Pattern matching techniques for correcting low confidence OCR words in a known context

Glenn Ford, Susan Hauser[*], Daniel X. Le, George R. Thoma

National Library of Medicine, Bethesda, Maryland 20894

ABSTRACT

A commercial OCR system is a key component of a system developed at the National Library of Medicine for the automated extraction of bibliographic fields from biomedical journals. This 5-engine OCR system, while exhibiting high performance overall, does not reliably convert very small characters, especially those that are in italics. As a result, the “affiliations” field that typically contains such characters in most journals, is not captured accurately, and requires a disproportionately high manual input. To correct this problem, dictionaries have been created from words occurring in this field (e.g., university, department, street addresses, names of cities, etc.) from 230,000 articles already processed. The OCR output corresponding to the affiliation field is then matched against these dictionary entries by approximate string-matching techniques, and the ranked matches are presented to operators for verification. This paper outlines the techniques employed and the results of a comparative evaluation.

Keywords: Automated Data Extraction, Scanning, OCR, NLM, MARS, Pattern Matching

1. INTRODUCTION

The Lister Hill National Center for Biomedical Communications, a research and development division of the National Library of Medicine (NLM), has developed a distributed system to automate the entry of bibliographic citation data into the National Library of Medicine's MEDLINE® database. The system, called MARS for Medical Article Record System, has evolved to include scanning and OCR conversion of the entire first page of each journal article that is to be indexed for MEDLINE. After OCR conversion, additional modules automatically group characters into zones, identify those zones pertinent to the citation (Title, Author, Affiliation and Abstract), check the spelling of words containing low-confidence characters, and reformat certain fields to conform to NLM conventions. The system also includes specially designed workstations for operators to manually enter certain data that does not always appear on the first page, and to perform a final check of all of the data.

The success of the automated processes following the OCR conversion depends to a great extent on the quality of the OCR output. Although the commercial OCR system used by MARS performs well in general, accuracy is often poor for small fonts or italic characters. In particular, authors' affiliations frequently appear as small and/or italic characters, resulting in many incorrect characters in the Affiliation field. Consequently, the final check and correction of the Affiliation field requires a disproportionate amount of human labor compared to other fields extracted by our automated system. This paper describes research toward automatically correcting many of the incorrect words in the Affiliation field prior to presenting the data to the operator, thus reducing labor costs and increasing overall system reliability.

2. BACKGROUND

In addition to the ASCII text, the data output by the commercial OCR system employed in MARS includes a confidence level for each character, from 1 to 9. The OCR system is conservative about assigning high confidence. Thus, any character with the highest confidence level of 9 is almost always correct. Characters with confidence lower than 9 are highlighted in red for the final check, drawing operator attention to words that may be incorrect. For about one in five affiliations, there are so many highlighted words that operators prefer to type the entire affiliation rather than examine and correct each word. This is

- Correspondence: Email: hauser@nlm.nih.gov time consuming and represents a potential source of error in the completed bibliographic record since at the verification stage, manually entered fields are not double-keyed.

Words that frequently appear in the Affiliation field in biomedical journals are drawn from a relatively small vocabulary that denote institutions and their divisions, such as University and Department, the various branches of medicine and biology, such as Pathology and Biophysics, and names of cities, states and countries. We hypothesize that if OCR has correctly converted some of the characters in these words, it may be possible to determine the correct word through partial string matches or other matching techniques. For example, if the word as output by the OCR system is "Univlersitv" with v, l, and v being low confidence characters, then "University" is very probably the correct word.

We investigated several approximate string-matching techniques to explore the possibility of reliably substituting a correct word in the Affiliation field for a word that had been incorrectly converted by the OCR system. The goal is to find one correct word and automatically make the substitution, or find a short list of candidate words from which the operator can select the correct word. For this system it is better to err on the side of caution than to introduce incorrect words into the text. If the operators learn that they cannot trust the accuracy of words in the affiliation field, they will be compelled to examine each word, even if it is not highlighted, thus offsetting our objective of saving operator time.

3. METHODS

Correct words and a count of their occurrences were extracted from the final, corrected Affiliation field of approximately 230,000 journal articles that had already been processed by the MARS system. There were 96,982 unique words of 2 or more characters that occurred one or more times in this historical data. This list of words is the basis for various dictionaries that are searched for words to substitute for the word containing low confidence characters. Initially, two dictionaries were created for further evaluation. The "full dictionary" contained the entire set of 96,982 words. The "small dictionary" contained 10,263 unique words that occur at least 10 times. The total number of words from the historical data was 3,423,465 with 86,719 words dropped because they were single characters (length = 1). The sum of word occurrences for the small dictionary of about ten thousand words was 3,264,915, or 95.37% of the total word occurrence in the historical data, supporting our premise that the bulk of the words in the affiliation field come from a relatively small vocabulary.

Next, three sets of Ground Truth data were generated to evaluate string-matching techniques. From a set of 5,551 journal articles processed by the MARS system, we extracted both the original OCR data for the affiliation field and the corresponding final, corrected data. From these, two sets of Ground Truth data were generated for testing. Ground Truth set number 1 was created by human operators viewing both the OCR output and the final text, and assigning a correct word from the final text to each of the low confidence words in the OCR text. This set contained 15,609 pairs of words. Human judgment was able to pair some of the extremely poor OCR text with an appropriate word from the final text. Ground Truth set number 2 was created by a computer program using an edit distance algorithm to pair low confidence words with a correct word in the final text. This set contained 21,800 words. Although larger, this second set does not contain word pairs for some of the very poor OCR words because the computer program was unable to match words that had a large percent of the characters substituted.

A third Ground Truth set was created from low confidence words in the affiliation field from 30 newly processed journals. These 30 journals were selected because they had a particularly high number of low confidence words in the affiliation field, the overall average being 24% low confidence words. Human operators viewing the OCR output and the final text generated Ground Truth set number 3. This set contained 6877 word pairs.

All three Ground Truth sets were used to evaluate five approximate string-matching techniques: Partial Match, Near Neighbor[1,2,3,4], Bi-gram, Soundex and Probability Matching.

Partial Matching can find words in the reference dictionary where the OCR word has one or more character errors, but whose length is correct. Partial Matching uses a ternary tree[4] to store the dictionary words. Ternary trees provide extremely fast access and allow searching on wild cards. In this technique, the "wild card" character '.' is substituted for one or more of the low confidence characters in the OCR word. For example, if we have an OCR word, Deparlmemt, with confidence values, 9699878956, a match would be found for Depar.me.t or D.par.me.t, but not for D.parlme.t or D.par.memt. This is because ternary search trees combine the time efficiency of digital tries with the space efficiency of binary search trees.

Near-Neighbor matching finds all of the words in the dictionary that are within a given Hamming distance of the OCR word. The Near-Neighbor matching algorithm searches the ternary tree developed by the partial match process. The Hamming distance is a measure of the difference between two messages, each consisting of a finite string of characters, expressed by the number of characters that need to be changed to obtain one from the other. For example, “Deparlmemt” and “Department” has a Hamming distance of two, whereas "Butter" and "ladder" are four characters apart

The Bi-gram techniques were adapted from in-house software to suggest spelling alternatives for online Library clients[5]. For Bi-gram searches, each word in the reference dictionary is searched for all possible bi-grams in the OCR word. The Bi-gram algorithm takes all dictionary words and breaks them into two letter word chunks. For example, Department contains 9 bi-grams: De, ep, pa, ar, rt, tm, me, en, nt. Dictionary words containing multiple bi-grams in the OCR word are possible matches. The matching algorithm takes as input a word from the OCR output and the possible corresponding words in the dictionary. It then proceeds to match 2 letter chunks of the OCR word with the corresponding dictionary words. The bi-gram prunes out unlikely candidate words through an algorithm that inputs a dictionary word length, the OCR word length and the number of word chunks that match.

Soundex matching finds words in the reference dictionaries that are phonetically similar to the OCR word. This technique proved to have a number of difficulties in overcoming character substitutions caused by the OCR system. For example, the word Department would often be interpreted as Deparlment by the OCR engine. These two words are phonetically quite different since the letter ‘t’ and the letter ‘l’ in the English language do not have similar sounds. This is a simple example of many problems using phonetic matching of a word dictionary with OCR output. However, future research may be envisioned toward the design of an OCR phonetic word-matching algorithm. A study of the most common OCR character substitutions could be incorporated into the OCR engine to produce several permutations of the OCR word. These words could then all use the soundex algorithm to produce multiple words to rank as possible word matches.

The fifth technique, Probability Matching, developed in-house[6], compares the OCR word to each word in the dictionary using an edit distance based on OCR character substitution frequencies and the frequency of occurrence of the dictionary word, calculating the probability that the OCR word was produced by each dictionary word. The most probable words are then ranked using an extension of the Probability Match method. Words are returned with a confidence and ranking value, where the confidence is equivalent to the calculated probability.

All of these search methods can produce multiple matches. When Partial Match, Near Neighbor, Bi-gram or Soundex matching produce multiple matches, the Probability Matching technique is used to assign confidence and ranking values to the candidate words.

4. EVALUATION AND RESULTS

The first four approximate string matching techniques, with various parameter values, were tested with the ground truth word pairs from sets 1 and 2 to find those that would result in a high percentage of successful matches and a very low number of false matches.

Near-Neighbor matching had a false match rate of 2.7% and performed poorly for words in which the OCR word contained a different number of characters than in the correct word. Soundex matching had a false match rate of 5.6%. Its poor performance is not unexpected since OCR errors are not related to phonetics. These two methods were not investigated further.

Good overall results (obtaining a high percent of words matched, a very low percent of false positives, and high computational efficiency) for Ground Truth sets 1 and 2 were achieved by combining full word matching using Partial Match and two of the approximate string matching techniques in a three step process. The first step is to search for the complete OCR word in the full dictionary. If the word is not found, the second step is partial matching with wild card substitutions, using the small dictionary. If no matches are found here, the third step is the bi-gram search with the small dictionary. The results of this method are:

Correct word in match list	93.2%
Correct word ranked number one	89.4%
False positive (match list with no correct word)	2.8%

False negative (no match, correct word in dictionaries)	0.0%
True negative (no match, correct word not in dictionaries)	4.1%

Applying this three-step process to Ground Truth set 3 yielded the following results:

Correct word in match list	78%
Correct word ranked number one	77%
False positive (match list with no correct word)	11%
No matches	11%

Because data in set 3 were from particularly difficult journals, the lower percentage of matches was not unexpected, and returning the correct word 78% of the time is valuable.

Our efforts now turned toward reducing the number of false positives. We observed that many of the false positives were words with a very low frequency in the dictionary, for example, Universiy, Shool and tde, each a clearly wrong word that occurred only once. We surmise that these either slipped by the final correction process or are fragments of email addresses. Rather than undertake the daunting task to manually edit the full dictionary, four smaller dictionaries were created based solely on word frequency. These four dictionaries contained all words with a frequency of 2 or more, 10 or more, 50 or more and 100 or more.

Of the 6877 words pairs in Ground Truth set 3, the OCR word was correct for 4403 words, or 64% of the time, leaving 2474 words needing correction. The remaining evaluation concentrated on those 2474 words. We obtained results for three matching methods and the four dictionaries. Method A is the three step process described earlier, followed by the Probability Matching function applied to any words left unmatched. Method B is an initial search for the complete OCR word (the first step of the three step process), followed by the Probability Matching function for words not found. Method C is the Probability Matching function alone. Figure 1 shows the results of the 12 scenarios using Ground Truth set 3. The horizontal axis labels identify the Method and Dictionary. For example, B50 identifies the results of the test using method B for matching with the dictionary containing words with a frequency of 50 or more.

For each of the 2474 OCR Incorrect words, either a correct match was included in the list of words returned from the matching method, no correct match was included in the list of words, or no words were returned. The number of words in each category are indicated by the vertical bars in Figure 1.

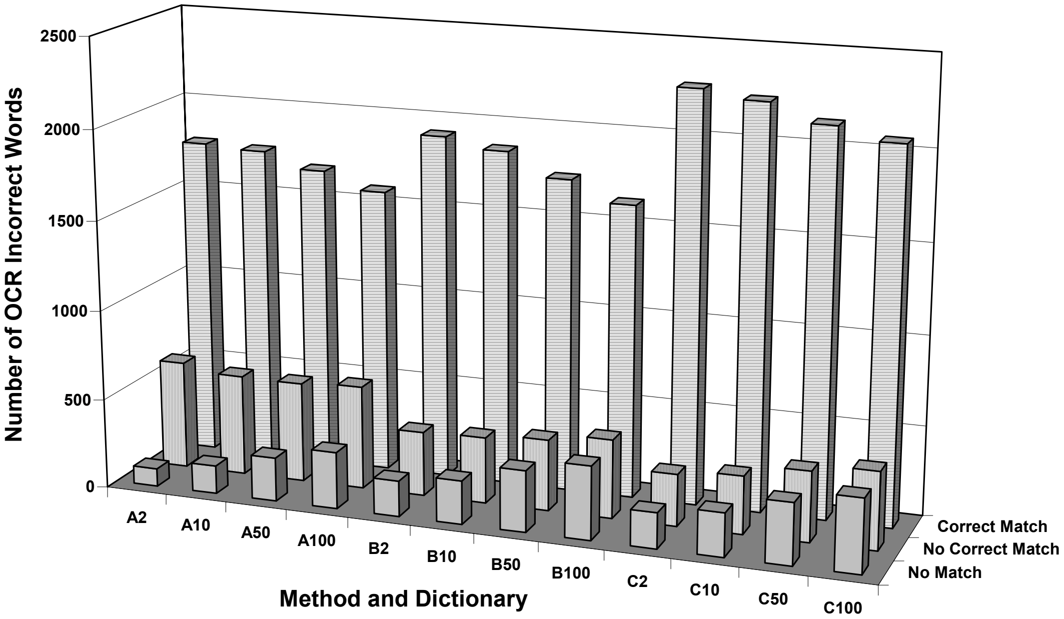


Figure 1. Results of Matching with Three Methods and Four Dictionaries

As shown in Figure 1, for all three matching methods, smaller dictionaries yielded a larger number of incorrect matches (false positives) rather than a smaller number. For any dictionary tested, method C produced the largest number of correct matches, and the smallest number of incorrect matches. To compare with other results, the results for test C2 are expressed in table format for the 2474 OCR incorrect words:

Correct word in match list	80%
Correct word ranked number one	72%
False positive (match list with no correct word)	12%
No matches	8%

And for the entire set of 6877 low confidence words:

Correct word in match list	82%
Correct word ranked number one	78%
False positive (match list with no correct word)	13%
No matches	5%

Although the percent of correct words returned is higher, so is the percent of false positives, which was not reduced by using a smaller dictionary. The next approach to reducing false positives was to explore the confidence value returned by the Probability-Matching algorithm using the dictionary of words with a frequency of 2 or more. The objective was to determine a confidence threshold that would eliminate a larger number of false positives than correct matches. For the 2474 OCR incorrect words, Figure 2, shows the cumulative fraction a) of cases with a correct word returned in the match list as a function of the confidence of the correct word, and b) of cases with no correct word returned in the match list as a function of the confidence of the first word in the list.

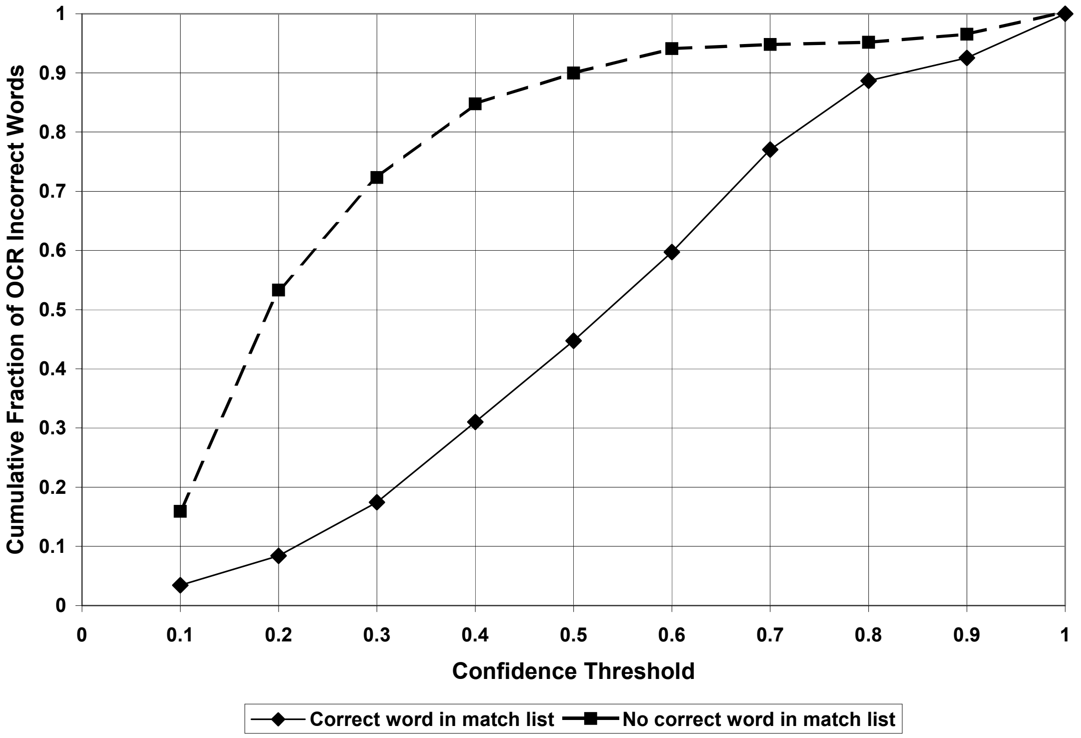


Figure 2. Probability-Matching Confidence of Correct Words and False Positives: Cumulative Fraction.

The chart shows that by eliminating words with a confidence less than 0.3, for example, we could remove over 70% of the false positives while removing less than 20% of the correct matches. However, there are many more correct matches to begin with. Figure 3 is similar to Figure 2, but shows the cumulative count for each case, rather than the percentage.

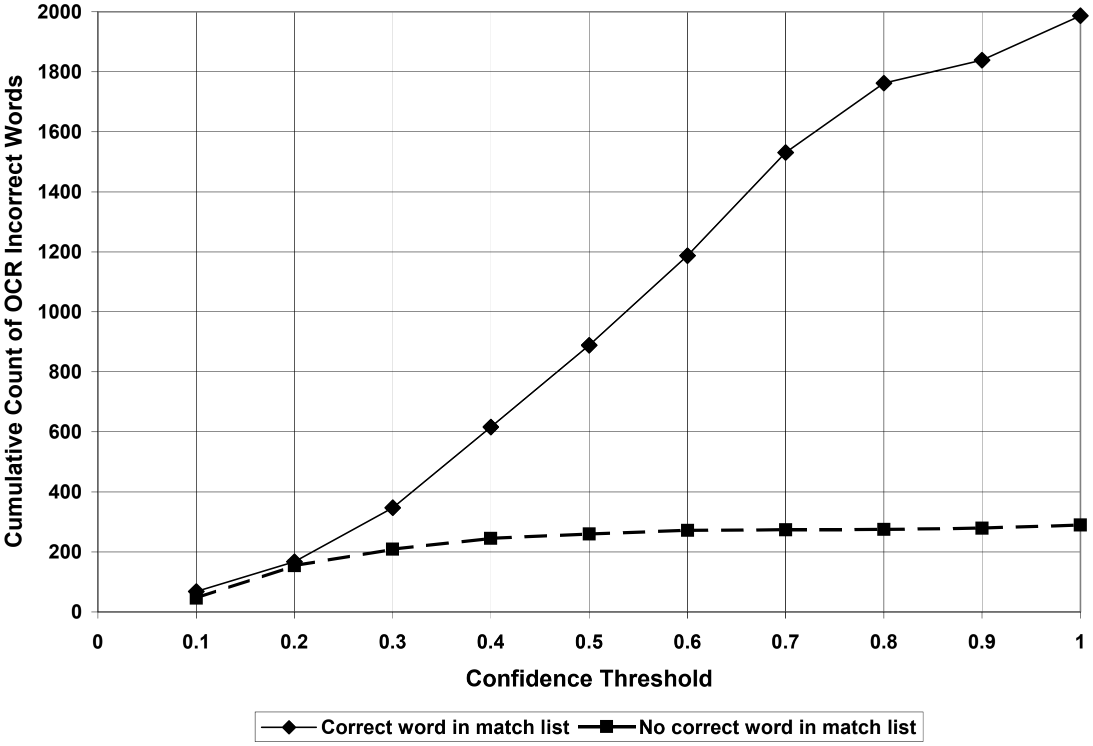


Figure 3. Probability-Matching Confidence of Correct Words and False Positives: Cumulative Count.

This chart shows that for any confidence threshold, there are more correct words returned from the matching algorithm than incorrect words. So the cost of eliminating false positives with low confidence values is eliminating an even larger number of correct matches.

5. DISCUSSION

Some of the performance differences among methods A, B and C can be attributed to the whole word matching process that is part of A and B but not C. In whole word matching, a single OCR error or omission can result in a word that is found in the dictionary and thus returned as a single match. For example, several non-English variations of the word University are included in the dictionaries, including Universit, Universita, Universitaire, Universitat, and Universite. If the OCR word is "Universit", simple matching will return "Universit" as the correct word even though the actual word was "University" or one of its other variations. For the same example, Probability Matching returns several similar words, including the correct one.

Another difference between method A and C is the Bi-gram matching process that is part of A and not C. Although Bi-gram matching contributes 93% of the correct matches found by method A it also contributes 73% of the incorrect matches. Bi-gram matching appears to be sensitive to OCR words with one or two incorrect letters. If the incorrect letter results in Bi-grams that occur in other words, those words can appear in the match list rather than the correct word. Because Probability Matching uses both OCR-error based edit distances and word frequencies to select candidate words from the dictionary, it is less sensitive to a few OCR errors or omissions. The table below illustrates the difference with four examples of words for which no matches were returned by the whole word search or partial matching.

OCR word	**Correct word**	**Matches returned by Bi-gram search**	**Matches returned by Probability Matching**
Vieteritt1rv	Veterinary	Kettering	Veterinary
Soulternii	Southern	Coulter	Southern
Departiltenwut	Department	Departement	Department
		Departimento	
Depaltrtmncilt	Department	(none)	Department

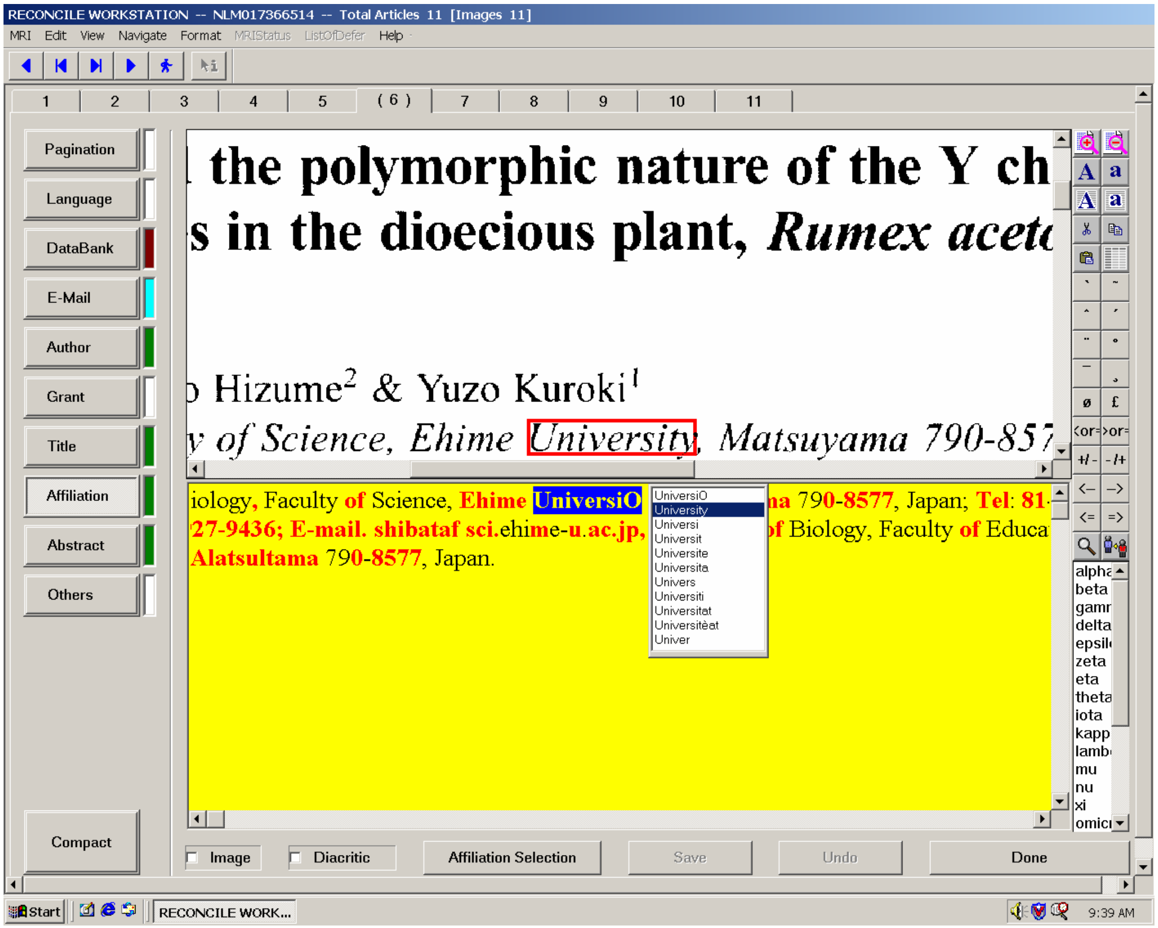
The impressive results of Probability Matching are offset by its processing time. Because it compares the word to be matched with every word in the dictionary, it can take up to 1 second per word depending on word length, dictionary size and computer speed. Because bigger dictionaries have been shown to yield better results, we would use a large dictionary, thus incurring long processing times. Further research is indicated toward an acceptable multi-stage process in which "easy" words are matched reliably by a faster process, and the remaining words are matched using Probability Matching.

6. IMPLEMENTATION

Software has been developed to implement any of the matching algorithms within the existing design of the MARS database and workflow[7]. The program invokes word matching for each OCR word in the affiliation field that contains low confidence characters. If one or more words are found, they are inserted into the affiliation record following the original word, which is left unchanged.

The application that presents the affiliation text to the operator for checking will highlight the first word in the list and show the other words in a drop down list from which the operator can select a different word if necessary. The output OCR word, as seen in the lower half of Figure 4, is *UniversiO*. The upper half of the window, separated by a movable splitter displays the scanned image and has a red box around the word containing low confidence characters to be corrected. In the example presented in Figure 4 the word matching process found 10 words that could possibly match UniversiO. The first word in the list (*UniversiO)* is the original OCR word, the default highlighted word (*University*) is the highest ranked word match, the third word in the list (*Universi*) is the second highest ranked word match, and so on. The user has the option to hit escape and leave the original word highlighted, hit tab and substitute the original OCR word with the highest ranked word (as in this case), or highlight any of the words in the list using their mouse or keyboard. If using a mouse the user can double click the correct word. If using a keyboard the user can use the arrow keys and hit enter or tab to select the highlighted word. Thus the ease of selection is relative to the need for correction: selecting the original OCR word or the first candidate word in the match list is accomplished with a single keystroke. These cases account for approximately 90% of the words containing low confidence characters. Once a user has taken action the application will move on to the next low confidence encounter. In many cases this is a single low confidence character in the word. Only if a word match exists will a word be highlighted.

We anticipate that the use of approximate string matching techniques to find candidate words plus a user interface that accommodates easy viewing and selection of those candidate words will reduce operator labor and increase system reliability.



l the polymorphic nature of the Y ch
s in the dioecious plant, *Rumex acet*

Hizume2 & Yuzo Kuroki1
y of Science, Ehime University, Matsuyama 790-857

iology, Faculty of Science, Ehime UniversiO na 790-8577, Japan; Tel: 81-
27-9436; E-mail. shibataf sci.ehime-u.ac.jp, of Biology, Faculty of Educa
Alatsultama 790-8577, Japan.

Figure 4. Graphical User Interface for correction of low confidence OCR words.

REFERENCES

1. JL Bentley , R. Sedgewick, "Fast algorithms for sorting and searching strings", *Proc. 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, Jan 1997.
2. S. Baase, *Computer Algorithms*, Addison-Wesley, 1988.
3. P. Hall, G.R. Dowling, "Approximate String Matching", *ACM Computing Surveys*, **12**, pp 381- 401, 1980.
4. J. Bentley, B Sedgewick, *Ternary Search Trees*, Dr. Dobb's Journal, pp 20-25, 1998.
5. G. Divita, "A Spelling Suggestion Technique for Terminology Servers". to appear in *American Medical Informatics Association Fall Symposium,* 2000.
6. T. Lasko, et. al. "Approximate string matching algorithms for limited-vocabulary OCR output correction". *Proceedings of SPIE*, Vol. 4307, Document Recognition and Retrieval VIII, 2000.
7. G.R. Thoma, "Automating data entry into MEDLINE", *Proceedings of the 1999 Symposium on Document Image Understanding Technology*, pp. 217-18, 1999.