The Maximum Entropy Method for Analyzing Retrieval Measures

Javed A. Aslam*, Emine Yilmaz, Virgiliu Pavlu
College of Computer and Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115
{jaa,emine,vip}@ccs.neu.edu

ABSTRACT

We present a model, based on the maximum entropy method, for analyzing various measures of retrieval performance such as average precision, R-precision, and precision-at-cutoffs. Our methodology treats the value of such a measure as a constraint on the distribution of relevant documents in an unknown list, and the maximum entropy distribution can be determined subject to these constraints. For good measures of overall performance (such as average precision), the resulting maximum entropy distributions are highly correlated with actual distributions of relevant documents in lists as demonstrated through TREC data; for poor measures of overall performance, the correlation is weaker. As such, the maximum entropy method can be used to quantify the overall quality of a retrieval measure. Furthermore, for good measures of overall performance (such as average precision), we show that the corresponding maximum entropy distributions can be used to accurately infer precision-recall curves and the values of other measures of performance, and we demonstrate that the quality of these inferences far exceeds that predicted by simple retrieval measure correlation, as demonstrated through TREC data.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Performance evaluation*

General Terms

Theory, Measurement, Experimentation

Keywords

Evaluation, Maximum Entropy, Average Precision

*We gratefully acknowledge the support provided by NSF grant CCF-0418390.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

1. INTRODUCTION

The efficacy of retrieval systems is evaluated by a number of performance measures such as average precision, R-precision, and precisions at standard cutoffs. Broadly speaking, these measures can be classified as either *system-oriented* measures of overall performance (e.g., average precision and R-precision) or *user-oriented* measures of specific performance (e.g., precision-at-cutoff 10) [3, 12, 5]. Different measures evaluate different aspects of retrieval performance, and much thought and analysis has been devoted to analyzing the quality of various different performance measures [10, 2, 17].

We consider the problem of analyzing the quality of various measures of retrieval performance and propose a model based on the maximum entropy method for evaluating the quality of a performance measure. While measures such as average precision at relevant documents, R-precision, and 11pt average precision are known to be good measures of overall performance, other measures such as precisions at specific cutoffs are not. Our goal in this work is to develop a model within which one can numerically assess the overall quality of a given measure based on the reduction in uncertainty of a system's performance one gains by learning the value of the measure. As such, our evaluation model is primarily concerned with assessing the relative merits of system-oriented measures, but it can be applied to other classes of measures as well.

We begin with the premise that the quality of a list of documents retrieved in response to a given query is strictly a function of the sequence of relevant and non-relevant documents retrieved within that list (as well as R, the total number of relevant documents for the given query). Most standard measures of retrieval performance satisfy this premise. Our thesis is then that given the assessed value of a "good" overall measure of performance, one's uncertainty about the sequence of relevant and non-relevant documents in an unknown list should be greatly reduced. Suppose, for example, one were told that a list of 1,000 documents retrieved in response to a query with 200 total relevant documents contained 100 relevant documents. What could one reasonably infer about the sequence of relevant and non-relevant documents in the unknown list? From this information alone, one could only reasonably conclude that the likelihood of seeing a relevant document at any rank level is uniformly $1/10$. Now suppose that one were additionally told that the average precision of the list was 0.4 (the maximum possible in this circumstance is 0.5). Now one could reasonably conclude that the likelihood of seeing relevant documents at low numerical ranks is much greater than the likelihood of seeing relevant documents at high numerical ranks. One's uncertainty about the sequence of relevant and non-relevant documents in the unknown list is greatly reduced as a consequence of the strong constraint that such an average precision places on lists in this situation. Thus, average precision is highly informative. On the other hand, suppose that one were instead told that the precision of the documents in the rank range $[100, 110]$ was 0.4. One's uncertainty about the sequence of relevant and non-relevant documents in the unknown list is not appreciably reduced as a consequence of the relatively weak constraint that such a measurement places on lists. Thus, precision in the range $[100, 110]$ is not a highly informative measure. In what follows, we develop a model within which one can *quantify* how informative a measure is.

We consider two questions: (1) What can reasonably be inferred about an unknown list given the value of a measurement taken over this list? (2) How accurately do these inferences reflect reality? We argue that the former question is properly answered by considering the maximum entropy distributions subject to the measured value as a constraint, and we demonstrate that such maximum entropy models corresponding to good overall measures of performance such as average precision yield accurate inferences about underlying lists seen in practice (as demonstrated through TREC data).

More specifically, we develop a framework based on the maximum entropy method which allows one to infer the most "reasonable" model for the sequence of relevant and non-relevant documents in a list given a measured constraint. From this model, we show how one can infer the most "reasonable" model for the unknown list's entire precision-recall curve. We demonstrate through the use of TREC data that for "good" overall measures of performance (such as average precision), these inferred precision-recall curves are accurate approximations of actual precision-recall curves; however, for "poor" overall measures of performance, these inferred precision-recall curves do not accurately approximate actual precision-recall curves. Thus, maximum entropy modeling can be used to quantify the quality of a measure of overall performance.

We further demonstrate through the use of TREC data that the maximum entropy models corresponding to "good" measures of overall performance can be used to make accurate predictions of other measurements. While it is well known that "good" overall measures such as average precision are well correlated with other measures of performance, and thus average precision could be used to reasonably predict other measures of performance, we demonstrate that the maximum entropy models corresponding to average precision yield inferences of other measures even more highly correlated with their actual values, thus validating both average precision and maximum entropy modeling.

In the sections that follow, we first describe the maximum entropy method and discuss how maximum entropy modeling can be used to analyze measures of retrieval performance. We then describe the results of applying our methodology using TREC data, and we conclude with a summary and future work.

2. THE MAXIMUM ENTROPY METHOD

The concept of entropy as a measure of information was first introduced by Shannon [20], and the Principle of Maximum Entropy was introduced by Jaynes [7, 8, 9]. Since its introduction, the Maximum Entropy Method has been applied in many areas of science and technology [21] including natural language processing [1], ambiguity resolution [18], text classification [14], machine learning [15, 16], and information retrieval [6, 11], to name but a few examples. In what follows, we introduce the maximum entropy method through a classic example, and we then describe how the maximum entropy method can be used to evaluate measures of retrieval performance.

Suppose you are given an unknown and possibly biased six-sided die and were asked the probability of obtaining any particular die face in a given roll. What would your answer be? This problem is under-constrained and the most seemingly "reasonable" answer is a uniform distribution over all faces. Suppose now you are also given the information that the average die roll is 3.5. The most seemingly "reasonable" answer is still a uniform distribution. What if you are told that the average die roll is 4.5? There are many distributions over the faces such that the average die roll is 4.5; how can you find the most seemingly "reasonable" distribution? Finally, what would your answer be if you were told that the average die roll is 5.5? Clearly, the belief in getting a 6 increases as the expected value of the die rolls increases. But there are many distributions satisfying this constraint; which distribution would you choose?

The "Maximum Entropy Method" (MEM) dictates the most "reasonable" distribution satisfying the given constraints. The "Principle of Maximal Ignorance" forms the intuition behind the MEM; it states that one should choose the distribution which is least predictable (most random) subject to the given constraints. Jaynes and others have derived numerous entropy concentration theorems which show that the vast majority of all empirical frequency distributions (e.g., those corresponding to sequences of die rolls) satisfying the given constraints have associated empirical probabilities and entropies very close to those probabilities satisfying the constraints whose associated entropy is maximal [7].

Thus, the MEM dictates the most random distribution satisfying the given constraints, using the entropy of the probability distribution as a measure of randomness. The entropy of a probability distribution $\vec{p} = \{p_1, p_2, \ldots, p_n\}$ is a measure of the uncertainty (randomness) inherent in the distribution and is defined as follows

$$H(\vec{p}) = -\sum_{i=1}^{n} p_i \lg p_i.$$

Thus, maximum entropy distributions are probability distributions making no additional assumptions apart from the given constraints.

In addition to its mathematical justification, the MEM tends to produce solutions one often sees in nature. For example, it is known that given the temperature of a gas, the actual distribution of velocities in the gas is the maximum entropy distribution under the temperature constraint.

We can apply the MEM to our die problem as follows. Let the probability distribution over the die faces be $\vec{p} = \{p_1, \ldots, p_6\}$. Mathematically, finding the maximum entropy distribution over die faces such that the expected die roll is

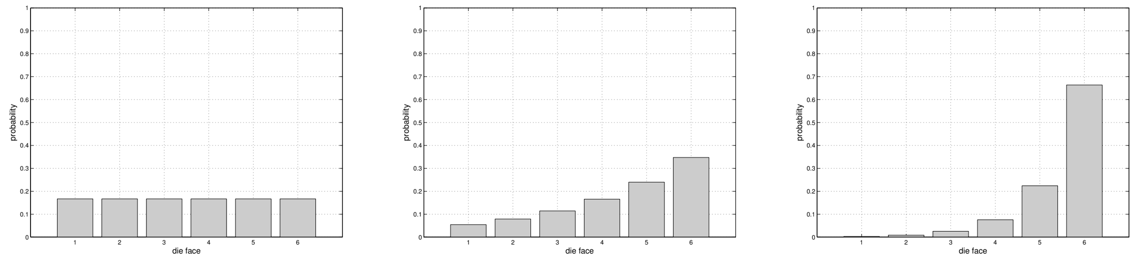

Figure 1: Maximum entropy die distributions with mean die rolls of 3.5, 4.5, and 5.5, respectively.

d corresponds to the following optimization problem:

Maximize: $H(\vec{p})$

Subject to:

1. $\sum_{i=1}^{6} p_i = 1$
2. $\sum_{i=1}^{6} i \cdot p_i = d$

The first constraint ensures that the solution forms a distribution over the die faces, and the second constraint ensures that this distribution has the appropriate expectation. This is a constrained optimization problem which can be solved using the method of Lagrange multipliers. Figure 1 shows three different maximum entropy distributions over the die faces such that the expected die roll is 3.5, 4.5, and 5.5, respectively.

2.1 Application of the Maximum Entropy Method to Analyzing Retrieval Measures

Suppose that you were given a list of length N corresponding to the output of a retrieval system for a given query, and suppose that you were asked to predict the probability of seeing any one of the 2^N possible patterns of relevant documents in that list. In the absence of any information about the query, any performance information for the system, or any *a priori* modeling of the behavior of retrieval systems, the most "reasonable" answer you could give would be that all lists of length N are equally likely. Suppose now that you are also given the information that the expected number of relevant documents over all lists of length N is R_{ret}. Your "reasonable" answer might then be a uniform distribution over all $\binom{N}{R_{ret}}$ different possible lists with R_{ret} relevant documents. But what if apart from the constraint on the number of relevant documents retrieved, you were also given the constraint that the expected value of average precision is ap? If the average precision value is high, then of all the $\binom{N}{R_{ret}}$ lists with R_{ret} relevant documents, the lists in which the relevant documents are retrieved at low numerical ranks should have higher probabilities. But how can you determine the most "reasonable" such distribution? The maximum entropy method essentially dictates the most reasonable distribution as a solution to the following constrained optimization problem.

Let $p(r_1, ..., r_N)$ be a probability distribution over the relevances associated with document lists of length N, let $rel(r_1, ..., r_N)$ be the number of relevant documents in a list, and let $ap(r_1, ..., r_N)$ be the average precision of a list. Then the maximum entropy method can be mathematically formulated as follows:

Maximize: $H(\vec{p})$

Subject to:

1. $\sum_{r_1,\cdots,r_N} p(r_1, \ldots, r_N) = 1$
2. $\sum_{r_1,\cdots,r_N} ap(r_1, \ldots, r_N) \cdot p(r_1, \ldots, r_N) = ap$
3. $\sum_{r_1,\cdots,r_N} rel(r_1, \ldots, r_N) \cdot p(r_1, \ldots, r_N) = R_{ret}$

Note that the solution to this optimization problem is a distribution over possible lists, where this distribution effectively gives one's *a posteriori* belief in any list given the measured constraint.

The previous problem can be formulated in a slightly different manner yielding another interpretation of the problem and a mathematical solution. Suppose that you were given a list of length N corresponding to output of a retrieval system for a given a query, and suppose that you were asked to predict the probability of seeing a relevant document at some rank. Since there are no constraints, all possible lists of length N are equally likely, and hence the probability of seeing a relevant document at any rank is $1/2$. Suppose now that you are also given the information that the expected number of relevant documents over all lists of length N is R_{ret}. The most natural answer would be a R_{ret}/N uniform probability for each rank. Finally, suppose that you are given the additional constraint that the expected average precision is ap. Under the assumption that our distribution over lists is a product distribution (this is effectively a fairly standard independence assumption), we may solve this problem as follows. Let

$$p(r_1, \ldots, r_N) = p(r_1) \cdot p(r_2) \cdots p(r_N)$$

where $p(r_i)$ is the probability that the document at rank i is relevant. We can then solve the problem of calculating the probability of seeing a relevant document at any rank using the MEM. For notational convenience, we will refer to this product distribution as the *probability-at-rank* distribution and the probability of seeing a relevant document at rank i, $p(r_i)$, as p_i.

Standard results from information theory [4] dictate that if $p(r_1, \ldots, r_N)$ is a product distribution, then

$$H(p(r_1, \ldots, r_N)) = \sum_{i=1}^{N} H(p_i)$$

where $H(p_i)$ is the binary entropy

$$H(p_i) = -p_i \lg p_i - (1 - p_i) \lg(1 - p_i).$$

Furthermore, it can be shown that given a product distribution $p(r_1, \ldots, r_N)$ over the relevances associated with docu-

Maximize: $\sum_{i=1}^{N} H(p_i)$
Subject to:

1. $\frac{1}{R}\sum_{i=1}^{N}\left(\frac{p_i}{i}\left(1+\sum_{j=1}^{i-1} p_j\right)\right) = ap$
2. $\sum_{i=1}^{N} p_i = R_{ret}$

Figure 2: Maximum entropy setup for average precision.

Maximize: $\sum_{i=1}^{N} H(p_i)$
Subject to:

1. $\frac{1}{R}\sum_{i=1}^{R} p_i = rp$
2. $\sum_{i=1}^{N} p_i = R_{ret}$

Figure 3: Maximum entropy setup for R-precision.

Maximize: $\sum_{i=1}^{N} H(p_i)$
Subject to:

1. $\frac{1}{k}\sum_{i=1}^{k} p_i = PC(k)$
2. $\sum_{i=1}^{N} p_i = R_{ret}$

Figure 4: Maximum entropy setup for precision-at-cutoff.

ment lists of length N, the expected value of average precision is

$$\frac{1}{R}\sum_{i=1}^{N}\left(\frac{p_i}{i}\left(1+\sum_{j=1}^{i-1} p_j\right)\right). \tag{1}$$

(The derivation of this formula is omitted due to space constraints.) Furthermore, since p_i is the probability of seeing a relevant document at rank i, the expected number of relevant documents retrieved until rank N is $\sum_{i=1}^{N} p_i$.

Now, if one were given some list of length N, one were told that the expected number of relevant documents is R_{ret}, one were further informed that the expected average precision is ap, and one were asked the probability of seeing a relevant document at any rank under the independence assumption stated, one could apply the MEM as shown in Figure 2. Note that one now solves for the maximum entropy product distribution over lists, which is equivalent to a maximum entropy probability-at-rank distribution. Applying the same ideas to R-precision and precision-at-cutoff k, one obtains analogous formulations as shown in Figures 3 and 4, respectively.

All of these formulations are constrained optimization problems, and the method of Lagrange multipliers can be used to find an analytical solution, in principle. When analytical solutions cannot be determined, numerical optimization methods can be employed. The maximum entropy distributions for R-precision and precision-at-cutoff k can be obtained analytically using the method of Lagrange multipliers. However, numerical optimization methods are required to determine the maximum entropy distribution for average precision. In Figure 5, examples of maximum entropy probability-at-rank curves corresponding to the measures average precision, R-precision, and precision-at-cutoff 10 for a run in TREC8 can be seen. Note that the probability-at-rank curves are step functions for the precision-at-cutoff and R-precision constraints; this is as expected since, for example, given a precision-at-cutoff 10 of 0.3, one can only reasonably conclude a uniform probability of 0.3 for seeing a relevant document at any of the first 10 ranks. Note, however, that the probability-at-rank curve corresponding to average precision is smooth and strictly decreasing.

Using the maximum entropy probability-at-rank distribution of a list, we can infer the maximum entropy precision-recall curve for the list. Given a probability-at-rank distribution $\vec{p}$, the number of relevant documents retrieved until rank i is $REL(i) = \sum_{j=1}^{i} p_j$. Therefore, the precision and recall at rank i are $PC(i) = REL(i)/i$ and $REC(i) = REL(i)/R$. Hence, using the maximum entropy probability-

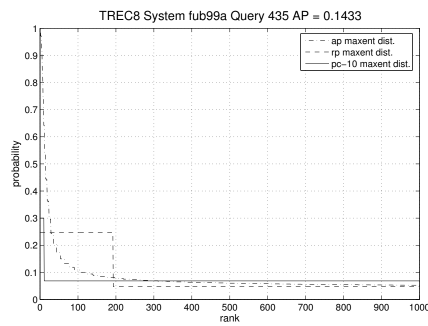

Figure 5: Probability-at-rank distributions.

at-rank distribution for each measure, we can generate the maximum entropy precision-recall curve of the list. If a measure provides a great deal of information about the underlying list, then the maximum entropy precision-recall curve should approximate the precision-recall curve of the actual list. However, if a measure is not particularly informative, then the maximum entropy precision-recall curve need not approximate the actual precision-recall curve. Therefore, noting how closely the maximum entropy precision-recall curve corresponding to a measure approximates the precision-recall curve of the actual list, we can calculate how much information a measure contains about the actual list, and hence how "informative" a measure is. Thus, we have a methodology for evaluating the evaluation measures themselves.

Using the maximum entropy precision-recall curve of a measure, we can also predict the values of other measures. For example, using the maximum entropy precision-recall curve corresponding to average precision, we can predict the precision-at-cutoff 10. For highly informative measures, these predictions should be very close to reality. Hence, we have a second way of evaluating evaluation measures.

3. EXPERIMENTAL RESULTS

We tested the performance of the evaluation measures average precision, R-precision, and precision-at-cutoffs 5, 10, 15, 20, 30, 100, 200, 500 and 1000 using data from TRECs 3, 5, 6, 7, 8 and 9. For any TREC and any query, we chose those systems whose number of relevant documents retrieved was at least 10 in order to have a sufficient number of points on the precision-recall curve. We then calculated the maximum entropy precision-recall curve subject to the given measured constraint, as described above. The maximum entropy precision-recall curve corresponding to an average precision constraint cannot be determined analytically; therefore, we used numerical optimization[1] to find the maximum entropy distribution corresponding to average precision.

We shall refer to the execution of a retrieval system on a particular query as a *run*. Figure 6 shows examples of maximum entropy precision-recall curves corresponding to average precision, R-precision, and precision-at-cutoff 10 for three different runs, together with the actual precision-recall curves. We focused on these three measures since they are perhaps the most commonly cited measures in IR. We also provide results for precision-at-cutoff 100 in later plots and detailed results for all measures in a later table. As can be seen in Figure 6, using average precision as a constraint, one can generate the actual precision-recall curve of a run with relatively high accuracy.

In order to quantify how good an evaluation measure is in generating the precision-recall curve of an actual list, we consider two different error measures: the root mean squared error (RMS) and the mean absolute error (MAE). Let $\{\pi_1, \pi_2, \ldots, \pi_{R_{ret}}\}$ be the precisions at the recall levels $\{1/R, 2/R, \ldots, R_{ret}/R\}$ where R_{ret} is the number of relevant documents retrieved by a system and R is the number of documents relevant to the query, and let $\{m_1, m_2, \ldots, m_{R_{ret}}\}$ be the estimated precisions at the corresponding recall levels for a maximum entropy distribution corresponding to a measure. Then the MAE and RMS errors are calculated as follows.

$$RMS = \sqrt{\frac{1}{R_{ret}} \sum_{i=1}^{R_{ret}} (\pi_i - m_i)^2}$$

$$MAE = \frac{1}{R_{ret}} \sum_{i=1}^{R_{ret}} |\pi_i - m_i|$$

The points after recall R_{ret}/R on the precision-recall curve are not considered in the evaluation of the MAE and RMS errors since, by TREC convention, the precisions at these recall levels are assumed to be 0.

In order to evaluate how good a measure is at inferring actual precision-recall curves, we calculated the MAE and RMS errors of the maximum entropy precision-recall curves corresponding to the measures in question, averaged over all runs for each TREC. Figure 7 shows how the MAE and RMS errors for average precision, R-precision, precision-at-cutoff 10, and precision-at-cutoff 100 compare with each other for each TREC. The MAE and RMS errors follow the same pattern over all TRECs. Both errors are consistently and significantly lower for average precision than for the other measures in question, while the errors for R-precision are consistently lower than for precision-at-cutoffs 10 and 100.

Table 1 shows the actual values of the RMS errors for all measures over all TRECs. In our experiments, MAE and RMS errors follow a very similar pattern, and we therefore omit MAE results due to space considerations. From this table, it can be seen that average precision has consistently lower RMS errors when compared to the other measures. The penultimate column of the table shows the average RMS errors per measure averaged over all TRECs. On average, R-precision has the second lowest RMS error after average precision, and precision-at-cutoff 30 is the third best measure in terms of RMS error. The last column of the table shows the percent increase in the average RMS error of a measure when compared to the RMS error of average precision. As can be seen, the average RMS errors for the other measures are substantially greater than the average RMS error for average precision.

We now consider a second method for evaluating how informative a measure is. A highly informative measure should properly reduce one's uncertainty about the distribution of relevant and non-relevant documents in a list; thus, in our maximum entropy formulation, the probability-at-rank distribution should closely correspond to the pattern of relevant and non-relevant documents present in the list. One should then be able to accurately predict the values of other measures from this probability-at-rank distribution.

Given a probability-at-rank distribution $p_1, p_2, \ldots, p_N$, we can predict average precision, R-precision and precision-at-cutoff k values as follows:

- $ap = \frac{1}{R} \sum_{i=1}^{N} \left(\frac{p_i}{i} \left(1 + \sum_{j=1}^{i-1} p_j \right) \right)$
- $rp = \frac{1}{R} \sum_{i=1}^{R} p_i$
- $PC(k) = \frac{1}{k} \sum_{i=1}^{k} p_i$

The plots in the top row of Figures 8 and 9 show how average precision is actually correlated with R-precision, precision-at-cutoff 10, and precision-at-cutoff 100 for TRECs 6 and 8, respectively. Each point in the plot corresponds to a system and the values of the measures are averaged over all queries. Using these plots as a baseline for comparison, the plots in the bottom row of the figures show the correlation between the actual measures and the measures predicted using the average precision maximum entropy probability-at-rank distribution. Consider predicting precision-at-cutoff 10 values using the average precision maximum entropy distributions in TREC 6. Without applying the maximum entropy method, Figure 8 shows that the two measures are correlated with a Kendall's τ value of 0.671. However, the precision-at-cutoff 10 values inferred from the average precision maximum entropy distribution have a Kendall's τ value of 0.871 when compared to actual precisions-at-cutoff 10. Hence, the predicted precision-at-cutoff 10 and actual precision-at-cutoff 10 values are much more correlated than the actual average precision and actual precision-at-cutoff 10 values. Using a similar approach for predicting R-precision and precision-at-cutoff 100, it can be seen in Figures 8 and 9 that the measured values predicted by using average precision maximum entropy distributions are highly correlated with actual measured values.

We conducted similar experiments using the maximum entropy distributions corresponding to other measures, but since these measures are less informative, we obtained much smaller increases (and sometimes even decreases) in inferred correlations. (These results are omitted due to space considerations.) Table 2 summarizes the correlation improvements possible using the maximum entropy distribution corresponding to average precision. The row labeled τ_{act} gives the actual Kendall's τ correlation between average precision and the measure in the corresponding column. The row labeled τ_{inf} gives the Kendall's τ correlation between the

[1]We used the TOMLAB Optimization Environment for Matlab.

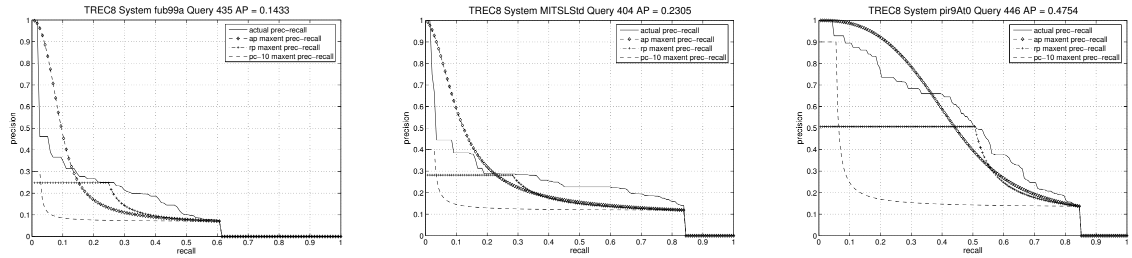

Figure 6: Inferred precision-recall curves and actual precision-recall curve for three runs in TREC8.

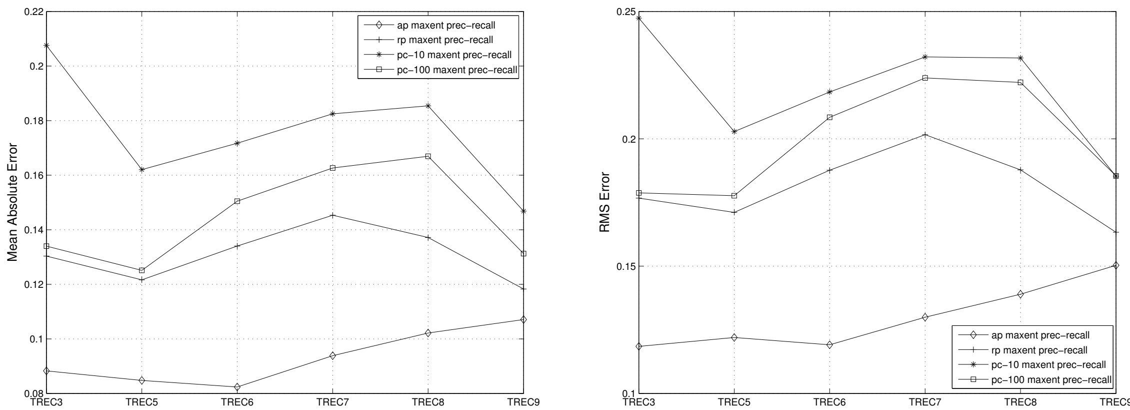

Figure 7: MAE and RMS errors for inferred precision-recall curves over all TRECs.

	TREC3	TREC5	TREC6	TREC7	TREC8	TREC9	AVERAGE	%INC
AP	0.1185	0.1220	0.1191	0.1299	0.1390	0.1505	0.1298	–
RP	0.1767	0.1711	0.1877	0.2016	0.1878	0.1630	0.1813	39.7
PC-5	0.2724	0.2242	0.2451	0.2639	0.2651	0.2029	0.2456	89.2
PC-10	0.2474	0.2029	0.2183	0.2321	0.2318	0.1851	0.2196	69.1
PC-15	0.2320	0.1890	0.2063	0.2132	0.2137	0.1747	0.2048	57.8
PC-20	0.2210	0.1806	0.2005	0.2020	0.2068	0.1701	0.1968	51.6
PC-30	0.2051	0.1711	0.1950	0.1946	0.2032	0.1694	0.1897	46.1
PC-100	0.1787	0.1777	0.2084	0.2239	0.2222	0.1849	0.1993	53.5
PC-200	0.1976	0.2053	0.2435	0.2576	0.2548	0.2057	0.2274	75.2
PC-500	0.2641	0.2488	0.2884	0.3042	0.3027	0.2400	0.2747	111.6
PC-1000	0.3164	0.2763	0.3134	0.3313	0.3323	0.2608	0.3051	135.0

Table 1: RMS error values for each TREC.

	TREC3			TREC5			TREC6		
	RP	PC-10	PC-100	RP	PC-10	PC-100	RP	PC-10	PC-100
τ_{act}	0.921	0.815	0.833	0.939	0.762	0.868	0.913	0.671	0.807
τ_{inf}	0.941	0.863	0.954	0.948	0.870	0.941	0.927	0.871	0.955
%Inc	2.2	5.9	14.5	1.0	14.2	8.4	1.5	29.8	18.3
	TREC7			TREC8			TREC9		
	RP	PC-10	PC-100	RP	PC-10	PC-100	RP	PC-10	PC-100
τ_{act}	0.917	0.745	0.891	0.925	0.818	0.873	0.903	0.622	0.836
τ_{inf}	0.934	0.877	0.926	0.932	0.859	0.944	0.908	0.757	0.881
%Inc	1.9	17.7	3.9	0.8	5.0	8.1	0.6	21.7	5.4

Table 2: Kendall's τ correlations and percent improvements for all TRECs.

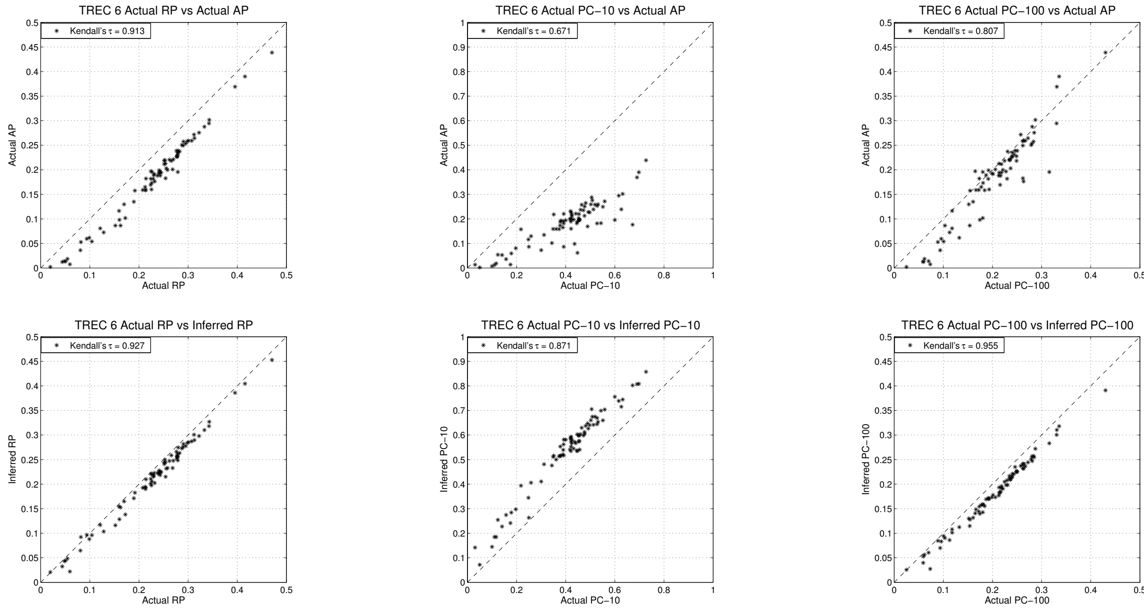

Figure 8: Correlation improvements, TREC6.

measure inferred from the maximum entropy distribution corresponding to average precision and the measure in the corresponding column. The row labeled %Inc gives the percent increase in correlation due to maximum entropy modeling. As can be seen, maximum entropy modeling yields great improvements in the predictions of precision-at-cutoff values. The improvements in predicting R-precision are noticeably smaller, though this is largely due to the fact that average precision and R-precision are quite correlated to begin with.

4. CONCLUSIONS AND FUTURE WORK

We have described a methodology for analyzing measures of retrieval performance based on the maximum entropy method, and we have demonstrated that the maximum entropy models corresponding to "good" measures of overall performance such as average precision accurately reflect underlying retrieval performance (as measured by precision-recall curves) and can be used to accurately predict the values of other measures of performance, well beyond the levels dictated by simple correlations.

The maximum entropy method can be used to analyze other measures of retrieval performance, and we are presently conducting such studies. More interestingly, the maximum entropy method could perhaps be used to help develop and gain insight into potential new measures of retrieval performance. Finally, the predictive quality of maximum entropy models corresponding to average precision suggest that if one were to estimate some measure of performance using an incomplete judgment set, that measure should be average precision—from the maximum entropy model corresponding to that measure alone, one could accurately infer other measures of performance.

Note that the concept of a "good" measure depends on the purpose of evaluation. In this paper, we evaluate measures based on how much information they provide about the overall performance of a system (a system-oriented evaluation). However, in different contexts, different measures may be more valuable and useful, such as precision-at-cutoff 10 in web search (a user-oriented evaluation). R-precision and average precision are system-oriented measures, whereas precision-at-cutoff k is typically a user-oriented measure. Another important conclusion of our work is that one can accurately infer user-oriented measures from system-oriented measures, but the opposite is not true.

Apart from evaluating the information captured by a single measure, we could use the MEM to evaluate the information contained in *combinations* of measures. How much does knowing the value of precision-at-cutoff 10 increase one's knowledge of a system's performance beyond simply knowing the system's average precision? Which is more informative: knowing R-precision and precision-at-cutoff 30, or knowing average precision and precision-at-cutoff 100? Such questions can be answered, in principle, using the MEM. Adding the values of one or more measures simply adds one or more constraints to the maximum entropy model, and one can then assess the informativeness of the combination. Note that TREC reports many different measures. Using the MEM, one might reasonably be able to conclude which are the most informative combinations of measures.

5. REFERENCES

[1] A. L. Berger, V. D. Pietra, and S. D. Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22:39–71, 1996.

[2] C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40. ACM Press, 2000.

[3] W. S. Cooper. On selecting a measure of retrieval effectiveness. part i. In *Readings in information retrieval*, pages 191–204. Morgan Kaufmann Publishers Inc., 1997.

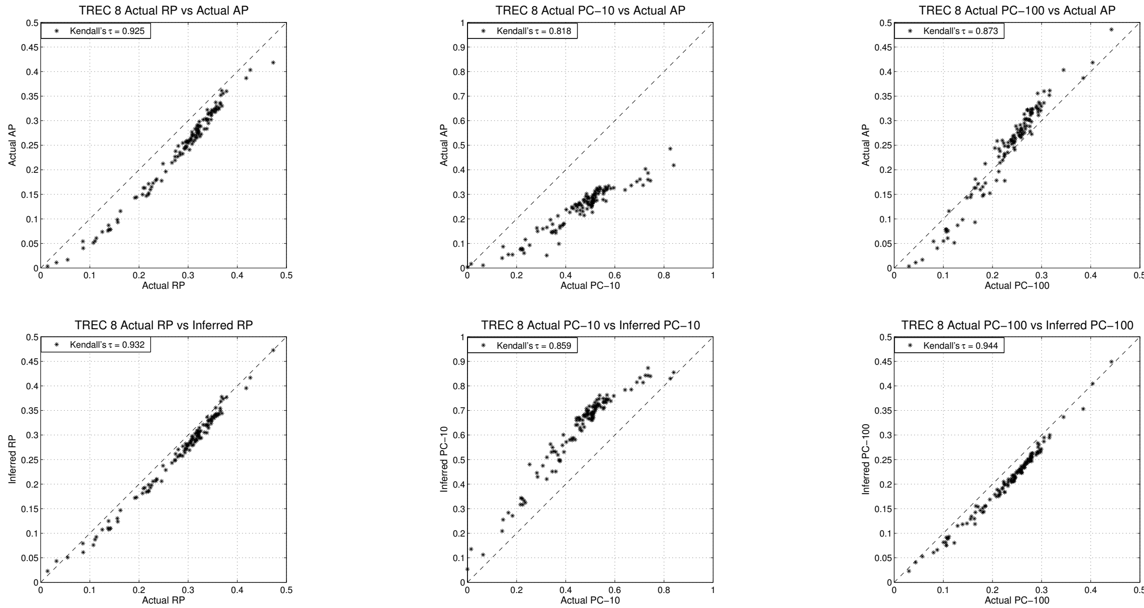

Figure 9: Correlation improvements, TREC8.

[4] T. M. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & sons, 1991.

[5] B. Dervin and M. S. Nilan. Information needs and use. In *Annual Review of Information Science and Technology*, volume 21, pages 3–33, 1986.

[6] W. R. Greiff and J. Ponte. The maximum entropy approach and probabilistic ir models. *ACM Trans. Inf. Syst.*, 18(3):246–287, 2000.

[7] E. Jaynes. On the rationale of maximum entropy methods. In *Proc.IEEE*, volume 70, pages 939–952, 1982.

[8] E. T. Jaynes. Information theory and statistical mechanics: Part i. *Physical Review 106*, pages 620–630, 1957a.

[9] E. T. Jaynes. Information theory and statistical mechanics: Part ii. *Physical Review 108*, page 171, 1957b.

[10] Y. Kagolovsky and J. R. Moehr. Current status of the evaluation of information retrieval. *J. Med. Syst.*, 27(5):409–424, 2003.

[11] P. B. Kantor and J. Lee. The maximum entropy principle in information retrieval. In *SIGIR '86: Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 269–274. ACM Press, 1986.

[12] D. D. Lewis. Evaluating and optimizing autonomous text classification systems. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 246–254. ACM Press, 1995.

[13] R. M. Losee. When information retrieval measures agree about the relative quality of document rankings. *J. Am. Soc. Inf. Sci.*, 51(9):834–840, 2000.

[14] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[15] D. Pavlov, A. Popescul, D. M. Pennock, and L. H. Ungar. Mixtures of conditional maximum entropy models. In T. Fawcett and N. Mishra, editors, *ICML*, pages 584–591. AAAI Press, 2003.

[16] S. J. Phillips, M. Dudik, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *ICML '04: Twenty-first international conference on Machine learning*, New York, NY, USA, 2004. ACM Press.

[17] V. Raghavan, P. Bollmann, and G. S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3):205–229, 1989.

[18] A. Ratnaparkhi and M. P. Marcus. Maximum entropy models for natural language ambiguity resolution, 1998.

[19] T. Saracevic. Evaluation of evaluation in information retrieval. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–146. ACM Press, 1995.

[20] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal 27*, pages 379–423 & 623–656, 1948.

[21] N. Wu. *The Maximum Entropy Method*. Springer, New York, 1997.