DOE/NV/10872--T187

UNLV
Information Science
Research Institute

Quarterly Progress Report

T. A. Nartker
March 31, 1995

Table of Contents

UNLV Information Science Research Institute
Quarterly Progress Report

T. A. Nartker
March 31, 1995

I. Board of Advisors Activity

A written report of both the Spring'94 and Fall'94 meetings is not yet available. Ms. Donna Harman, who acted as secretary for these meetings, has promised these reports soon.

The next Board meeting will be on Wednesday April 26th, after the SDAIR'95 Symposium.

II. Symposium Activity

The 1995 SDAIR Symposium is scheduled for Monday, Tuesday, & Wednesday, April 24-26 at the Desert Inn Hotel in Las Vegas. The four invited talks will be delivered by Dr. Richard Casey, Dr. Karen Sparck Jones, Dr. Abe Bookstein, and Dr. Hiromichi Fujisawa. Twenty five contributed papers and nine poster papers have been accepted. The Proceedings are being printed at this time. The Final Program is shown in Appendix A.

III. Staff Activity

Recruiting

None

Travel/ Meetings

In February, K. Taghva and J. Kanai traveled to San Jose, to present papers at SPIE'95. Kazem presented a paper titled "Evaluation of an Automatic Markup System" and Junichi presented a paper titled "Preliminary Evaluation of Histogram Based Binarization Algorithms."

In March, Shahram Latifi traveled to Snowbird Utah to present a paper "Compressed Image Manipulation," at the 1995 Data Compression Conference.

Papers accepted or presented

A paper by K. Taghva, J. Borsack, and A. Condit titled, "Effects of OCR Errors on Ranking and Feedback Using the Vector Space Model," was accepted for publication in Information Processing & Management.

A chapter for a Handbook of OCR and DIA titled "Information Retrieval and OCR" was submitted by K. Taghva, J. Borsack, and A. Condit.

IV. Document Analysis Program

OCR Devices

Testing for the ISRI fourth annual report of OCR accuracy has been completed. Eight organizations have submitted an OCR device(s) for this test. In addition to their binary engines, both Caere and HP have submitted greyscale engines. The international reputation of our annual tests is reflected in the number of devices submitted from foreign countries.

OCR Test system

Version #5 of our OCR experimental environment is complete and operational. It was used to conduct this years tests.

OCR Databases/GT1

All new datasets were completed and used for the fourth annual test of OCR accuracy. These include the DOE Sample#3 from the LSS prototype, the Magazine Sample, a Business Letter Sample, an English Newspaper Sample, and a Spanish Newspaper Sample.

Page 1 of Appendix B shows the ISRI methodology for preparing this ground-truth test data. Page 2 shows the status of all ISRI datasets on March 31. (The foreign language test data preparation shown is part of our DoD contract.)

OCR Databases/Foreign languages

As part of our contract with Fort Meade, we are preparing ground-truth text for a collection of newspaper articles printed in several foreign languages. We are collecting newspapers in Spanish, Chinese, and Japanese (as well as English).

As mentioned above, Appendix B gives an overview of our current ground-truth data preparation activities. Both the Spanish and the Chinese datasets are complete.

OCR Experiments

Testing, result analysis, and report writing are complete for our 1995 OCR Technology Assessment Report. This report is being printed at this time.

OCR Technical reports/thesis

Reports on "Prediction of OCR Accuracy" and on "Adaptive Restoration of Document Text" have been prepared and are included in our 1995 Annual Research Report.

Interaction with OCR vendors

None

Interaction with OCR research organizations

None

V. Text-Retrieval Program

TR Databases

We are continuing to collect relevance judgments for the queries associated with the residue (called GT1) of the LSS prototype database. We expect to complete all needed relevance judgments sometime this coming summer.

TR Experiments/Projects

Our text retrieval group has completed a prototype of the MANICURE OCR post-processing system. This system is available for testing by DOE (or TRW) staff. A report on MANICURE has been prepared and is being included in our 1995 Annual Research report.

TR Technical reports/thesis

"Autotag, A Tool for the Logical Markup of Documents" Allen Condit, M.S. Computer Science, May 1995. There are several other graduate students pursuing TR 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.

VI. Institute Activity

LSS working group meetings & report

During the Fall quarter, T. Nartker and K. Taghva participated with the LSS Technical Working group.

Institute visitors

Our most important visits during the Winter quarter were representatives from Argonne National Labs. They expressed interest in using Manicure to recover the text from a DOE database of documents involving human radiation experiments. They also expressed interest in building a team, including ISRI staff, to build the LSS.

Institute visitors this quarter:

Date	Visitor	Agency
2/09/95	Bob Kero & Lucian Russell	Argonne National Labs
2/14/95	Representatives from ARMA	Las Vegas Chapter, Am. Records Mgt. Assn.

Institute seminars

None

New agency contacts/ new research proposals

Our foreign language work for Fort Meade has been suspended. We are awaiting continued funding for this project.

VII. Goals Achieved/Goals for Next Quarter

Goals from last quarter:

1) The final program for SDAIR'95 is shown in Appendix A. The Proceedings have been sent to the printer.
2) No new Affiliate program members were added during the Winter quarter.
3) Our 1995 Annual Report is complete and has gone to the printer.
4) All preparation of data samples is complete. (see Appendix C) We expect to complete archival of these new datasets during the Spring quarter.
5) A version of MANICURE is complete and ready for initial testing.
6) Work has continued in collecting relevance judgments.

Goals for next quarter:

1) Select committee members and invited speakers for SDAIR'96.
2) Continue to recruit new members for the affiliates program.
3) Select datasets for our 1996 Annual OCR Test.
4) Continue work with the LSS Technical Working Group.
5) Investigate the use of MANICURE in the RDMS.
6) Continue work on relevance judgments for queries associated with the documents in GT1.

APPENDIX A.

SDAIR'95: Final Program

Appendix A removed.
Conf. Brochure

APPENDIX B.

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 Ground-Truth Databases (As of 3/31/95)

	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**
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	...	...	...									
Industry Annual Reports	**Yes**	**75**	**100**	**Yes**	**300**	**100**	**Yes**	**200**	**2,3,4b 3gs**	**67**						