DOE/NV/10872--T231

UNLV
Information Science
Research Institute

Quarterly Progress Report

T. A. Nartker
September 30, 1995

Table of Contents

I. Board of Advisors Activity 1

II. Symposium Activity 1

III. Staff Activity 1

Recruiting
Travel
Papers accepted or presented

IV. Document Analysis Program 2

OCR Devices
OCR Test system
OCR Databases/GT1
OCR Databases/Foreign languages
OCR Experiments
OCR Technical reports/thesis
Interaction with OCR vendors
Interaction with OCR research organizations

V. Text-Retrieval Program 4

TR Software systems
TR Databases
TR Experiments/Projects
TR Technical reports/thesis
Document Routing Project

VI. Institute Activity 5

Proposals for Continued and New Support
Institute visitors
Institute seminars
New agency contacts

VII. Goals Achieved This Quarter/Goals for Next Quarter 6

Appendix A. Status of ground-truth preparation activities
Appendix B. Test of OCR systems using DOE documents
Appendix C. Proposed agreement with the DFKI

UNLV Information Science Research Institute
Quarterly Progress Report

T. A. Nartker
September 30, 1995

I. Board of Advisors Activity

Although a Fall board meeting is normally held in October, it has been canceled this year due to budgetary constraints.

The next Board meeting will be held at our Symposium in April of 1996.

II. Symposium Activity

The 1996 SDAIR Symposium will be held on April 15, 16, & 17, 1996 at the Alexis Park Hotel. At this time, Dr. Juergen Schuermann, Dr. H. P. Frei, and Dr. Mike Lesk have accepted our invitation to give an Invited Talk. We are still searching for a fourth invited speaker.

III. Staff Activity

Recruiting

None

Travel/ Meetings

In July, T. Nartker attended the LSSARP meeting in Green Bay, Wisconsin.

In July, K. Taghva, A. Condit, & J. Borsack attended the SIGIR meeting in Seattle, Washington. Kazem was the moderator for a panel discussion on information retrieval education and one of the organizers of an associated workshop on information retrieval education.

In August, T. Nartker attended the International Workshop on Graphics Recognition at Penn State University.

Also in August, J. Kanai and T. Nartker attended the ICDAR'95 conference in Montreal Canada. Tom served on the program committee for this conference and chaired a session on handwriting recognition. Junichi presented two papers.

Papers accepted or presented

A paper by T. Nartker, S. Rice, and F. Jenkins titled "OCR Accuracy: UNLV's Fourth Annual Test," was published in the July issue of INFORM magazine.

At ICDAR'95, Junichi presented a paper titled "Prediction of OCR Accuracy using Simple Image Features," co-authored with L. Blando and T. Nartker and a paper titled "Adaptive Image Restoration of Text Images that Contain Touching or Broken Characters," co-authored with P. Stubberud and V. Kalluri.

IV. Document Analysis Program

New OCR Devices

None

OCR Test system

Steve Rice has completed several improvements to our OCR experimental environment. The string-matching algorithm that was used to compute word accuracy in Version 5.0 was sub optimal, that is, it was not guaranteed to find the minimum number of misrecognized words. An optimal string-matching algorithm has been implemented and tested in August as part of the development of Version 5.1.

An optimal algorithm for word accuracy finds a correspondence between OCR-generated and correct strings of words in which the number of word insertions and substitutions needed to correct the OCR-generated string is minimized. We showed in a previous paper that any longest common subsequence (LCS) algorithm could be used to find this correspondence.

We have implemented Myers's LCS algorithm together with a preprocessing step in which words that occur in one but not both of the input strings are removed. Since these words cannot be part of a LCS, it is safe to remove them and takes only linear time to do so. The LCS algorithm, which is quadratic in the worse case, runs faster by operating on smaller strings.

Getting a precise measurement of how much the optimal string-matching algorithm affects the word error counts produced by its sub optimal predecessor proved to be difficult because the new algorithm required a different treatment of words adjacent to wildcards (tildes). To illustrate, suppose a page image contains the word "Discover" followed by a trademark symbol. The ground-truth representation for this would be "Discover~". If an OCR system generates "Discovery" for this, the old algorithm would have been tolerant and concluded that the word "Discover" had been recognized correctly; however, the new algorithm is more conservative and considers the word to be misrecognized. Indeed, if "Discovery" were stored in a text retrieval system, one would not get a "hit" when searching for "Discover". We estimate that the number of misrecognized words decreases by about 1% when using the new algorithm, but depending on the number of words adjacent to wildcards in the sample, the net effect could be an increase of a few percent due to the more stringent treatment of wildcards.

Other improvements in Version 5.1 include:

Custom Stopword Lists: The "wordacc" program now accepts an arbitrary list of stopwords. Stopword and non-stopword accuracy are now computed depending upon this list. If no list is specified, then the default set of 110 stopwords from BASISplus is used as before.

Word Accuracy Confidence Intervals: The "wordaccci" program has been developed and is analogous to the "accci" program for character accuracy. "Wordaccci" accepts two or more word accuracy reports and computes an approximate 95% confidence interval for word accuracy using the jackknife estimator.

Non-stopword Accuracy as a Function of the Number of Stopwords: Given a word accuracy report and a list of stopwords in order by decreasing frequency of occurrence, the new "nonstopacc" program computes the non stopword accuracy as a function of the number of stopwords. The non-stopword accuracy is determined by excluding the K most common stopwords, and is computed for all values of K ranging from 0 up to the number of stopwords in the list. By plotting the results, non-stopword accuracy is no longer a single, isolated point, but is represented by a curve. This sort of display is more meaningful to information retrieval researchers, who tend to use different numbers of stopwords.

OCR Databases/GT1

We met with Dave Warriner to discuss a new DOE sample (to be used in 1997), selected from documents in the current RIS. Although no dataset has been defined, Dave expressed a desire that we test microfilm images next year. We plan a test of microfilm images using Sample#3 ground-truth.

Page 1 of Appendix A shows the ISRI methodology for preparing ground-truth test data. Page 2 shows the status of all ISRI English Language datasets on September 30.

OCR Databases/Foreign languages

As mentioned above, Appendix A gives an overview of our current ground-truth data preparation activities. Page 3 shows the status of foreign language test datasets prepared as part of our Fort Mead contract. Current work is focused on a new sample of German business letters and on the Japanese horizontal and vertical samples.

OCR Experiments

We have prepared a special accuracy test of all available OCR technologies using DOE Sample#3. This report indicates which OCR technologies would be best for both LSS and RIS use if Sample #3 contained pages typical of these systems. We request that this report, shown as Appendix B, not be made public.

OCR Technical reports/thesis

G.S. Rajarathinam has completed work for his MS in EE. His thesis concerned "Adaptive Sorting Algorithms for Automatic Zoning Evaluation". There are several other graduate students pursuing document analysis research projects. In our quarterly reports, we provide a summary of completed thesis projects but we do not provide interim progress reports for these projects.

Interaction with OCR vendors

Both AEG Electrocom, in Lake Constance, Germany and Hewlett Packard Laboratories, in Kawasaki, Japan have joined our Industrial Affiliate program.

Interaction with OCR research organizations

Our sample of German business letters from the DFKI in Germany is complete. We expect to utilize this sample in next years test. A copy of our English business letter sample has been sent to them.

A proposed agreement (and cooperative project) in which ISRI will prepare the ground-truth text of approximately 2000 additional German business letters, has been sent to the DFKI (see Appendix C).

V. Text-Retrieval Program

TR Databases

We are cutting back our effort to collect relevance judgments for the queries associated with the LSS prototype database. As of September, we have completed relevance judgments for 1190 of a total of 1370 documents. Although this collection is an excellent mini-LSS (WITH SAMPLE QUERIES AND RELEVANCE JUDGMENTS) we do not have adequate funds to complete this database. We plan to seek funding from other agencies to complete this database and make it publicly available.

TR Experiments/Projects

Our text retrieval group has begun the process of preparing documentation for the MANICURE OCR post-processing system. Although MANICURE is completely operational, we have devoted all previous effort to improving its performance.

TR Technical reports/thesis

None

VI. Institute Activity

Proposals for Continued and New Support

Much of our activities in August and September, and much of our current activity, is focused on writing up our work and preparing proposals for funding to different agencies.

In August and September, we submitted three proposals to continue our work in text-retrieval. These were:

1. A proposal to the DEPSCOR program to build a new text-retrieval system that incorporates all we have learned from our research with the LSS. This new retrieval system would be built to anticipate noisy/OCR document input and would exploit logical mark-up in the stored documents. No existing text retrieval system contains these features.
2. A proposal to NSF to design university level curricula in IR.
3. A proposal to the "Distance Education/Senate bill 204" program to make available two Junior High instructional programs on the WEB. The programs are in Mathematics and in the Social Sciences.

Several other proposals are currently being prepared.

LSS working group meetings & report

Although the LSS Technical Working group did not meet during the Fall quarter, all members attended the LSSARP meeting in Green Bay, Wisconsin.

Institute visitors

Institute visitors this quarter:

Date	Visitor	Agency
7/17/95	Dr. W. Eckstein	Lehrstuhl
8/18/95	Mr. Ishitazi &. Mr. Sakai	Toshiba
8/29/95	Mr.'s R. Irish, B. Scheidler, & T. Barchi & John Gandy	NRC/Inspector general's office DOE/OCRWM

Institute seminars

None.

New agency contacts

None

VII. Goals Achieved/Goals for Next Quarter

Goals from last quarter:

1) Work to install components from MANICURE in the RDMS has stopped. We have been led to believe that all DOE work on both the RDMS and the LSS will be terminated.

2) Three speakers have been invited for SDAIR'96.

3) We have completed documentation on the OCR Experimental Environment, version#5.

4) Both AEG Electrocom, in Lake Constance, Germany and Hewlett Packard Laboratories, in Kawasaki, Japan have joined our Industrial Affiliate program.

5) The industry annual report data sample is about 75% complete. The mystery sample is complete through the scanning and zoning step. No progress has been made in defining source data for DOE Sample #4.

6) We believe the LSS Technical Working Group has disbanded.

7) We have reduced our effort to complete the set of relevance judgments for queries associated with the LSS prototype documents.

Goals for next quarter:

1) Prepare an ISRI final report to be delivered to the DOE in January We propose that our next quarterly report (to be delivered in January) be a final report of work on this contract.

2) We plan to continue to seek funding to support our work in OCR & text-retrieval.

3) Continue to recruit new members for the affiliates program.

4) Complete the annual report data sample for our 1996 Annual OCR Test. Continue work on the mystery sample.

APPENDIX A.

ISRI Methodology for Preparing Ground-Truth Test Data and Status of Ground-Truth Data Preparation Activities

ISRI METHODOLOGY
FOR PREPARING GROUND-TRUTH OCR DATA

1. **Acquire document sample**
   - Choose document class
   - Identify source for documents
   - Choose selection strategy and method of acquisition
   - Obtain documents

2. **Select page sample**
   - Choose page (or partial page) sampling strategy
   - Sample and prepare pages
   - Choose zone types

3. **Train data entry personnel**
   - Select & acquire tools
   - Train data entry staff

4. **Scan pages**
   - Determine scanning variables (i.e. threshold & page placement)
   - Scan all pages (usually at 200, 300, 400, &GS)
   - Verify images
   - Quality control

5. **Archive document & page collections**

6. **Manually zone images**

7. **Prepare Truth text**
   - Prepare multiple manual entries
   - Prepare text from ISRI voting algorithm
   - Resolve multiple manual entry & voting differences

8. **Archive images,**
   zone information,
   all manual & voting truth input, &
   ground-truth text

ISRI English Ground-Truth Databases

(As of 9/30/95)

	Document Sample			Page Sample			T.S.	Scan Pages			Arch.	Zone Images		Prepare Truth		Arch.
Name	Defined	# Docs.	% Com	Defined	# Pages	% Com	Com.	# Imgs	Resol.	% Com		# Zones	% Com	# Character	% Com	
DOE Sample1	Yes	250	100	Yes	240	100	Yes	132	3b	100	Yes	242	100	278,786	100	Yes
DOE-Sample1 Synthesized	Same	250	100	Synthe-sized	132	100	Yes	242 (x 9)	Synthe. At 3b	100	Yes	242 (x 9)	100	278,786 (x 9)	100	Yes
DOE-Sample1 SynPrintedScand	Same	250	100	Syn-Printed	132 (x 9)	100	Yes	242 (x 9)	3b	100	Yes	242 (x 9)	100	278,786 (x 9)	100	Yes
DOE Sample2	Yes	2500	100	Yes	460	100	Yes	460	3b	100	Yes	1313	100	817,946	100	Yes
DOE Sample3	Yes	2500	100	Yes	800	100	Yes	800	2,3,4b 3gs	100	Yes	2431	100	1,463,512	100	Yes
Magazine Sample1	Yes	100	100	Yes	200	100	Yes	200	2,3,4b 3gs	100	Yes	1414	100	666,134	100	Yes
US-Newspaper Sample1	Yes	50	100	Yes	200	100	Yes	200	2,3,4b 3gs	100	Yes	758	100	492,080	100	Yes
BusinessLetter Sample1	Pages	200	100	Yes	200	100	Yes	200	2,3,4b 3gs	100	Yes	1508	100	319,756	100	Yes
DOJ-Microfilm Sample1	Film Images	2000	100	Yes	200	100	Yes	200	2b	100	Yes	704	100	471,755	100	Yes
FAX1 BusinessLetters	Same	200	100	Yes	200	100	Yes	200	Fine & Stand.	100	Yes	1357	100	319,756	100	Yes
Industry Annual Reports	Yes	75	100	Yes	300	100	Yes	300	2,3,4b 3gs	100	No	1703	100		**75**	No
Magazine Sample2	Yes	75	100	Yes	**300**	**100**	Yes	**300**	**2,3,4b 3gs**	**100**	No	**2325**	**100**		**25**	No
96 Mystery Sample	Yes	**60**	**100**	**Yes**	**200**	**100**	**Yes**	**200**	**2,3,4b 3gs**	**100**	**No**	**548**	**100**			
Commerce Business Daily	Yes															
Supreme Court Slips	Yes	10	100	Yes	100	100	Yes	100	2,3,4b 3gs	100	No	219	100	186710	**85**	No

BOLD - indicates a change from last report

ISRI Foreign Language Ground-Truth Databases

(As of 9/30/95)

	Document Sample			Page Sample			T.S.	Scan Pages			Arch.	Zone Images		Prepare Truth		Arch.
Name	Defined	# Docs.	% Com	Defined	# Pages	% Com	Com.	# Imgs	Resol.	% Com		# Zones	% Com	# Character	% Com	
Chinese-test Sample	Yes	N,M, B,GA	100	Yes	57	100	No	57	3b 3gs	100	...	57	100	48,378	100	
Japanese-test Sample	Yes	27	100	Yes	73	100	No	4	2,3,4b 3gs	5	...	29	...	4,400	...	
Span-Newspaper Sample1	Yes	36	100	Yes	144	100	Yes	144	2,3,4b 3gs	100	Yes	569	100	348,091	100	Yes
China-Newspaper Sample1	Yes	19	100	Yes	95	100	Yes	95	3b 3gs	100	No	614	100	81,080	100	
Japan-Newspaper Sample1.Verticle	Yes	30	100	Yes	60	100	Yes			0	No				30	
Japan-Newspaper Sample1.Horizon	Yes (same)	30	100	Yes	57	100	Yes			0	No				0	
Cyrillic-Newspaper Sample1	No	...	...	No	...	...	...									
Farsi-Newspaper Sample1	Yes	45	100	No	...	...	...									
GermanBusLetter Sample1	**Yes**	**183**	**100**	**Yes**	**200**	**75**	**Yes**	**200**	**2,3,4b 3gs**	**100**	**No**	**1813**	**100**	**310454**	**75**	**No**

BOLD - indicates a change from last report

APPENDIX B.

Test of OCR Systems using DOE Documents

Test of OCR Systems Using DOE Documents

Stephen V. Rice and Thomas A. Nartker

Information Science Research Institute
University of Nevada, Las Vegas
4505 Maryland Parkway
Box 454021
Las Vegas, NV 89154-4021

September 1, 1995

1 Introduction

With support from the U.S. Department of Energy (DOE), the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas, conducts a program of applied research in optical character recognition (OCR) and information retrieval. As part of this program, ISRI conducts an annual test of the leading OCR systems. In April 1995, the results of the "Fourth Annual Test of OCR Accuracy" were presented at the Fourth Annual Symposium on Document Analysis and Information Retrieval in Las Vegas.

The scope of the annual test has expanded considerably since its inception in 1992, but every year there has been considerable focus on the accuracy of OCR systems when processing DOE documents. In the fourth annual test, the largest DOE sample to date was utilized: 785 pages containing 213,552 words and 1,463,512 characters. These pages were selected at random from the 100,000-page collection of scientific and technical documents that were gathered for the Licensing Support System (LSS) prototype.

This report supplements the fourth annual test report by focusing specifically on the performance of five leading OCR systems on the DOE sample, as well as the performance of the ISRI Voting Machine, which combines the outputs from these five systems to produce a more accurate output.

2 OCR Systems

The five OCR systems evaluated in this report are:

- Caere OCR Version 138.1,
- ExperVision RTK Version 3.0,

- MAXSOFT-OCRON Recore Version 3.2,
- Recognita OCR Version 3.0, and
- XIS OCR Engine Version 10.5.

The Caere and XIS systems were judged to be "first-tier" systems, and the MAXSOFT and Recognita systems were rated as "second-tier" systems in the fourth annual test. ExperVision did not submit a version for the fourth annual test; however, the version tested here was the best overall system in the third annual test.

The ISRI Voting Machine Version 5.0 is also evaluated in this report. This system processes a page image by submitting it to each of the five "participating" OCR systems, and then matches the five resulting text files in such a way as to find agreements and disagreements among the participants. Where there is a disagreement, the identity of a character is determined essentially by taking a majority vote among the participants. Tests of earlier versions of the ISRI Voting Machine are described in the first and third annual test reports.

The Caere and XIS systems were operated on a Sun SPARCstation. The MAXSOFT and Recognita systems were tested under PC Windows, and the ExperVision system executed under PC DOS. The ISRI Voting Machine operates on a Sun SPARCstation, although three of its participants were executed remotely on PCs.

3 Test Methodology

Page images were produced using a Fujitsu M3096G scanner. In this report, only tests involving binary images are discussed (see the fourth annual test report for some results using gray scale images). For all tests described in this report, the resolution of the binary images is 300 dots per inch (dpi), except the test of the effect of resolution, in which 200 and 400 dpi images were also utilized.

Each page image was manually "zoned," i.e., the text regions of the page were delineated and ordered. The coordinates of these "zones" were supplied to the OCR system in each test, with the exception of the automatic zoning test. Correct text was carefully prepared for each zone, and used to evaluate the accuracy of OCR-generated text. The ISRI OCR Experimental Environment Version 5.0 is a collection of software tools that was used to compare OCR-generated text with correct text and compute the performance measures that are presented in this report.

4 Character Accuracy

Table 1 shows the character accuracy of each system when processing the DOE Sample. The number of errors corresponds to the number of edit operations (character insertions, substitutions, and deletions) needed to correct the OCR-generated text.

Graph 1 displays approximate 95% confidence intervals for character accuracy. Non-overlapping intervals indicate a statistically significant difference in accuracy.

	Errors	% Accuracy
ISRI Voting Machine	25,749	98.24
ExperVision RTK	32,017	97.81
XIS OCR Engine	34,644	97.63
Caere OCR	37,503	97.44
MAXSOFT-OCRON Recore	56,746	96.12
Recognita OCR	57,713	96.06

Table 1: Character Accuracy

Thus, the ExperVision and XIS systems demonstrated that they are significantly more accurate than the MAXSOFT and Recognita systems when processing the DOE pages.

Graph 2 shows that reducing the resolution of images from 300 to 200 dpi causes a dramatic decline in accuracy, yet increasing the resolution to 400 dpi offers little or no advantage.

The median character accuracy achieved by several OCR systems when processing a particular page is a good measure of the quality of the page. Using this measure, the pages of the DOE Sample were divided into five "Page Quality Groups" of approximately equal size. Group 1 contains the pages having the best quality, and Group 5 contains the pages having the worst quality. Graph 3 indicates the character accuracy within each group. Roughly 75% of the total number of errors are made on the worst 20% of the pages (i.e., Group 5). The challenges presented by these pages include broken and touching characters, difficult tables, and lines of text that are skewed (rotated) or curved (from photocopying a page of a bound book).

5 Word Accuracy

Word accuracy is defined as the percentage of words that are correctly recognized, where a word is any sequence of one or more letters. It is useful to distinguish between stopwords and non-stopwords. Stopwords are common words such as *the*, *of*, *in*, etc. which are normally not indexed by a text retrieval system. Thus, it is non-stopword accuracy, i.e., the percentage of non-stopwords that are correctly identified, that is especially of interest in a text retrieval application. Table 2 gives the number of misrecognized words, and the word, stopword, and non-stopword accuracy for each system. Graph 4 plots the non-stopword accuracy within each Page Quality Group.

In text retrieval, users search not only for specific terms (non-stopwords), but also phrases. Graph 5 displays the phrase accuracy, as a function of phrase length, for each system.

	Misrec. Words	Word Accuracy	Stopword Accuracy	Non-stopword Accuracy
ISRI Voting Machine	5,872	97.25	98.86	96.35
ExperVision RTK	8,331	96.10	98.20	94.93
XIS OCR Engine	9,239	95.67	98.44	94.13
Caere OCR	9,386	95.60	98.05	94.24
MAXSOFT-OCRON Recore	15,451	92.76	96.49	90.70
Recognita OCR	16,674	92.19	95.69	90.25

Table 2: Word Accuracy

6 Marked Character Efficiency

An OCR system places flags in the output, known as reject characters and suspect markers, to attract the attention of the user to potential errors in the OCR-generated text. The process of examining these flags, or "marked characters," and correcting the identified errors, is efficient if a large percentage of the flags do in fact point to errors. But if many of the flags point to characters that are correct, then the user must spend much time verifying the correctness of these characters.

Graph 6 presents marked character efficiency curves, which reflect the efficiency of the correction process for each system. Each curve shows how the character accuracy of the generated text increases as the user examines more and more marked characters and corrects the identified errors.

7 Automatic Zoning

In the test of automatic zoning, we evaluate the ability of the OCR system to locate the text regions on a page and determine their correct reading order. For a page containing multiple columns of text, each column should be regarded as one zone, and the resulting generated text should be "de-columnized." However, for a page containing a table, all columns of the table should be placed in a single zone so that the structure of the table is not lost.

The cost of correcting automatic zoning errors can be expressed in terms of edit operations: insertions to enter missing blocks of text, and move operations to correct the ordering of blocks. Using a conversion factor, each move operation is converted into an equivalent number of insertions so that the cost of correction is ultimately expressed solely in terms of insertions.

Graph 7 shows the cost of correcting automatic zoning errors for a range of conversion factors. The XIS OCR Engine performed the best on this test.

8 Analysis of Voting

Overall, the ISRI Voting Machine produced 20% fewer errors, and 30% fewer misrecognized words, than the best of its participants, which was the ExperVision system. But the rate of reduction varied depending on page quality (see Tables 3 and 4). A much greater reduction was achieved when processing the better quality pages (Groups 1 to 4).

	Page Quality Group					
	1	2	3	4	5	Total
ExperVision RTK	479	1,042	1,471	4,738	24,287	32,017
ISRI Voting Machine	57	291	621	2,048	22,732	25,749
Reduction	*88%*	*72%*	*58%*	*57%*	*6%*	*20%*

Table 3: Reduction in the Number of Errors Due to Voting

	Page Quality Group					
	1	2	3	4	5	Total
ExperVision RTK	141	312	533	1,334	6,011	8,331
ISRI Voting Machine	28	92	245	685	4,822	5,872
Reduction	*80%*	*71%*	*54%*	*49%*	*20%*	*30%*

Table 4: Reduction in the Number of Misrecognized Words Due to Voting

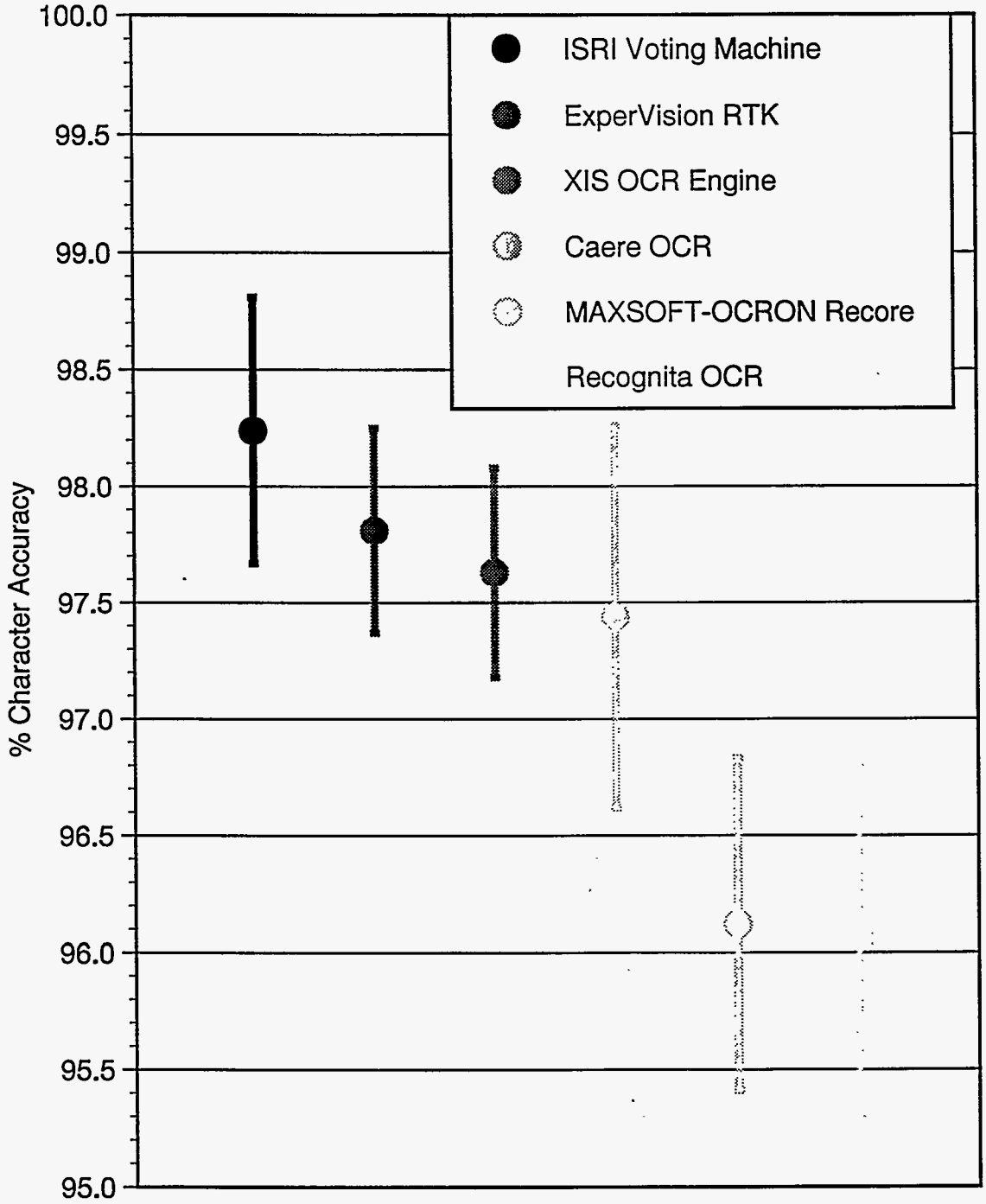
9 Conclusion

ISRI has conducted a comprehensive evaluation of the accuracy of the leading OCR systems and the ISRI Voting Machine when processing DOE documents. Overall, errors were made on about 2% of the characters. However, the distribution of these errors was not uniform across the set of pages due to varying page quality. About 75% of the errors were made on the worst 20% of the pages.

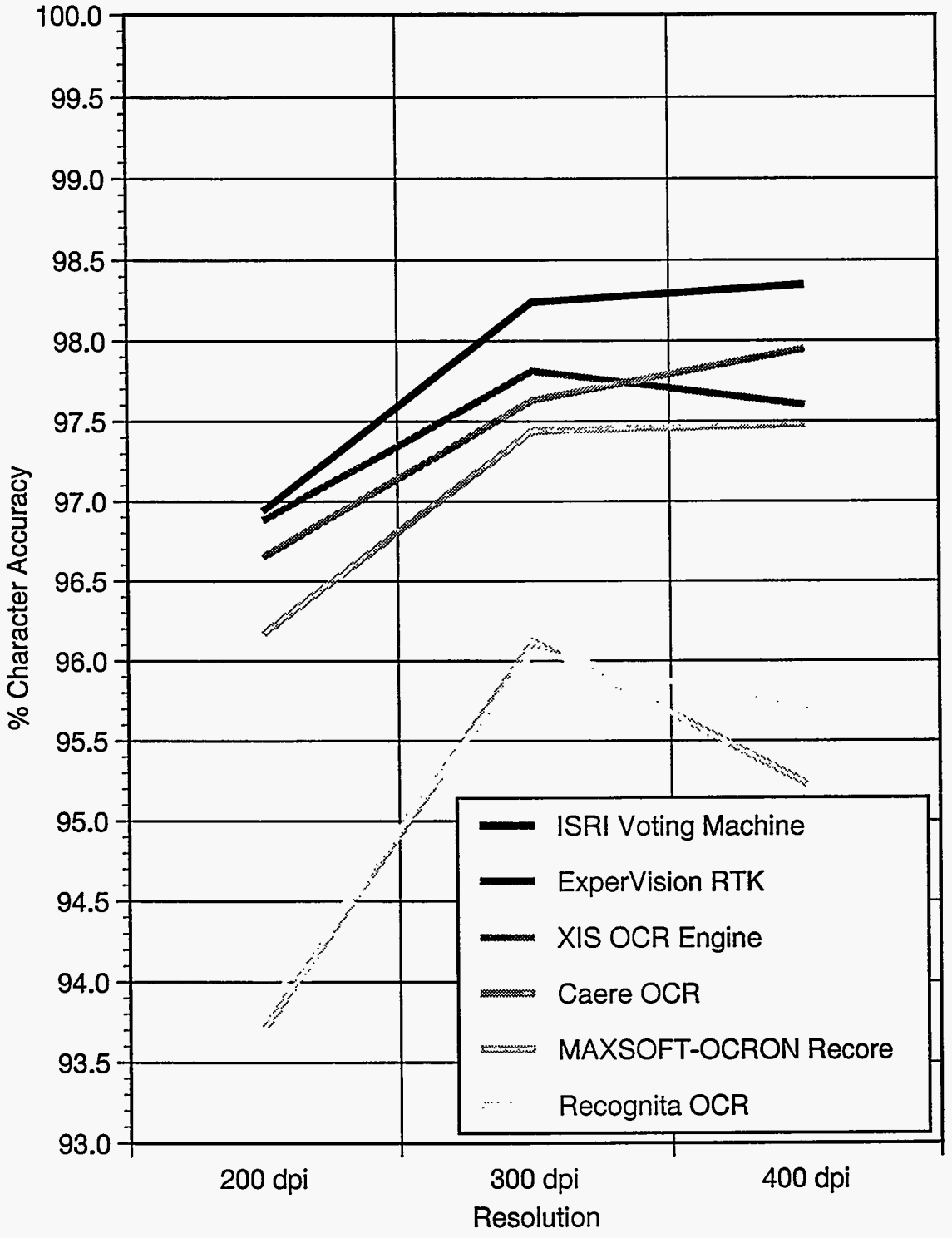
If OCR-generated text will be searched using information retrieval techniques, then it is important to consider the non-stopword accuracy of the text. Roughly one out of every 20 occurrences of non-stopwords was missed, but again, these misrecognized words were not distributed uniformly throughout the sample.

The ISRI Voting Machine is a tool for the automatic correction of OCR errors. In this test, it demonstrated that it is very effective for all but the worst quality pages.

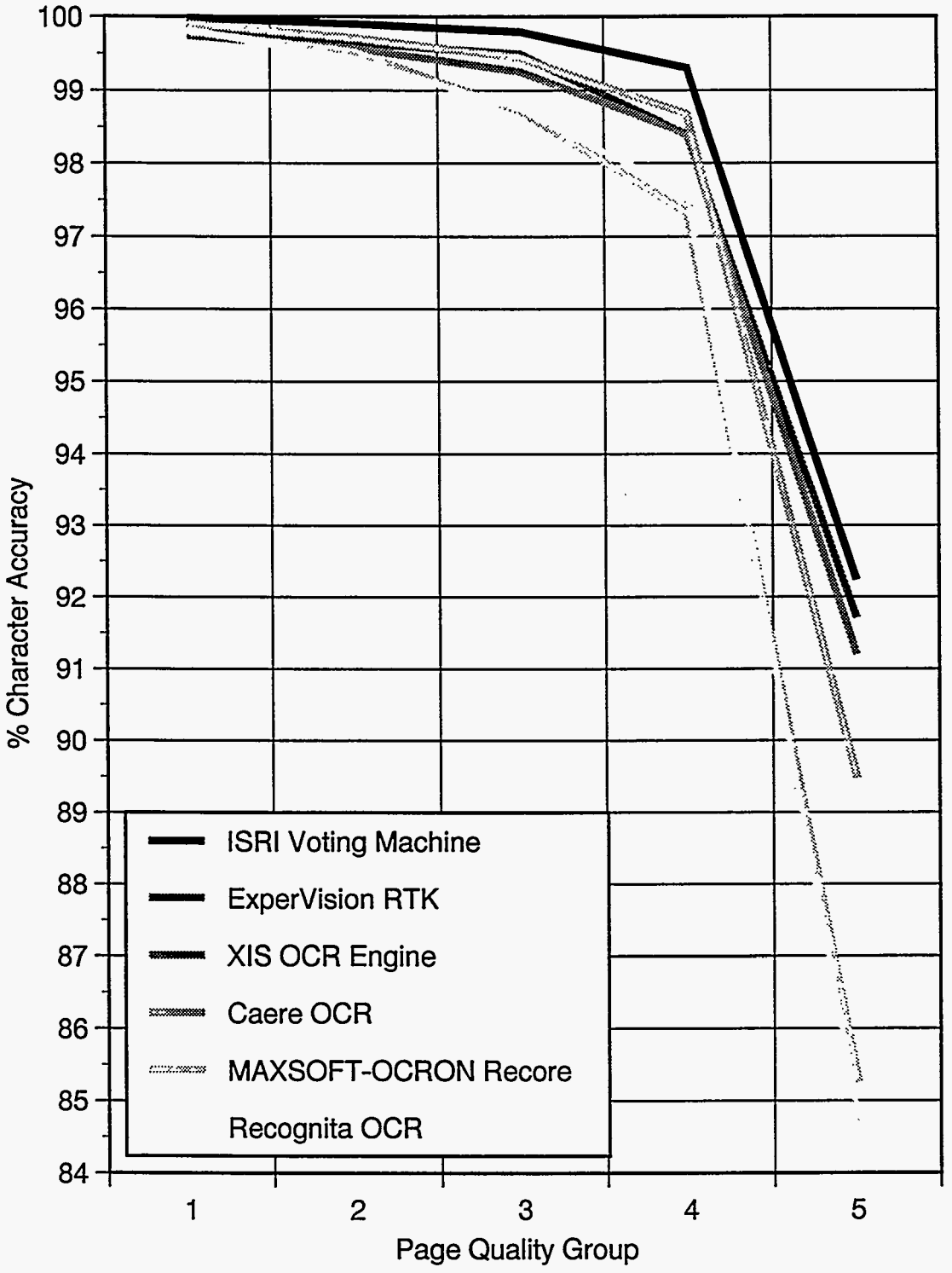
Graph 1: Character Accuracy



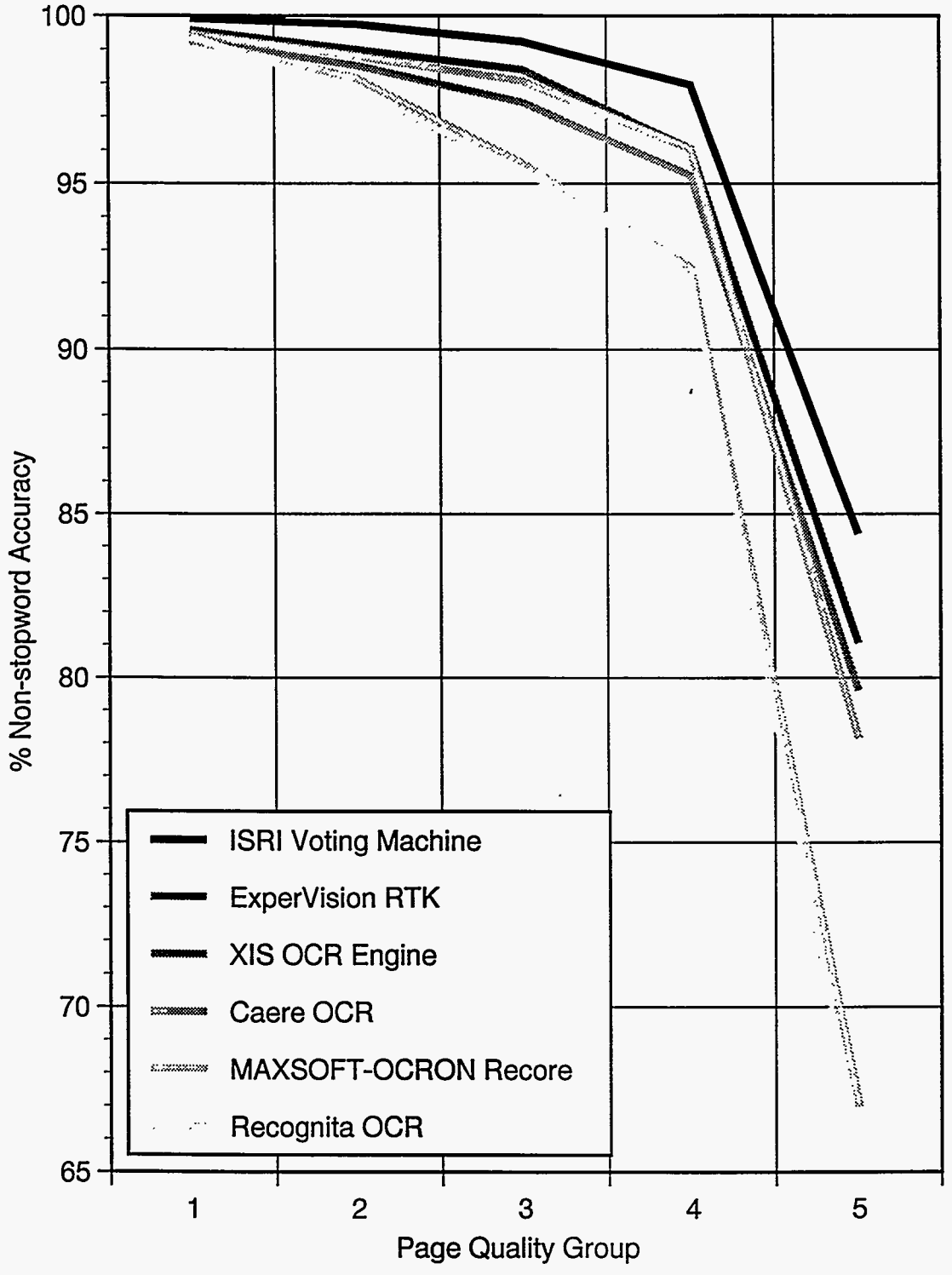
Graph 2: Effect of Resolution



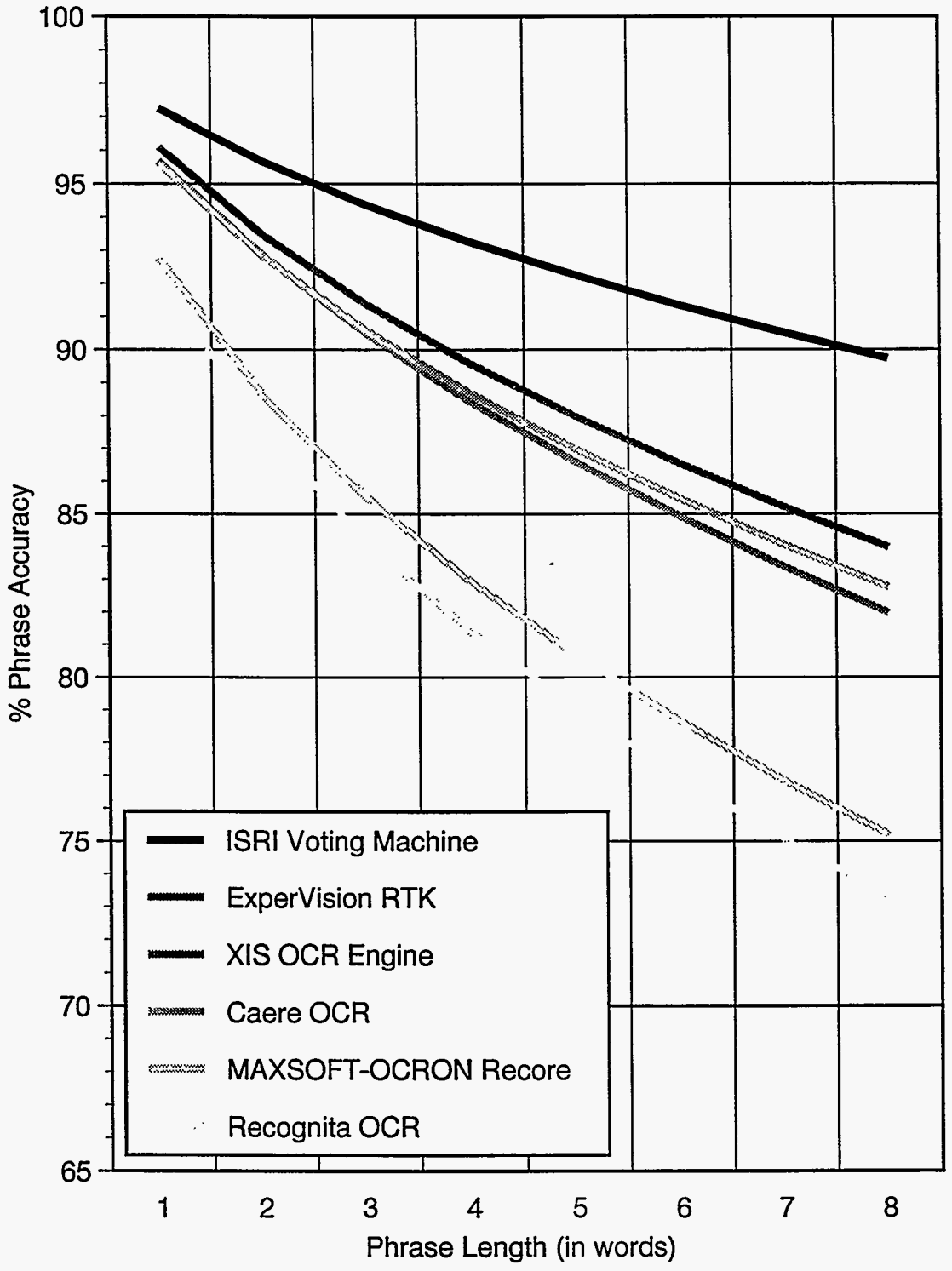
Graph 3: Character Accuracy vs. Page Quality



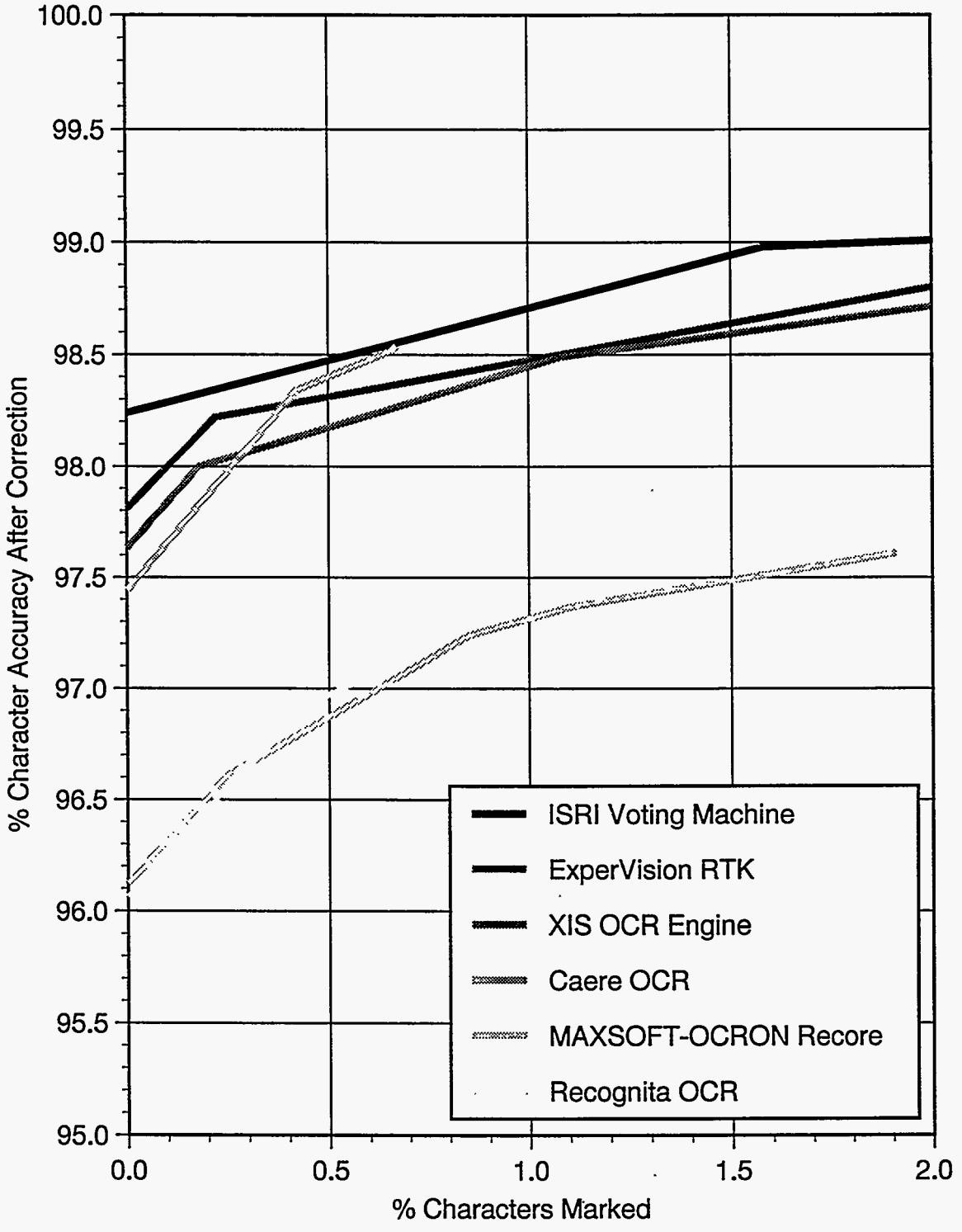
Graph 4: Non-stopword Accuracy vs. Page Quality



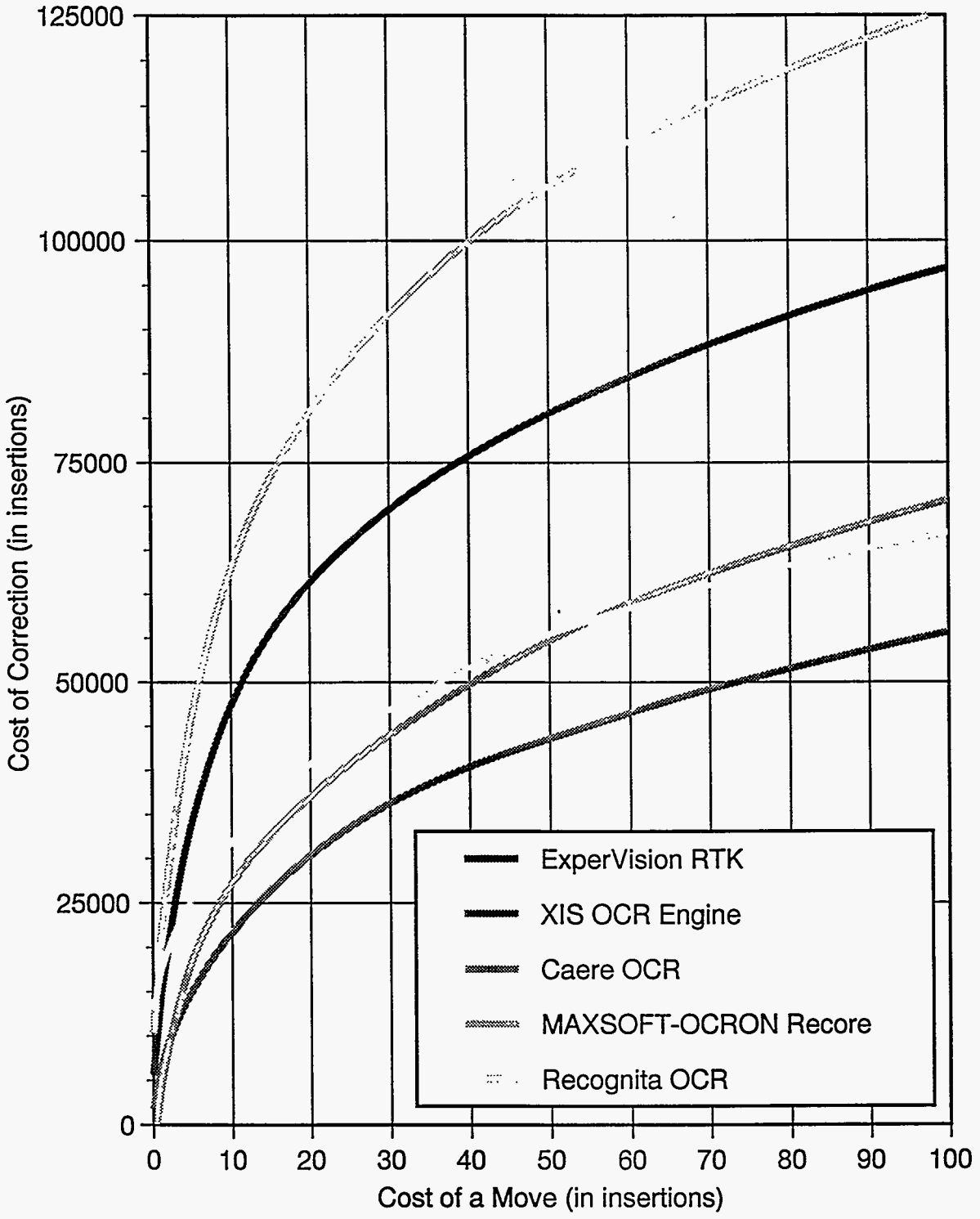
Graph 5: Phrase Accuracy



Graph 6: Marked Character Efficiency



Graph 7: Automatic Zoning



APPENDIX C.

Proposed agreement with the DFKI

AGREEMENT

As a part of an existing agreement between the

Information Science Research Institute (ISRI)
at the University of Nevada, Las Vegas
Las Vegas, NV, USA

and the

Deutsches Forschungszentrum fur Kunstliche Intelligenz GmbH
(DFKI - German Research Center for Artificial Intelligence)
Kaiserslautern, Germany

these parties also agree to the following terms of a cooperative project in the automatic analysis of German business letters.

DFKI will obtain a random sample of at most 2000 pages of German business letters and will scan (at 300dpi) and zone all images. DFKI will provide both a photocopy of the pages as well as an electronic copy of the images and zones to ISRI.

ISRI will undertake the task of preparing machine readable ground-truth text from these pages. ISRI will conduct up to three keyboard entries of each page and up to two synchronizing steps and will provide an electronic copy of the ground-truth of all zoned text to DFKI.

DFKI will reimburse ISRI for the cost of preparing this ground-truth text. The ISRI cost will be $15.00US/page.

DFKI and ISRI will each retain a hardcopy page for each invoice as well as an electronic copy of each page-image and the associated ground-truth text for in-house research.

It is agreed that all pages will be treated as confidential and, with the exception of small snippets of text that illustrate technical problems, will not be made public.

DFKI and ISRI will each perform independent research in OCR and text retrieval using these data and will share the results of their efforts.

For DFKI:

Prof. Dr. A. Dengel
Wissenschaftlicher Direktor

For ISRI:

Dr. T. A. Nartker
Director

For DFKI:

Dr. Andreas Harder
Kaufmaennischer Leiter

For the Board of Regents, University and Community College System of Nevada, on behalf of the University of Nevada, Las Vegas:

Dr. William E. Schulze
Director of Sponsored Programs