An Evaluation of Information Retrieval Accuracy with Simulated OCR Output

W.B. Croft[†], S.M. Harding[†],
K. Taghva[‡], and J. Borsack[‡]
[†]Computer Science Department
University of Massachusetts, Amherst
[‡]Information Science Research Institute
University of Nevada, Las Vegas

Abstract

Optical Character Recognition (OCR) is a critical part of many text-based applications. Although some commercial systems use the output from OCR devices to index documents without editing, there is very little quantitative data on the impact of OCR errors on the accuracy of a text retrieval system. Because of the difficulty of constructing test collections to obtain this data, we have carried out evaluations using simulated OCR output on a variety of databases. The results show that high quality OCR devices have little effect on the accuracy of retrieval, but low quality devices used with databases of short documents can result in significant degradation.

1 Introduction

Text-based information systems have become increasingly important in business, government, and academia. In many applications, the source of the text is not documents from word processors, but instead documents in their original paper form. Although imaging systems provide a simple means of storing these documents and retrieving them through manually assigned keywords, full-text access will in general be much more effective. In order to get from paper documents to full-text retrieval, OCR will be a crucial part of the process.

For printed documents, OCR techniques can recognize words with a high level of accuracy. The number of errors, however, is such that a substantial amount of human editing is required to make the text output suitable for archival and display. The cost of this editing is a major factor in most current OCR applications. For applications that focus on automatic indexing and retrieval, it is possible that the raw word accuracy of the OCR output may be sufficient, and that expensive editing can be avoided. Some text retrieval systems have taken this approach, combining OCR for indexing and imaging for display.

From an information retrieval point of view, the main issue is the impact of OCR indexing errors on the accuracy or effectiveness of the system. The accuracy of an IR system is typically measured using precision and recall[1] with a test collection

[1]Precision is the percentage of retrieved documents that are relevant, and recall is the percentage of relevant documents that are retrieved, for a

consisting of a document database, queries, and relevance judgements for those queries [6]. Despite the fact that there are commercial retrieval systems that use OCR input, the lack of availability of test collections means that there is very little published data about the effect on retrieval accuracy. In a recent study, Taghva, Borsack, Condit and Erva [8] did a comparison of the output of a retrieval system using a document database created using scanning and OCR, and the same database with errors removed by editing. The comparison was done by comparing the overlap of the retrieved documents for a set of test queries. The results showed that the output was very similar, but the study was limited by the small size of the database, the lack of relevance judgements, and the use of a Boolean logic retrieval system.

What is really needed is data showing the effect on recall and precision of OCR indexing with a range of databases, and with a retrieval system that produces ranked output. Ranking systems have clear advantages relative to Boolean logic systems in terms of average effectiveness, and can use simple query formulations without Boolean operators. The fact that they are based on partial matching may in fact make them less susceptible to OCR errors.

The problem with obtaining this data is that it is extremely expensive to build test collections, and even more expensive to build them for OCR experiments. In this paper, we describe our first approach to obtaining accuracy data using simulated OCR output for a range of databases. The simulation is done using data about word error rates for a variety of devices tested at the UNLV Information Science Research Institute (ISRI) [5]. Although the simulation is not completely accurate, it is the first study about OCR and retrieval effectiveness where the results have some basis on actual OCR data.

In the next section, we describe how the simulation was done. The third section gives details on the test collections used, their characteristics, and the experiments that were performed. The results of the experiments are summarized in the fourth section, and the final section suggests future directions for this work.

2 The OCR Simulation

The data that was used for the simulation was a study of character and word error rates for a range of OCR devices and software [5]. The study was done using a sample of 460 document pages from a Department of Energy test database. The word error rates that were reported in this study are not uniformly distributed throughout the document. In fact, error rates are summarized by device, by page type, by word type, and by word length.

The word types distinguished were stopwords and non-stopwords, where stopwords are simple function words such as "and", "the", "of". Pages were divided into groups based on the number of OCR errors on them. Some pages, presumably those with high-quality initial images, had virtually no errors on them, whereas others, which either had poor quality originals or poor quality scans, had large numbers of errors. Statistics were reported for the percentage of pages in each group, and for word error rates by word length within each page group. Table 1 shows some of this data.

To produce the simulated OCR test collections, we assumed that the statistics reported in this study would apply to all the document types in the collections we used. We also assumed that word length and type (stopword or non-stopword) were the only factors in determining the chance of an OCR error in a particular page group. A refinement would be to give higher prob-

particular query.

Table 1: Page quality groups defined for simulating OCR error rates on text retrieval performance. Average accuracy by page group for the two OCR systems used as the basis for the simulations are in final two columns (OCR1 and OCR2). The standard page size used for the simulation runs was 1778 characters/page.

Page Quality Group	Number of Pages	Number of Characters	Accuracy OCR1 (%)	Accuracy OCR2 (%)
1	80	165,110	98.8	99.9
2	77	163,019	96.7	99.0
3	85	162,367	93.1	98.3
4	96	163,176	85.5	96.7
5	122	164,274	62.1	88.3
Total	460	817,946		

abilities of error to those words which contain character strings that are commonly confused by OCR devices. Some data about these common confusions is available, but we decided to ignore this factor in our initial experiments.

Two other important assumptions were made. The first is that all OCR errors result in a corrupted word that is discarded and not indexed. In actual OCR data, valid words are sometimes transformed by errors into other valid words, although it is unlikely with longer words and difficult to simulate accurately. More importantly, OCR errors would typically result in words that are similar to spelling errors that would be indexed. Because the retrieval process is driven by matches with un-corrupted query words, there not much difference between discarding a word and generating a misspelling. There are, however, some effects. The indexes generated in a real OCR environment will contain a large number of these "misspelled" words. When data is presented on index sizes in this study, those words are ignored. In addition, there is some chance that a commonly misrecognized word will affect the term frequencies that are used to normalize probability estimates in the retrieval system. This is more likely to happen with incorrect zoning, which leads us to the next assumption.

In this study, we ignore errors caused by incorrect zoning, that is, attempting to do OCR on figures, maps, etc. A recent study has shown that the type of OCR errors generated by incorrect zoning can have an impact on retrieval performance [7]. In other words, the zoning accuracy of an OCR system is an important factor in determining retrieval performance independent of the word accuracy of the system. This happens because zoning errors can generate large numbers of "misspelled" words that affect a retrieval system's probability estimates. The study reported here focuses on the impact of the word accuracy rates of the OCR system.

To generate a simulated OCR database, then, an IR test database is indexed using standard techniques such as tokenization, stemming, and stopword removal [6, 1]. During this process, the text of a document is randomly assigned to page groups, and index words are randomly discarded according to the error rates for that page group and word length. The result of the OCR simulation is a database in which documents may be indexed by fewer terms than the original database.

More specifically, two sets of statistics were used, representing the best and the worst OCR performance observed in the UNLV tests. The five page group classes, representing different levels of page quality, were assigned randomly to the text input stream during the indexing process. The page group and the corresponding set of character recognition error rates remained in effect for the duration of a page. The probability of being in any particular page group was determined from the total number of characters for the page group, divided by the total number of characters for all page groups, which was close to 1 in 5, but not exactly so.

Page size was a constant and determined from a calculation dividing the total number of characters in the data set by the total number of pages. The character counts for each page group were a part of the UNLV data. A random number generator producing values between 0 and 1 determined page group assignment when a page full of characters had been read.

Simulation of OCR word errors was done by a randomly assigned number between 0 and 1 reflecting the probability of error for a word of its length and page group. If the number fell in the error range, it was discarded, otherwise, processed as usual. Word positions, which are used in proximity operators, were counted whether discarded or not.

The results of this process on four test collections are given in the next section.

3 The Experiments

The experiments were done using the INQUERY information retrieval system developed at the University of Massachusetts [2]. INQUERY is based on a probabilistic model of retrieval, has a number of advanced features, and has consistently achieved excellent results at the ARPA-sponsored TREC and TIPSTER evaluations (see [4] for an overview of the TREC evaluation). For the purposes of these experiments, the main features of INQUERY are that it does automatic indexing and produces ranked lists of documents in response to a query. These are features that are common to many recent information retrieval systems. All ranking systems use weighting or estimation functions to determine the relative importance of words in the query and document. As the results discussed later show, the form of these estimation functions can be important in an OCR environment.

Four test collections were used in these experiments. The collections were selected to represent a range of sources, document sizes, and query sizes. The CACM collection is a small collection of Computer Science Abstracts [3] and has been a standard benchmark for a number of years. NPL is a larger collection of short documents and short queries that has been used in a variety of IR experiments. WEST is a collection of long, full-text, legal information, specifically case law. The WSJ collection is the largest number of documents, which are moderate length, full-text articles from the Wall St. Journal. The WSJ queries are also the longest of any collection. The WSJ collection is a subset of the TIPSTER collection described in [4].

In general, we would expect OCR errors to have more impact on the collections of short documents, since long documents would have much more redundant information. This is one of the factors that is tested in the experiments.

Table 2 gives statistics for two of the collections (NPL and WEST) showing the number of word tokens assigned to each page quality group and the number of OCR errors generated in each group. Note that

Table 2: Summary of total words and OCR errors generated for two test collections, NPL and WEST. Numbers represent word tokens encountered for STD (no OCR errors), and worst and best OCR simulations (OCR1 and OCR2) from UNLV tests. These collections contain the shortest and longest average documents respectively.

NPL

	STD	OCR1		OCR2	
Page Group	Total	Total	Errors	Total	Errors
1	479,163	96,815	1,254	96,666	412
2		95,237	4,569	95,926	1,665
3		95,562	8,737	94,772	2,255
4		95,293	16,714	95,508	5,196
5		96,269	40,370	96,291	13,993

WEST

	STD	OCR1		OCR2	
Page Group	Total	Total	Errors	Total	Errors
1	39,549,976	7,984,570	92,881	7,979,432	35,898
2		7,887,047	360,866	7,881,652	141,677
3		7,847,856	674,372	7,849,518	189,767
4		7,882,217	1,336,320	7,886,171	442,373
5		7,948,286	3,190,063	7,953,203	1,147,189

Table 3: Summary statistics for the three versions of four collections used to evaluate the effect of OCR errors on retrieval performance. STD refers to the original collection. OCR1, OCR2 are the worst and the best OCR systems, respectively, from UNLV tests. The dictionary term counts represent the number of unique word stems in the version dictionary. All indexed terms are the number of word stems encountered during the indexing of the text excluding stopwords.

Collection	Collection Size	Document Cnt	Average Chars/Doc
CACM	1,639,440	3,204	512
NPL	3,748,316	11,429	327
WEST	297,501,776	11,953	24,889
WSJ	279,249,494	98,735	2,828

Collection	Dictionary Terms			All Indexed Terms		
	STD	OCR1	OCR2	STD	OCR1	OCR2
CACM	5,998	5,644	5,903	115,294	96,282	110,386
NPL	7,689	7,144	7,558	275,517	229,786	264,258
WEST	155,542	144,294	152,891	22,817,834	19,353,353	21,830,212
WSJ	197,255	182,341	193,508	24,454,116	20,797,586	23,448,131

Table 4: Statistics on standard query sets for each of four collections used to evaluate OCR errors on retrieval performance.

Collection	Total Queries	Number of Words/Query			Average Unique Words/Query
		Min	Mean	Max	
CACM	50	2	14.24	49	13.0
NPL	93	3	7.26	12	7.1
WEST	34	5	11.05	20	9.6
WSJ	50	13	32.68	118	29.3

these numbers do not represent indexed terms. Many of the words were stopwords and thus discarded. Also, many of these word forms were conflated to single unique indexed stems, as stemming was used in the indexing runs.

Table 3 shows the results of the OCR simulation on the indexing of all four collections. It gives the figures for the original collection and the two OCR simulations. It should be noted again that "misspelled" words generated by OCR errors are not included in these figures. The table shows that the number of unique terms is reduced considerably in the case of OCR1 (consistently about 7%) and much less in the case of OCR2. From this we would certainly expect to see more impact on the retrieval performance of OCR1.

Table 4 gives the statistics for the queries associated with these collections. The main feature here is the length of the Wall St Journal queries. Long queries are another form of redundancy that may offset the effect of OCR errors. From this point of view, the NPL collection has the worst combination of characteristics in that it has both short documents and short queries. We should emphasize, however, that the error generation process is only applied to the document texts, not the queries.

4 Summary of Results

The following tables show the results of the retrieval experiments using the three versions of each of the four test collections. Table 5 gives the overall results using the average precision over all recall levels. The CACM entry is a result for one query set. Combined results for 100 runs using four query sets showed average -6.4 and -1.1 percent performance degradation in retrieval performance between standard and OCR1 and OCR2 simulations.

The results appear to support the view that collections with short documents and short queries will be affected the most by OCR errors. The collection with the biggest degradation in average precision is NPL. This is also the only collection where the better OCR system (OCR2) caused a significant loss in precision compared to the original collection. The CACM collection, which also has many short documents, had the next largest degradation in performance. The WEST collection, which has very large documents, had the lowest degradation for both OCR systems. From these results, it can also be concluded that using the best OCR system for input to a text retrieval system will generally not significantly affect retrieval performance for databases with long documents. This conclusion is with respect to word errors other than those generated by zoning errors. It is worth mentioning here that the general "rule of thumb" used in IR experiments is that a change in average precision of less than 5% is not significant, and a change of around 10% is very significant. For the NPL collection, then, even the best OCR input resulted in a significant loss in performance.

In order to look at these results in more depth, tables 5 through 9 contain standard recall-precision tables, which show the average precision figures at standard recall points. These tables show that the highest losses in accuracy generally occur at higher recall levels (i.e. further down the ranking). This is what would be expected in that documents which contain many query terms will be less affected by the loss of one of those terms, and these are typically the terms at the top end of the ranking.

To study the effect of random variation, we did a large number of retrieval runs for the CACM collection. The only factor that varied between these runs was the random effect of the OCR errors. Table 10 shows that although performance degradations are generally consistent, occasional runs can result in performance im-

Table 5: Retrieval performance for four standard text collections showing effects of two levels of simulated OCR error rates. Values are average precision over 10 standard recall points from 10 to 100 percent. Percentage differences are given in parentheses. Results for CACM are for one of four query sets. Average performance loss for 100 simulation runs for CACM was -6.4 and -1.1 percent for OCR1 and OCR2 respectively.

Collection	Average Precision				
	STD	OCR1		OCR2	
CACM	34.9	32.5	(-6.9%)	34.3	(-1.7%)
NPL	25.8	23.2	(-10.1%)	23.5	(-9.1%)
WEST	48.2	46.2	(-4.0%)	48.0	(-0.4%)
WSJ	39.9	38.1	(-4.5%)	39.3	(-1.5%)

Table 6: The standard recall-precision table for the CACM collection for one of 25 runs using query set 2.

Recall	Precision (93 queries)				
	STD	OCR1		OCR2	
10	66.9	64.7	(- 3.3)	70.4	(+5.3)
20	53.5	53.0	(- 0.9)	55.8	(+4.4)
30	47.0	45.6	(- 3.1)	46.9	(-0.3)
40	40.0	37.4	(- 6.5)	41.0	(+2.5)
50	34.7	30.0	(-13.3)	34.4	(-0.7)
60	28.8	24.5	(-14.8)	26.6	(-7.5)
70	20.3	17.4	(-14.5)	19.3	(-5.3)
80	15.9	12.4	(-22.0)	15.5	(-2.3)
90	10.6	7.5	(-29.9)	10.1	(-5.4)
100	8.1	5.3	(-33.8)	7.3	(-8.8)
avg	32.6	29.8	(- 8.6)	32.7	(+0.5)

Table 7: The standard recall-precision table for the NPL collection.

Recall	Precision (93 queries)				
	STD	OCR1		OCR2	
10	57.4	52.8	(-8.1)	55.8	(-2.9)
20	48.5	45.9	(-5.2)	46.0	(-5.2)
30	40.3	35.2	(-12.9)	35.2	(-12.8)
40	33.3	27.9	(-16.1)	29.2	(-12.1)
50	26.2	22.9	(-12.6)	22.5	(-14.1)
60	18.1	16.1	(-11.1)	16.5	(-9.0)
70	13.7	12.3	(-10.2)	12.2	(-11.1)
80	10.5	9.6	(-7.9)	9.5	(-9.1)
90	6.8	6.1	(-10.5)	5.2	(-22.7)
100	3.6	3.5	(-1.5)	2.8	(-22.1)
avg	25.8	23.2	(-10.1)	23.5	(-9.1)

Table 8: The standard recall-precision table for the WEST collection.

Recall	Precision (34 queries)				
	STD	OCR1		OCR2	
10	78.1	77.0	(-1.4)	77.9	(-0.3)
20	73.8	72.5	(-1.6)	73.8	(+0.0)
30	71.9	70.3	(-2.3)	71.8	(-0.2)
40	62.0	58.9	(-5.0)	61.6	(-0.6)
50	54.9	52.0	(-5.3)	54.9	(+0.0)
60	45.3	43.4	(-4.2)	44.7	(-1.3)
70	37.3	35.2	(-5.7)	37.2	(-0.4)
80	29.7	28.5	(-4.0)	29.1	(-2.0)
90	17.9	16.3	(-8.9)	17.9	(+0.1)
100	10.7	8.4	(-21.7)	10.7	(+0.4)
avg	48.2	46.2	(-4.0)	48.0	(-0.4)

Table 9: The standard recall-precision table for the WSJ collection.

Recall	Precision (50 queries)				
	STD	OCR1		OCR2	
10	68.3	67.7	(-0.7)	67.5	(-1.0)
20	60.2	60.3	(+0.1)	60.4	(+0.2)
30	53.6	53.1	(-0.9)	53.3	(-0.6)
40	48.2	47.1	(-2.3)	47.4	(-1.6)
50	42.0	40.0	(-4.7)	42.2	(+0.4)
60	37.8	35.1	(-7.1)	37.3	(-1.3)
70	32.9	30.1	(-8.4)	32.2	(-1.9)
80	27.4	23.6	(-13.9)	26.2	(-4.4)
90	19.9	16.9	(-14.9)	18.9	(-5.0)
100	8.7	7.2	(-17.5)	7.6	(-12.6)
avg	39.9	38.1	(-4.5)	39.3	(-1.5)

provements, even with OCR1. Significant changes between runs, as occurs sometimes in OCR1, are more likely to happen with small collections where the recall and precision for a particular query can be significantly affected by changes to just a few documents. There are two ways in which discarding terms at random can improve retrieval performance. One is that the documents that were penalized by the OCR errors were documents that contained query terms but were not relevant. Making those documents hard, or even impossible, to retrieve results in better performance. Another significant, and more common, effect is that OCR errors can change the frequencies used to calculate the relative importance of words. The most important of these is the maximum frequency for any word in a particular document. This frequency is used for normalization and changes to this frequency can have significant results, particularly in collections with small document sizes. We are currently developing new estimation techniques for word importance that are much less sensitive to the types of errors introduced by OCR.

5 Future Work

The simulations described above could be made more accurate by taking into account which characters are commonly confused by OCR devices. By using knowledge of what types of characters are generated in error, we could also attempt to simulate the generation of valid index terms and the generation of misspelled words. The most difficult aspect of the simulation to improve would be to generate errors arising from poor zoning. In most real applications of OCR to archiving and retrieving information, zoning will be an important issue.

The value of the experiments given here is to give some quantitative data that is reasonably accurate. This data shows that even though the output of the best OCR devices can be adequate for automatic indexing and retrieval of databases of longer full-text documents, for collections of very short documents, OCR errors can have a significant impact on retrieval performance. The most important types of errors will not be random, but rather when a relevant document is made unretrievable by poor quality scanning or word recognition. Regard-

Table 10: Average precision results at 10 standard recall levels for each of 25 repeated indexing runs using CACM query set 2. Numbers in parentheses represents percent difference with standard collection.

Run	CACM query set 1				
	STD	OCR1		OCR2	
1	32.6	29.8	(-8.6)	32.7	(+0.5)
2	32.6	29.6	(-9.2)	32.6	(-0.1)
3	32.6	28.8	(-11.7)	32.7	(+0.4)
4	32.6	33.1	(+1.4)	32.5	(-0.2)
5	32.6	30.0	(-8.0)	32.8	(+0.7)
6	32.6	31.2	(-4.2)	31.6	(-2.9)
7	32.6	30.2	(-7.4)	31.2	(-4.2)
8	32.6	30.7	(-5.9)	32.5	(-0.2)
9	32.6	31.4	(-3.8)	32.2	(-1.0)
10	32.6	29.7	(-8.8)	32.4	(-0.6)
11	32.6	30.2	(-7.3)	31.8	(-2.4)
12	32.6	29.9	(-8.1)	32.5	(-0.3)
13	32.6	30.4	(-6.7)	31.6	(-3.1)
14	32.6	30.1	(-7.7)	32.9	(+0.9)
15	32.6	32.6	(+0.2)	32.0	(-1.7)
16	32.6	29.9	(-8.3)	31.7	(-2.6)
17	32.6	29.1	(-10.6)	32.3	(-1.0)
18	32.6	29.8	(-8.6)	33.3	(+2.3)
19	32.6	29.7	(-8.8)	32.7	(+0.4)
20	32.6	30.3	(-7.0)	32.4	(-0.4)
21	32.6	29.9	(-8.4)	31.6	(-3.1)
22	32.6	30.0	(-7.9)	32.5	(-0.2)
23	32.6	30.3	(-6.9)	33.1	(+1.6)
24	32.6	31.0	(-4.9)	31.7	(-2.9)
25	32.6	31.5	(-3.2)	32.5	(-0.3)

less of the source of the errors, the results suggest the need for more research on automatic correction schemes, even when the OCR output is not needed for display.

As mentioned above, the other area of research that is being pursued is to develop estimation functions for the retrieval system that are less sensitive to OCR errors. In general, however, it does appear that ranking systems provide a degree of robustness in an environment where documents contain errors.

Acknowledgments

This research was supported in part by the NSF Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst, and by the Information Science Research Institute, University of Nevada, Las Vegas.

References

[1] N. J. Belkin and W.B. Croft. Information filtering and information retrieval: Two sides of the same coin. *Communications of the ACM*, 35(12):29–38, 1992.

[2] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

[3] Edward A. Fox. Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts. Technical Report 83-561, Department of Computer Science, Cornell University, Ithaca, NY 14853, September 1983.

[4] D. Harman. Overview of the first TREC conference. In *Proceedings of the 16th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 36–47, 1993.

[5] S. Rice, J. Kanai, and T. Nartker. An evaluation of OCR accuracy. In *UNLV Information Science Research Institute Annual Report*, pages 9–20, 1993.

[6] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[7] K. Taghva. Results of applying probabilistic ir to ocr text. pages 1–20, 1994.

[8] K. Taghva, J. Borsack, A. Condit, and S. Erva. The effects of noisy data on text retrieval. In *UNLV Information Science Research Institute Annual Report*, pages 71–80, 1993.