Quantifying Query Ambiguity

Steve Cronen-Townsend and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
{crotown, croft}@cs.umass.edu

ABSTRACT

We develop a measure of a query with respect to a collection of documents with the aim of quantifying the query's ambiguity with respect to those documents. This measure, the *clarity score*, is the relative entropy between a query language model and the corresponding collection language model. We substantiate that the clarity score measures the coherence and specificity of the language used in documents likely to satisfy the query. We also argue that it provides a suitable quantification of the (lack of) ambiguity of a query with respect to a collection of documents and has potential applications throughout the field of information retrieval. In particular, the clarity score is shown to correlate positively with average precision in evaluations using TREC test collections. Hence, as one example, the clarity score could serve as a predictor of query performance. Systems would then be able to identify vague information requests and respond differently than they would to clear and specific requests.

1. INTRODUCTION

An important challenge for information retrieval systems is dealing gracefully with ambiguous queries. For example, suppose a user interested in the competitors in the 1988-1989 soccer World Cup gives the query "World Cup" against the AP88 collection of news articles. If those two words are the only evidence the system has about what the user means, it is simply impossible for the system to return the soccer articles consistently higher in the ranked list than the articles about World Cup chess tournaments, that are, in fact, predominant in the chosen collection among the articles that use the query terms frequently. 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 occasional soccer articles farther down the list. Hence, for this collection, the query "World Cup" is significantly more ambiguous (less clear) than than the query "World Cup soccer." However, if there were no articles about World Cup chess tournaments in the collection, the two queries would be similar in ambiguity. Hence ambiguity is a matter of degree, and, furthermore, since the degree of ambiguity depends of the details of the the documents present in the collection all queries are ambiguous to some degree.

Clearly information retrieval systems must act on more user information than just the query words in cases of high query ambiguity. One common approach has been treating all queries the same initially but then refining the document list in response to user feedback on the initial list. We envisage a new approach where vague queries will be handled differently than clear ones from the beginning. In this paper, we address the important first step of automatically identifying queries that are vague with respect to a given collection of documents.

Firstly, we motivate the definition of a clarity score and show how to compute it and give examples and visualizations. Despite the possibility of examining clarity score contributions due to individual terms, we have not developed clarity scores as a term-weighting scheme and show that there is no discernible correlation with a traditional inverse document frequency measure. Then we substantiate that the clarity score measures the coherence and specificity of the language associated with documents likely to satisfy the query and argue that it therefore serves as a suitable quantification of query ambiguity. We then show the positive association between clarity score and performance of a query using test collections, examine related work, and conclude.

2. DEFINING CLARITY SCORES

2.1 Motivation

Starting from a language modeling viewpoint, quantifying the ambiguity of a query rests on the notion of a query language model. By a query language model, we mean a language model representing the collection word usage that is associated with the query. The simplest version of this is a unigram language model, which is an estimated probability distribution for the occurrence of single terms in documents associated with the query.

Counting the number of topics in this language model seems like a sensible quantification of ambiguity, since it captures the idea that a query is ambiguous if it returns documents about many different topics. However, the notion of topic is not mathematically well-defined, since topics are overlapping and depend on a chosen level of topic generality. Some advances have been made in this area by trying

HLT '02 San Diego, CA
.

to capture latent semantic information[7], but we have not pursued this approach.

Considering the entropy of the query model is a potentially interesting approach, since the entropy of a probability distribution measures directly how much the distribution specifies certain values. So, in this case, the entropy measures how strongly the distribution specifies certain words.

In this approach it is important to remove, as much as possible, the large and fluctuating contributions to the entropy from generic words. This can be done by computing the relative entropy between the query model and the overall collection language model, instead of the entropy of the query language model alone. These considerations lead to a measure where a query whose language model looks like the language model for the whole collection receives a low score, and a query whose language model is very different from the the collection language model receives a high score. Thus we are measuring something like the lack of ambiguity, and we call the measure the *clarity score*.

We believe that a typical way a query language model proves similar to the collection model is that the query selects documents that are about many different topics. The blurring of the individual topic models together creates an overall distribution that is smoother and more similar to the collection language model than the model for a query that selects documents about a single topic. The single-topic model will have large spikes at topic words. Thus, the similarity between the query language model and the collection language model is related to the coherence of the retrieved set, or the degree to which the favored documents are about the same topic.

Query ambiguity, as compared to single word ambiguity, is relatively simple. We believe that it can be fairly well characterized by a single number. Since the notions of clarity and ambiguity are themselves rather vague, we take our measure as the definition of the clarity (lack of ambiguity) of a query with respect to a certain collection of documents. Thus query ambiguity, in our conception, becomes strictly one-dimensional. We now make the definition of clarity scores mathematically precise, leaving the issue of its degree of correspondence to the general connotations of the terms ambiguity and clarity for Section 3.

2.2 Definition

In order to compute a clarity score, one must first create a query language model. We have investigated both of the methods put forward by Lavrenko and Croft[11] for estimating such models[1]. The best all-around way of estimating the query model for the purposes of computing clarity scores is Lavrenko and Croft's *Method 1*[4]. In this approach one, in effect, assumes that the query terms and the words in the documents are sampled identically and independently from the query model unigram distribution.

The query model unigram distribution is estimated by

$$P(w|Q) = \sum_{D \in R} P(w|D)P(D|Q), \tag{1}$$

where w is any term, D is a document, Q is the query, and R is the set of all documents containing at least one query term. Inside conditional probabilities, D refers to a language model estimated from the corresponding single document.

[1]Lavrenko and Croft refer to these as *relevance models.*

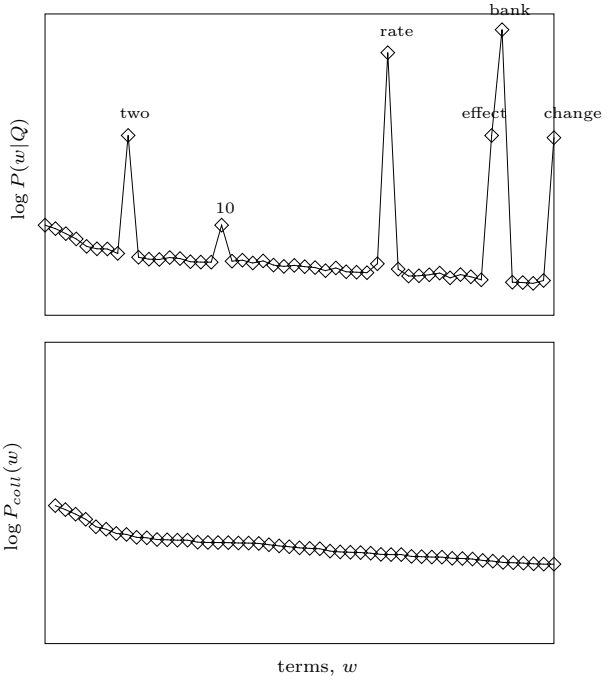

Figure 1: The query language model for query A, "Show me any predictions for changes in the prime lending rate and any changes made in the prime lending rates" in TREC disk 1 and the collection language model for that corpus. The top 50 terms are plotted in order of collection probabilities.

The likelihood of a query given a document is estimated using the language modeling approach of [17] as

$$P(Q|D) = \prod_{q \in Q} P(q|D), \tag{2}$$

and $P(D|Q)$ is obtained by Bayesian inversion with uniform prior probabilities for documents in R and a zero prior for documents that contain no query terms.

The probabilities $P(w|D)$ and $P(q|D)$ in (1) and (2) are estimated with linear smoothing[12] given by

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

where $P_{ml}(w|D)$ is simply the relative frequency of term w in documents D, $P_{coll}(w)$ is the relative frequency of the term in the collection as a whole, and $\lambda = 0.6$ throughout this study. Figure 1 shows a query language model for a clear query and the corresponding collection model.

The clarity score for the query is simply the relative entropy, or Kullback-Leibler divergence[3], between the query language model and the collection language model, given by

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

where V is the entire vocabulary of the collection.

In practice the most efficient way to calculate clarity scores is to estimate the query model by sampling documents to produce a query language model using equation (1). Unless otherwise noted, all clarity scores in this paper are calculated by sampling until a limit of 500 unique documents is reached. Alternatively to equation (4), one can compute

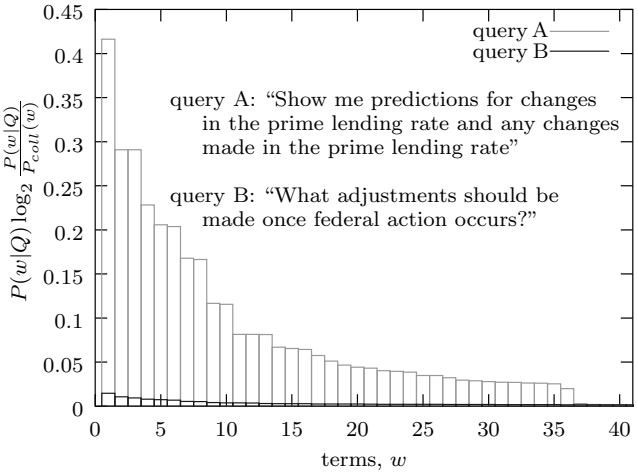

Figure 2: A clear versus a vague query in TREC disk 1 with the top 40 terms in 4) plotted separately. The top six contributions, in order, for query A are for the terms "bank," "hong," "kong," "rate," "lend," and "prime." The top six contributing terms for query B are "adjust," "federal," "action," "land," "occur," and "hyundai."

the divergence with the roles of the query and collection language models reversed. One can also use either the query model of equation (1) or Lavrenko and Croft's *Method 2*-type models. Of the 4 combinations, the definition given above performs slightly better than the others on large collections. See [4] for more details.

2.3 Examples

Consider two queries, A and B against the TREC disk 1 collection of documents. Query A is "Show me any predictions for changes in the prime lending rate and any changes made in the prime lending rates" (see Figure 1 for the query language model). Query B is "What adjustments should be made once federal action occurs?". These queries happen to be two query variants for TREC topic 56 from the query track[2]. Figure 1 shows the language models used in computing the clarity score for query A, which is 2.85 bits. In terms of these figures, the clarity score is the top graph minus the bottom graph, times 2 to the power of the top graph and summed over terms. Just the difference between the two graphs makes a function with spikes at the most topical words and the final step of taking the expectation value of this quantity using the top distribution enhances the peaks further still. This procedure can also be interpreted as computing the number of bits wasted, on average, when one encodes term-occurrence events sampled from the query language model with a code optimally designed for the collection language model.

The best way to visualize the difference between clear and vague queries is to plot the clarity contributions on a term-by-term basis as in Figure 2. Here $P(w|Q) * \log \frac{P(w|Q)}{P_{coll}(w)}$ is plotted for each of the top 40 contributing terms w, sorted in descending order of contribution. In this representation, the clarity score is the total area of the bars leading to a clarity score of 2.85 for query A (grey bars) and 0.37 for query B (black bars). The contrast between the high-clarity query A and the low-clarity query B is evident.

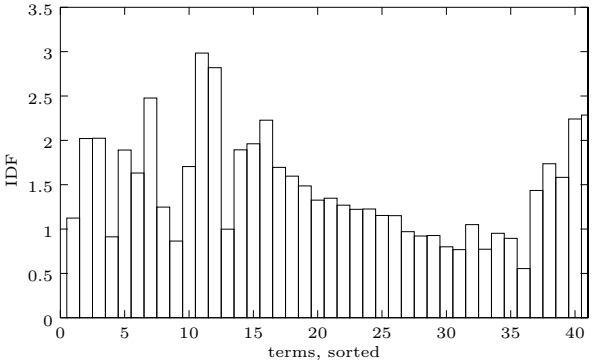

Figure 3: The IDFs of the top 40 contributing terms to the clarity score of query A with the terms plotted in the same order as in Figure 2.

It is important to note that we have not developed this method as a weighting scheme for terms. Although the contributions of individual terms to the clarity score of an entire query may be examined (as in Figure 2) these scores do not, for example, correlate with traditional inverse document frequency(IDF) measures. Consider, for example, IDF, defined as

$$IDF(w) = \log_{10} \frac{\text{number of docs}}{\text{number of docs containing } w}. \quad (5)$$

This measure is plotted in Figure 3 for the top 40 clarity score contributing terms for query A. Comparing this graph to the grey bars in Figure 2 shows evidence for only a possible small correlation between IDF and clarity score contributions to this query. Any correlation disappears quickly as one considers more and more terms since the additional terms have uniformily small clarity contributions but often have large IDF values. The leftmost bar, corresponding to "bank", and the fourth bar corresponding to "rate," show relatively modest IDF scores for these two terms, despite their large contributions to the clarity score of query A. A term's clarity score contribution reflects the degree to which that term's probability in the query model rises above its background level. A term's IDF, however, depends on the collection but not on the query under consideration, making it a less specific measure.

Further examples from the TREC-7 ad hoc track are shown in Table 1. The first entry indicates that query that had rank 1 (lowest) score was the title of TREC topic 379, namely "mainstreaming" and it has a clarity score of 0.311. Similarly the last entry indicates that the highest scoring query in the set of 50 queries was "anorexia nervosa bulimia", the title of TREC topic 369, with a score of 2.921.

Some insight into our method can be gained by considering the query at rank 46 "sick building syndrome" where none of the terms individually seem very specific but through documents where they co-occur they generate a fairly spiky language model. On the flip side, the query "drug legalization benefits" scores poorly, at rank 7, even though the terms seem reasonable. In this case, the term "benefit" probably helps the query marginally because it occurs in so many documents and does not probably co-occur extremely strongly with the other two terms. Moreover, the specific meaning

of legalization of drugs is lost here by the stemming, which causes "legalization" to match any other words with the stem "legal". This effect might also be exacerbated by a limited number of documents about the legalization of drugs in the collection. Documents that only match one of 'legal" or "drug" or match both but are not about legalization of drugs are about many topics and make a language model similar to the over collection language model, leading to a low clarity score.

rank	topic	query	score
1	379	mainstreaming	0.311
2	354	journalist risks	0.315
3	391	R&D drug prices	0.372
4	359	mutual fund predictors	0.401
5	381	alternative medicine	0.428
6	395	tourism	0.464
7	360	drug legalization benefits	0.481
8	376	World Court	0.490
9	371	health insurance holistic	0.492
10	366	commercial cyanide uses	0.498
41	382	hydrogen fuel automobiles	1.630
42	355	ocean remote sensing	1.700
43	351	Falkland petroleum exploration	1.797
44	353	Antarctica exploration	1.801
45	365	El Nino	1.832
46	396	sick building syndrome	1.882
47	374	Nobel prize winners	1.977
48	368	in vitro fertilization	2.031
49	356	postmenopausal estrogen Britain	2.083
50	369	anorexia nervosa bulimia	2.921

Table 1: The clarity scores of the ten lowest and the ten highest clarity scoring title queries in TREC-7. The query at rank 42 ("ocean remote sensing") made such a peaked $\mathbf{P(D|Q)}$ that only 89 unique documents appeared in 250,000 samples and were used to construct the query language model. The query at rank 48 ("in vitro fertilization") was modelled with all the 63 documents that contained either terms matching "vitro" or "fertilize." Similarly, the query at rank 10, "commercial cyanide uses" was modelled with all of the 414 documents that matched the two query terms after the stop word "uses" was removed. All other queries were modelled with the usual 500 documents.

3. QUERY AMBIGUITY AND CLARITY

Figure 4 shows a set of interrelated queries and their clarity scores in the TREC disk 1 collection. Since our system turns all text to lower case and morphologically normalizes it[9], the query "jobs" is considered equivalent to "Jobs" and would match documents mentioning the co-founder of Apple Computer or mentioning employment.

The right side from top to bottom shows a typical series of transformations where additional terms are added to reduce the ambiguity of the query. The original query "jobs" would rank highly documents about many topics, whereas "textiles jobs" more narrowly focuses on documents about jobs related to cloth. Finally "denim textiles jobs" narrows down the meaning of the query even further with correspondingly

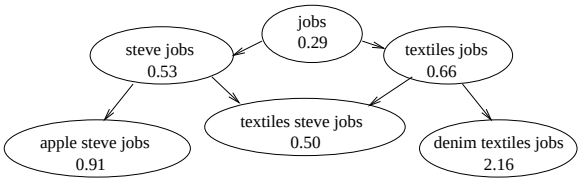

Figure 4: Clarity scores of some queries in the TREC disk 1 collection. Arrows show adding a query term.

Collection	Queries	Number	R	P-value
AP88+89	$101-200$	100	0.368	1.2×10^{-4}
TREC 7	$351-400$	50	0.577	2.7×10^{-5}
TREC 8	$401-450$	50	0.494	2.7×10^{-4}
TREC7+8	$351-450$	100	0.536	4.8×10^{-8}

Table 2: The correlation of clarity scores with average precision in several TREC test collections. The queries are the titles of the TREC topics (usually a few words).

higher clarity scores at each step. The left side of the figure follows a similar progression. The middle query "textiles steve jobs" is an ambiguous combination of the two streams of meaning and its clarity score is lower then either of the two word queries "steve jobs" or "textiles jobs." These examples suggest that the clarity score is related, to some degree, to everyday notions of the ambiguity of a query as a whole, in terms of the number of topics it could be referring to.

The middle query "textiles steve jobs" offers a way to probe this relationship even further. By hand-labeling the documents and computing the clarity score on the subsets we see some interesting trends. Using just the top 20 documents to estimate the query model, the clarity score of this ambiguous query is 1.39[2]. Using just the top 20 documents mentioning the co-founder of Apple Computer, Steve Jobs, the clarity score is 1.42. Using just the top 20 documents that mention textiles the clarity score is 1.54, while the clarity score using the top 20 mentioning neither is 1.26. In this simple example mixing documents from disparate topics reduces the clarity score, and those documents that are in the miscellaneous category are the lowest scoring subset, presumably since they are a mix of many different subjects. Documents in the miscellaneous category are mainly contain "by Steve Cook" (which increases $P(D|Q)$ since "steve" is a query term) and happen to be about a variety of computer-related topics, or are about employment but do not mention textiles. Experiments like this provide evidence for our assertion that clarity scores provide a suitable quantification of query ambiguity.

4. CLARITY AND RETRIEVAL PRECISION

In order to assess the potential for using clarity scores to predict the performance of a query, we measure the correlation between clarity scores and average precision scores for

[2]This score is higher than that reported in Figure 4 since the artificial certainty created by estimating the query language model with 20 as compared to 500 documents raises the relative entropy.

various TREC test collections. Since the distributions are unknown, an appropriate test is the Spearmann rank correlation test[6]. The scores are replaced by rankings, and one computes what appears to be a correlation coefficient from elementary statistics between the two rankings. A score of 1 indicates perfect agreement in the rankings and a score of -1 indicates opposite rankings. This is a distribution-free statistic whose null distribution (if the two rankings are unassociated) is well-approximated by a normal for sample sizes as large as 50, our smallest sample size, thus it is straightforward to compute the p-values, or probabilities that results as extreme or more extreme occurred by chance. See Table 2 for some results. More details and data may be found in [4].

5. RELATED WORK

Ambiguity has long been of interest in information retrieval[5] and in the larger field of language technologies(see [12]). The information retrieval work focuses on the improvements possible through retrieving documents based on unambiguous word senses(see [10], [15], and [16]). In operation, this approach requires word sense disambiguation, the automatic determination of word senses, when the documents are indexed. The present work is different in that it measures ambiguity without attempting to resolve it, and that it measures the query as a whole.

Another focus of research attention started with the automatic choice of index terms[1] and led to automatic identification of stop words[19] and content-bearing terms[8]. This stream has produced two studies closely related to ours. Pirkola and Jarvelin[13] examine individual term contributions to the retrieval effectiveness of queries and have some success at identifying the most important query term when there is no information as to the actual relevance of the documents to the query. In seeking to classify questions as easy or hard, Sullivan[18] models very long question text directly and compares questions in a sophisticated way to an existing set of questions whose effectiveness at retrieving relevant documents (when viewed as information retrieval-style queries) has been measured.

Additionally, a nearly identical mathematical approach to ours has been used to model selectional preferences in natural language[14].

6. CONCLUSIONS

We have shown that the query clarity score, as defined, corresponds to certain connotations of the lack of ambiguity. The example in Section 3 suggests that queries whose likely documents are a mix of documents from disparate topics receive a lower score than if they resulted in a topically-coherent retrieved set. It has also been shown that the measure correlates strongly and significantly with average precision. Clarity score computation does not refer to the notion of the information need behind the query and may be used to make predictions about the performance of a query in general information retrieval systems, where there is no information as to the actual relevance of documents to certain requests. Based on these initial results, query clarity promises to be a useful tool in many areas of information retrieval and related language technologies.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912 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.

8. REFERENCES

[1] R. Bookstein and D. R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25:312–318, 1974.

[2] C. Buckley. The trec-9 query track. In E. Voorhees and D. Harman, editors, *Proceedings of the Ninth Text REtrieval Conference(TREC-9)*, 2000. NIST Special Publication 500-249.

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

[4] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. Submitted to SIGIR 2002.

[5] L. Earl. Use of word government in resolving syntactic and semantic ambiguities. *Information Storage and Retrieval*, 9:639–664, 1973.

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

[7] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual international ACM SIGIR Conference*, pages 50–57, 1999.

[8] W. Kim and W. J. Wilbur. Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology*, 52(3):247–259, 2001.

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

[10] R. Krovetz and W. B. Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141, 1992.

[11] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proc. of the 24th Annual ACM SIGIR Conference*, pages 120–127, September 2001.

[12] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

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

[14] P. Resnik. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159, 1996.

[15] M. Sanderson and K. van Rijsbergen. The impact on retrieval effectiveness of skewed frequency distributions. *ACM Transactions on Information Systems*, 17(4):440–465, 1999.

[16] H. Schütze and J. Pederson. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and*

Information Retrieval, pages 161–175, 1995.

[17] F. Song and W. B. Croft. A general language model for informatioon retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference*, pages 279–280, 1999.

[18] T. Sullivan. Locating question difficulty through explorations in question space. In *Proceedings of the 1st ACM/IEEE Joint Conference on Digital Libraries*, pages 251–252, 2001.

[19] W. J. Wilbur and K. Sirotkin. The automatic identification of stop words. *Journal of Information Science*, 18(1):45–55, 1992.