A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation

Cyril Goutte and Eric Gaussier

Xerox Research Centre Europe
6, chemin de Maupertuis
F-38240 Meylan, France

Abstract. We address the problems of 1/ assessing the confidence of the standard point estimates, precision, recall and F-score, and 2/ comparing the results, in terms of precision, recall and F-score, obtained using two different methods. To do so, we use a probabilistic setting which allows us to obtain posterior distributions on these performance indicators, rather than point estimates. This framework is applied to the case where different methods are run on different datasets from the same source, as well as the standard situation where competing results are obtained on the same data.

1 Introduction

Empirical evaluation plays a central role in estimating the performance of natural language processing (NLP) or information retrieval (IR) systems. Performance is typically estimated on the basis of synthetic one-dimensional indicators such as the precision, recall or F-score. Even when multi-dimensional performance indicators are used, such as the recall-precision curve, synthetic indicators, such as the average precision at standard recall levels, are derived from it and used for comparison. One-dimensional performance measures, however, do not tell the full story, especially when they are estimated on the basis of little data, and are therefore intrinsically highly variable. This raises the following questions: Given a system and its results on a particular collection, how confident are we on the computed precision, recall and F-score? Do these measures tell us anything about the behavior of the system in general? The use of bootstrap [1, 2] allows one to partly answer these questions, by deriving approximate confidence intervals for the different point estimates. However, the summary statistics we consider here (precision, recall and F-score) do not always correspond to sample means or medians (as is the case for the summary statistics considered in [2]), and the bootstrap method may fail to give accurate confidence intervals. In this contribution, we adopt a different probabilistic point of view that allows us first to estimate the distribution of three indicators, precision, recall and F-score, and then to provide answers to the above questions.

The final version of this paper will appear in:
D.E. Losada and J.M. Fernández-Luna (eds) *Proceedings of the European Colloquium on IR Resarch (ECIR'05)*, LLNCS 3408 (Springer), pp. 345–359.

A related and crucial point is the comparison of experiments on the same dataset. Such a comparison is usually performed by resorting to paired statistical tests, as the paired t-test, the Wilcoxon test and the sign-test (see for example [3, 4]), or the bootstrap method or ANOVA. Some of these methods (e.g. the paired t and Wilcoxon tests) are not adapted to the three main indicators we retain, while others can be used (as the sign test or the bootstrap in some instances). The framework we rely on allows us to propose an additional tool for comparing two systems by providing an answer to the question: "What is the probability that sytem A outperforms (in terms of precision, recall and F-score) system B ?"

In the following section, we introduce the probabilistic framework we retained, and show how we infer distributions for precision, recall and F-score, as well as how such distributions can be used to compare two different systems. We then proceed (section 3) to the case of paired comparison of experimental outcomes, which may be used when systems are tested on the same dataset. These models are tested in section 4 on the outcomes of text categorisation experiments. Finally, we discuss the implication and perspectives of this work and conclude.

2 Precision, Recall and F-score

An arguably complete view of a system's performance is given by the precision-recall curve, which is commonly summarised in a single indicator using the average precision over various standard recall levels or number of documents. Other scores may be defined to reflect the performance, such as the break-even point, the scaled utility used at TREC[5], etc. Synthetic one-dimensional performance measures, however, do not allow to take into account the intrinsic variability in the scores, especially when calculated on little data. Note that this does not mean that evaluations performed on large collections are imune to this problem. At the 2002 TREC filtering track, for example, query 151 had only 22 relevant documents out of 723,141 test documents. This means that a variation in the assignment of one of the 22 relevant documents yields a variation of around 5% on recall. In the remainder of this paper, we focus on three standard performance indicators, namely precision, recall and F-score, and will first try to infer distributions that account for their intrisic variability.

For illustration purposes, we consider the following simple setting: each object is associated with a binary label ℓ which accounts for the correctness of the object with respect to the task at hand. In addition, the system produces an assignment z indicating whether it believes the object to be correct (or relevant) or not. The experimental outcome may be conveniently summarised in a confusion table:

		Assignment z	
		+	-
Label ℓ	+	TP	FN
	-	FP	TN

where + and - stand for relevant and non relevant, TP (resp. TN) stands for true positive (resp. negative) and FP (resp. FN) for false positive (resp. negative). From these counts, one can compute the precision (p) and recall (r):

$$p = \frac{TP}{TP + FP} \qquad r = \frac{TP}{TP + FN} \tag{1}$$

Taking the (weighted) harmonic average of precision and recall leads to the F-score ([6]):

$$F_\beta = (1 + \beta^2)\frac{p\,r}{r + \beta^2\,p} = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP} \tag{2}$$

Both precision and recall have a natural interpretation in terms of probability. Indeed, precision may be defined as the probability that an object is relevant given that it is returned by the system, while the recall is the probability that a relevant object is returned:

$$p = P(\ell = +|z = +) \qquad r = P(z = +|\ell = +) \tag{3}$$

This may seem like a trivial reformulation. However, there is a big semantic difference: in the original formulation, p and r are just formulas calculated from the observed data; in the probabilistic framework, the data $\mathcal{D} = (TP, FP, FN, TN)$ actually arises from p and r, which are parameters of a (primitive) generative model. Thus, the usual expressions (1) arise only as estimates of these unknown parameters.

2.1 Probabilistic model

Each system divides a particular collection into four distinct sets, corresponding to the true and false positives and negatives. The actual counts TP, FP, FN and TN can thus be seen as the results of independently drawing elements from these four sets. This view justifies the following simple assumption:

Assumption 1 *Observed TP, FP, FN and TN counts follow a multinomial distribution with parameters π_{TP}, π_{FP}, π_{FN}, π_{TN}:*

$$P\,(\mathcal{D} = (TP,\ FP,\ FN,\ TN)) = \frac{n!}{TP!\ FP!\ FN!\ TN!}\pi_{TP}^{TP}\pi_{FP}^{FP}\pi_{FN}^{FN}\pi_{TN}^{TN} \tag{4}$$

This is denoted by $\mathcal{D}|\pi \sim \mathcal{M}\,(n; \pi)$, with the multinomial parameter $\pi \equiv (\pi_{TP}, \pi_{FP}, \pi_{FN}, \pi_{TN})$, and $\pi_{TP} + \pi_{FP} + \pi_{FN} + \pi_{TN} = 1$. Using the property that marginals and conditionals of a multinomial-distributed vector follow binomial distributions, it can be shown that (see Appendix A):

Property 1 *The distribution of TP given the number of returned objects $M_+ = TP + FP$ is a binomial with parameters M_+ and p.*

Property 2 *The distribution of TP given the number of relevant objects $N_+ = TP + FN$ is a binomial with parameters N_+ and r.*

From property 1, we can write the likelihood of p as:

$$L(p) = P(\mathcal{D}|p) \propto p^{TP}(1-p)^{FP} \tag{5}$$

Inference on p can then be performed using Bayes' rule:

$$P(p|\mathcal{D}) \propto P(\mathcal{D}|p)P(p) \tag{6}$$

where $P(p)$ is the prior distribution. A natural choice for the priori distribution of a binomial distribution is the conjugate Beta distribution ([7, 8]). As there is no reason to favour high vs. low precision, we use a symmetric Beta prior:

$$p \sim Be(\lambda, \lambda) \quad : \qquad P(p) = \frac{\Gamma(2\lambda)}{\Gamma(\lambda)^2}\, p^{\lambda-1}(1-p)^{\lambda-1} \tag{7}$$

where $\Gamma(\lambda) = \int_0^{+\infty} u^{\lambda-1}\exp(-u)\,\mathrm{d}u$ is the Gamma function. Combining equations 5, 6 and 7 we get:

$$P(p|\mathcal{D}) \propto p^{TP+\lambda-1}(1-p)^{FP+\lambda-1} \tag{8}$$

that is, $p|\mathcal{D} \sim Be(TP+\lambda, FP+\lambda)$. The posterior distribution for the precision is therefore a Beta distribution that depends on TP, FP and the prior parameter λ. The expectation and mode for $P(p|\mathcal{D})$ are:

$$\overline{p} = \frac{TP+\lambda}{TP+FP+2\lambda}, \qquad \mathrm{mode}(p) = \frac{TP+\lambda-1}{TP+FP+2\lambda-2} \tag{9}$$

For $TP + FN < 2 - 2\lambda$ or $TP < 1 - \lambda$, the mode is either 0 or 1.

The Beta distribution offers a lot of flexibility on $[0; 1]$, and subsumes two interesting cases: $\lambda = 1/2$, Jeffrey's non-informative prior, and $\lambda = 1$, the uniform prior. Jeffrey's non-informative prior has the nice theoretical property that it is invariant through re-parameterisation [8]. This means that the non-informative prior for an arbitrary transformation $p' = f(p)$ is the transformation of the non-informative prior for p using the usual change-of-variable rule (which is not the case for a uniform prior). For $\lambda = 1$, we get the maximum likelihood estimate $\mathrm{mode}(p) = TP/(TP+FP)$. It turns out to be the usual formula for the precision (eq. 1). Note, however, that the expected value of p is a smoothed estimate $\overline{p} = (TP+1)/(TP+FP+2)$, aka Laplace smoothing. Obviously, using Property 2, a similar development yields the posterior distribution for the recall: $r|\mathcal{D} \sim Be(TP+\lambda, FN+\lambda)$, with the expectation and mode as in eq. 9 (replacing FP by FN).

Confidence intervals for p and r can easily be obtained from Beta tables, or through numerical approximations of (the integral of) the Beta distribution[1]. Estimating the probability that the precision/recall of a system is greater than the one of another system can be done through sampling strategies. We won't detail them here, as they are described for the F-score below.

[1] Standard mathematical packages usually provide such approximations.

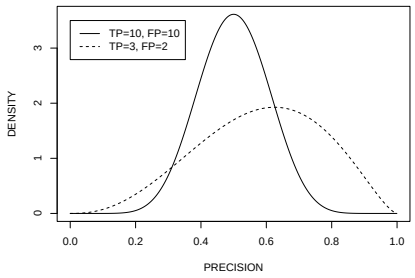

Fig. 1. Distribution of the precision for 2 systems with different outcomes (section 2.2). Although system 1 (solid) does worse on average, it is much less variable, and actually outperforms system 2 (dashed) in as much as 35% of cases.

Two cases of particular practical interest are the situations where $TP+FP = 0$, that is, the system does not return anything, and $TP + FN = 0$, no objects are relevant in the test set. In such cases, the traditional expression (1) is not valid. On the other hand, with the probabilistic model, the posterior is equal to the prior and the expectation (9) gives an estimate of $\overline{p} = 1/2$. This seems intuitively reasonable as the fact that the system does not return any object does not mean it will never do so in the future. In addition, the evidence from our experiment does not allow to favour low or high precision, suggesting that 50% may be a reasonable guess for p.

2.2 Example

Let us consider an example where system 1 returns 10 true positives and 10 false positives, while system 2 returns 3 true positives and 2 false positives. Using only the traditional formula for precision (1), system 2 ($p = 3/5$) seems largely superior to system 1 ($p = 1/2$). The probabilistic view tells another story. Assuming Jeffrey's prior, system 2 seems better on average ($\overline{p} = 58\%$, mode $= 63\%$) than system 1 (mode $= \overline{p} = 50\%$), but has a much larger variability, as shown in figure 1. As a consequence, the probability that system 2 outperforms system 1 with respect to precision is actually only around 65%, which implies that it is not significant at any reasonable level.

2.3 *F*-score

In order to combine our results on precision and recall, we now consider the distribution of the F_1 score, given by eq. 2, with $\beta = 1$: $F_1 = \frac{2pr}{p+r}$. Given two variables with Gamma distributions $X \sim \Gamma(\alpha, h)$ and $Y \sim \Gamma(\beta, h)$, with identical shape parameter h, then three interesting properties hold:

(1) $\forall c > 0$, $c.X \sim \Gamma(\alpha, c.h)$; (2) $X + Y \sim \Gamma(\alpha + \beta, h)$; (3) $\frac{X}{X+Y} \sim Be(\alpha, \beta)$

Property 3 allows us to postulate that the posterior distributions of p and r, which are Beta distributions (8), arise from the combination of independent Gamma variates:

$$p = \frac{X}{X+Y}, \quad r = \frac{X}{X+Z} \quad \text{with} \begin{cases} X \sim \Gamma(TP+\lambda, h) \\ Y \sim \Gamma(FP+\lambda, h) \\ Z \sim \Gamma(FN+\lambda, h) \end{cases} \tag{10}$$

Combining these in the F-score expression, and using the fact that $U = 2X$ is a Gamma variate (Property 1) and that $V = Y + Z$ is also a Gamma variate (Property 2), we get:

$$F_1 = \frac{U}{U+V} \qquad \text{with} \begin{cases} U \sim \Gamma(TP+\lambda, 2h) \text{ and} \\ V \sim \Gamma(FP+FN+2\lambda, h). \end{cases} \tag{11}$$

In order to compare two systems with different experimental outcomes $D^{(1)} = \left(TP^{(1)}, FP^{(1)}, FN^{(1)}, TN^{(1)}\right)$ and $D^{(2)} = \left(TP^{(2)}, FP^{(2)}, FN^{(2)}, TN^{(2)}\right)$, we wish to evaluate the probability $P(F_1^{(1)} > F_1^{(2)})$, that is, since $F_1^{(1)}$ and $F_1^{(2)}$ are independent:

$$\int_0^1 \int_0^1 \mathrm{I}\left(F_1^{(1)} > F_1^{(2)}\right) \, P\left(F_1^{(1)}\right) P\left(F_1^{(2)}\right) \, \mathrm{d}F_1^{(1)} \mathrm{d}F_1^{(2)} \tag{12}$$

where $\mathrm{I}(\cdot)$ is the indicator function which has value 1 iff the enclosed condition is true, 0 otherwise. As the distributions of $F_1^{(1)}$ and $F_1^{(2)}$ are not known analytically, we cannot evaluate (12) exactly, but we can estimate whether $P(F_1^{(1)} > F_1^{(2)})$ is larger than any given significance level using Monte Carlo simulation. This is done by creating large samples from the distributions of $F_1^{(1)}$ and $F_1^{(2)}$, using Gamma variates as shown in equation 11. Let us write these samples $\left\{f_i^{(1)}\right\}_{i=1\ldots L}$ and $\left\{f_i^{(2)}\right\}_{i=1\ldots L}$. The probability $P(F_1^{(1)} > F_1^{(2)})$ is then estimated by the empirical proportion:

$$\widehat{P}(F_1^{(1)} > F_1^{(2)}) = \frac{1}{L} \sum_{i=1}^{L} \mathrm{I}\left(f_i^{(1)} > f_i^{(2)}\right) \tag{13}$$

Note that the reliability of the empirical proportion will depend on the sample size. We can use inference on the probability $P(F_1^{(1)} > F_1^{(2)})$, and obtain a Beta posterior from which we can assess the variability of our estimate. Lastly, the case $\beta \neq 1$ is similar, although the final expression for F_β is not as simple as (11), and involves three Gamma variates. Comparing two systems in terms of F_β is again done by Monte Carlo simulation, in a manner exactly equivalent to what we have described for F_1.

3 Paired comparison

In the previous section, we have not made the specific assumption that the two competing systems were to be tested on the same dataset. Indeed, the inference

that we presented is valid if two systems are tested on distinct datasets, as long as they are sampled from the same (unknown) distribution. When two systems are tested on the same collection, it may be interesting to consider the paired outcomes on each object. Typically, a small difference in performance may be highly consistent and therefore significant. This leads us to consider now the following situation: on a single collection of objects $\{d_j\}_{j=1...N}$, with relevance labels ℓ_j, we observe experimental outcomes for two systems: $\left\{z_j^{(1)}\right\}_{j=1...N}$ and $\left\{z_j^{(2)}\right\}_{j=1...N}$.

For each object, three cases have to be considered: 1. System 1 gives the correct assignment, system 2 fails; 2. System 2 gives the correct assignment, system 1 fails; 3. Both system yield the same assignment. Let us write π_1, π_2 and π_3 the probability that a given object falls in either of the three cases above. Given that the assignments are independent, both accross systems and accross examples, and following the same reasoning as the one behind *assumption 1*, we assume that the experimental outcomes follow a multinomial distribution with parameters N and $\pi = (\pi_1, \pi_2, \pi_3)$. For a sequence of assignments $Z = \left\{z_j^{(1)}, z_j^{(2)}\right\}$, the likelihood of π is:

$$P(Z|\pi) \propto \pi_1^{N_1} \pi_2^{N_2} \pi_3^{N_3} \tag{14}$$

with N_1 (resp. N_2) the number of examples for which system 1 (resp. 2) outperforms system 2 (resp. 1), and $N_3 = N - N_1 - N_2$. The conjugate prior for π is the Dirichlet distribution, a multidimensional generalisation of the Beta distribution, $\pi|\alpha \sim D(\alpha_1, \alpha_2, \alpha_3)$:

$$P(\pi|\alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \pi^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} \pi_3^{\alpha_3 - 1} \tag{15}$$

with $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ the vector of hyper-parameters. Again, the uniform prior is obtained for $\alpha = 1$ and the non-informative prior for $\alpha = 1/2$[2]. From equations 15 and 14, applying Bayes rule we obtain the posterior $P(\pi|Z, \alpha) \propto P(Z|\pi)P(\pi|\alpha)$:

$$P(\pi|Z, \alpha) = \frac{\Gamma(N + \sum_k \alpha_k)}{\prod_k \Gamma(N_k + \alpha_k)} \prod_k \pi_k^{N_k + \alpha_k - 1} \tag{16}$$

which is a Dirichlet $D(N_1 + \alpha_1, N_2 + \alpha_2, N_3 + \alpha_3)$.

The probability that system 1 is superior to system 2 is

$$P(\pi_1 > \pi_2) = \mathrm{E}_{\pi|Z,\alpha}(\mathrm{I}(\pi_1 > \pi_2)) \tag{17}$$

[2] Although other choices are possible, it seems that if no prior information is available about which system is best, it is reasonable to impose $\alpha_1 = \alpha_2$. The choice of α_3 may be different, if the two competing systems are expected to agree more often than they disagree.

$$= \int_0^1 \left(\int_0^{min(\pi_1, 1-\pi_1)} P(\pi_1, \pi_2, 1-\pi_1-\pi_2|Z, \alpha) d\pi_2 \right) d\pi_1 \tag{18}$$

which implies integrating over the incomplete Beta function. This can not be carried out analytically, but may be estimated by sampling from the Dirichlet distribution. Given a large sample $\left\{\pi^j\right\}_{j=1\ldots L}$ from the posterior (16), equation 17 is estimated by:

$$\widehat{P}(M1 > M2) = \frac{\#\left\{j|\pi_1^j > \pi_2^j\right\}}{L} \tag{19}$$

Other ways of comparing both systems include considering the difference $\Delta = \pi_1 - \pi_2$ and the (log) odds ratio $\rho = \ln(\pi_1/\pi_2)$. Their expectations under the posterior are easily obtained:

$$\mathrm{E}_{\pi|Z,\alpha}(\Delta) = \frac{N_1 + \alpha_1 - N_2 - \alpha_2}{N + \alpha_1 + \alpha_2 + \alpha_3} \tag{20}$$

$$\mathrm{E}_{\pi|Z,\alpha}(\rho) = \Psi(N_1 + \alpha_1) - \Psi(N_2 + \alpha_2) \tag{21}$$

with $\Psi(x) = \Gamma'(x)/\Gamma(x)$ the Psi or Digamma function. In addition, the probability that either the difference or the log odds ratio is positive is $P(\Delta > 0) = P(\rho > 0) = P(\pi_1 > \pi_2)$.

Note: This illustrates the way this framework updates the existing information using new experimental results. Consider two collections $\mathcal{D}^1 = \left\{d_j^1\right\}_{1\ldots N^1}$ and $\mathcal{D}^2 = \left\{d_j^2\right\}_{1\ldots N^2}$. Before any observation, the prior for π is $D(\alpha)$. After testing on the first dataset, we obtain the posterior:

$$\pi|\mathcal{D}^1, \alpha \sim D\left(N_1^1+\alpha_1, N_2^1+\alpha_2, N_3^1+\alpha_3\right)$$

This may be used as a prior for the second collection, to reflect the information gained from the first collection. After observing all the data, the posterior becomes:

$$\pi|\mathcal{D}^2, \mathcal{D}^1, \alpha \sim D\left(N_1^2+N_1^1+\alpha_1, N_2^2+N_2^1+\alpha_2, N_3^2+N_3^1+\alpha_3\right)$$

The final result is therefore equivalent to working with a single dataset containing the union of D^1 and D^2, or to evaluating first on D^2, then on D^1. This property illustrates the convenient way in which this framework updates our knowledge based on additional incoming data.

4 Experimental results

In order to illustrate the use of the above probabilistic framework in comparing experimental outcomes of different systems, we use the text categorisation

	F_1 score		Prob		F_1 score (nnn)		Prob		F_1 score (ltc)		Prob	
Category	ltc	nnn	ltc>nnn	+/-	lin	poly	lin>poly	+/-	lin	poly	lin>poly	+/-
earn	98.66	98.07	93.71	0.24	98.07	96.36	99.96	0.02	98.66	98.80	34.68	0.48
acq	94.70	93.88	82.09	0.38	93.88	79.31	100.00	0.00	94.70	94.73	49.39	0.50
money-fx	76.40	75.12	63.31	0.48	75.12	64.40	99.64	0.06	76.40	73.90	75.45	0.43
crude	86.96	86.15	61.53	0.49	86.15	71.12	100.00	0.00	86.96	86.26	60.33	0.49
grain	89.61	88.22	69.55	0.46	88.22	73.99	99.99	0.01	89.61	86.57	85.66	0.35
trade	75.86	77.47	34.07	0.47	77.47	75.11	71.77	0.45	75.86	77.13	38.76	0.49
interest	73.98	77.93	17.03	0.38	77.93	68.24	98.83	0.11	73.98	72.27	64.72	0.48
wheat	79.10	80.79	36.81	0.48	80.79	77.27	75.14	0.43	79.10	80.92	36.45	0.48
ship	75.50	80.25	17.99	0.38	80.25	66.22	99.44	0.07	75.50	73.47	63.78	0.48
corn	83.17	83.64	46.73	0.50	83.64	65.35	99.76	0.05	83.17	82.00	58.15	0.49
dlr	78.05	74.23	69.34	0.46	74.23	25.17	100.00	0.00	78.05	72.97	74.18	0.44
oilseed	61.90	62.37	48.58	0.50	62.37	42.35	98.77	0.11	61.90	58.23	65.28	0.48
money-sup	70.77	67.57	64.53	0.48	67.57	70.13	38.19	0.49	70.77	74.19	36.06	0.48
Micro-avg	90.56	89.86	89.12	0.31	89.86	79.44	100.00	0.00	90.56	90.28	67.69	0.47

Table 1. F-score omparison for 1/ 'ltc' versus 'nnn' weighting scheme (left), 2/ linear versus polynomial kernel on 'nnn' weighting (centre), and 3/ linear versus polynomial kernel on 'ltc' weighting (right).

task defined on the Reuters21578 corpus with the ModApte split. We are here interested in evaluating the influence of two main factors: the term weighting scheme and the document similarity. To this end, we consider two term weighting schemes:

nnn (no weighting): term weight is the raw frequency of the term in the document, no inverse document frequency is used and no length normalisation is applied;

ltc (log-tf-idf-cosine): the term weight is obtained by taking the log of the frequency of the term in the document, multiplied by the inverse document frequency, and using cosine normalisation (which is equivalent to dividing by the euclidean norm of the unnormalised values).

For 13 categories with the largest numbers of relevant documents, we train a Support Vector Machine (SVM) categoriser ([9]) on the 9603 training document from the ModApte split, and test the models on the 3299 test documents. In SVM categorisers, the document similarity is used as the kernel. Here, we compare the linear and quadratic kernels. We wish to test whether 'ltc' is better than 'nnn' weighting, and whether the quadratic kernel *significantly* outperforms the linear kernel. Most published results on this collection show a small but consistent improvement with higher degree kernels.

We first compared the results we obtained by comparing the $F1$-score distributions, as is outlined in section 2 (using $\lambda = \frac{1}{2}$, Jeffrey's prior, and $h = 1$). This comparison is given in table 1. As one can note, with the linear kernel, the

'ltc' weighting scheme seems marginally better than 'nnn'. The most significant difference is observed on the largest category, 'earn', where the probability that 'ltc' is better almost reaches 94%. Note that despite the fact that the overall number of documents is the same for all categories, a half percent difference on 'earn' is more significant than a 5% difference on 'ship'. This is due to the fact that, as explained above, the variability is actually linked to the number of *relevant* and *returned* documents, rather than the overall number of documents in the collection.

The centre columns in table 1 show that the 'nnn' weighting scheme has devastating effects on the quadratic kernel. The micro-averaged F-score is about 10 point lower, and all but three categories are significantly worse (at the 98% level) using the quadratic kernel. This is because when no IDF is used, the similarity between documents is dominated by the frequent terms, which are typically believed to be less meaningful. This effect is much worse for the quadratic kernel, where the implicit feature space works with *products* of frequencies. Finally, the right columns of table 1 investigate the effect of the kernel, when using the 'ltc' weighting. In that case, both kernels yield similar results, which are not significant at any reasonable level. Thanks to IDF and normalisation, we do not observe the dramatic difference in performance that was apparent with 'nnn' weighting.

The results in table 1 do not take into account the fact that all models are run on the same data. Using the paired test detailed in section 3, we get more sensitive results (table 2; we have used here $\alpha_1 = \alpha_2 = \frac{1}{2}$, the value for α_3 being of no importance for our goal). The 'ltc' weighting seems significantly better than 'nnn' on the three biggest categories, and the performance improvement seems weakly significant on two additional categories: 'dlr' and 'money-sup'. Small but consistent differences (8 to 3 in favour of 'ltc' in 'money-sup') may actually yield better scores than large, but less consistent, disagreements (for example, 21 to 15 in 'crude'). The middle rows of table 2 confirm that, with 'nnn' weighting, the linear kernel is better than the quadratic kernel. 'trade', 'wheat' and 'money-sup' do not show a significant difference and, surprisingly, category 'money-fx' shows a weakly significant difference, whereas the F-score test (table 1) gave a highly significant difference (we will discuss this below). The rightmost rows in table 2 show that, with 'ltc' weighting, the polynomial kernel is significantly better than the linear kernel for three categories ('trade', 'wheat' and 'money-sup') and significantly worse for 'grain'. The polynomial kernel therefore significantly outperforms the linear kernel more often than the reverse, although the micro-averaged F-score given in table 1 is slightly in favour of the linear kernel.

	linear kernel		Prob		'nnn' weighting		Prob		'ltc' weighting		Prob	
Category	ltc>	nnn>	ltc>nnn	+/-	lin>	p2>	lin>p2	+/-	lin>	p2>	lin>p2	+/-
earn	17	4	99.77	0.05	48	12	100.00	0.00	1	4	9.76	0.30
acq	43	28	96.41	0.19	282	14	100.00	0.00	6	7	39.68	0.49
money-fx	39	23	97.90	0.14	58	43	93.76	0.24	11	6	88.70	0.32
crude	21	15	84.32	0.36	62	21	100.00	0.00	4	2	78.66	0.41
grain	17	11	87.32	0.33	46	10	100.00	0.00	9	2	98.48	0.12
trade	23	22	55.75	0.50	28	28	49.19	0.50	2	7	4.36	0.20
interest	24	24	49.73	0.50	38	21	98.61	0.12	5	3	76.29	0.43
wheat	10	9	59.56	0.49	14	13	57.87	0.49	0	3	3.39	0.18
ship	6	11	11.38	0.32	22	4	99.97	0.02	3	1	83.71	0.37
corn	6	5	61.45	0.49	19	2	99.99	0.01	1	0	82.31	0.38
dlr	13	6	94.62	0.23	191	2	100.00	0.00	5	3	75.81	0.43
oilseed	12	9	74.31	0.44	24	10	99.23	0.09	3	2	66.66	0.47
money-sup	8	3	93.43	0.25	10	11	41.22	0.49	0	3	3.49	0.18

Table 2. Paired comparison of 1/ 'ltc' and 'nnn' weighting schemes (left), 2/ linear ('lin') and quadratic ('p2') kernel, on 'nnn' weighting, and 3/ linear and quadratic kernel on 'ltc' weighting.

5 Discussion

There has been a sizeable amount of work in the Information Retrieval community targetted towards proposing new performance measures for comparing the outcomes of IR experiments(two such recent atempts can be found in ([10, 11]). Here, we take a different standpoint. We focus on widely used measures (precision, recall, and F-score), and infer distributions for them that allow us to evaluate the variability of each measure, and assess the significance of an observed difference. Although this framework may not be applicable to arbitrary performance measures, we believe that it can also apply to other ones, such as the TREC utility. In addition, using Monte-Carlo simulation, it is possible to sample from simple distributions, and combine the samples in non-trivial ways (cf the F-score comparison in section 2). The only alternative we are aware of to compute both confidence intervals for the three measures we retained and assess the significance of an observed difference is the bootstrap method. However, as we already mentioned, this method may fail for the statistics we retained. Note, nevertheless, that the bootstrap method is very general and may be used for other statistics than the ones considered here. It might also be the case that, based on the framework we developed, a parametric form of the bootstrap can be used for precision, recall and F-score. This is something we plan to investigate.

In the case where two systems are compared on the same collection, approaches to performance comparison typically rely on standard statistical tests such as the paired t-test, the Wilcoxon test or the sign test [4], or variants of them [12, 13]. Neither the t-test nor the Wilcoxon test directly apply to binary (relevant/not relevant) judgements (which are at the basis of the computation

of precision, recall and F-score)[3]. Both the sign test and, again, the bootstrap method seem to be applicable to binary judgements (the same objections as above hold for the bootstrap method). Our contribution in this framework is to have put at the disposal of experimenters an additional tool for system evaluation. A direct comparison with the aforementioned methods still has to be conducted.

Experimental results provided in section 4 illustrate some differences between the two tests we propose. The F-score test assesses differences in F-score, and may be applied to results obtained on different collections (for example random splits from a large collection). On the other hand, the paired test must be applied to results obtained on the same dataset, and seems more sensitive to small, but consistent differences. In one instance ('money-fx', linear vs. quadratic on 'nnn', table 2), the F-score test was significant, while the paired test was not. This is because although a difference in F-score necessary implies a difference in disagreement counts (58 to 43 in that case), this disagreement may not be consistent, and therefore yield a larger variabliliy. In that case, an 8 to 3 difference would give the same F-score difference, and be significant for the paired test.

6 Conclusion

We have presented in this paper a new view on standard Information Retrieval measures, namely precision, recal, and F-score. This view, grounded on a probabilistic framework, allows one to take into account the intrinsic variability of performance estimation, and provides, we believe, more insights on system performance than traditional measures. In particular, it helps us answer questions like: "Given a system and its results on a particular collection, how confident are we on the computed precision, recall and F-score?", or "Can we compare (in terms of precision, recall and F-score) two systems evaluated on two different datasets from the same source?" and lastly "What is the probability that system A outperforms (in terms of precision, recall and F-score) system B when compared on the same dataset?".

To develop this view, we have first shown how precision and recall naturally lead to probabilistic interpretation, and how one can derive probabilistic distributions of them. We have then shown how the F-scores could be rewritten in terms of Gamma variates, and how to compare F-scores obtained by two systems based on Monte-Carlo simulation. In addition, we have presented an extension to paired comparison, which allows one to perform a deeper comparison between

[3] One possibility to use them would be to partition the test data into subsets, compute say precision on each subset and compare the distributions obtained with different systems. However, this may lead to poor estimates on each subset, hence to poor comparison (furthermore, the intrinsic variability of each estimate is not taken into account).

two systems run on the same data. Lastly, we have illustrated the new approach to performance evaluation we propose on a standard text categorisation task, with binary judgements (in class/not in class) for which several classic statistical tests are not well suited, and discussed the relations of our approach to existing ones.

References

1. Efron, B.E.: The Jacknife, the Bootstrap and Other Resampling plans. Volume 38 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM (1982)
2. Savoy, J.: Statistical inference in retrieval effectiveness evaluation. Information Processing & Management **33** (1997) 495–512
3. Tague-Sutcliffe, J., Blustein, J.: A statistical analysis of the TREC-3 data. In Harman, D., ed.: Proceedings of the third Text Retrieval Conference (TREC). (1994) 385–398
4. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of SIGIR'93, ACM, Pittsburg, PA (1993) 329–338
5. Robertson, S., Soboroff, I.: The TREC 2002 filtering track report. In: Proc. Text Retrieval Conference. (2002) 208–217
6. van Rijsbergen, C.J.: Information Retrieval. 2nd edition edn. Butterworth (1979)
7. Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis. Wiley (1973)
8. Robert, C.: L'Analyse Statistique Bayesienne. Economica (1992)
9. Joachims, T.: Making large-scale svm learning practical. In Schölkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods — Support Vector Learning, MIT Press (1999)
10. Mizzaro, S.: A new measure of retrieval effectiveness (or: What's wrong with precision and recall). In: In T. Ojala editor, International Workshop on Information Retrieval (IR'2001). (2001)
11. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems **20** (2002)
12. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of COLING'00, Saarbrücken, Germany (2000)
13. Evert, S.: Significance tests for the evaluation of ranking methods. In: Proceedings of COLING'04, Geneva, Switzerland (2004)

A Proof of property 1 and 2

The available data is uniquely characterised by the true/false positive/negative counts TP, FP, TN, FN. Let us note $\mathcal{D} \equiv (TP, FP, FN, TN)$. Our basic modelling assumption is that for sampled of fixed size n, these four counts follow a multinomial distribution. This assumption seems reasonable and arises for example if the collection is an independant identically distributed (i.i.d.) sample from the document population. If we denote by n_1, n_2, n_3 and n_4 (with $n_1 +$

$n_2 + n_3 + n_4 = n$) the actual counts observed for variables TP, FP, FN and TN, then:

$$P\left(\mathcal{D} = (n_1, n_2, n_3, n_4)\right) = \frac{n!}{n_1!\, n_2!\, n_3!\, n_4!} \pi_1^{n_1} \pi_2^{n_2} \pi_3^{n_3} \pi_4^{n_4} \tag{22}$$

With $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$ and $\pi \equiv (\pi_1, \pi_2, \pi_3, \pi_4)$ is the parameter of the multinomial distribution: $\mathcal{D} \sim \mathcal{M}(n; \pi)$. We will use the following two properties of multinomial distributions:

Property 3 (Marginals) *Each component i of $\mathcal{D}$ follows a binomial distribution $\mathcal{B}(n; \pi_i)$, with parameters n and identical probability π_i.*

Property 4 (Conditionals) *Each component i of $\mathcal{D}$ conditionned on another component j follows a binomial distribution $\mathcal{B}\left(n - n_j; \frac{\pi_i}{1-\pi_j}\right)$, with parameters $n - n_j$ and probability $\frac{\pi_i}{1-\pi_j}$.*

It follows from these properties that:

$$(TP + FP) \sim \mathcal{B}(n; \pi_1 + \pi_2) \qquad \text{and} \qquad (TP + FN) \sim \mathcal{B}(n; \pi_1 + \pi_3) \tag{23}$$

Proof. From the above properties, $TP \sim \mathcal{B}(n; \pi_1)$ and $FP|TP \sim \mathcal{B}\left(n - TP; \frac{\pi_2}{1-\pi_1}\right)$, hence:

$$\begin{aligned}
P(TP + FP = k) &= \sum_{x=0}^{k} P(TP = x) P(FP = k - x | TP = x) \\
&= \sum_{x=0}^{k} \binom{n}{x} \pi_1^x (1 - \pi_1)^{n-x} \binom{n-x}{k-x} \left(\frac{\pi_2}{1-\pi_1}\right)^{k-x} \\
&\quad \left(1 - \frac{\pi_2}{1-\pi_1}\right)^{n-k} \\
&= \binom{n}{k} (1 - (\pi_1 + \pi_2))^{n-k} \sum_{x=0}^{k} \binom{k}{x} \pi_1^x \pi_2^{k-x}
\end{aligned} \tag{24}$$

and as $\sum_{x=0}^{k} \binom{k}{x} \pi_1^x \pi_2^{k-x} = (\pi_1 + \pi_2)^k$ (the binomial theorem), the distribution of $TP + FP$ is indeed binomial with parameters n and $\pi_1 + \pi_2$ (and similarly for $TP + FN$). □

Using eq. 23 and the fact that $TP \sim \mathcal{B}(n; \pi_1)$, we obtain the conditional distribution of TP given $TP + FP$:

$$TP|(TP + FP) \sim \mathcal{B}\left(M_+; \frac{\pi_1}{\pi_1 + \pi_2}\right) \tag{25}$$

with $M_+ = TP + FP$.

Proof. The conditional probability of TP given $TP + FP$ is obtained as:

$$P(TP{=}k|TP + FP{=}M_+) = \frac{P(TP{=}k)P(FP{=}M_+{-}k|TP{=}k)}{P(TP + FP{=}M_+)}$$

As all probabilities involved are known binomials, we get:

$$P(TP = k|TP + FP = M_+) = \frac{\binom{n}{k}\pi_1^k(1-\pi_1)^{n-k}.\binom{n-k}{M_+-k}\left(\frac{\pi_2}{\pi_1+\pi_2}\right)^{M_+-k}\left(1-\frac{\pi_2}{\pi_1+\pi_2}\right)^{n-M_+}}{\binom{n}{M_+}(\pi_1+\pi_2)^{M_+}(1-\pi_1+\pi_2)^{n-M_+}} \tag{26}$$

Using the fact that $\frac{\binom{n}{k}\binom{n-k}{M_+-k}}{\binom{n}{M_+}} = \binom{M_+}{k}$ and after some simple algebra, this simplifies to:

$$P(TP = k|TP + FP = M_+) = \binom{M_+}{k}\left(\frac{\pi_1}{\pi_1+\pi_2}\right)^k\left(\frac{\pi_2}{\pi_1+\pi_2}\right)^{M_+-k} \tag{27}$$

in other words, eq. 25. □

Now remember the definition of the precision:

$$p = P(z = +|l = +) = \frac{P(z = +, l = +)}{P(l = +)} \tag{28}$$

Notice that $P(z = +, l = +)$ is by definition π_1, the probability to get a "true positive", and $P(l = +)$ is $\pi_1 + \pi_2$, the probability to get a positive return (either false or true positive). This shows that $p = \pi_1/(\pi_1 + \pi_2)$, and that the distribution of TP given $TP + FP$ (eq. 25) is a binomial $\mathcal{B}(M_+, p)$. The justification for $TP|(TP + FN)$ goes the same way.