Evaluating and Optimizing Autonomous Text Classification Systems

David D. Lewis AT&T Bell Laboratories Murray Hill, NJ 07974; USA

lewis@research.att.com

(Appeared (w/ same pagination) in SIGIR 95: Proceedings of the Eighteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, July, 1995, pp. 246–254. One typo on p. 249 has been corrected.)

Abstract

Text retrieval systems typically produce a ranking of documents and let a user decide how far down that ranking to go. In contrast, programs that filter text streams, software that categorizes documents, agents which alert users, and many other IR systems must make decisions without human input or supervision. It is important to define what constitutes good effectiveness for these autonomous systems, tune the systems to achieve the highest possible effectiveness, and estimate how the effectiveness changes as new data is processed. We show how to do this for binary text classification systems, emphasizing that different goals for the system lead to different optimal behaviors. Optimizing and estimating effectiveness is greatly aided if classifiers that explicitly estimate the probability of class membership are used.

1 Introduction

Ranked retrieval is the information retrieval (IR) researcher's favorite tool for dealing with information overload. Ranked retrieval systems display documents in order of probability of relevance or some similar measure. Users see the best documents first, and decide how far down the ranking to go in examining the available information. The central role played by ranking in this approach has led researchers to evaluate IR systems primarily, often exclusively, on the quality of their rankings. (See, for instance, the TREC evaluations [1].)

In some IR applications, however, ranking is not enough:

- A company provides an SDI (selective dissemination of information) service which filters newswire feeds. Relevant articles are faxed each morning to clients. Interaction between customer and system takes place infrequently. The cost of resources (tying up phone lines, fax machine paper, etc.) is a factor to consider in operating the system.
- A text categorization system assigns controlled vocabulary categories to incoming documents as they are stored in a text database. Cost cutting has eliminated manual checking of category assignments.

Permission to make digital/hard copies of all or part of this material without fee is granted provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery, Inc. (ACM). To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires specific permission and/or fee. SIGIR '95 Seattle WA USA 1995 ACM 0-89791-714-6/95/07.\$3.50

 An "agent" program monitors low content text streams (e.g. Usenet newsgroups) and alerts a user when a relevant message appears. On most days, the user is not alerted and has no interaction with the agent.

A ranked retrieval system is a tool for people who are actively pursuing information. Implicit in its design is the assumption that a user wants to examine at least some of the available information. Many users, however, are actively avoiding information. They want to spend no time with, and have no awareness of, a particular information source unless highly relevant material becomes available. The decision of which items, if any, are grounds for disturbing the user becomes critical.

Increasingly, IR systems are autonomous in this fashion, making decisions without immediate human oversight or awareness. We propose the following method for constructing and operating autonomous IR systems:

- Define an effectiveness measure that, when applied to a set of decisions made by the system, computes a score indicating how good those decisions were.
- 2. Tune the system so that the expected effectiveness of its decisions will be the best possible.
- Have the system estimate the effectiveness of its decisions in an ongoing fashion, and notify an appropriate party when these estimates indicate a problem with handling new data.

This method would seem too obvious to present, except that few IR systems are currently constructed in this fashion. In this paper we present the details of this approach for a class of simple autonomous IR systems: systems that decide whether or not a document belongs to a single class. While simplistic, this binary classification approach covers a wide range of useful decisions: to fax an article or not, to assign a category or not, to alert a user or not, and so on.

Section 2 discusses how the effectiveness of a binary classifier can be measured when the correct classification for each document is known. Section 3 then defines the expected effectiveness of a binary classifier in an operational setting, where the correct classifications are not known. Section 4 discusses rules for making the expected effectiveness of a classifier the best possible, and points out that the proper rule will vary with the particular effectiveness measure chosen. This variation in classification rules is strikingly exhibited in Sections 5 to 7, which present three different effectiveness measures and the classification rules appropriate for each. The concluding sections discuss the implications of our analysis for classifier design, and point out some directions for future research.

	Expert Says Yes	Expert Says No	
System Says Yes	a	b	a+b=k
System Says No	c	d	c + d = n - k
	a+c=r	b+d=n-r	a+b+c+d=n

Figure 1: A set of n classification decisions can be represented as a contingency table.

2 Measuring Effectiveness

The simplest and most common classification task is a binary one, where a system must decide whether or not an item belongs to a single class. Assume that a set of n documents has been classified by a binary text classification system and, separately, by an expert who judges the true classification. We can summarize the relationship between the system classifications and the expert judgments in a contingency table (Figure 1). Each entry in the table specifies the number of documents with the specified outcome. For instance, a is the number of times the system decided Yes, and Yes was in fact the correct answer.

The effectiveness measures most widely used in IR are defined in terms of the contingency table:

- recall = a/(a+c)
- precision = a/(a+b)
- fallout = b/(b+d)

In words, recall is the proportion of class members that the system assigns to the class, and precision is the proportion of documents assigned to the class by the system that really are class members. An ideal system would have recall and precision of 1. Fallout is the proportion of nonclass members that the system assigns to the class, and is an alternative to precision. An ideal system would have fallout of 0.

Perfect recall can be achieved by a system that puts every document in the class, while perfect precision and fallout can be achieved by a system that puts no documents in the class, so using just one of these measures does not provide an interesting evaluation. One solution is to consider recall and precision, or recall and fallout, together and look at how the quantities trade off against one another under different parameter settings for the system. This is the usual approach in evaluating ranking systems, which do not have to commit to classification decisions.

Another approach is to define a single effectiveness measure that takes into account both errors of commission (b) and errors of omission (c). One such measure is error rate:

• error rate = (b+c)/(a+b+c+d)

A wide range of effectivenes measures have been defined. As early as 1973, Cooper commented that "too many ingenious and superficially plausible measures have been invented already" [2]. Our goal in this paper will be not to propose new measures, or even to recommend one measure over another. Instead we will consider several families of single effectiveness measures, examining how the performance of systems on each can be estimated and optimized.

3 Estimating Effectiveness

The simplest approach to estimating the future effectiveness of a classifier is to compare the classifier's decisions to an expert's decisions on a test set of documents, and assume that effectiveness on documents encountered in the future will stay the same. This is a questionable assumption when the classifier is applied to timevarying data such as news feeds or electronic mail. We would like to estimate the classifier's effectiveness on new data as it is encountered, without the expense of constantly acquiring new expert judgments.

Our key assumption is that an unknown expert judgment of class membership can be modeled by a Bernoulli (0/1) random variable, Z_i , with parameter p_i giving the probability that Z_i will take on the value 1 [3, p. 87]. The event $Z_i = 1$ occurs if the expert judges that document i belongs to the class, and $Z_i = 0$ occurs otherwise. We assume here and throughout the paper that each document is judged independently of all others.

Since the decisions of the classifier are under its control, they are represented by simple variables, s_i , rather than random variables. A value of $s_i = 1$ represents the classifier assigning document i to the class, while $s_i = 0$ represents non-assignment.

Under this model the contingency table entries are random variables A, B, C, and D defined in terms of the random judgment variables $\vec{Z} = Z_1, ..., Z_n$ and the system decision variables $\vec{s} = s_1, ..., s_n$ (Figure 2). Given an effectiveness measure g(a, b, c, d) defined in terms of the normal contingency table, we can define a corresponding function of random variables g(A, B, C, D). There is in turn an equivalent function $h(\vec{s}, \vec{Z})$ in terms of the system decision and the expert judgment variables.

Following Robertson [6], we define the *expected effectiveness* of a text classification system on a test set to be the expected value, $E[h(\vec{s}, \vec{Z})]$, of the effectiveness function. Using the definition of expected value [3, p. 69] we have:

$$E[h(\vec{s}, \vec{Z})] = \sum_{\vec{z} \in \{0,1\}^n} P(\vec{Z} = \vec{z}) h(\vec{s}, \vec{z})$$

$$= \sum_{\vec{z} \in \{0,1\}^n} (\prod_{i=1}^n p_i^{z_i} (1 - p_i)^{(1 - z_i)}) h(\vec{s}, \vec{z}) \quad (1)$$

In words, the expected effectiveness is the sum of the effectiveness values that would be obtained for each of the 2^n possible judgments, weighted by the probability of each judgment. For particular

¹The notion that class membership in general, and relevance to users in particular, has a probabilistic character is controversial [4], [5, pp. 127–129]. The reader who is disturbed by the notion is free to assume that p_i is an IR system's estimate of the probability of class membership [6], rather than an actual property of the document/class relationship.

	Expert Says Yes	Expert Says No	
System Says Yes	$A = \sum_{i=1}^{n} s_i Z_i$	$B = \sum_{i=1}^{n} s_i (1 - Z_i)$	k
System Says No	$C = \sum_{i=1}^{n} (1 - s_i) Z_i$	$D = \sum_{i=1}^{n} (1 - s_i)(1 - Z_i)$	n-k
	$R = \sum_{i=1}^{n} Z_i$	$n - R = \sum_{i=1}^{n} (1 - Z_i)$	n

Figure 2: When class labels for documents are not known, the contingency table entries can be viewed as random variables.

effectiveness measures $h(\vec{s}, \vec{Z})$ we can sometimes find simpler expressions for the expected effectiveness.

4 Optimizing Expected Effectiveness

Many text classifiers can be tuned to control their behavior. Once an effectiveness measure is defined, we want to tune the classifier such that the expected effectiveness will be the best possible. More formally, we want a policy that, when applied to a set of documents d_1, \ldots, d_n with corresponding judgment variables Z_1, \ldots, Z_n produces categorization decisions s_1, \ldots, s_n such that $\mathrm{E}[h(\vec{s}, \vec{Z})]$ is maximized or minimized as appropriate.

Different assignment policies will be necessary to optimize different effectiveness measures. A guideline that is often useful in devising assignment strategies was developed by Cooper in the context of text retrieval systems:

The Probability Ranking Principle: If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for this purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data [W. Cooper in [6]].

The PRP is a guideline, not a theorem, and may or may not hold for any particular effectiveness measure. Interestingly, while the PRP is stated in terms of rankings, it is usually applied in showing that a binary classification, not a ranking, is optimized when the binary classification is produced by taking the topmost items of a probability-based ranking. We offer the following variant of the PRP, which makes this notion explicit:

The Probability Ranking Principle for Binary Classifications: For a given set of items presented to a binary classification system, there exists a classification of the items such that the probability of class membership for all items assigned to the class is greater than or equal to the probability of class membership for all items not assigned to the class, and the classification has optimal expected effectiveness.

When the PRP holds for an effectiveness measure, finding an optimal classification is greatly simplified. The items may be sorted on the probability of class membership, p_i , and only the n+1 classifications consisting of the top k items for some $0 \le k \le n$

need be considered. This is a great improvement over considering all 2^n possible classifications.

The PRP does not, however, say how to choose the best of the n+1 valid classifications, nor does it say how to classify individual items in the absence of the entire set of items on which effectiveness will be measured. A strengthening of the PRP that addresses these limitations is:

The Probability Thresholding Principle for Binary Classifications: For a given effectiveness measure, there exists a threshold p, $0 \le p \le 1$, such that for any set of items, if all and only those items with probability of class membership greater than p are assigned to the class, the expected effectiveness of the classification will be the best possible for that set of items.

The PTP strengthens the PRP in two ways. It asserts that a classification decision can be made for each document individually while still optimizing effectiveness on whatever set of documents the system is evaluated on. Secondarily, it says that documents with the same probability of relevance can be treated identically.

When the PTP holds, optimizing a classifier's effectiveness is trivial. A threshold on probability of relevance can be set once, and the system is guaranteed to operate optimally in the future, no matter what the distribution of probabilities of relevance encountered is. (Of course, the system must be able to accurately estimate those probabilities of relevance.)

It is easy to see that if the PTP holds, then the PRP will also hold. The reverse is not necessarily true. In the next three sections we will present three families of effectiveness measures, two traditional and one novel but of obvious interest. One will satisfy both the PRP and the PTP. A second will satisfy the PRP only. The third will also satisfy the PRP only, but will actually best be optimized by not doing binary classification at all.

5 Decision-Theoretic Loss Measures

One approach to measuring effectiveness comes from assigning a numeric penalty or *loss* [7, p. 14] to each of the four possible contingency table outcomes. A loss may be positive, indicating that the outcome is undesirable, or negative, indicating that the outcome is desirable (since a negative loss is a gain). Cooper was an early proponent of this approach to measuring the effectiveness of IR systems [2].

Let λ_{ij} be the loss associated with making decision i when the correct decision is j, where decision 1 is putting the document in the class, and decision 2 is not putting the document in the class. We can define an effectiveness measure, L_{λ} , where $\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})$, corresponding to the mean loss incurred by the classifier on a set of

n items:

$$L_{\lambda} = (\lambda_{11}a + \lambda_{12}b + \lambda_{21}c + \lambda_{22}d)/n \tag{2}$$

As discussed earlier, we can define a corresponding function of the random variables for the class membership judgments:

$$\begin{split} L_{\lambda}(\vec{s}, \vec{Z}) &= (\lambda_{11}A + \lambda_{12}B + \lambda_{21}C + \lambda_{22}D)/n \\ &= \frac{1}{n} \sum_{i=1}^{n} \lambda_{11}s_{i}Z_{i} + \lambda_{12}s_{i}(1 - Z_{i}) \\ &+ \lambda_{21}(1 - s_{i})Z_{i} + \lambda_{22}(1 - s_{i})(1 - Z_{i}) \\ &= \frac{1}{n} \sum_{i=1}^{n} (\lambda_{m}s_{i} + (\lambda_{21} - \lambda_{22}))Z_{i} + (\lambda_{12} - \lambda_{22})s_{i} + \lambda_{22}(3) \end{split}$$

where $\lambda_m = \lambda_{11} - \lambda_{12} - \lambda_{21} + \lambda_{22}$. The expected value of the effectiveness function, $E[L_{\lambda}(\vec{s}, \vec{Z})]$, equals:

$$\frac{1}{n} \sum_{i=1}^{n} (\lambda_{m} s_{i} + (\lambda_{21} - \lambda_{22})) p_{i} + (\lambda_{12} - \lambda_{22}) s_{i} + \lambda_{22}$$
 (4)

Since the Z_i are assumed to be independent, then by properties of the variance of random variables we also have that the variance, $\text{var}[L_{\lambda}(\vec{s}, \vec{Z})]$, of the effectiveness function equals:

$$\frac{1}{n^2} \sum_{i=1}^{n} (\lambda_m s_i + (\lambda_{21} - \lambda_{22}))^2 p_i (1 - p_i)$$
 (5)

If the p_i are not too close to 0 or 1, and $k = \sum_{i=1}^n s_i$ is relatively large, then $L_{\lambda}(\vec{s}, \vec{Z})$ will have an approximately normal distribution, giving the following 95% confidence interval for expected effectiveness:

$$E[L_{\lambda}(\vec{s}, \vec{Z})] \pm 1.96 \sqrt{\text{var}[L_{\lambda}(\vec{s}, \vec{Z})]}$$
 (6)

Turning to optimization, it is easy to show [7, p. 15] that $E[L_{\lambda}(\vec{s}, \vec{Z})]$ is minimized if s_i is 1 exactly when

$$p_i > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11}) + (\lambda_{12} - \lambda_{22})}. (7)$$

Since expected loss can be minimized by comparing the probability of class membership of each test item to a threshold based solely on the effectiveness measure being used, the PTP applies, and thus the PRP as well. The threshold value depends on λ , which in turn depends on what users want from the system. There is no single threshold that is best for all purposes.

When $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$, L_{λ} is just error rate, and we want

$$p_i > \frac{1}{1+1} = 0.5 \ . \tag{8}$$

So, unsurprisingly, the number of errors a classifier makes is minimized by assigning to the class exactly those documents with a better than even chance of class membership.

As another example, Cooper [2, 8] and McCarn [9, 10] have studied loss-based effectiveness measures that take into account only retrieved documents. In this case $\lambda_{21} = \lambda_{22} = 0$, and we have

$$p_i > \frac{\lambda_{12}}{\lambda_{12} - \lambda_{11}} \tag{9}$$

Test set p_i 's: 0.4 0.4 0.2				
	Assigned	Exact $E[F_1]$	Approx. $E[F_1]$	
	none ($C_F = 0$)	0.0000	0.0000 ± 0	
	none ($C_F = 1$)	0.2880	0.2880 ± 0	
	0.4	0.3253	0.4000 ± 0.3560	
	0.4 0.4*	0.4576	0.5333 ± 0.1943	
	0.4 0.4 0.2	0.4392	0.5000 ± 0.1200	
Test set p _i 's: 0.9 0.9 0.4				
	Assigned	Exact $E[F_1]$	Approx. $E[F_1]$	
	none ($C_F = 0$)	0.0000	0.0000 ± 0	
	none ($C_F = 1$)	0.0060	0.0060 ± 0	
	0.9	0.5640	0.5625 ± 0.0790	
	0.9 0.9*	0.8532	0.8571 ± 0.0550	
	0.9 0.9 0.4	0.8264	0.8462 ± 0.0466	

Figure 3: Exact and approximate expected values of Van Rijsbergen's F_1 measure for two small test sets, and several choices of which documents to assign to the class. The optimal assignments are *'ed

01

$$p_i \lambda_{11} + (1 - p_i) \lambda_{12} < 0. (10)$$

In other words, when only retrieved documents are evaluated, a document should be assigned to the class exactly when the expected loss for doing so is negative, i.e. some gain is expected.

It is important to point out (as did Cooper in his discussion of "Objection 10" [2]) that adding up losses associated with individual documents is only the most simplistic application of decision-theoretic ideas to IR system evaluation. A simple addition suggested by McCarn [9, 10] is to associate a loss with using the system at all, in addition to the losses associated with individual documents. This is plausible, for instance, in an alerting application. Much more sophisticated modifications are possible, for instance associating diminishing gains with each additional relevant document.

6 Van Rijsbergen's E and F Measures

Another family of effectiveness measures, satisfying certain measurement theoretic properties, was defined by Van Rijsbergen [5, pp. 168–176]. The *E-measure* combines recall and precision into a single score:

$$E_{\beta} = 1 - \frac{1}{\alpha(\frac{1}{P}) + (1 - \alpha)\frac{1}{R}}$$

$$= 1 - \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$
(11)

Here R is recall, P is precision, $\alpha=1/(\beta^2+1)$, and β ranges from 0 to infinity. Lower values of E correspond to higher effectiveness, so the F-measure

$$F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{12}$$

is often used instead. F_0 is the same as precision, F_∞ is recall, and values of β between 0 and ∞ give varying weights to recall and precision. When the E or F measures are used in experimental IR, results are often presented for β values of 0.5 (recall half as important as precision), 1.0 (equal weight on recall and precision), and 2.0 (recall twice as important as precision).

Replacing R and P with their contingency table definitions gives:

$$F_{\beta} = \frac{(\beta^2 + 1)a}{(\beta^2 + 1)a + b + \beta^2 c}$$
 (13)

but makes clear that the case a=b=c=0 must be handled specially. Van Rijsbergen did not define what value his measure should have when a=b=c=0, but if we are to retain the measure's lack of dependence on cell d (inherited from recall and precision), then F_{β} in this case should be a constant value, C_F . We therefore define F_{β} to be:

$$F_{\beta} = \begin{cases} C_F & \text{if } a+b=0 \text{ and } a+c=0\\ 0 & \text{if } a+b=0 \text{ and } a+c>0\\ \frac{(\beta^2+1)a}{(\beta^2+1)a+b+\beta^2c} & \text{otherwise} \end{cases}$$
(14)

where the second case could be combined with the third, but separating them will clarify our analysis. As with the loss measure, we can define a corresponding function of \vec{s} and \vec{Z} :

$$F_{\beta}(\vec{s}, \vec{Z}) = \begin{cases} C_F & \text{if } \vec{s} = \vec{0} \text{ and } \vec{Z} = \vec{0} \\ 0 & \text{if } \vec{s} = \vec{0} \text{ and } \vec{Z} \neq \vec{0} \\ \frac{(\beta^2 + 1) \sum_{i=1}^n s_i Z_i}{\sum_{i=1}^n s_i + \beta^2 Z_i} & \text{otherwise} \end{cases}$$
(15)

In contrast to loss, the expected value, $\mathrm{E}[F_{\beta}(\vec{s},\vec{Z})]$ does not have a simple form. When n is small (up to perhaps 20), it is practical to compute $\mathrm{E}[F_{\beta}(\vec{s},\vec{Z})]$ directly using the definition of expected value (Equation 1):

$$\sum_{\substack{z \in \{0,1\}^n \\ \vec{z} \neq \vec{0}}} P(\vec{Z} = \vec{z}) \frac{(\beta^2 + 1) \sum_{i=1}^n s_i z_i}{\sum_{i=1}^n s_i + \beta^2 z_i} \quad \text{if } \vec{s} \neq \vec{0}$$

$$P(\vec{Z} = \vec{0}) C_F \qquad \qquad \text{if } \vec{s} = \vec{0}$$
(16)

For larger n, only approximations will be practical to use. The simplest one:

$$E[F_{\beta}(\vec{s}, \vec{Z})] \approx \begin{cases} C_F \prod_{i=1}^{n} (1 - p_i) & \text{if } \vec{s} = \vec{0} \\ \frac{(\beta^2 + 1) \sum_{i=1}^{n} s_i p_i}{\sum_{i=1}^{n} s_i + \beta^2 p_i} & \text{otherwise} \end{cases}$$
(17)

is exact for $\vec{s} = \vec{0}$ and has a maximum error of

$$\frac{1}{\mathrm{E}[Y]^2} \left(K \mathrm{var}[Y] + \sqrt{\mathrm{var}[X] \mathrm{var}[Y]} \right) \tag{18}$$

when $\vec{s} \neq \vec{0}$. (The relevant quantities are specified in Appendix A.) As to the value of C_F , the score a system gets for assigning no documents when no documents are relevant, cases can be made for setting it to 0, 1, or some value in between. We prefer the value $C_F = 1$, on the reasoning that a system that assigns no documents to the class when there are in fact no class members is operating perfectly.

Figure 3 compares the approximation to $\mathrm{E}[F_{eta}(\vec{s},\vec{Z})]$ from Equation 17 with the exact value from Equation 16, for various classifications \vec{s} of two small hypothetical test sets. The values are similar, though not identical. Note that the exact value is always within the error bounds on the approximate value. (Regrettably, the bounds appear quite loose when $\sum_{i=1}^n s_i$ and $\sum_{i=1}^n p_i$ are small.) In Figure 3 we presented data only for classifications \vec{s} that

In Figure 3 we presented data only for classifications \vec{s} that satisifed the Probability Ranking Principle. Appendix B proves that this suffices—the classification maximizing the expected value of F_{β} for a particular testset will always satisfy the PRP.

On the other hand, F_{β} does *not* satisfy the Probability Thresholding Principle. Figure 3 shows the set of p_i values for two small test sets. For the first set, where no documents have a high probability of class membership, documents with p_i 's of 0.4 should be assigned

Test set p_i 's: 0.5 0.5			
	Assigned	Exact E[SEC]	
	none	1.5	
	0.5	0.5	
	0.5 0.5	1.5	
Test set p_i 's: 0.7 0.6			
	Assigned	Exact E[SEC]	
	none	2.14	
	0.7	0.54	
	0.7 0.6	0.94	
Test set p_i 's: 0.5 0.3			
	Assigned	Exact E[SEC]	
	none	1.1	
	0.5	0.5	
	0.5 0.3	1.9	

Figure 4: Expected values of SEC (squared error in counting) for various classifications of a small test set.

to the class to optimize $\mathrm{E}[F_1(\vec{s}, \vec{Z})]$. For the second set, where there are documents with a very high probability of class membership, documents with a p_i of 0.4 should not be put in the class. Yet the effectiveness measure F_1 is identical in the two cases—only the test data has changed. This shows that no fixed threshold can be used to optimize F_β if the distribution of p_i 's is not known.

How then should one set up a text classification system to optimize F_{β} ? If it is possible to classify documents in batches, then explicitly estimating $\mathrm{E}[F_{\beta}(\vec{s},\vec{Z})]$, and choosing the classification that maximizes this value, should give the highest effectiveness. If documents cannot be batched, but must be classified as they are received, then classifying each document taking into account the probabilities of the last n documents received, for some reasonable value of n, should also give good performance. If the classifier used does not estimate probabilities, then the best that can be done is to tune the classify to optimize the observed $F_{\beta}(\vec{s}, \vec{z})$ on some test set, and hope that data encountered in the future will have similar properties.

These complexities are not unique to F_{β} . How a user wants an IR system to treat documents with a 40% chance of being relevant often does depend on whether or not documents with higher probabilities of relevance are available. Any effectiveness measure that captures this contextual aspect of user preferences (including, for instance, decision theoretic approaches more sophisticated than that of the previous section) will present the same difficulties as F_{β} .

7 Effectiveness of Counting

Suppose one wants to know how many documents from a set belong to a class. An obvious strategy is to build a classifier, run it on the set of documents, and count how many times the class is assigned. But how should such a classifier be tuned?

From the contingency table (Figure 1) we see that the true number of class members is a+c while the number of documents assigned to the class is a+b. A plausible effectiveness measure to minimize is the *squared error in counting (SEC)*, that is the square of the difference between the true number of class members and the assigned number of class members:

SEC =
$$(\sum_{i=1}^{n} z_i - \sum_{i=1}^{n} s_i)^2$$

= $((a+c) - (a+b))^2$
= $(b-c)^2$ (19)

Minimizing SEC minimizes the difference between the number

of documents belonging to the class and the number of documents assigned to the class. The corresponding function of \vec{s} and \vec{Z} is:

$$SEC(\vec{s}, \vec{Z}) = ((A+C) - (A+B))^{2}$$
$$= (\sum_{i=1}^{n} Z_{i} - \sum_{i=1}^{n} s_{i})^{2}$$
(20)

Figure 4 shows that, like F_{β} , SEC does not satisfy the PTP. The first test set shows that items with the same p_i should sometimes not be classified identically, violating the PTP. In addition, the best possible threshold for the second test set lies between 0.6 and 0.7, while for the third test set it lies between 0.3 and 0.5, another violation of the PTP.

SEC does satisfy the PRP, but only because all assignments with the same value of $k = \sum_{i=1}^n s_i$ have the same SEC. Since only the number of documents assigned to the class, not which documents are assigned to the class, matters, we can just as well define a variant of SEC to take a real number e instead of a classifiction \vec{s} as its first argument:

$$sec(e, \vec{Z}) = ((\sum_{i=1}^{n} Z_i) - e)^2$$
 (21)

The expected value of this is

$$E[sec(e, \vec{Z})] = E[((\sum_{i=1}^{n} Z_i) - e)^2]$$
 (22)

which is minimized when [11, p. 107]:

$$e = E[\sum_{i=1}^{n} Z_i] = \sum_{i=1}^{n} p_i$$
 (23)

Note that $\sum_{i=1}^{n} p_i$ need not be an integer, while $\sum_{i=1}^{n} s_i$ must be. So not only is it unnecessary to actually assign items to a class in order to count them, it may actually be impossible to produce an optimal estimate in that fashion.

When $e=\mathrm{E}[\sum_{i=1}^n Z_i]$, $sec(e,\vec{Z})$ is simply the variance of $\sum_{i=1}^n Z_i$:

$$sec(E[\sum_{i=1}^{n} Z_i], \vec{Z}) = var[\sum_{i=1}^{n} Z_i] = \sum_{i=1}^{n} p_i(1 - p_i)$$
 (24)

When n is large, and the p_i are not too near 0 or 1, $\sum_{i=1}^{n} Z_i$ will be approximately normally distributed, and a 95% confidence interval for the number of category members will be:

$$\sum_{i=1}^{n} p_i \pm 1.96 \sqrt{\sum_{i=1}^{n} p_i (1 - p_i)}$$
 (25)

To summarize, if our goal is to count class members, and if we have estimates of the probability of class membership, we should use the estimates directly to estimate the number of class members, rather than use them to classify documents. Of course, our estimates of the p_i 's will have some bias and variance themselves, so the confidence interval given in Equation 25 will be somewhat optimistic.

Our approach to counting is similar in spirit to that of Thomas and colleagues [12]. They estimate p = (A + C)/n by stratified sampling from a large set of test documents, with strata defined by word combinations. A small sample of documents is judged and p is estimated, using a variety of sophisticated methods, without classifying the remainder of the test set.

SEC can be contrasted with the average squared error measure used by Fuhr, Hüther, and Pfeifer [13, 14]:

$$s^{2} = \frac{1}{n} \sum_{i=1}^{n} (z_{i} - p_{i})^{2}.$$
 (26)

While also using a notion of squared error, this is a measure of the quality of the probability estimates p_i , rather than a measure of the effectiveness of binary classifications, as is SEC, or of an estimated count, as is SEC.

8 Discussion

Text classifiers are increasingly used in an autonomous fashion, but rarely in an optimal fashion. Whether built by hand or by automated training, they are often not evaluated. When evaluated, the effectiveness measure may not be clearly defined or optimized. In other cases, the effectiveness measure may not be appropriate, as when ranking-oriented measures (e.g. recall-precision curves), are used to evaluate autonomous text categorization systems. Even when an appropriate effectiveness measure is defined and optimized, this is typically done by tuning to a particular data set, without recognition that effectiveness, and the appropriate setting for optimizing effectiveness, may vary over time².

These problems can be avoided by the strategy presented in the introduction and elaborated in the paper: define an effectiveness measure, compute its expected value as a function of classification decisions and probability of class membership, and choose a classification that optimizes that expected value. As a side effect, the system can constantly monitor its expected effectiveness, so that action can be taken when effectiveness changes. While we have presented this as a strategy for building autonomous text classification systems, it of course will be useful even when there is human checking of the system's decisions.

Our approach relies on classifiers that produce accurate estimates of the probability of class membership. While there has been recent progress in producing probabilistic text classifiers [14, 15, 16], more work is needed. We also need a better understanding of the bias and variance of the estimates produced by these classifiers, and of how these impact, for instance, confidence intervals on effectiveness measures.

All current IR methods that produce reasonable probability estimates rely on relevance feedback or other procedures for training on judged data. This does not seem an undue burden when setting up a system that will act autonomously over a long period of time. The amount of judged training data necessary to produce probabilistic classifiers can be reduced by active learning [15], though this complicates estimating the bias and variance of the resulting probability estimates.

9 Future Work

Our approach to estimating and optimizing the effectiveness of autonomous text classification systems raises new questions. In particular, the fact that the F-measure can not be optimized without knowing the context in which a document is classified suggests several directions for research.

First, we need more research into which effectiveness measures best capture what users want autonomous classifiers to do. In a rare study of this sort, McCarn [9, 10], analyzing data of Pollitt [17] on searches of bibliographic databases, found that a loss-based effectiveness measure was highly predictive of the amount of money a user stated they would be willing to pay for the search result. This is encouraging, since loss-based measures are easy to estimate and optimize. On the other hand, the wide use of precision and recall in both the research and operational IR communities, and the almost complete avoidance of loss-based measures, suggest that loss is unlikely to be a complete solution.

If user needs are best captured by effectiveness measures that can only be optimized in context, then in setting up a system we need to consider over what interval effectiveness will be measured.

 $^{^2}$ Many of these problems can be found in the author's own past work. For instance, Lewis and Gale [15] categorized documents so as to minimize $\mathrm{E}[L_{0,1,1,0}]$ (expected error rate), but then evaluated those categorizations using F_1 .

Do users want effectiveness to be optimized on a daily, weekly, or some other basis, and how do we determine this?

Earlier we presented possible approaches to optimizing contextsensitive effectiveness measures such as F_{β} . Empirical research on the actual distribution of p_i values will determine how carefully this must be done. Other approaches are possible, for instance algorithms that approximately estimate and optimize effectiveness without knowing the full context, by modeling how past contexts have changed over time. Another avenue of investigation is algorithms that are allowed to delay their decisions, but that can only save a limited fraction of the documents encountered before making a decision. What is the relationship between the size of this buffer and the ability to approximate optimal effectiveness?

It should also be possible to generalize the approach taken in this paper to choices among more than two alternatives, and to measures of average effectiveness across multiple classification decisions.

10 Conclusion

If we are to trust an IR system to classify documents with little or no human oversight, we want the system to operate as well as possible, and we want to know how well it is operating. We propose that this be done by explicitly choosing an effectiveness measure, computing the expected value of this measure as a function of the classification decisions and probability of relevance of the documents, and making those classification decisions that optimize expected effectiveness. This should be done in an ongoing fashion, with the system adapting to and monitoring changes in the data encountered.

Optimizing one effectiveness measure will not in general optimize others, so there is no single best system for any text classification task. The choice among plausible alternatives should be made with clear goals in mind.

11 Acknowledgements

This work has benefited from conversations with many colleagues, including Jason Catlett, Ken Church, William Cohen, Norbert Fuhr, Yoav Freund, Bill Gale, David Hull, David Ittner, Mark Jones, Paul Kantor, Michael Kearns, Doug McIlroy, Rich Pennenga, Larry Rafsky, Rob Schapire, and Chris Watkins. The comments of the anonymous referees were also of great help, as were many energetic and enlightening discussions with members of the TREC-1 through TREC-4 program committees, with particular thanks to Chris Buckley, Sue Dumais, Donna Harman, and Steve Robertson.

References

- D. K. Harman, editor. The Second Text REtrieval Conference (TREC-2), Gaithersburg, MD 20899, 1994. National Institute of Standards and Technology. Special Publication 500-215.
- [2] William S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24:87–100, March–April 1973.
- [3] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. Introduction to the Theory of Statistics. Mcgraw-Hill, New York, 3rd edition, 1974.
- [4] Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of* the American Society for Information Science, pages 321–343, November–December 1975.

- [5] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.
- [6] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, December 1977.
- [7] Richard O. Duda and Peter E. Hart. Pattern Classification and Scene Analysis. Wiley-Interscience, New York, 1973.
- [8] William S. Cooper. On selecting a measure of retrieval effectiveness. Part II. Implementation of a philosophy. *Journal of the American Society for Information Science*, 24:413–424, November–December 1973.
- [9] Davis B. McCarn. Online systems—techniques and services. In Martha E. Williams, editor, *Annual Review of Information Science and Technology*, Vol. 13, pages 85–124. Knowledge Industry Publications, Inc., 1978.
- [10] Davis B. McCarn and Craig M. Lewis. A mathematical model of retrieval system performance. *Journal of the American Society for Information Science*, 41(7):495–500, October 1990.
- [11] Bernard W. Lindgren. Statistical Theory. Chapman & Hall, New York, 4th edition, 1993.
- [12] T. Thomas, C. Kruger, C. Scovel, and J. Shumate. Text to information: Sampling uncertainty in an example from physician/patient encounters. In *Symposium on Document Analysis* and Information Retrieval, 1995. To appear.
- [13] Norbert Fuhr and Hubert Hüther. Optimum probability estimation from empirical distributions. *Information Processing and Management*, pages 493–507, 1989.
- [14] Norbert Fuhr and Ulrich Pfeifer. Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumptions. ACM Transactions on Information Systems, 12(1):92–115, January 1994.
- [15] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In W. Bruce Croft and C. J. van Rijsbergen, editors, SIGIR 94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 3–12, London, 1994. Springer-Verlag.
- [16] Wm. S. Cooper, Aitao Chen, and Frederic C. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text Retrieval Conference (TREC-2)*, pages 57–65, Gaithersburg, MD, 1994. National Institute of Standards and Technology. NIST Special Publication 500-215.
- [17] Arthur S. Pollitt. CANCERLINE evaluation project: Final report. Technical report, Medical Library, School of Medicine, The University of Leeds, Leeds, England, 1977.
- [18] Richard von Mises. Mathematical Theory of Probability and Statistics. Academic Press, New York, 1964.

A Error Bounds on An Approximation to $E[F_{\beta}]$

We want to approximate $E[F_{\beta}(\vec{s}, \vec{Z})]$ where :

$$F_{\beta}(\vec{s},\vec{Z}) = \left\{ \begin{array}{ll} C_F & \text{if } \vec{s} = \vec{0} \text{ and } \vec{Z} = \vec{0} \\ 0 & \text{if } \vec{s} = \vec{0} \text{ and } \vec{Z} \neq \vec{0} \\ \frac{(\beta^2 + 1) \sum_{i=1}^n s_i Z_i}{\sum_{i=1}^n s_i + \beta^2 Z_i} & \text{otherwise} \end{array} \right.$$

When $\vec{s} = \vec{0}$, $E[F_{\beta}(\vec{s}, \vec{Z})] = C_F P(\vec{Z} = \vec{0})$. When $\vec{s} \neq \vec{0}$, we want to know

$$E\left[\frac{X}{Y}\right] = E\left[\frac{(\beta^2 + 1)\sum_{i=1}^{n} s_i Z_i}{\sum_{i=1}^{n} s_i + \beta^2 Z_i}\right]$$
(27)

We suggest approximating E[X/Y] by E[X]/E[Y]:

$$E[\frac{X}{Y}] \approx \frac{E[X]}{E[Y]} = \frac{(\beta^2 + 1) \sum_{i=1}^{n} s_i p_i}{\sum_{i=1}^{n} s_i + \beta^2 p_i}$$
(28)

Under the assumptions that $E[Y] \neq 0$ and that the maximum value taken on by |X/Y| is less than or equal to K, a general formula for the error bounds on an approximation of E[X/Y] by E[X]/E[Y] is [18, pp. 400–401]:

$$\left| \operatorname{E}\left[\frac{X}{Y}\right] - \frac{\operatorname{E}[X]}{\operatorname{E}[Y]} \right| \le \frac{1}{\operatorname{E}[Y]^2} \left(K \operatorname{var}[Y] + \sqrt{\operatorname{var}[X] \operatorname{var}[Y]} \right) \tag{29}$$

The necessary quantities are easy to derive:

$$X = (\beta^{2} + 1) \sum_{i=1}^{n} s_{i} Z_{i}$$

$$E[X] = (\beta^{2} + 1) \sum_{i=1}^{n} s_{i} p_{i}$$

$$var[X] = (\beta^{2} + 1)^{2} \sum_{i=1}^{n} s_{i} p_{i} (1 - p_{i})$$

$$Y = \sum_{i=1}^{n} s_{i} + \beta^{2} Z_{i}$$

$$E[Y] = \sum_{i=1}^{n} s_{i} + \beta^{2} p_{i}$$

$$var[Y] = \beta^{4} \sum_{i=1}^{n} p_{i} (1 - p_{i})$$

Finally, $|X/Y| \le 1$ for all possible values of X and Y (given $\vec{s} \ne \vec{0}$), so we can set K to 1. So the maximum error of the approximation is:

$$\frac{1}{(\sum_{i=1}^{n} s_i + \beta^2 p_i)^2} \beta^2 \left((\beta^2 \sum_{i=1}^{n} p_i (1-p_i)) + (\beta^2 + 1) \sqrt{(\sum_{i=1}^{n} s_i p_i (1-p_i))(\sum_{i=1}^{n} p_i (1-p_i))} \right)$$

It is possible that a better approximation and/or better error bounds for this approximation can be found.

B F_{β} Satisfies the PRP

In the following we state formally and prove that the Probability Ranking Principle applies to Van Rijsbergen's F-measure.

Theorem: Let $Z_1, ... Z_n$ be independent Bernoulli variables with parameters $p_1, ... p_n$ respectively. Let β be a real number greater than or equal to 0. Assume $\vec{s} \in \{0,1\}^n$ is such that

$$E[F_{\beta}(\vec{s}, \vec{Z})] \ge E[F_{\beta}(\vec{t}, \vec{Z})] \text{ for all } \vec{t} \in \{0, 1\}^n$$
 (30)

Then for all $1 \le j, k \le n$, if $s_j = 1$ and $s_k = 0$, then $p_j \ge p_k$.

Proof: Suppose false and let $\vec{s} \in \{0,1\}^n$ be such that $E[F_{\beta}(\vec{s},\vec{Z})]$ takes on its maximum, $s_j = 1$, $s_k = 0$, and $p_j < p_k$. Let $\vec{s'}$ be such that $s'_j = 0$, $s'_k = 1$, and $s'_i = s_i$ for all other s'_i . Then by the optimality of \vec{s} :

$$E[F_{\beta}(\vec{s}, \vec{Z})] - E[F_{\beta}(\vec{s'}, \vec{Z})] \ge 0$$

$$\left(\sum_{\vec{z}\in\{0,1\}^n} P(\vec{Z}=\vec{z}) \frac{(\beta^2+1)\sum_{i=1}^n s_i z_i}{\sum_{i=1}^n s_i + \beta^2 z_i}\right) - \left(\sum_{\vec{z}\in\{0,1\}^n} P(\vec{Z}=\vec{z}) \frac{(\beta^2+1)\sum_{i=1}^n s_i' z_i}{\sum_{i=1}^n s_i' + \beta^2 z_i}\right) \ge 0$$

$$\sum_{\vec{z} \in \{0,1\}^n} \ P(\vec{Z} = \vec{z}) \left(\frac{(\beta^2 + 1) \sum_{i=1}^n s_i z_i}{\sum_{i=1}^n s_i + \beta^2 z_i} - \frac{(\beta^2 + 1) \sum_{i=1}^n s_i' z_i}{\sum_{i=1}^n s_i' + \beta^2 z_i} \right) \ge 0$$

where the substitutions for $\mathrm{E}[F_{\beta}(\vec{s},\vec{Z})]$ and $\mathrm{E}[F_{\beta}(\vec{s'},\vec{Z})]$ are justified since $\vec{s} \neq \vec{0}$ and $\vec{s'} \neq \vec{0}$. For all $\vec{z} \in \{0,1\}^n$ the denominators of the two terms are identical and greater than 0, so we can drop them:

$$\sum_{\vec{z} \in \{0,1\}^n} P(\vec{Z} = \vec{z}) \left((\beta^2 + 1) \sum_{i=1}^n s_i z_i - (\beta^2 + 1) \sum_{i=1}^n s_i' z_i \right) \ge 0$$

$$\sum_{\vec{z} \in \{0,1\}^n} P(\vec{Z} = \vec{z}) \sum_{i=1}^n (s_i - s_i') z_i \ge 0$$

Since $s_i = s'_i$ except when i = j or i = k, we have:

$$\sum_{\vec{z} \in \{0,1\}^n} P(\vec{Z} = \vec{z}) \left((s_j - s_j') z_j + (s_k - s_k') z_k \right) \ge 0$$

$$\sum_{\vec{z} \in \{0,1\}^n} P(\vec{Z} = \vec{z})(z_j - z_k) \ge 0$$

This can be rewritten as

$$\sum_{\vec{z} \in \{0,1\}^n} \left(\prod_{i=1}^n p_i^{z_i} (1-p_i)^{(1-z_i)} \right) (z_j - z_k) \ge 0$$

$$\sum_{\substack{w_{jk}^{z} \in \{0,1\}^{n-2} \\ i \neq j,k}} \left(\prod_{\substack{i=1 \\ i \neq j,k}}^{n} p_{i}^{z_{i}} (1-p_{i})^{(1-z_{i})} \right) \sum_{z_{j}=0}^{1} \sum_{z_{k}=0}^{1} \left(p_{j}^{z_{j}} (1-p_{j})^{(1-z_{j})} p_{k}^{z_{k}} (1-p_{k})^{(1-z_{k})} \right) (z_{j}-z_{k}) \geq 0$$

where $w_{jk}=(z_1,...z_{j-1},z_{j+1},...,z_{k-1},z_{k+1},...,z_n)$. Expanding the last two summations:

$$\sum_{\substack{w_{jk}^{-} \in \{0,1\}^{n-2} \\ i \neq j,k}} \left(\prod_{\substack{i=1 \\ i \neq j,k}}^{n} p_{i}^{z_{i}} (1-p_{i})^{(1-z_{i})} \right) (p_{j}(1-p_{k}) - p_{k}(1-p_{j})) \geq 0$$

$$\sum_{\substack{w_{jk}^{z} \in \{0,1\}^{n-2} \\ i \neq j, k}} \left(\prod_{\substack{i=1 \\ i \neq j, k}}^{n} p_i^{z_i} (1 - p_i)^{(1 - z_i)} \right) (p_j - p_k) \ge 0$$

$$(p_j - p_k) \sum_{\substack{w_{jk}^- \in \{0,1\}^{n-2} \\ i \neq j,k}} \left(\prod_{\substack{i=1 \\ i \neq j,k}}^n p_i^{z_i} (1 - p_i)^{(1-z_i)} \right) \ge 0$$

$$(p_j - p_k) \sum_{\vec{w_{jk}} \in \{0,1\}^{n-2}} P(\vec{W_{jk}} = \vec{w_{jk}}) \ge 0$$

But the summation is over all possible values of $W_{jk} = (Z_1, ..., Z_{j-1}, Z_{j+1}, ..., Z_{k-1}, Z_{k+1}, ..., Z_n)$ and so must add to 1, so:

$$p_i - p_k \geq 0$$

But this contradicts our assumption that $p_j < p_k$.