gaussians of standard deviation 0.2[2] are shown in Figure 4. The lowest clarity-score that was observed was 0.89, so the non-zero probability estimates for clarity scores slightly below that is due to our smoothing of the observed data.

The results of our evaluation are shown in Table 2. The average observed clarity score was 1.92, with the *predicted effective* questions significantly higher in average clarity score and the *predicted ineffective* questions significantly lower in average clarity score.

Since the distribution of clarity scores is not known and clearly depends on the queries, we turned to the average rank of the two classes (when all questions are ranked 1 through 445 in terms of clarity score) as a meaningful, distribution free statistic. We did 1 million trials of a sampling experiment where all questions were randomly assigned to the two classes with the observed probability of 0.5[4] We then estimated the P-values as the relative frequency of randomly tagged groups having mean ranks as or more extreme than the average ranks in the human-tagged groups. Therefore, we have established that clarity scores are significantly correlated with one information researcher's predictions, as compared to random tagging.

Considering examples provides a little more appreciation for the correlations, however. Consider the question "Where can I purchase an inexpensive computer?" from our sample. This question was tagged as *predicted ineffective*, since the non-common terms "purchase", "inexpensive," and "computer" in the question are likely to co-occur in web pages with very diverse language use. This question, in fact, received a clarity score of 0.89 which is the lowest observed in our set of 445 questions searching WT10g. At the other extreme, consider the question "Where can I find lyrics to Eleanor Rigby?" This question contains the terms "lyrics," "eleanor," and "rigby" that seem likely to co-occur in a unique context (a page giving the lyrics to a popular Beatles song) and the tagger predicted it was specific and focused enough to retrieve consistent answers on the Web. It received a clarity score of 8.08, the highest clarity score of the 445 questions. Agreements like these lead to the observed correlation. Understanding some examples can give us a little more confidence in the observed correlations despite weakness of having to compare it to random tagging to assess statistical significane.

by effectively causing the question model to be estimated from a few documents that also contain the same misspelling

[4]The tagger was given no instructions regarding what proportion of the questions to predict as effective. The numbers the tagger put in the two classes, however, were essentially equal.

6. RELATED WORK

The main related work in the area of passage retrieval is the submissions to the TREC-9 question answering track in the 250 character passage category[13][5]. The area of predicting the performance of queries and now of questions is a new area, that was largely opened up by the recent document-level work[6, 7].

7. FUTURE WORK

Future work on the passage retrieval system will focus on leveraging the power of relevance models[11] as well as models of answers to questions of the various types, called answer models. For example, the answer to a time question should contain some sort of date or time reference. Using an answer model, passages that did not contain a date or time reference could be discounted in a principled way, even if they scored reasonably on likelihood of generating the question.

With the clarity calculator integrated into the passage retrieval system, the clarity calculator can use the same estimates of $P(S|Q)$ as the question answering system and continue to predict the quality of the ranked passage lists well as the passage retrieval system becomes increasingly sophisticated.

Additionally, slight modifications to the clarity score concept that allow the comparison of two retrieval techniques for the same query (e.g. with query expansion and without) are showing promise at allowing systems to automatically select (on average) the best technique for a given query. By allowing researchers to break away, in a principled way, from treating all queries/questions the same way we believe that clarity score techniques will eventually significantly improve the retrieval performance of real systems.

8. CONCLUSIONS

We have demonstrated a method of predicting the quality of ranked lists of passages in passage retrieval systems. The method is based on computing the clarity scores for the questions over the appropriate set of candidate passages. The clarity score measure predicts both how difficult any subsequent extraction of short answers from the passages will be and how well the retrieval puts passages containing answers in high ranks. These are vital issues for question answering systems.

We also studied queries logged on the web and performed an evaluation that is suggestive that document-level clarity scores might be useful in predicting if a question is focused and specific enough to retrieve pages from the Web containing consistent answers.

In extending the previous work on clarity scores into the new domains of QA and web retrieval, we have provided a meaningful measure a questions' suitability for a given collection. With the development of the measure as an important first step, we believe it will eventually be useful in systems that treat various questions differently. Such uses will involve automatically identifying questions that are ambiguous with respect to the collection and could be clarified by dialogue with the user, and the automatic selection of the best performing retrieval technique for a given question.

[5]Papers available through
`http://trec.nist.gov/pubs/trec9/index.track.html`

9. ACKNOWLEDGEMENTS

We thank Yun Zhou, David Fisher, Fang-Fang Feng, and Don Metzler for technical help and advice. We are also indebted to James Allan for insightful comments and to Margaret Connell for query tagging. This work was supported in part by the Center for Intelligent Information Retrieval, in part by SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-02-1-8903, in part by Advanced Research and Development Activity under contract number MDA904-01-C-0984, and in part by NSF grant IIS-9907018. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

10. REFERENCES

[1] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for Web retrieval experiments. *Information Processing and Management*, to appear.

[2] A. Bowman and A. Azzilini. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, New York, 1997.

[3] C. Buckley. `trec_eval` IR Evaluation Package. Available from `ftp://ftp.cs.cornell.edu/pub/smart`.

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

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

[6] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proc. of Human Language Technology 2002*, pages 94–98, March 2002.

[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, August 2002.

[8] S. Dumais, M. Banko, E.Brill, J. Lin, and A. Ng. Web question answering: Is more always better? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, August 2002.

[9] J. D. Gibbons and S. Chakraborty. *Nonparametric Statistical Inference, 3rd ed.* Marcel Dekker, New York, New York, 1992.

[10] R. Krovetz. Viewing morpholgy as an inference process. In *Proc. of the 16th Annual ACM SIGIR Conference*, pages 191–202, June–July 1993.

[11] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Research and Development in Information Retrieval*, pages 120–127, 2001.

[12] T. Morton, 2001. Personal Communication.

[13] E. M. Voorhees. Overview of the TREC-9 question answering track. In E. Voorhees and D. Harman, editors, *Proceedings of the Ninth Text REtrieval Conference(TREC-9)*, 2000. NIST Special Publication 500-249.