Measuring the Accuracy of the OCR in the Making of America

A report prepared by Douglas A. Bicknese, completed in fulfillment of Directed Field Experience, Winter 1998, University of Michigan, School of Information

OVERVIEW

This project examined the text files created by Optical Character Recognition (OCR) software used in the University of Michigan Making of America project. The goal of this project was to develop a method for distinguishing accurate OCR files from OCR files with an unacceptable number of errors, without having to examine each file. The ability will enable the Digital Library Production Service to put online those OCR files with a high probability of accuracy and to estimate the amount of "clean-up" required to correct pages with an unacceptable number of errors.

CONTENTS

BACKGROUND

During the summer of 1996, the University of Michigan and Cornell University undertook a large scale digitization project entitled The Making of America. Through the use of an outside vendor, this project captured approximately 1,600 books and 50,000 journal articles as TIFF images, which are now accessible to the public through the web interface for the project. The University of Michigan's Digital Library Production Service developed a process to run the approximately 630,000 images through an OCR program with a minimal amount of human intervention. The OCR program created text files linked to the TIFF

images of the page. This allows the user to conduct full-text searches, greatly increasing the usability of MOA. The initial OCR work was done using the Xerox Imaging Systems OCR program ScanWorx, but was later redone for improved accuracy using the program Prime OCR.[1]

OCR accuracy varies from page-to-page depending on a number of variables. Many of the original book and journal pages have become brittle, faded, and foxed, while others have been written on or otherwise marred. Any of these factors can cause the program to mistake one letter for another. Ornamental text, illustrations and tables also cause problems for the OCR program, since it tries to make words out of the lines and dots in the illustration. Errors during scanning production, such as skew, also cause errors in OCR accuracy. While the number of errors on any given page may be few and are easily overlooked by the human eye, they cause problems for full-text searches. For example, a smudge on the page may cause the OCR to read “slavery” as “slavory", and thus the OCR will miss a match and the search results will not be comprehensive.

Examining the large number of text files generated by the OCR program line-by-line was too expensive and time consuming to be considered feasible by DLPS. This project sampled a portion of the OCR files created from the MOA monographs developed a method to easily separate the files felt to be “reasonably accurate” from ones requiring further “clean-up". This was done in three phases:

Phase I: The Sample Strategy

Since the time scale of this project precluded examining every image, a sample of OCR files was taken instead. In order to ensure that samples were taken from a variety of different books while still keeping the sample size manageable, the sample was stratified as follows:

5% of all titles less than 100 pages in length:

2% of all titles 100-300 pages in length

1% of all titles 301-500 pages in length

.5% of all titles over 500 pages in length

A list of all of the titles in each page range was then printed and assigned a number. Microsoft Excel’s random number generator feature was then used to randomly pick titles. For each title selected, five percent of the pages within each selected title were then sampled using Excel once again. Pages containing front matter, tables, illustrations, or advertisements were excluded from this sample. This decision was made due to the predictably adverse effect these factors had on OCR accuracy. If the random page selected contained one of these features, the next available page was chosen instead. By the end of the project, a total of 165 pages had been sampled and examined.

The selected pages were then printed and then compared to the original word-by-word. Time to examine a page once printed averaged about five minutes. Although we considered examining the pages on-line and using a spell checker, the added eye strain was taxing enough to justify the added effort of printing the

pages .[2]

Phase II: Developing a method for determining accuracy

When Prime OCR creates an OCR file, it also generates a number indicating the program's *confidence* of the OCR (hereafter referred to as Prime Score). The numbers generated range from 100 (total failure) to 900 (perfect). The staff at the University of Michigan's DLPS began with the assumption that the Prime Scores were reliable, but further investigation was needed before those scores could be relied on for any accuracy predictions. Other numbers generated by the program that proved useful were the character and word counts for the pages, as they helped to identify pages that contained an image or a blank page.

While the random pages were being selected, any pages in the sampled title with a low Prime Score were also noted. For this project, a low Prime Score was felt to be 750 or less. These pages were examined to identify what may have caused the problem. In most cases, the low score was due to the presence of an illustration or blank page and not due to poor OCR accuracy of any actual text on the page. These pages were usually easy to identify due to the low character count which accompanied the low Prime Score. Based on this examination, we concluded that a page with fewer than 100-200 characters and a low Prime Score probably contains an illustration or a blank page. Pages that had a low Prime Score and above 200 characters often proved to be tables, advertisements with layouts differing from the rest of the book, or pages that contained images and text in combination. However, some pages containing only text and having a large numbers of OCR errors were identified as well.

After each page was examined, the number of character and word errors was noted and compared with the character and word count for the page to determine the percent of accurate characters and words for that page. Since running heads were not deemed to be significant for retrieval, they were not examined in this project; consequently,.the number of words and characters in the page headers was subtracted from the total counts for the page before determining the percent accuracy. These figures were then stored in a spreadsheet and examined using simple features in the statistics program SPSS.

Phase III - Analyzing the results

The University of Michigan's Digital Library Production Service aims to identify pages it considers "reasonably accurate." For the purpose of this study, "reasonably accurate" was felt to be 99.0% or better character accuracy for pages which did not contain an illustration, blank page, table, or advertisement. Of the 165 images examined, 147 (89.1%) met this criteria. The mean accuracy of all of the pages felt to be "reasonably accurate" was 99.8511% for character accuracy and 99.3447% for word accuracy.

The next task was to identify the relationship between a page's Prime Score and its actual OCR accuracy. The goal was to identify a Prime Score above which it can be assumed that pages are "reasonably accurate." The tables below illustrate the level of accuracy for Primes Scores above 850, 860, 870, 880, and 890. [3]

Prime Score	Number of images less than 99.0% accurate	Number of images 99.0% accurate or better	TOTAL	Mean Accuracy	Hypothetical Errors
Less than 850	15 (88.3%)	2 (1.4%)	17 (10.3%)		
850 or above	3 (16.7%)	145 (98.6%)	148 (89.7%)	Char: 99.8147% Word: 99.2793%	Char: 4.6325 Word: 3.17108
TOTAL	18 (100%	147 (100%)	165 (100%)		

Prime Score	Number of images less than 99.0% accurate	Number of images 99.0% accurate or better	TOTAL	Mean Accuracy	Hypothetical Errors
Less than 860	16 (88.9%)	6 (4.1%)	22 (13.3%)		
860 or above	2 (11.1%)	141 (95.9%)	143 (86.7%)	Char: 99.8382% Word: 99.3663%	Char: 4.045 Word: 2.78828
TOTAL	18 (100%)	147 (100%)	165 (100%)		

Prime Score	Number of images less than 99.0% accurate	Number of images 99.0% accurate or better	TOTAL	Mean Accuracy	Hypothetical Errors
Less than 870	16 (88.9%)	11 (7.5%)	27 (16.4%)		

870 or above	2 (11.1%)	136 (92.5%)	138 (83.6%)	Char: 99.8439% Word: 99.3937%	Char: 3.9025 Word: 2.6672
TOTAL	18 (100%)	147 (100%)	165 (100%)		

Prime Score	Number of images less than 99.0% accurate	Number of images 99.0% accurate or better	TOTAL	Mean Accuracy	Hypothetical Errors
Less than 880	17 (94.4%)	20 (13.6%)	37 (22.4%)		
880 or above	1 (5.6%)	127 (86.4%)	128 (77.6%)	Char: 99.8627% Word: 99.4460%	Char: 3.4325 Word: 2.376
TOTAL	18 (100%)	147 (100%)	165 (100%)		

Prime Score	Number of images less than 99.0% accurate	Number of images 99.0% accurate or better	TOTAL	Mean Accuracy	Hypothetical Errors
Less than 890	18 (100%)	55 (37.4%)	73 (44.2%)		
890 or above	0 (0%)	92 (62.6%)	92 (55.8%)	Char: 99.8959% Word: 99.5856%	Char: 2.6025 Word: 1.82336

TOTAL	18 (100%)	147 (100%)	165 (100%)		

The results of the analysis of the sample used for this project confirmed a relationship between a high Prime Score and “reasonable accuracy.” When Prime Scores of 850 or above were examined, the number of pages with less than 99.0% accuracy was 16.7% of the sample. As the Prime Score cut off point was increased, the number of pages having less than 99% accuracy fell, with none being found when the Prime Score cut off was 890 or above. However, if the Prime Score is a predictor of OCR accuracy, an important point to consider is that increasing the Prime Score cut off point also increases the number of “false negatives.” A “false negative” in this case is when pages meeting the “reasonable accuracy” standard are removed from consideration because it’s Prime Score was less than the cut-off point. An example of this is when the Prime Score cut-off is set at 890: all of the cases with less than 99.0% character accuracy (18 cases, or 10.91% of the total sample) were removed from the sample. However, 37.4% of the pages which were felt to have “reasonable accuracy” were removed as well due to their Prime Score.

SUGGESTIONS FOR FUTURE WORK

The figures generated by the OCR software can be used to facilitate the quality control process for putting the MOA OCR text files on-line. Continued study needs to first factor in the use of the Prime Scores in identifying what pages should be corrected, and then factor in the use of the Prime Score in identifying which pages can be displayed.

Based on the information generated by this study, it appears to be reasonable to consider pages having a Prime Score of 880 or above as being “acceptable” to put on-line without spending time on further clean-up. In the case of this study, 128 of the 165 pages examined had a Prime Score of 880 or above (77.6% of the total sample). Of these, there was only one “false positive” which had a character accuracy of less than 99.0%. Using 880 as a Prime Score cut-off point excluded 20 “false negatives”; however, this is acceptable to help ensure that the number of “false positives” is minimal. The mean character percent accuracy using 880 as the Prime Score cut-off point was 99.8627% and the mean word percent accuracy was 99.4460%. Using these figures on a hypothetical page of 2500 characters and 440 words would result in 3.425 character errors and 2.376 word mistakes. While these figures are less than 100%, they are still reasonable enough to be used by most researchers.

If the OCR files are to be “cleaned-up” in the future, the numbers generated by the OCR software can be used to prioritize pages to be corrected. Pages with a Prime Score of less than 750 and fewer than 100 characters have a high probability of containing little or no text and cleanup could be postpone until later in the process. Goals could be set to correct pages within various Prime Score ranges (e.g. 850-870 for one time period, then 830-850 for the next, etc.). Since the pages with higher Prime Scores should contain few errors, clean-up should be relatively fast and would allow more pages to put on public display than would selecting the pages sequentially.

SUMMARY

The goal of this project was to develop a method to estimate OCR accuracy without having to examine every page and to easily separate "acceptable" and "unacceptable" pages. 165 pages were sampled from the Making of America's OCR files and a relationship was confirmed between the OCR program's predicted confidence and the actual character and word accuracy of the page. In the case of the MOA project, pages above a designated Prime Score cut off point could be assumed to be "reasonably accurate." This will greatly assist the "clean-up" effort for the project as it will help designate what pages should corrected as well as identify pages which contain little or no text and probably do not actually require any clean-up. The Prime Score cut off point is also potentially an important indicator in a process of identifying pages of text that can be shown to the end user.

1.For more specific information on the Making of America, see the "About MOA" webpage . *D-Lib Magazine* has also published two articles by key individuals in the project. Blumson and Shaw discuss the OCR conversion, SGML encoding, search engines, and interface presentation for the MOA project in Blumson, Sarr and Elizabeth J. Shaw, "Making of America, Online Searching and Page Presentation at the University of Michigan." *D-Lib Magazine*. July/ August 1997 on-line edition , viewed 4/22/98. John Price-Wilkin discusses MOA in his article on "on-the-fly" conversion in Price Wilkin, John. "Just-in-time Conversion, Just-in-case Collections, Effectively leveraging rich document formats for the WWW." *D-Lib Magazine*. on-line edition , viewed 4/22/98.

2. The latest generation of spell checkers, such as the one found on Microsoft Word 7.0, is more practical for on-line inspection than earlier generations of spell checkers. This type of spell checker displays the entire page and underlines the words not in the spell checker's dictionary. While there is still the added eye strain of on-line inspection, use of this type of spell checker may make on-line inspection of OCR files more practical in future projects.

3. Based on pages of 2500 characters/440 words, which we determined as an approximate number of characters/ words per page in the sample.