Computing Precision and Recall with Missing or Uncertain Ground Truth

Bart Lamiroy, Tao Sun

Computing Precision and Recall with Missing or Uncertain Ground Truth

Bart Lamiroy[1] and Tao Sun[2]

[1] Université de Lorraine – LORIA, Nancy, France – Bart.Lamiroy@loria.fr
[2] Lehigh University – Computer Science and Engineering, Bethlehem, PA, USA

Abstract. In this paper we present a way to use precision and recall measures in total absence of ground truth. We develop a probabilistic interpretation of both measures and show that, provided a sufficient number of data sources are available, it offers a viable performance measure to compare methods if no ground truth is available. This paper also shows the limitations of the approach, in case a systematic bias is present in all compared methods, but shows that it maintains a very high level of overall coherence and stability. It opens broader perspectives and can be extended to handling partial or unreliable ground truth, as well as levels of prior confidence in the methods it aims to compare.

1 Introduction

Performance evaluation of information retrieval methods in a broad sense, *i.e.* globally any process associating high level information to a collection of weakly structured data often relies on comparing the output of the methods under evaluation to selected and verified data, for which the expected outcome of the methods is known (*cf.* [20] in graphical document analysis, for instance). These data are usually referred at as *ground truth*.

As long as the retrieval goals can be correctly captured and the scope of the data on which the methods must operate remains controllable, relying on ground-truth is possible [2, 7]. However, when the size of the potential data space becomes unmanageable of when it becomes more controversial to fully formalize the required outcome of the methods under investigation, fixing or obtaining ground truth becomes problematic to impossible. In some cases, especially when the data sets grow to a significant size, en when the retrieval process tends to favor *precision* rather than *recall* (*cf.* next section for definitions) performance evaluation approaches may rely on sampling and statistical extrapolation [8], rather than exhaustive validation. This still requires as sufficiently large set of ground-truthed data, however. Other approaches use higher level knowledge to assess coherence patterns in classified data [3].

In this paper we approach the problem differently, by making the assumption that there is either no ground truth available, or that the available ground truth may be unreliable (for instance, coming from crowd-sourced annotation processes, for which no post-processing has been done, or scenarios where human

feedback interferes with pre-established ground truth [19]). We show that by reformulating classical performance metrics like precision and recall in probabilistic terms we can establish a ranking between competing approaches that is comparable to the one that would be obtained in presence of reliable ground-truth. In that aspect, it shares some very interesting similarities with work related to classifier fusion using majority voting [11, 4]. This similarity will be addressed in Section 4.3.

Before that, and after a brief recall of the definitions of Precision and Recall in Section 2, we develop the theoretical framework of our approach in Section 3. Section 4 provides a series of experimental validations of our method and exposes some of its limitations. Further work and extensions are provided in Section 5.

2 Precision and Recall

2.1 General Definitions and Notation

Precision Pr and Recall Rc (and often associated F-measure or ROC curves) are standard metrics expressing the *quality* of Information Retrieval methods [15]. They are usually expressed with respect to a query q (or averaged over a series of queries) over a data set Δ such that:

$$Pr_q^{\Delta} = \frac{\left|\mathcal{P}_q^{\Delta} \cap \mathcal{R}_q^{\Delta}\right|}{\left|\mathcal{R}_q^{\Delta}\right|} \tag{1}$$

$$Rc_q^{\Delta} = \frac{\left|\mathcal{P}_q^{\Delta} \cap \mathcal{R}_q^{\Delta}\right|}{\left|\mathcal{P}_q^{\Delta}\right|} \tag{2}$$

where $\mathcal{P}_q^{\Delta}$ is the set of all documents in Δ, relevant to query q, and where $\mathcal{R}_q^{\Delta}$ is the set of documents actually retrieved by q. Although we can make a safe assumption by considering $\mathcal{R}_q^{\Delta}$ known (*i.e.* the query q can actually be executed, and returns a known, manageable set of results), the same assumption does not always hold for $\mathcal{P}_q^{\Delta}$, as will be shown later. For ease of reading we will refer to respectively Pr, $\mathcal{P}$, Rc, and $\mathcal{R}$, when there is no ambiguity on Δ and q.

Often both are combined in the F_β measure, where

$$F_\beta = \left(1 + \beta^2\right) \frac{PrRc}{\beta^2 Pr + Rc} \tag{3}$$

and where β expresses the importance of recall with respect to precision. Generally, $\beta = 1$, so that both are considered of equal importance.

2.2 Other Interpretations and Frameworks

Precision, Recall and the F-measure can also be defined with respect to *true positives* τ_p, *false positives* ϕ_p, *true negatives* τ_n and *false negatives* ϕ_n. In that

case, the corresponding formulas are:

$$Pr = \frac{\tau_p}{\tau_p + \phi_p} \tag{4}$$

$$Rc = \frac{\tau_p}{\tau_p + \phi_n} \tag{5}$$

$$F_\beta = \frac{\left(1 + \beta^2\right) \tau_p}{\left(1 + \beta^2\right) \tau_p + \beta^2 \phi_n + \phi_p} \tag{6}$$

Here again, it is necessary to know the values of τ_p, ϕ_p, τ_n and ϕ_p (as, previously, the sets $\mathcal{P}$ and $\mathcal{R}$) in order to be able to do the computations.

It is also possible to give probabilistic interpretations to Pr and Rc. In that case, Pr would be the probability that a random document retrieved by the query is relevant, and Rc that a random relevant document be retrieved by the query (taking as assumption that documents have uniform distributions). This is the interpretation we are going to use in the next sections.

3 Absence of Ground Truth

Previously enumerated metrics all made the assumption that the returns of queries can, in some way be qualified as "good" or "bad". Most often, there even is the assumption that this can actually be quantified: belonging to set $\mathcal{P}$, τ_p, *etc.* This implies that there is some absolute knowledge of *ground truth* or an *oracle* function available for the assessment of these quantities. While it is very convenient to rely on established truth to further train or evaluate methods, it is often very costly to obtain in many cases, and even impossible in others. Furthermore, it generally requires some human intervention or validation of some sorts, which makes the ground-truthing process both difficultly scalable and error prone, and therefore costly.

This paper presents a way to estimate precision and recall using a probabilistic model, allowing either to compare algorithms operating on the same data, without the requirement of establishing ground truth, or, to leverage crowdsourcing to establish ground truth in presence of noise, errors and mistakes. In order to achieve this, we shall first establish the underlying assumptions to our approach, in section 3.1, defining the context in which we have conceived our model. We then develop the mathematical foundations and tools in section 3.2.

3.1 General Assumptions

In what follows we are assuming that the following general conditions and notations apply:

1. We are considering generic system $\mathcal{S}$ that, given a query q, partitions[3] a set of documents $\Delta = \{\delta_i\}_{i=1\ldots d}$ into $\mathcal{S}^{q+}$ and $\mathcal{S}^{q-}$.

[3] For the absent-minded reader, "*partitioning*" Δ into $\mathcal{S}^+$ and $\mathcal{S}^-$ entails that $\Delta = \mathcal{S}^+ \cup \mathcal{S}^-$ and $\mathcal{S}^+ \cap \mathcal{S}^- = \emptyset$

The partitioning function $\mathcal{S}^q$ is defined as

$$\begin{aligned} \mathcal{S}^q : \Delta &\to \{+, -\} \\ \delta_i &\mapsto \mathcal{S}^q(\delta_i) \end{aligned} \tag{7}$$

$\mathcal{S}^{q+}$ (resp. $\mathcal{S}^{q-}$) is defined as the inverse image of $\{+\}$ (resp. $\{-\}$).

2. Other systems, similar to $\mathcal{S}^q$ exist and their partitioning results are available. It is assumed that these systems operate in the same semantic context, and therefore aim to achieve the same partitioning as $\mathcal{S}^q$. We shall refer to the set of these systems as $\Sigma^q = \{\mathcal{S}_i^q\}_{i=1\ldots s}$

In what follows, and where it is obvious, parameter q will be omitted. Table 1 gives an example overview of what three different systems could produce for a given query over a particular document set Δ.

Δ	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$
δ_1	+	+	+
δ_2	+	+	+
δ_3	-	+	-
δ_4	+	-	-
δ_5	+	-	-
δ_6	-	-	+
δ_7	-	-	-

$\mathcal{S}_1^+ = \{\delta_1, \delta_2, \delta_4, \delta_5\}$
$\mathcal{S}_1^- = \{\delta_3, \delta_6, \delta_7\}$
$\mathcal{S}_2^+ = \{\delta_1, \delta_2, \delta_3\}$
$\mathcal{S}_2^- = \{\delta_4, \delta_5, \delta_6, \delta_7\}$
$\mathcal{S}_3^+ = \{\delta_1, \delta_2, \delta_6\}$
$\mathcal{S}_3^- = \{\delta_3, \delta_4, \delta_5, \delta_7\}$

Table 1. Example of query systems $\mathcal{S}_i$ operating on document set Δ

3.2 Performance Evaluation

The question that arises now is how to compare different $\mathcal{S}_i$ and decide which one performs best. Traditionally, one would take an evaluation test set $\Delta_\star$ for which the ground truth of a query $q_\star$ is known and available. We shall refer to this ground truth as $\Delta_\star^+$ and $\Delta_\star^-$ (*i.e.* $\Delta_\star^+$ is the partition of $\Delta_\star$ containing the documents corresponding to $q_\star$, $\Delta_\star^-$ its complement). This knowledge then allows to compute precision and recall values, as described in Section 2, for all $\mathcal{S}_i$ and establish a performance metric adapted to the context under consideration.

When $\Delta_\star^+$ and $\Delta_\star^-$ are unavailable, it is less obvious to compare the results of the different $\mathcal{S}_i$. One well documented approach is to use statistical estimators by considering each $\mathcal{S}_i(\Delta)$ as the outcome of some random variable. What we are going to develop here, is very similar, but particularly focused on the expression of precision and recall.

Simplified Case First we're making the assumption that all $\mathcal{S}_i$ are of equal importance, and that there is no *a priori* knowledge available allowing to presume

some of the systems are more reliable than others. This assumption will be alleviated in later work. We also assume all documents have equal frequency and occurrence probability.

For the arguments developed next, we need to introduce two "virtual" query systems, $\mathcal{S}_\top$ and $\mathcal{S}_\bot$. $\mathcal{S}_\top$ always returns all documents for any given query, $\mathcal{S}_\bot$ never returns any. In other terms,

$$\mathcal{S}_\top^+ = \Delta, \mathcal{S}_\top^- = \emptyset \tag{8}$$

$$\mathcal{S}_\bot^+ = \emptyset, \mathcal{S}_\bot^- = \Delta \tag{9}$$

We are also slightly reconsidering the partitioning function defined in equation (7), such that it returns values in $\{1, 0\}$ rather than in $\{+, -\}$.

Under these hypotheses, the probability that a document δ_i belongs to $\Delta_\star^+$ is

$$P(\delta_i) = \frac{1}{s+2} \sum_{k=1\ldots s,\bot,\top} S_k(\delta_i) \tag{10}$$

The results of the application of this to the example in Table 1, is represented in Table 2.

Δ	$P(\delta_i)$	$\mathcal{S}_\top$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_\bot$
δ_1	0.8	1	1	1	1	0
δ_2	0.8	1	1	1	1	0
δ_3	0.4	1	0	1	0	0
δ_4	0.4	1	1	0	0	0
δ_5	0.4	1	1	0	0	0
δ_6	0.4	1	0	0	1	0
δ_7	0.2	1	0	0	0	0

Table 2. Example

Given the hypothesis of equidestribution of all documents δ_i in Δ and given the probabilistic definition of precision in Section 2.2, stating that Pr "is the probability that a random document retrieved by a query is relevant", we can now define $Pr(\mathcal{S}_k)$:

$$Pr(\mathcal{S}_k) = \frac{\sum_{i=1\ldots d} P(\delta_i) S_k(\delta_i)}{\sum_{i=1\ldots d} S_k(\delta_i)} \tag{11}$$

Similarly, Rc was defined as "the probability for a random relevant document to be retrieved by the query". In our case, however relevancy has no longer a binary value, but has been replaced by $P(\delta_i)$. By reformulating this conditional probability and using Bayes' theorem (and using the fact that the inverse

conditional of Rc is Pr), things smooth out elegantly.

$$\begin{aligned} Rc\left(\mathcal{S}_k\right) &= Prob\left(\text{retrievedBy}_{\mathcal{S}_k}\left(\delta_i\right) \,\middle|\, \text{isRelevant }_{(\delta_i)}\right) \\ &= Prob\left(\text{isRelevant}\left(\delta_i\right) \,\middle|\, \text{retrievedBy}_{\mathcal{S}_k(\delta_i)}\right) \frac{Prob\left(\text{retrievedBy}_{\mathcal{S}_k}\left(\delta_i\right)\right)}{Prob\left(\text{isRelevant}\left(\delta_i\right)\right)} \\ &= Pr\left(\mathcal{S}_k\right) \frac{\frac{1}{d}\sum_{i=1..d} S_k\left(\delta_i\right)}{\frac{1}{d}\sum_{i=1..d} P\left(\delta_i\right)} \\ &= \frac{\sum_{i=1...d} P\left(\delta_i\right) S_k\left(\delta_i\right)}{\sum_{i=1...d} S_k\left(\delta_i\right)} \frac{\sum_{i=1..d} S_k\left(\delta_i\right)}{\sum_{i=1..d} P\left(\delta_i\right)} \\ &= \frac{\sum_{i=1...d} P\left(\delta_i\right) S_k\left(\delta_i\right)}{\sum_{i=1..d} P\left(\delta_i\right)} \end{aligned} \tag{12}$$

It is interesting to notice the resemblance between equations (1) and (11) as well as between (2) and (12). Table 3 shows the values obtained when applied to the examples of Table 2.

Δ	$P\left(\delta_i\right)$	$\mathcal{S}_\top$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_\perp$
δ_1	0.8	1	1	1	1	0
δ_2	0.8	1	1	1	1	0
δ_3	0.4	1	0	1	0	0
δ_4	0.4	1	1	0	0	0
δ_5	0.4	1	1	0	0	0
δ_6	0.4	1	0	0	1	0
δ_7	0.2	1	0	0	0	0
Sum	3.4	7	4	3	3	0
$\sum P\mathcal{S}_k$		3.4	2.4	2	2	0
Pr		0.49	0.6	0.67	**0.67**	∞
Rc		1	**0.71**	0.59	0.59	0

Table 3. Example of precision and recall computations without established ground truth.

4 Experimental Validation

In order to experimentally validate the model developed we have taken two contexts. One consists in taking the results of experiments reported in [10] related to comparing standard symbol recognition techniques. A second is related to evaluation of binarization algorithms on downstream treatment and is very similar to the experiments conducted in [13].

4.1 Symbol Recognition

In this section we use the experimental results reported in [10]. In this paper, the authors compare 5 different symbol recognition methods on a set of electrical wiring diagrams. Since their dataset has no known ground truth, they use a panel of human annotators to select and determine which ground truth corresponds to which query.

Since the authors in [10] report retrieval efficiency, as defined in [9], we have resampled their raw experimental data to extract precision and recall. The results, with respect to the human-defined ground-truth reported by the authors is shown in Figure 1.

Figure 2 reproduces the precision and recall values obtained using our method on the exact same data. It is interesting to note that, with one noteworthy exception, the ordering of the tested methods, with respect to precision or recall (*i.e.* when ordering methods from high precision/recall to low) is respected. Although not reproduced here, this also holds for the F-measure. What is even more compelling, is that the methods 'SC' and 'GFD' maintain their similarity in both cases, with and without consideration of ground truth.

The one exception is the 'ARG' method. While considered as a tie with 'SC' and 'GFD' with our method, it significantly outperforms all other approaches according to the ground truth. This is a very interesting result, and is currently under investigation.

4.2 Document Binarization

The data used in this second study are the historical images collected from the Library of Congress on-line data set[1]. A total of 60 TIF format images with a resolution of 300 dpi. Various genres from official documents to private letters are included. The degraded quality of these images, such as uneven illumination, bleeding-through, handwritten marks, *etc*, are be a great challenge for recognition algorithms. In this case, we are going to try and use our approach to evaluating binarization quality to downstream recognition, as in [13]. The document image analysis pipeline consists of three stages: binarization – OCR – named entity recognition.

Binarization is the first stage, and three thresholding methods are used in this stage respectively. They are Otsu [14], Sauvola [16] and Wolf [21]. Otsu's method is a global thresholding method while the latter two are local thresholding methods. After all the images are converted into binary images, the resultant binary images were converted to ASCII texts by the Tesseract-3.00 [17] open source software package in the second stage. Finally, Stanford Named Entity Recognizer [5] is used in the third stage. To sum up, we have three different pipelines this way. Although our method aims to calculate precision and recall without ground truth, we still need ground truth to evaluate if our method can achieve the goal proposed in Section 3.2. Since the ground truth of the historical images are not directly available, we generate the ground truth ourselves by manual typing the text and carefully proofreading.

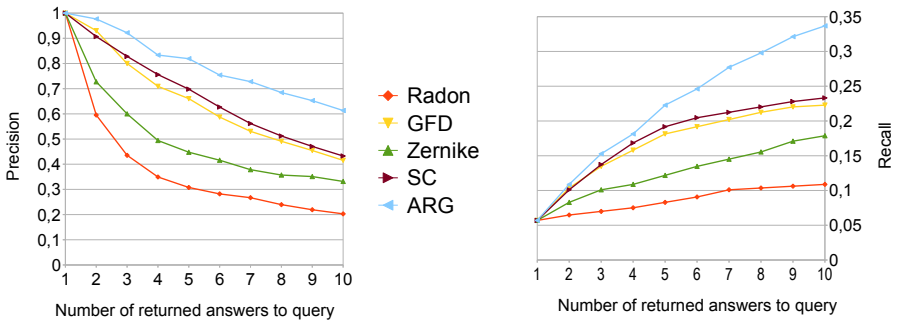

Fig. 1. Precision and Recall as reported in [10]

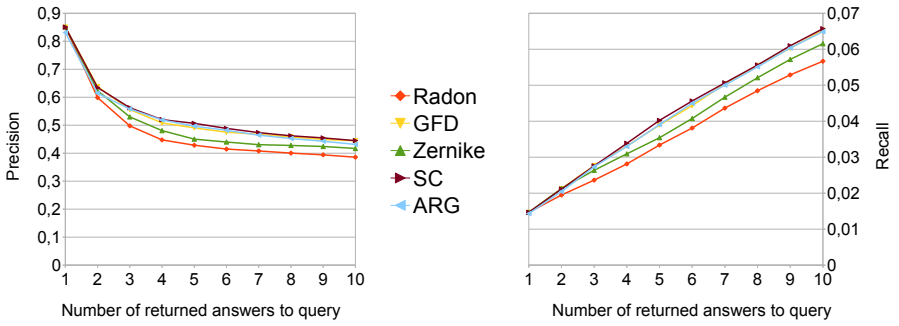

Fig. 2. Precision and Recall as computed without ground truth

Since the three different pipelines depend on three different thresholding methods, we use the names of them to stand for the three pipelines, respectively. The calculation of average precision and recall is based on the outputs of these pipelines, which are the named entity extraction results. When evaluating our method, we use two different ways to process the outputs of the three pipelines. Method I considers all the recognized named entities as 'bag-of-words", so they are organized in an alphabetical way. While Method II uses a multiple sequence alignment algorithm [18] to align the three outputs first, the original positions of these named entities are kept this way. The experiment results are shown in the following tables. From Table 4 we can see that Sauvola and Wolf beat Otsu thresholding method. The reason is obvious. Only one threshold is determined for the whole image by Otsu, while for the other two methods, different thresholds are calculated according to the grey distribution of their corresponding local windows. Table 5 and Table 6 show the results of our ground-truthless precision and recall measures using each of the metrics described before (Method I and

	Otsu	Sauvola	Wolf
Precision	0.6223	0.7715	0.7533
Recall	0.5915	0.7281	0.7230

Table 4. Average Recognition Accuracies with Ground Truth

	$S^\top$	Otsu	Sauvola	Wolf	$S_\perp$
Precision	0.4000	0.6327	0.6757	0.6722	∞
Recall	1.0000	0.5153	0.5660	0.5662	0

Table 5. Method I: Average Recognition Accuracies without Ground Truth

	$S^\top$	Otsu	Sauvola	Wolf	$S_\perp$
Precision	0.5733	0.6035	0.6450	0.6416	∞
Recall	1.0000	0.6550	0.6988	0.6957	0

Table 6. Method II: Average Recognition Accuracies without Ground Truth

II). We can see again the performance of Sauvola and Wolf is better than that of Otsu, while recognition accuracies between Sauvola and Wolf are similar. Both of them indicate that even if without ground truth, the precision and recall computed by our method is similar to those computed with ground truth.

4.3 Limitations

It would be an error to consider the approach developed in this paper as a complete and equivalent replacement of ground truth. Since the approach consists in finding an overall consensus between the tested methods, it is sensitive to collective bias. This is illustrated in the following example, taken from the raw data of the ICDAR 2011 contest described in [13].

The contest setup is quite similar than the one used in the previous section where its general aim is concerned. The difference lies in the fact that 24 different 4-stage pipelines are compared to one another. The document analysis pipelines consist in binarization – text segmentation – OCR and named entity detection, using 3 different binarization algorithms, 4 text segmentation methods and 2 OCRs.

As reported in [13], the tested pipeline is very sensitive to the quality of the used OCR engine. The results obtained using the 24 different execution paths, where every other path uses one of the 2 tested OCR engines, show that one of them clearly outperforms the other.

In order to compare these results with the approach developed in this paper we are not going to use raw F-Measure values, since the previous results have shown that there may be a significant difference in range. Instead, we are going to look at the ranking of the different methods with respect to their decreasing F-Measure. Using the Method I of the previous section, we obtain the results represented in Fig. 3.

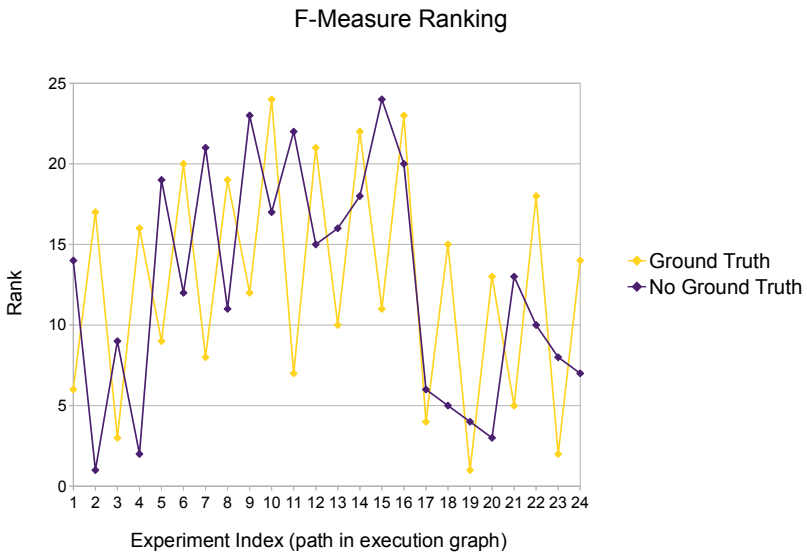

Fig. 3. Comparison of F-Measure ranking between ground truth based and ground truth-less measures.

There are two observations to be made regarding these results. The first, quite puzzling one, is that although both curves follow the same global trend, they are in complete opposite phase with respect to the oscillation induced by the OCR quality. Second, a closer look at the figures shows that there is an averaging effect operating. Since both OCR engines are consistent in their errors, they introduce a bias in the consensus values computed by our method, thus pulling the F-Measures toward an average value.

By separating the results in function of the OCR, we observe that we obtain much more coherent, and more encouraging results, in line with what we observed in the previous sections. Fig. 4 shows that the overall ranking pattern is preserved when projecting the F-Measures by OCR. It is clear, on the other hand, that there is no total equivalence between the ranking obtained with ground truth and the one obtained without. However, global ranking (top – middle – bottom tiers) is very consistent.

These results very much recall the experiments reported in [11] in the case of classifier fusion. Although there are some fundamental difference in combining binary classifiers by majority voting and the approach developed here, the underlying formalism is very much the same. The main differences are that one the one hand, we are not applying a full majority vote, in our case. Although the probability of an individual document being relevant depends on the number of systems having classified it as such, and therefore relates to a voting system, this probability is not truncated to either 0 or 1, as it would have been, in the

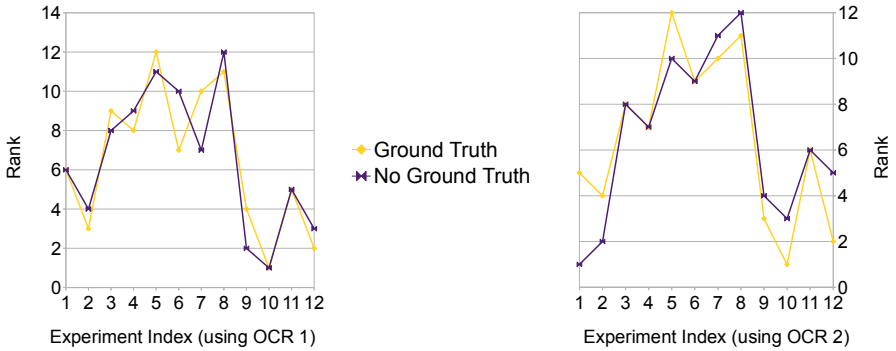

Fig. 4. Comparison of F-Measure ranking between ground truth based and ground truth-less measures in function of the underlying OCR method.

case of majority voting. On the other hand, the goal of classifier fusion is to obtain a new classifier, performing better than its individual contributors. This is not the aim in our case, where we just want to express a ranking between the different classifiers. One may argue, however, that the classifier obtained by majority voting may provide a theoretical boundaries to the reliability of the probabilistic Precision and Recall values presented in the previous sections. The math behind this assumption needs to be further developed and assessed.

5 Extensions

The probabilistic model developed in section 3.2 makes the simplifying assumption that both all data and all methods have uniform confidence values: no method is considered more reliable than the others, and all data either belongs or does not belong to the query results.

5.1 Method Weighting

Our model is capable of integrating ground truth, and may even handle uncertain ground truth (*e.g.* coming from reliable, but not fully verified human annotations). To that avail, the ground truth can be integrated as being the result of some "oracle" system $S_{\mathcal{O}}$, and the probability of a document δ_i belonging to $\Delta_\star^+$, as expressed in (10) should be slightly modified.

$$P(\delta_i) = \sum_{k=1\ldots s,\perp,\top,\mathcal{O}} S_k(\delta_i)\,\kappa_{S_k} \tag{13}$$

Where κ_{S_k} is the confidence value associated to system S_k, and $\sum_k \kappa_{S_k} = 1$. In the case we previously developed, all systems had equal confidence, and

$\kappa_{S_k} = \frac{1}{s+2}$. In case of one or more oracle systems $S_{\mathcal{O}}$, its confidence value can be adapted consequently. Setting $\kappa_{S_{\mathcal{O}}} = 1$ would be equivalent to the commonly admitted use of (undisputed) ground truth. Moreover, in cases where multiple versions of reference interpretations exist [12] it now becomes possible to handle varying degrees of ground "truth"[4] by attributing appropriate values to the corresponding oracle systems.

5.2 Confidence Voting

Similarly, it is now possible to extend the approach beyond binary attribution of documents to queries, since systems can very well express their confidence of a document being relevant to a query with a probability value. All formulae and tables developed in section 3 remain valid in this context, and the probabilistic precision and recall computations are directly transposable to the case where individual documents for a given query have a probability of pertinence rather than a binary valuation. Furthermore, this can be combined with the method weighting expressed in the previous section.

6 Conclusion and Future Work

In this study we have presented how to compute precision and recall without presence of formally identified ground truth. Results indicate that this measure is coherent with real, ground truth based precision and recall measures, although it can obviously not infer ground truth and achieve the exact same performance as if ground truth were actually available.

On the other hand, the mathematical framework supporting the computation of probabilistic precision and recall has the interesting property to handle a continuum of situations ranging from perfectly known and available ground truth, over uncertain ground truth to total absence of it.

The major condition for this method to work, however, is that it has access to a number of competing systems that are providing multiple possible answers to the same queries, each of them supposedly trying to achieve the best possible result. This is particularly well suited for large scale performance evaluation contexts like the one experimented in [13] and formally developed in [12]. Its use in larger scale experiments will also contribute in further establishing the exact differences between full use of ground truth and the approximation presented in this paper.

Further work and development will consist in establishing how to rank or take into account user-contributed "partial" ground truth, especially considering "yes/no/unknown" information. Currently, our framework makes the assumption that all systems operate on the exact same set of queries and documents. There exist models that are capable of integrating overlapping or dissimilar

[4] Since there cannot exist varying degrees in truth, we prefer the term of "interpretation".

query and document sets [6]. It would be interesting to confront them to our approach and to study how partial ground truth (for instance, resulting from crowd-sourced contributions) can be integrated and improve overall performance of our approach.

Acknowledgements

The authors would like to acknowledge Dr. Santosh K.C. for having provided the experimental data, used in Section 4.1. They also thank Prof. Dan Lopresti for having pointed them to voting approaches in classifier fusion.

Bart Lamiroy was a visiting scientist at Lehigh University in 2010–2011. This work was conducted at the Computer Science and Engineering Department at Lehigh University and was supported in part by a DARPA IPTO grant administered by Raytheon BBN Technologies.

References

1. Library of congress, http://memory.loc.gov/
2. Antonacopoulos, A., Karatzas, D., Bridson, D.: Ground truth for layout analysis performance evaluation. In: Bunke, H., Spitz, A. (eds.) Document Analysis Systems VII, Lecture Notes in Computer Science, vol. 3872, pp. 302–311. Springer Berlin / Heidelberg (2006)
3. Baraldi, A., Bruzzone, L., Blonda, P.: Quality assessment of classification and cluster maps without ground truth knowledge. Geoscience and Remote Sensing, IEEE Transactions on 43(4), 857 – 873 (april 2005)
4. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. MACHINE LEARNING 36, 105–139 (1999)
5. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL. The Association for Computer Linguistics (2005)
6. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. Advances in Information Retrieval 27th European Conference on IR Research ECIR 2005 Santiago de Compostela Spain March 2123 2005 Proceedings 3408, 345–359 (2005)
7. Grosicki, E., Carree, M., Brodin, J.M., Geoffrois, E.: Results of the rimes evaluation campaign for handwritten mail processing. In: Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on. pp. 941 –945 (july 2009)
8. Hauff, C., Hiemstra, D., de Jong, F., Azzopardi, L.: Relying on topic subsets for system ranking estimation. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 1859–1862. CIKM '09, ACM, New York, NY, USA (2009)
9. Kankanhalli, M.S., Mehtre, B.M., Wu, J.K.: Cluster-based color matching for image retrieval. Pattern Recognition 29, 701–708 (1995)
10. K.C., S., Lamiroy, B., Wendling, L.: Spatio-structural symbol description with statistical feature add-on. In: The Ninth International Workshop on Graphics Recognition (2011)
11. Kuncheva, L., Whitaker, C., Shipp, C., Duin, R.: Limits on the majority vote accuracy in classifier fusion. Pattern Analysis & Applications 6, 22–31 (2003)

12. Lamiroy, B., Lopresti, D., Korth, H., Jeff, H.: How carefully designed open resource sharing can help and expand document analysis research. In: Agam, G., Viard-Gaudin, C. (eds.) Document Recognition and Retrieval XVIII. SPIE Proceedings, vol. 7874. SPIE, San Francisco, CA USA (January 2011)
13. Lamiroy, B., Lopresti, D., Sun, T.: Document Analysis Algorithm Contributions in End-to-End Applications. In: 11th International Conference on Document Analysis and Recognition - ICDAR 2011. International Association for Pattern Recognition, Beijing, China (Sep 2011)
14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics 9(1), 62–66 (Jan 1979)
15. van Rijsbergen, C.J.: Information Retrieval. Butterworth (1979)
16. Sauvola, J.J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recognition 33(2), 225–236 (2000)
17. Smith, R.: An overview of the tesseract ocr engine. In: ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition. pp. 629–633. IEEE Computer Society (2007), http://www.google.de/research/pubs/archive/33418.pdf
18. Thompson, J.D., Higgins, D.G., Gibson, T.J.: Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22(22), 4673–80 (1994)
19. Tombre, K., Lamiroy, B.: Pattern recognition methods for querying and browsing technical documentation. In: Ruiz-Shulcloper, J., Kropatsch, W. (eds.) Progress in Pattern Recognition, Image Analysis and Applications, Lecture Notes in Computer Science, vol. 5197, pp. 504–518. Springer Berlin / Heidelberg (2008)
20. Valveny, E., Dosch, P., Winstanley, A., Zhou, Y., Yang, S., Yan, L., Wenyin, L., Elliman, D., Delalandre, M., Trupin, E., Adam, S., Ogier, J.M.: A general framework for the evaluation of symbol recognition methods. International Journal on Document Analysis and Recognition 9, 59–74 (2007)
21. Wolf, C., Doermann, D.S.: Binarization of low quality text using a markov random field model. In: ICPR (3). pp. 160–163 (2002)