Triage of OCR Results Using 'Confidence' Scores

Prateek Sarkar[a] Henry S. Baird[a]
John Henderson[b]

[a]Document Image Decoding, Xerox Palo Alto Research Center, Palo Alto, CA, USA
[b]Document Imaging Service Center, Xerox Business Solutions, Mitcheldean, UK

ABSTRACT

We describe a technique for modeling the character recognition accuracy of an OCR system – treated as a "black box" – on a particular page of printed text based on an examination only of the output top-choice character classifications and, for each, a "confidence score" such as is supplied by many commercial OCR systems. Latent conditional independence (LCI) models perform better on this task, in our experience, than naive uniform thresholding methods. Given a sufficiently large and representative dataset of OCR (errorful) output and manually "proofed" (correct) text, we can automatically infer LCI models that exhibit a useful degree of reliability. A collaboration between a PARC research group and a Xerox legacy conversion service bureau has demonstrated that such models can significantly improve the productivity of human proofing staff by "triaging" – that is, selecting to bypass manual inspection – pages whose estimated OCR accuracy exceeds a threshold chosen to ensure that a customer-specified per-page accuracy target will be met with sufficient confidence. We report experimental results on over 1400 pages. Our triage software tools are running in production and will be applied to more than 5 million pages of multi-lingual text.

Keywords: Triage, Optical character recognition, OCR quality, scan conversion, service bureau, latent conditional independence model

1. INTRODUCTION

The cost of conversion of paper documents, through image acquisition (scanning) and recognition (OCR), into machine-legible coded form (such as ASCII or Unicode) is often dominated by the expense of manual post-OCR correction. This occurs because the present state of the art of OCR can only rarely yield uniformly high accuracies across collections of dissimilar documents, for example, those containing a variety of typefaces, languages, layout formats, and image qualities.

If it were possible automatically to decide which of the documents emerging from OCR possess acceptable accuracy, then these documents could skip manual correction, lowering production costs. We call this a *triage* decision, in analogy with the practice of emergency medical rescue staff who are trained to classify injured individuals into those who need immediate medical attention and those who do not. Medical triage tries to maximize the number of survivors given finite care facilities; OCR triage tries to maximize the number of pages that skip correction given a fixed uniform accuracy target.

We describe a method for triage based on the availability of "confidence scores" attached to each character interpretation reported by the OCR system: these scores are, typically, integers with a narrow range (in our case [0,255]) of values which bear some relation to the accuracy of that interpretation. We treat the OCR system as a "black box": that is, we do not rely on any prior knowledge of the internal algorithms and heuristics that it employs, nor upon any documented or rumored motivation for the scores. In particular we do not interpret them as posterior probabilities of correctness. Experience has led us, in fact, to be so loathe to depend on prior models of their behavior that we view them as categorical rather than numerical. In other words our models do not assume any explicit functional relationship between the numerical value of confidence scores and probabilities of error.

Our principal goal is to build triage tools which can be applied fully automatically and at high speed in a high-throughput service bureau setting. The tools should be automatically trainable given sets of OCR results and corresponding high-quality "ground truthed" files. The tools should also assist production supervisors in

the crucial choice of operating points that trade off cost versus quality: that is, the rate at which documents are triaged versus the risk that a triaged page does not meet a given accuracy target. As far as we are aware there have been no previously published studies of triage in our sense. We are aware that triage has been used in a number of OCR service bureaus.[1]

Our technical approach is motivated by latent conditional independence models (LCI models) (see, for example,[2]). A bivariate LCI model is described in detail in Section 2. The application of this model to triage is described in Section 3.

We have applied experimental prototypes of our triage tools to over 1400 pages of patent literature documents provided by the European Patent Office (in three languages: English, French, and German) to the Xerox Business Services Document Imaging Services Center (DISC) in Mitcheldean, U.K. These experiments are described in Section 4. A few possible extensions and conclusions are in Sections 5 and 6.

2. BIVARIATE LATENT CONDITIONAL INDEPENDENCE MODEL

A discrete bivariate distribution can be represented in general by a two-dimensional array $P = [p_{m,n}]$ whose elements are non-negative and add up to one. If the two discrete random variables can assume M and N distinct values respectively (*i.e.*, $1 \leq m \leq M$, and $1 \leq n \leq N$), such a representation comprises of $M \times N$ parameters bound by one linear constraint ($\sum_{m,n} p_{m,n} = 1$). It is often helpful to describe discrete bivariate distributions (DBD) with a parsimonious model so that, in applications, fewer parameters need to be estimated. One way to achieve this is a latent conditional independence (LCI) model where

$$
\begin{aligned}
P &= [\boldsymbol{u}_1 \ldots \boldsymbol{u}_K]\ \mathrm{diag}(a_1, \ldots, a_K)\ [\boldsymbol{v}_1 \ldots \boldsymbol{v}_K]^T = \sum_{k=1}^{K} a_k \boldsymbol{u}_k \boldsymbol{v}_k^T \\
&\text{where } \sum_{k=1}^{K} a_k = 1, \qquad 0 \leq a_k \leq 1 \\
&\boldsymbol{u}_k = [u_{k,1} u_{k,2} \ldots u_{k,M}]^T, \qquad 0 \leq u_{k,m} \leq 1, \qquad \sum_{m=1}^{M} u_{k,m} = 1 \\
&\boldsymbol{v}_k = [v_{k,1} v_{k,2} \ldots v_{k,N}]^T, \qquad 0 \leq v_{k,n} \leq 1, \qquad \sum_{n=1}^{N} v_{k,n} = 1
\end{aligned}
\tag{1}
$$

This model has $K(M + N + 1)$ parameters bound by $2K + 1$ linear constraints. If K is much smaller than both M and N, then the number of parameters in an LCI model is much smaller than the general DBD representation. As an example, in our experiments M and N were 156 and 162 respectively, while K was set to 5.

It may assist intuition to consider a generative interpretation of the LCI model. Let $P = [p_{m,n}]$ describe a source that randomly picks a "factor" k (latent or unobserved) with probability a_k , where the distribution is over K possible factors. The source then independently generates two numbers (observed) – m ($1 \leq m \leq M$) with probability $u_{k,m}$, and n ($1 \leq n \leq N$) with probability $v_{k,n}$. Conditioned on any factor, k, the joint distribution of (m, n) is an independent combination (outer product) of the marginals represented by $u_{k,m}$ and $v_{k,n}$. However, the overall joint distribution of (m, n) given by P is not an independent combination of the overall marginals represented by $\sum_{k=1}^{K} a_k u_{k,m}$ and $\sum_{k=1}^{K} a_k v_{k,n}$ respectively. m and n are therefore not independent, but generated according to a convex sum of distributions each of which is a simple independent combination of marginals.

In probability notation, we can write $u_{k,m}$ as $p(m|k)$, $v_{k,n}$ as $p(n|k)$, and a_k as $p(k)$ and observe that the LCI model specifies

$$
p_{m,n} = p(m, n) = \sum_{k=1}^{K} p(k) p(m, n|k) = \sum_{k=1}^{K} p(k) p(m|k) p(n|k)
$$

Although latent conditional independence can be applied to model joint distributions of more than two random variables, they have been most popularly applied to bivariate distributions in contingency table analysis,[2] statistical exploration of grammatical relationships of words and semantic disambiguation,[3,4] probabilistic latent semantic indexing.[5] While LCI models have a probabilistic framework, similar factorizations have been explored with linear algebra techniques in positive matrix factorization that has been applied to face recognition,[6] and environmental studies.[7] If we do not restrict model parameters to be positive, methods such as singular value decomposition[8–10] can be applied for dimension reduction. An advantage of LCI is the intuitive appeal of the probability parameters as relative frequencies.

3. TRIAGING WITH LCI MODELS

Our OCR system generates, for each page in the input document, a string of pairs (m, n) where m is a character label, and n is a confidence score. Each pair is either "correct" (label m matches the ground-truth) or "error". Given the observation we compute the a posteriori probability of error $p(\text{error}|m, n)$.

$$p(\text{error}|m,n) = \frac{p(m,n|\text{error})p(\text{error})}{p(m,n|\text{error})p(\text{error}) + p(m,n|\text{correct})p(\text{correct})} \tag{2}$$

Each of the two conditional joint distributions over (m, n) (conditioned on "correct" and "error" respectively) are modeled as latent conditionally independent with K factors, where K is chosen experimentally to avoid over-fitting. The LCI model, in principle, allows for different groups of character labels (lower case, numeric, punctuation, etc.) to have different distributions of confidence score, where both the character groups and confidence-score statistics can be learned automatically from training data.

The average a posteriori probability of error computed over all (m, n) pairs in page of OCR output is used as a OCR quality score for the page. We choose a threshold score and triage all pages with lower average a posteriori probability of error than the threshold, *i.e.*, declare them as "good quality" pages. The true error rate of a page can be computed by manual labeling of erroneous output or by comparison to ground-truth using dynamic programming (see for example[11]). The average a posteriori probability of error is an estimated "quality score" which is correlated, albeit not perfectly, with the true error rate. By choosing a threshold on this score, we can trade off the truly good quality pages (true error rate below specified threshold) that are not triaged (false alarms) for the bad quality pages that are triaged (false hits).

Training: The LCI models that represent the class-conditional distributions $p(m, n|\text{correct})$ and $p(m, n|\text{error})$ have to be trained from samples of (m, n) pairs labeled as "correct" or "error". The labeling is done by aligning OCR output to ground-truth at the character level using a string matching algorithm (see, for example,[11]). For each category, "correct" and "error", we find an LCI-model that fits the data best in the maximum likelihood sense. The training process is formulated in terms of the Expectation Maximization (EM) algorithm.[3,12] The sufficient statistics for estimating the multinomial LCI model are the counts (number of occurrences) of each (m, n) pair for each category ("correct"/"error"). In practice, we run 50 iterations of the EM algorithm without directly measuring convergence criteria.

Validation: For a target true error rate, we can experiment on held out labeled data (a validation set) to generate an operating characteristic that empirically quantifies the trade off between false hit and triage rates. This facilitates the choice of a threshold to be used during triage that will maximize the number of pages triaged consistent with a reasonably low risk that too many pages will exceed the error target. The error target is set by the customer of the service bureau, and the acceptable risk level is chosen by the service bureau managers.

4. EXPERIMENTS AND RESULTS

1413 scanned pages provided to Xerox by the European Patent Office were processed with the TextBridge OCR system at DISC. The text on the pages was manually keyed in by DISC staff to provide ground-truth data. The OCR output (character labels and confidence scores) for each page was then aligned to ground-truth text using dynamic programming (such as Unix `diff`) to obtain "error" or "correct" labels for each (character label, confidence score) pair.

Table 1: Operating characteristics on validation data : LCI model based triaging method.

Operating threshold	Triage rate (% of total)	False hit rate rate%	Good and triaged	Bad but triaged ('false hits')	Good but manually corrected ('false alarms')	Bad and manually corrected	Total tested
0.002	0.00	0.00	0	0	469	238	707
0.004	3.25	0.00	23	0	446	238	707
0.006	11.60	0.00	82	0	387	238	707
0.008	22.21	0.00	157	0	312	238	707
0.010	35.08	0.85	242	6	227	232	707
0.011	40.88	1.27	280	9	189	229	707
0.012	47.38	2.55	317	18	152	220	707
0.014	55.73	4.95	359	35	110	203	707
0.016	67.33	8.35	417	59	52	179	707
0.018	74.26	12.16	439	86	30	152	707

The 1413 pages were then randomly permuted and partitioned into two samples of 706 and 707 pages respectively. The first sample was used for training the LCI models for "error" and "correct" groups, each model having 5 factors ($K = 5$).* The second sample was used for validation.

Table 1 shows the operating characteristics (false-hit rate vs. triage rate) on the validation set. The false hit and triage rates were computed for different values of the threshold average a posteriori probability of error, and a target quality of $\leq$1.5% OCR error.

Of the 707 pages in the validation set, OCR output on 469 pages (66%) met the target quality. With an operating threshold of .011 on the average a posteriori probability of error, 289 pages (41%) were triaged and bypass manual intervention. Of these, 280 were among the 469 "good quality" pages. Nine of the triaged pages did not actually meet target quality, representing a false hit rate of 1.3%, as an example of customer-specified tolerance.

Thus our triage method, applied to this set of page images, has been shown capable of bypassing manual correction for 41% of the document stream fully automatically and without compromising the quality goals established by the customer.

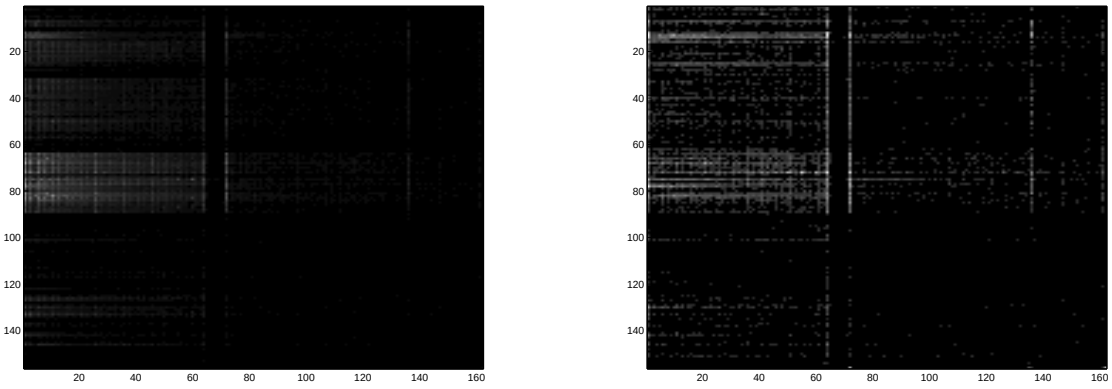
Table 2: Operating characteristics on validation data : simple "suspect" threshold based triaging method.

"Suspect" threshold = 64		
Operating threshold	Triage rate (% of total)	False hit rate rate%
100.00	0.00	0.00
99.80	38.19	6.51
99.60	46.25	9.34
99.40	51.06	11.60
99.20	54.46	13.44

"Suspect" threshold = 16		
Operating threshold	Triage rate (% of total)	False hit rate rate%
100.00	0.00	0.00
98.00	8.06	0.00
96.00	16.12	0.28
94.00	26.73	0.71
92.00	37.48	2.26
90.00	46.82	3.54

*It is not necessary, of course, that both distributions be modeled with the same number of factors.

Table 2 shows the results of triaging with a simpler estimate of OCR quality. Here we compare the confidence score accompanying each character in the OCR output to a preset threshold, and assume that characters with lower confidence are in error. The ratio of the number of such characters to the total number of characters in the OCR output is used as the simple measure of OCR quality. If this measure is treated as an estimated character error rate, very few pages (less than 8%) can be triaged. We can obtain better results by calibrating this measure using ground-truth data. As with our LCI model based method, we can compute operating characteristics (Table 2) on validation data (note that there is no training, per se, in this simpler method). On the left is the result of using the "suspect" threshold on the confidence score recommended by the vendor of the OCR system. On the right are the results of using the "suspect" threshold we found to best (qualitatively and empirically). Triaging with LCI models yields still better results.



(a) Correctly classified characters (b) Misclassified characters

Figure 1. Empirical distributions of confidence score (column index) conditioned on each character class (row index). Lighter cells indicate higher observed frequency.

Figure 1 illustrates the observation that motivated the use of LCI models for OCR quality estimation. The two images represent the joint (character-class, confidence score) distribution conditioned on the OCR proofing outcome ("correct" or "error"). The same "confidence score" (represented by a column in each figure) is associated with different relative frequencies of "correct" and "error" for different character classes. Each character class is represented by a row. We can consider a specific example with two character classes: lower case "a" and "," (comma). If we triage characters (rather than whole pages), on the basis of a chosen threshold on the confidence score (32, for this example), the difference between the two character classes is shown in the Table 3. Notice the stark difference in the numbers for the characters which are suspected to be erroneous. While close to 90% of these are actually correct for class "a", only about 50% are actually correct for class ",". This demonstrates the weakness of a using a fixed "suspect" threshold on the confidence score.

Table 3: Detecting errors with a fixed threshold on confidence score: differences between character classes

		Character class	
True OCR result	Suspected OCR result	"a"	","
Correct	Correct	94.8%	89.2%
Correct	Error	4.5%	3.8%
Error	Correct	0.3%	3.0%
Error	Error	0.4%	4.0%

4.1. Triage performance in production

At the time of this article going to press, our triage method is being applied, in full production, in an EPO contract covering six million pages of pre-scanned text documents. The circumstances under which the triage

tools are being used in production are somewhat different in detail from the earlier discussion. For example, they are affected by an evolution in the customer's quality formula, and DISC's implementation of it. Also the models were retrained progressively on more available data.

In the production batches of Patent Literature text recovery, approximately 800,000 pages have been processed through the triage tools. Of these, approximately 400,000 were processed using a "safe" threshold on the triage operating curve and achieved approximately 75% triage rate; where "safe" is $< 1\%$ false hit rate as measured on a 700 page sample.

A further 400,000 pages were processed using a "risky" threshold on the triage operating curve and achieved approximately 90% triage rate; where "risky" might be 8% false hit rate (based on the validation data for the model). The business reason for operating triage in this "risky" way is to concentrate a limited proofing resource on the most probable poorest quality OCR results.

Of the 800,000 pages processed in this way, some 1000 randomly selected pages were subjected to the customer's own quality checks and passed with acceptable text quality.

As we better understand (and can codify) some of the subjective leniency being applied during the customer's quality checks, we anticipate that we will be able to move our "safe" threshold on the triage operating curve to achieve greater triage rates (and our "risky" threshold will yield a lower false hit rate).

As we obtain more and more ground truth pages, we anticipate that the triage model will become an even better predictor, and also envisage that we will be able to generate language-specific triage models which we hope will be an even more reliable predictor (where the source language is known and has been modeled).

5. FUTURE EXTENSIONS

So far we have explored only a two-dimensional model, where the dimensions are character class and 'confidence' score. Our LCI model training and classification algorithms, however, can cope with higher dimensions. Thus they could exploit knowledge of any other characteristic of documents including typeface, language, content topic, and image quality. This will require either metadata or specialized automatic classification, which in some cases is immediately feasible.

Our triage training depends on minimum-cost string matching with simple per-character costs and no language model richer than character unigrams, and as a result it has some potential vulnerabilities. It cannot detect, and so cannot model, deletion errors, which could be significant in practice. Perhaps n-gram character language models can mitigate this. In our experiments so far, we seem not to suffer much from this obliviousness to deletions: perhaps we are protected by frequent co-occurrence of deletion errors with segmentation errors, and therefore substitution errors, which we *can* detect.

In this study we have triaged documents one full page at a time. Our existing method, with few or no refinements, may be sensitive enough for "on-line" triage in which, as the manual correction operator works down a page, overall page quality is continually re-estimated, so that the operator can be instructed to stop when the per-page accuracy target has, with sufficiently high statistical confidence, been reached. In addition, it is conceivable that a similar (but perhaps more refined) method could triage shorter passages, even individual words, well enough to direct manual correction to the most urgent corrections first. These two strategies are complementary and may be mutually enhancing.

6. CONCLUSION

We have shown that the per-character 'confidence' scores often provided by commercial OCR systems can be exploited fully automatically to 'triage' documents on which OCR has performed well, and so to reduce post-OCR manual correction costs significantly without sacrificing uniformly near-perfect accuracy across the document stream. Our approach, using latent conditional independence models, assumes that 'confidence' scores may behave differently for different character classes: in practice, we can identify subsets of character classes for which the scores behave similarly, allowing us to combine their statistics and so smooth otherwise uselessly sparse training data. Given a sufficiently large supply of accurate "ground-truth" text corresponding

to OCR results (which is commonly available in document conversion service bureaus such as Xerox's DISC), elementary string matching can label all the data needed to train the triage models.

The steady (but slow) improvement in overall accuracy of OCR systems offers some hope to users who are poorly served today. But no OCR system is able today, or, we believe, will be able in the foreseeable future, to provide uniformly near-perfect accuracy across a wide range of dissimilar documents. Thus, anyone who desires to scan and convert a large legacy document collection with uniformly high accuracy must expect to pay for post-OCR manual correction. This can be carried out successfully only by technically sophisticated service bureaus with trained staff. Service bureaus do not develop their own OCR systems; in fact, today they all use more or less comparable recognition technologies. Our triage method is one way for a particular conversion service bureau to achieve a significant efficiency advantage over its competitors without sacrificing high quality levels. Thus our triage method represents an important, and as we have argued, a possibly expanding, family of document image analysis methods able to relieve one of the most stubborn technical and commercial obstacles to high-quality large-scale legacy document-collection conversion.

Acknowledgments

We are grateful for the assistance of PARC colleagues Kris Popat, Tom Breuel, and Yoram Gat, and DISC colleagues Andrzej Zydron and Hugh Stabler for helpful discussions, assistance with experiments, the provision of crucial data, etc. We are also indebted to Mark Pettit, Mark Stefik, Julian Blackler, and Shaun Pantling for their vision, encouragement, patience, and material support.

REFERENCES

1. Roger Bradford, SAIC. Personal communication, September 2001.
2. M. J. Evans, Z. Gilula, and I. Guttman, "Latent class analysis of two way contingency tables by bayesian methods," *Biometrika* **76**(3), pp. 557–563, 1989.
3. M. Rooth, "Two-dimensional clusters in grammatical relations," in *Inducing lexicons with the EM algorithm, AIMS Report 4(3)*, 1998.
4. M. Rooth, S. Riezler, D. Prescher, G. Caroll, and F. Beil, "EM-based clustering for nlp applications," in *Inducing lexicons with the EM algorithm, AIMS Report 4(3)*, 1998.
5. T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the Twenty-second Annual International SIGIR Conference on Research and Development in Information Retrieval*, (Berkeley, California), 1999.
6. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience* **3**(1), pp. 71–86, 1991.
7. Y.-L. Xie, P. K. Hopke, P. Paatero, L. A. Barrie, and S.-M. Li, "Identification of source nature and seasonal variations of arctic aerosol by positive matrix factorization," *Journal of Atmospheric Science* **56**, pp. 249–260, 1999.
8. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science* **41**(6), pp. 391–407, 1990.
9. W. T. Freeman and J. B. Tenenbaum, "Learning bilinear models for two-factor problems in vision," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR '97)*, (Puerto Rico, USA), June 1997.
10. J. B. Tenenbaum and W. T. Freeman, "Separating style and content," in *Advances in Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan, and T. Petsche, eds., Morgan Kaufmann, San Mateo, CA, 1997.
11. R. A. Wagner and M. J. Fischer, "The string to string correction problem," *Journal of the Association of Computing Machinery* **21**, pp. 168–178, 1974.
12. A. P. Dempster, M. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society* **39**(1), pp. 1–38, 1977.