Boise State University
ScholarWorks

Electrical and Computer Engineering Faculty Publications and Presentations

Department of Electrical and Computer Engineering

1-19-2005

A Study of Style Effects on OCR Errors in the MEDLINE Database

Penny Garrison
Boise State University

Diane Davis
Boise State University

Tim Andersen
Boise State University

Elisa Barney Smith
Boise State University

Copyright 2005 Society of Photo-Optical Instrumentation Engineers. One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited. DOI: 10.1117/12.589408

THIS MATERIAL MAY BE PROTECTED BY US COPYRIGHT LAW (TITLE US CODE 17). THE REQUESTER IS LIABLE FOR INFRINGEMENT.

A Study of Style Effects on OCR Errors in the MEDLINE Database

Penny Garrison*, Diane Davis, Tim Andersen**, Elisa Barney Smith***
College of Engineering, Boise State University
Boise, Idaho, USA

ABSTRACT

The National Library of Medicine has developed a system for the automatic extraction of data from scanned journal articles to populate the MEDLINE database. Although the 5-engine OCR system used in this process exhibits good performance overall, it does make errors in character recognition that must be corrected in order for the process to achieve the requisite accuracy. The correction process works by feeding words that have characters with less than 100% confidence (as determined automatically by the OCR engine) to a human operator who then must manually verify the word or correct the error. The majority of these errors are contained in the affiliation information zone where the characters are in italics or small fonts. Therefore only affiliation information data is used in this research. This paper examines the correlation between OCR errors and various character attributes in the MEDLINE database, such as font size, italics, bold, etc. and OCR confidence levels. The motivation for this research is that if a correlation between the character style and types of errors exists it should be possible to use this information to improve operator productivity by increasing the probability that the correct word option is presented to the human editor. We have determined that this correlation exists, in particular for the case of characters with diacritics.

Keywords: OCR, style effects, error correction, OCR confidence

1. INTRODUCTION

The Lister Hill National Center for Biomedical Communications, a research division of the National Library of Medicine, has developed a system for the automatic extraction of bibliographic fields from scanned biomedical journal articles to populate their MEDLINE database [5,7,8]. This system, known as MARS or Medical Article Record System, initially populates the database by scanning and using OCR recognition. The results of the 5-engine OCR system are redirected through several algorithms that make corrections to the text. These algorithms were developed using dictionaries and previous OCR error research. After the data flows through these steps errors are still evident, requiring a human operator to correct the errors in order to ensure that the OCR accuracy is high enough. The operator is provided with an image of the word in question from the original scanned document and a list of potential words. The operator can scroll down through the list and select one, or type in the whole corrected word. If the correct word is not in the top 2-3 words the operator usually chooses to type in the word manually, which typically requires more time than selecting the correct word from the top of the list. Therefore, the ordering of the available word options directly affects the time and thus the cost of the OCR error correction step.

Much research has been done on how to deal with errors produced by OCR engines. The way the OCR engine searches for matching words can influence the accuracy. In [9] search terms are weighted probabilistically to increase search accuracy. Other approaches for searching OCR text are surveyed in [1]. These types of approaches work well during the search process, but are not appropriate for the case where the OCR text is going to be viewed by a human.

Another approach for improving OCR accuracy is to recognize which documents are likely to produce OCR errors, using such measures as white speckling, noise recognition, broken characters, and etc., and manually re-key only the documents that are determined to have problems [2]. The drawback of this approach is that recognition of noisy documents is not trivial, and generally requires that the system have some understanding of the type of data contained in the document. For example, documents that contain pictures are particularly difficult to label as noisy or degraded, unless you have some understanding of what the picture should look like.

* Penny Garrison is now with POWER Engineers, Inc., Boise, Idaho.
** tim@cs.boisestate.edu, http://coen.boisestate.edu/EBarneySmith/SP_lab
*** EBarneySmith@boisestate.edu, http://coen.boisestate.edu/EBarneySmith

Document Recognition and Retrieval XII, edited by Elisa H. Barney Smith, Kazem Taghva,
Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 5676 © 2005 SPIE and IS&T
0277-786X/05/$15

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms

Other approaches attempt to detect and correct OCR errors after they occur. The correction process can be at least partially automated as in [4,6,8,10], but may also have a manual step if accuracy requirements are particularly high. One method that MEDLINE employs is to segment the scanned document into zones, such as Author or Affiliation. After the zones are labeled this automated process verifies the spelling of low confidence words and then reformats the text to the National Library of Medicine's conventions.

This paper is focused on the task of detecting and classifying the types of OCR errors that currently occur in the MEDLINE MARS system. With most OCR engines, it is likely that a correlation exists between the character style, confidence level, and the errors that the OCR engine produces [3]. This research confirms that this correlation exists for the OCR engine used to process the MEDLINE database. Given this correlation, it should be possible to make a better guess as to the correct word and put it higher on the list of potential candidates, thus improving the efficiency of the manual correction process with a corresponding reduction in cost.

One of the complications that was encountered in doing this research was in trying to match the corrected text with the OCR output. In literature related to this subject, it was noted that many methods had adopted the approach of using dictionaries to help match commonly used words [1,6,8]. Unfortunately, the matching that we were attempting was not using common words, but rather people's names and their affiliations. Other research used historical datasets to match words that would not be found in a dictionary, but would be common for the type of document being processed. A historical dataset would have been helpful to correctly match some of the information, but was unavailable for this research.

A brief introduction and explanation of the types of OCR data and corrected data we encountered is described in Section 2. The string matching problem and our approaches to it are discussed in detail in Section 3. Section 4 gives the results of the experiments, and Section 5 gives the conclusion.

2. DATA

The data used in this analysis is from the Lister Hill National Center for Biomedical Communications. Over 25,000 records were used. An example of the format of one record is shown in Figure 1. The OCR engine result is OCR:TXT. OCR:CON lists the confidence of the OCR engine and OCR:ATT lists the style attribute value. The final results from the MARS operators are in FNL:TXT.

The OCR system assigns a confidence level and style attribute to each character that it processes. The confidence levels range from 1 to 9 with 9 indicating that the OCR system is 100% confident that the letter is correct. The style attributes are created by adding all appropriate style numbers together. Some common attributes in hexadecimal numbers are:

Normal	1	Italic	8
Bold	2	Superscript	10
Underline	4	Subscript	20

An attribute of 3 would indicate that a Normal, Bold font was used. The guidelines we were given on the attributes said that a Normal attribute would be included in every attribute, but when the data was analyzed, several instances of attributes appeared without it. One attribute we were especially interested in was 9, a Normal Italic Font, since more errors were expected to occur when the text is italicized.

OCR:TXT:Coronary Care Unit, Department of Cardiology, Hospital "S. Maria della Misericordia", Udine, Italy
OCR:CON:99999998999999888899999998999999999999999998888999999999999999989998889999999999989999999888888888
OCR:ATT:99999999999999999999999999999999999999999999999999999999999999II99999999999999999999999999II99999999999999
FNL:TXT:Department of Cardiology, Hospital S. Maria della Misericordia, Udine, Italy.

Figure 1: Example of data format.

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms

3. ANALYSIS

3.1 Manual Analysis

To better understand the matching that should occur between the OCR:TXT and the FNL:TXT we began by manually matching words and creating confusion matrices based only on the confidence level. In general, the process of matching the words output by an OCR engine with those in human corrected text is not trivial. But for this project the problem was exacerbated due to the instructions that the human operators were following in correcting OCR output. While there are much more complicated examples, the example in Figure 1 illustrates some of the difficulty that we encountered in matching OCR output with manually corrected text. In order to "reduce labor costs and increase overall system reliability" [4], the human operators were not required to correct all of the OCR output, but rather only required to correct those sections of the text deemed important. In addition, the order of the text was often changed to fit a standard ordering that had been predetermined. This greatly complicates the problem of matching OCR text with final text. Also punctuation and capitalization was changed to fit a standard. This introduces mismatches that are classified as "errors" by our automated matching process. While these are not per se OCR errors (since the OCR engine did not mis-recognize what was in the original document), it is difficult to automatically separate them from the true OCR errors, and they are in fact operator-entered, and therefore *desired* substitutions for the final text. Therefore, for the remainder of this paper they are considered together with the errors.

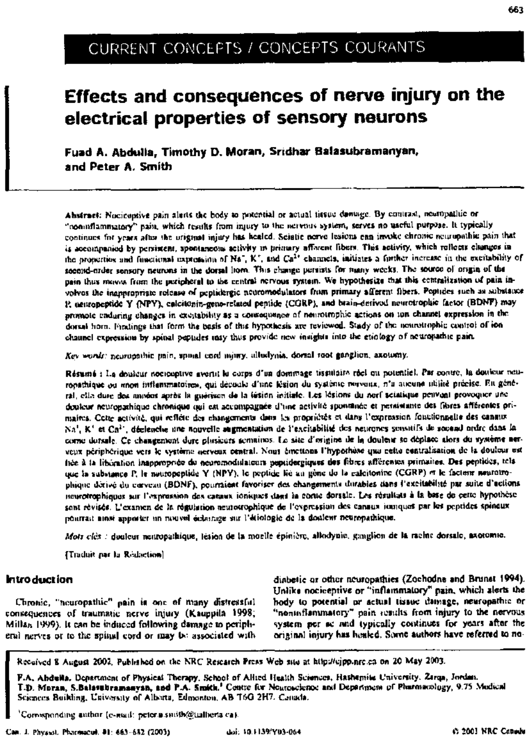
The first observation we made when analyzing the data was that there is a very high percentage of characters that have a confidence of 9, indicating that the OCR engine is in fact very accurate (or at least thinks that it is). However, it was noted that there were substitutions of characters with the confidence of 9. We had assumed that when the OCR engine had tagged a character as having a confidence of 9, then the OCR engine was 100% confident that the character was correct. However, there were 0.4% substitutions found with a confidence of 9. About 6% of these substitutions were problems where upper case letters had been changed to lower case letters (i.e. Ee, Oo, Bb). The majority of the remaining substitutions included removing an "s" or punctuation from words. While these were indicated as errors, they were probably caused by the reformatting of the text from the MEDLINE algorithms.

Some of the substitutions found included such things as extra characters at the beginning of some words, and several issues with diacritics. To gain a better understanding of these problems, and to better understand how the data changed during the entire process, for a couple of the errors we tracked down photocopies of the journal articles from which they were taken. Since the information we are presented with is minimal, finding examples was difficult. An example is shown in Figure 2. This process helped us to better understand the differences in the fonts and locations of the pertinent information.

Several of the errors resulted from non-standard writing practices. For instance, in several articles, the section beginnings were indicated by a bullet that resembled a cross, †, which was (understandably) recognized by the OCR process as a t. There were similar cases where bullets were used which the OCR recognized as an asterisk. Another issue we found was when a superscript number preceded the text. When the superscript was a 1 the OCR recognized it as a ', where as any other superscript number was translated to a normal character of that number. This type of error can be seen in Figure 2. These characters were also deleted before the FNL:TXT was used to populate the database. A further complication was seen in several articles from Korea, where there were several instances where an s was attached to the end of a word where it was not needed. For these cases the s had been deleted in the corrected text, even though it actually existed in the source document.

The final errors we manually investigated were the errors associated with diacritics. Examples of these errors are shown in Figure 3. These errors were also corrected either by algorithms or by the manual operators. To obtain meaningful results from our program it was necessary to take these errors into consideration while developing our string matching algorithms. Manual investigation of these errors enabled us to develop algorithms better suited to catch these errors and include them in our end results.

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms


663

CURRENT CONCEPTS / CONCEPTS COURANTS

Effects and consequences of nerve injury on the electrical properties of sensory neurons

Fuad A. Abdulla, Timothy D. Moran, Sridhar Balasubramanyan, and Peter A. Smith

[Traduit par la Rédaction]

Introduction

Chronic, "neuropathic" pain is one of many distressful consequences of traumatic nerve injury (Kauppila 1998; Millan 1999). It can be induced following damage to peripheral nerves or to the spinal cord or may be associated with diabetic or other neuropathies (Zochodne and Brunet 1994). Unlike nociceptive or "inflammatory" pain, which alerts the body to potential or actual tissue damage, neuropathic or "noninflammatory" pain results from injury to the nervous system per se and typically continues for years after the original injury has healed. Some authors have referred to no-

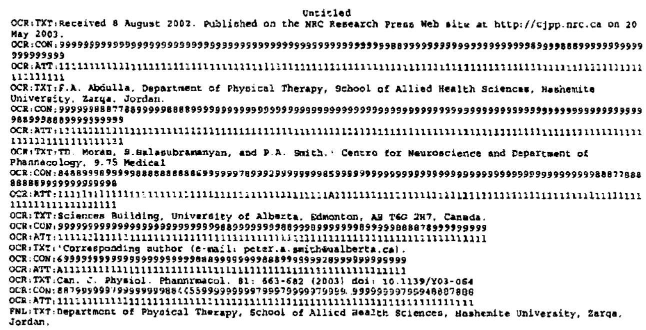
Received 8 August 2002. Published on the NRC Research Press Web site at http://cjpp.nrc.ca on 20 May 2003.

F.A. Abdulla. Department of Physical Therapy, School of Allied Health Sciences, Hashemite University, Zarqa, Jordan.
T.D. Moran, S.Balasubramanyan, and P.A. Smith.[1] Centre for Neuroscience and Department of Pharmacology, 9.75 Medical Sciences Building, University of Alberta, Edmonton, AB T6G 2H7, Canada.

[1]Corresponding author (e-mail: peter.a.smith@ualberta.ca).

Can. J. Physiol. Pharmacol. 81: 663–682 (2003) doi: 10.1139/Y03-064 © 2003 NRC Canada

(a)


Untitled
OCR:TXT:Received 8 August 2002. Published on the NRC Research Press Web site at http://cjpp.nrc.ca on 20 May 2003.
OCR:TXT:F.A. Abdulla, Department of Physical Therapy, School of Allied Health Sciences, Hashemite University, Zarqa, Jordan.
OCR:TXT:TD. Moran, S.Balasubramanyan, and P.A. Smith.' Centro for Neuroscience and Department of Phannacology, 9.75 Medical
OCR:TXT:Sciences Building, University of Alberta, Edmonton, AB T6G 2H7, Canada.
OCR:TXT:'Corresponding author (e-mail: peter.a.smith@ualberta.ca).
OCR:TXT:Can. J. Physiol. Phannrmacol. 81: 663-682 (2003) doi: 10.1139/Y03-064
FNL:TXT:Department of Physical Therapy, School of Allied Health Sciences, Hashemite University, Zarqa, Jordan.

(b)

Figure 2: Example of (a) actual document and (b) results returned from OCR.

True Text	OCR Text	Final Text
Génomes	G6nomes	G'enomes
Génétique	Genetique	G'en'etique

Figure 3: Examples of OCR errors involving diacritics.

3.2 Matching Algorithms

The automated matching was done with a three-tiered approach. The first matching algorithm developed compared the OCR:TXT data and the FNL:TXT data. The algorithm determined if there were any words that exactly matched. If words matched, then we determined if there was only one instance of that word in the parsed data. As we matched these unique and exact words, markers were set to use as placeholders for use in subsequent passes. These markers indicated that a match had occurred at that location.

The markers placed in the first unique and exact match algorithm were then used to invoke the unique and exact match algorithm again. Now instead of searching the entire line, the algorithm used the markers to locally search for unique and exact matches. Beginning at a marker and going until the next marker, the algorithm would search for new matches. Although these matches might not be unique when looking at the entire line, they could be unique if looking at a subset of that line. Because the OCR:TXT lines were for the most part quite accurate, the combination of the two runs of the unique and exact match algorithm resulted in a high number of matched words. In fact after the second run of the algorithm an average of 50% - 60% of the words were matched. By matching this high percentage of words using this simple algorithm, the speed of the process was increased by not having to use the minimum error matching algorithm on every word.

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms

Following the gathering of the unique and exact matches a last pass was required to determine high-confidence, inexact matches between words in the OCR.TXT and FNL.TXT (in other words, matches between words that contained character substitutions). By using the markers set in the unique and exact match algorithm we were able to significantly reduce the number of pair-wise word comparisons this step required. Because the number and order of the words in FNL:TXT could have been changed from the original OCR output, it was not possible to simply match words in a linear fashion based on the markers generated by the unique and exact word matching algorithms. Rather, the algorithm proceeded outward from the markers in a linear fashion matching those words that differed by a maximum of 25%. Thus, a word with 10 letters could have at most 2 errors in it. In order to accomplish this, we needed to compare letters within words rather than the entire word itself. The algorithm walked through every letter of the OCR:TXT word and compared it to letters in the FNL:TXT word. A word in the OCR:TXT data would be compared to most words in the FNL:TXT data. The comparison that yielded the most matches was flagged as the minimum error match. After this final pass of the algorithm, between 80 and 85% of the words in the FNL.TXT were matched to words in the OCR.TXT in a typical document.

If an insertion or deletion of a letter was found, it was treated as a substitution of a 2-character pair. During the string matching algorithms if there was not a one to one mapping of the letters; two or more characters could be mapped to one, or characters could even be mapped to NULL. For example, a "G" could be matched to "(;". In all, the algorithm had the ability to create 400 character combinations.

4. RESULTS

After the strings were matched, each matched OCR:TXT character was placed in a confusion matrix that corresponded to its attribute and its confidence level. If the data in a matrix was all on the primary diagonal, and no additional data was scattered throughout the remainder of the matrix, then the OCR system accurately read all the data for that confidence and that attribute. If the data was on a secondary diagonal, the diagonal that would be created when the OCR misidentified the case of a letter (i.e. Cc, Oo, Pp, etc.), or when the operator or MEDLINE's algorithms had changed the case. If however, data was not on the main or secondary diagonal than we could use these substitutions for further analysis.

Ten confidences were used, 1 through 9 and a null confidence. The NULL confidence occurred when a letter in the FNL:TXT did not match up with a letter in the OCR:TXT, such as if the final text has added extra words like "USA". Single character insertions would take the confidence of the adjacent character. After scanning the data a total of 34 style attribute combinations were discovered. Those combinations corresponded to the hexadecimal values of 1, 3, 5, 7, 9, 41, 43, 45, 49, 51, 53, 59, 69, 71, 73, 79, 2E, 3B, 3D, 3F, 4B, 4D, 6E, B1, B3, B5, B9, BB, C1, C3, C9, E2, EB, FC. A hexadecimal value of 41 would correspond to the attribute combination of fixed and normal. A null attribute combination was added to the scanned list of 34 to make a total of 35. Not all of these attributes and confidences occurred in parts of the text that were properly matched and further analyzed. With 10 confidences, 35 style attribute combinations and up to 400 characters or character pairs, a total of 350 400 X 400 matrices were created to analyze the data.

Following the creation of the original matrices, 45 new matrices were created based upon either attribute or confidence alone. From all the matrices, tables were built summarizing the distribution of errors by style and confidence. Table 1 summarizes the results from 35 matrices based upon attribute combinations only. Table 2 summarizes the results from the 10 matrices based upon confidence only. If the number of characters was not low and if the ratio of error/character appeared high then those matrices would be evaluated further. Based upon all of the following information we chose to further analyze the data in the confusion matrices of the following cases:

1) Attribute 1 (normal)
2) Attribute 9 (normal, italic)
3) Attribute 3B (normal, subscript, superscript, italic, bold)
4) Attribute 41 (normal, fixed)
5) Attribute 49 (normal, fixed, italic)
6) Attribute B1 (normal, subscript, superscript, sanserif)
7) Attribute B3 (normal, sanserif, italic, subscript, superscript)
8) Attribute C9 (normal, fixed, sanserif, italic)
9) Confidences 1-9

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms

Table 1 –Error Frequency by Style Attribute for Matched Characters

	# char	# errors	% errors		# char	# errors	% errors
Attribute 1	698427	3089	0.40%	Attribute 69	0	0	0.00%
Attribute 3	8609	41	0.50%	Attribute 6E	0	0	0.00%
Attribute 5	12	1	8.30%	Attribute 71	0	0	0.00%
Attribute 7	0	0	0.00%	Attribute 73	0	0	0.00%
Attribute 9	693431	3236	0.50%	Attribute 79	0	0	0.00%
Attribute 2E	0	0	0.00%	Attribute B1	156388	738	0.50%
Attribute 3B	2131	34	1.60%	Attribute B3	7905	84	1.10%
Attribute 3D	253	0	0.00%	Attribute B5	0	0	0.00%
Attribute 3F	0	0	0.00%	Attribute B9	132306	617	0.50%
Attribute 41	1431	28	2.00%	Attribute BB	1416	7	0.50%
Attribute 43	17	0	0.00%	Attribute C1	295	5	1.70%
Attribute 45	1	0	0.00%	Attribute C3	11	0	0.00%
Attribute 49	8982	105	1.20%	Attribute C9	1178	23	2.00%
Attribute 4B	76	6	7.90%	Attribute E2	0	0	0.00%
Attribute 4D	0	0	0.00%	Attribute EB	0	0	0.00%
Attribute 51	2	0	0.00%	Attribute FC	0	0	0.00%
Attribute 53	0	0	0.00%	Attribute NULL	30	4	13.30%
Attribute 59	0	0	0.00%	Total	1712901	8018	0.50%

Table 2 - Error Frequency by Confidence for Matched Characters

	# char	#errors	% errors
Conf 1	524	30	5.7%
Conf 2	2375	114	4.8%
Conf 3	2	0	0.0%
Conf 4	5204	256	4.9%
Conf 5	14316	212	1.5%
Conf 6	22093	205	0.9%
Conf 7	21931	140	0.6%
Conf 8	188985	1169	0.6%
Conf 9	1457471	5892	0.4%
NULL	0	0	0.0%
Total	1712901	8018	0.5%

4.1 Diacritics & Special Symbols

Letters, or combinations of letters, that are not common in written English often cause errors in the OCR process. Some of those symbols included:

§	'e	'a	'i	'o	'u	o'
"a	e'	a'	e`	´e	n`	^e

Many of the errors that occurred were due to diacritics. Since the diacritic is not usually used in English names and documents, an error would be detected. These differences in writing styles could be corrected before the data goes to the operator by adding a classifier to the OCR system that is able to recognize these characters, or adding a multicultural dictionary that could correct these errors.

One thing that complicates the analysis is that there is more than one character that can be used as a diacritic. As can be seen above 'e, e', ´e, and `e were all found in the text, however it is not clear in all cases whether the individual diacritics represents a grave or an acute. Another complication is that the diacritic is only found either in the final text or the OCR text, not both. This leaves a dilemma as to which character the diacritic should be attached. For example, the French word "Génétique" had the OCR output of "genetique", the operator changed the word to "g'en'etique" since they were given those instructions to compensate for the English standard keyboard. When trying to evaluate the error between "genetique" and "g'en'etique", a decision must be made on whether the algorithm should attached the diacritic to the "g" and "n" or the "e"s. It was decided that all diacritics would be attached to an adjacent vowel if one existed. This biased our results to highlighting errors associated with diacritics.

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms

Table 3: Occurrence of Substitution Errors by Attribute and Confidence

	OCR:FNL				
OCR:TXT	**Attr 1**	**Attr 9**	**Attr B1**	**Conf 4**	**Conf 8**
(;->G	100.0%	0.0%	0.0%	100.0%	0.0%
b->'e	0.0%	100.0%	0.0%	0.0%	100.0%
e->'e	28.6%	0.0%	0.0%	0.0%	0.0%
e->'o	0.0%	33.3%	0.0%	0.0%	0.0%
g->'e	33.3%	0.0%	0.0%	0.0%	0.0%
g->gy	66.7%	0.0%	0.0%	0.0%	100.0%
i->'a	17.6%	5.3%	0.0%	0.0%	11.1%
i->'e	70.6%	89.5%	90.0%	0.0%	88.9%
o->'a	50.0%	20.0%	0.0%	0.0%	50.0%
o->'e	50.0%	0.0%	0.0%	0.0%	50.0%
r->ry	40.0%	80.0%	0.0%	100.0%	75.0%
s->'e	90.0%	100.0%	0.0%	0.0%	0.0%

Table 4: Most Frequent Substitution Errors by Attribute and Confidence

	OCR:FNL				
OCR:TXT	**Attr 1**	**Attr 9**	**Attr B1**	**Conf 4**	**Conf 8**
(;	G			G	
b	B	'e			'e
e	'e	'o	'a	er/ee	ee
g	gy	gg	gy	gy	gy
G		g			g
i	l	'e	l	l	'e
l	i	L			'e
o	'e	'a			'a
r	ry/'i	ry	'e/rd	ry	ry
s	'e	'e			

4.2 Substitution Errors

The original hypothesis was that the style of the font would affect the substitution errors found. The frequently occurring substitutions were analyzed by attribute and confidence. In Table 3 some common OCR substitution pairs are shown. The leftmost column shows the output of the OCR engine followed by the correct character substitution. The remainder of the columns show the percentage of the errors that were from that given substitution. The percentage of substitution errors differ based upon attribute style. For example, 50% of the OCR errors were incorrectly identifying a character o as an `e when the character had an attribute of 1, while the same error never occurred under attribute 9 and B1.

If this trend is to be exploited to identify the best substitution for uncertain characters, the most frequently occurring substitution by style is needed. Table 4 shows some of the characters most frequently in error and their most common substitution by attribute and confidence level. Blanks indicate that the particular character was never found in error for that attribute or confidence class in the data analyzed. The top ranking substitution again varies by attribute and confidence level.

Several other common errors were also identified. These are discussed next.

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms

4.3 Punctuation

Punctuation was the cause of many substitutions in most attribute combinations. There were substitutions in which punctuation had been added, and also where punctuation had been removed. This resulted in some substitution pairs such as s → s, . It was unclear if these substitutions were occurring from the OCR engines, or if the MEDLINE algorithms or the operator had changed the format.

4.4 OCR Errors vs. Operator and Algorithm Alterations

Since the data we received had already been analyzed and changed by the MEDLINE algorithms and operators, it was very difficult to determine what the actual OCR error was. Removing or adding punctuations and rearranging words could be attributed to operator or algorithm alterations. But to isolate these changes from OCR errors was impossible without the MEDLINE algorithms and information on the format the operator is supposed to follow.

Of particular note is the fact that there were instances of the spelling being altered. In one case a word was spelled in the original document as Coloumbia, and the operator changed it to Columbia in the re-keyed text. We were unsure if a spelling error had occurred several times in the original documents, or if this was a cultural way to spell that word. Capitalization, as previously mentioned, also falls into this category.

Therefore, several of the substitutions that have been identified might actually be changes required by MEDLINE. For this particular application, the net result of showing the operator these substitutions as correction options is desirable whether they are from OCR errors or from MEDLINE specified formatting. Future research on the correlation between attribute and OCR substitutions should attempt to use raw data from the OCR engine and simple ground truth corrections. Data that has had the format changed should not be used.

4.5 Other

One error that is not prevalent in the results, but was still interesting, was the misrepresentation of a G as "(;" . The formation of the "characters" is very similar, and one could easily see how the OCR could make this mistake. This error happened multiple times but only with a normal attribute and would often have a confidence of 2. Although this is not an error with a high rate of occurrences, a simple replacement rule could be implemented before the operator is called upon to manually correct the OCR output, thus saving significant time.

5. CONCLUSION

Manual analysis of the data indicated some problems and additional manual analysis could have been helpful. Often when writing an analysis program, one writes for the general case since the specifics are unknown. With this in mind, many errors could have been misidentified. A case in point was the matching of the diacritics. Initially, we wrote the matching to look at confidence levels. But, after viewing the results, it became apparent that we needed to match directly to a vowel. With a general approach many extra letter errors may be incorrect. Although a system that analyzes the data in seconds is desirable, in some cases it could taint the data being returned. Except for diacritics, that was a tradeoff we were willing to accept. Continued retooling of the matching algorithm with the data and knowledge we have gathered would give us a much more accurate picture and minimize any errors caused by generalization. Further enhancement to the matching algorithm to allow better matching of low confidence words would also increase the number of substitution pairs seen and allow for better error analysis.

The amount of data contained in the confusion matrices created by the automatic categorizing program is overwhelming. Only 366 of the 400 available character pair possibilities were used in each matrix and still the results contained nearly 50 million data cells that needed to be analyzed. We chose to only analyze 17 matrices to see patterns in errors. This was still over 2 million data cells. Even with a large amount of data, we were able to see specific patterns when dealing with diacritics, punctuation, and special character pairs such as "(;". Using the information we gathered for these types of character substitutions, is should be possible for MARS to develop new algorithms that could help reduce operator correction costs.

Based on the analysis of substitution errors, it was determined that a correlation does exist between attribute class and substitution errors. A correlation also exists for confidence level, but the attribute class is more significant than the confidence level. Since the correct substitution for an OCR error depends on the font attribute(s) detected by the OCR

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms

engine, by taking these attributes into account it is possible to more accurately predict the correct character, which would reduce the costs of the manual OCR error correction step over an approach that does not use this information.

ACKNOWLEDGEMENTS

This work was supported by the Computing Research Association's Committee on the Status of Women in Computing Research (CRA-W), Collaborative Research Experience for Women in Undergraduate Computer Science and Engineering (CREW).

REFERENCES

[1] Beitzel, Steven M., Eric C. Jensen and David A. Grossman. "A Survey of Retrieval Strategies for OCR Text Collections." 2003 Symposium on Document Image Understanding Technologies, Greenbelt, MD, pp 145-151 (2003).

[2] Blando, Luis R., Junichi Kanai, Thomas A. Nartker and Juan Gonzalez. "Prediction of OCR Accuracy." Technical Report 92-02, Information Science Research Institute (1992).

[3] Chen, Su, Suresh Subramaniam, Robert M. Haralick and Ihsin T. Phillips. "Performance Evaluation of Two OCR Systems." Proceedings of Fourth Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, pp 299-317 (1994).

[4] Ford, Glenn, Susan Hauser, Daniel X. Le and George R. Thoma. "Pattern matching techniques for correcting low confidence OCR words in a known context." Document Recognition and Retrieval VIII, Proceedings of SPIE Volume 4307, pp 241-249 (2001).

[5] Ford, Glenn M., Susan E. Hauser and George R. Thoma "Automatic Reformatting of OCR Text from Biomedical Journal Articles." Proc. 1999 Symposium on Document Image Understanding Technology, College Park, MD, pp 321-325 (1999).

[6] Hauser, Susan, Jonathan Schaifer, Tehseen Sabir, Dina Demner-Fushman and George Thoma. "Correcting OCR Text by Association with Historical Datasets." Document Recognition and Retrieval X, Proceedings of SPIE Volume 5010, pp 84-93 (2003).

[7] Kim, Jongwoo, Daniel X. Le and George R Thoma. "Automated Labeling of Bibliographic Data Extracted From Biomedical Online Journals." Document Recognition and Retrieval X, Proceedings of SPIE Volume 5010, pp 47-56 (2003).

[8] Lasko, Thomas A. and Susan E. Hauser. "Approximate string matching algorithms for limited-vocabulary OCR output correction." Document Recognition and Retrieval VIII, Proceedings of SPIE Volume 4307, pp 232-240 (2001).

[9] Mittendorf, Elke, Peter Schäuble and Páraic Sheridan. "Applying Probabilistic Term Weighting to OCR Text in the Case of a Large Alphabetic Library Catalogue." Proceedings of the 18th annual international ACM SIGIR conference on retrieval, pp 328-335 (1995).

[10] Taghva, Kazem, Julie Borsack and Allen Condit. "An expert system for automatically correcting OCR output." Proceedings of the SPIE - Document Recognition, pp 270-278 (1994).

Downloaded from SPIE Digital Library on 06 Jun 2011 to 146.229.56.102. Terms of Use: http://spiedl.org/terms