Text verification in an automated system for the extraction of bibliographic data

George R. Thoma, Glenn Ford, Daniel Le, Zhirong Li

National Library of Medicine
Bethesda, Maryland

Abstract. An essential stage in any text extraction system is the manual verification of the printed material converted by OCR. This proves to be the most labor-intensive step in the process. In a system built and deployed at the National Library of Medicine to automatically extract bibliographic data from scanned biomedical journals, alternative means were considered to validate the text. This paper describes two approaches and gives preliminary performance data.

1 Introduction

The manual verification of the text produced by any OCR system, or one that is a sequence of document image analysis and understanding (DIAU) processes, is an essential stage to ensure the accuracy of the extracted text. At the R&D center of the National Library of Medicine, we have developed a DIAU-based system[1] that automatically extracts significant bibliographic data from scanned biomedical journals to populate MEDLINE®, the library's flagship database used worldwide by biomedical scientists and clinicians. The final step prior to accepting the bibliographic record thus created is a manual check for accuracy by human operators. Following a brief description of the overall data extraction system in Section 2, we discuss alternative approaches for the design of the workstation used by the text verification operators.

2 Overall system: brief description

The DIAU-based system, code-named MARS for *Medical Article Records System*, consists of both automated and operator-controlled subsystems as shown in Figure 1. The schematic shows automated processes as boxes with thin boundaries, and manual workstations with thick boundaries. The workflow is initiated at the CheckIn stage where a supervisor scans the barcode on a journal issue arriving at the production facility. This barcode number, called the "MRI", is routinely affixed to every journal issue by NLM staff. It therefore serves as a unique key to identify the issue, all the pages scanned in that issue, and indeed the outputs of all processes performed on

those page images. The scanning operator captures the first page of every article in the issue, since this page contains the fields we seek to extract automatically. The resulting TIFF images go into a file server and associated data into the MARS database for which the underlying DBMS is Microsoft's SQL Server. The OCR system accesses the TIFF images and produces the corresponding text as well as other data descriptive of the text characters such as bounding boxes, attributes (bold, italic, underlined), confidence level, font style and size, and others. The automatic zoning[2] (Autozone) module then blocks out the contiguous text using features derived from the OCR output data, followed by the automated labeling[3] (Autolabel) module that identifies the zones as the fields of interest (article title, author names, affiliations, abstract). The Autoreformat[4] module then reorganizes the syntax of the zone contents to adhere to MEDLINE conventions (e.g., author name *John A. Smith* becomes *Smith JA* for pre-2002 journal issues, and *Smith John A* for 2002 and later).

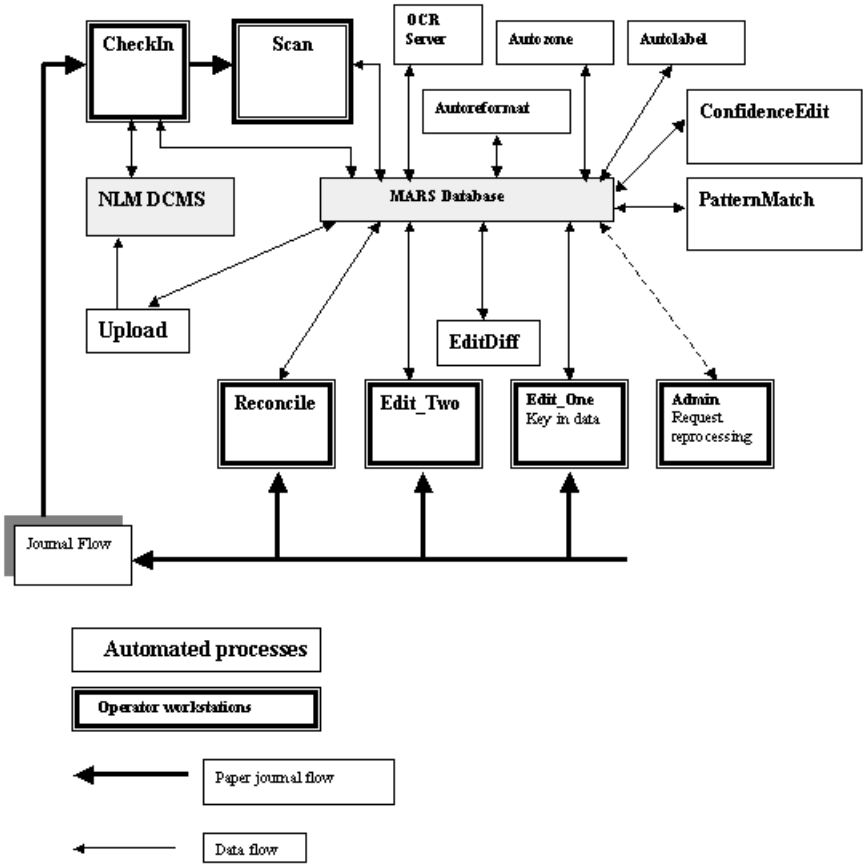
At this point, two lexicon-enabled modules operate on the data to reduce the burden on the operator performing the final checking and verification of the data: ConfidenceEdit that modifies the incorrect confidence levels assigned to the characters by the OCR system, and PatternMatch[5,6] that corrects institutional affiliations whose text, usually italicized and in small font size, is frequently recognized incorrectly by the OCR system. PatternMatch relies on a combination of two matching techniques: whole word matching (which compares the entire OCR output word with each word in a lexicon of words found in the affiliation fields from 230,000 journal articles) and 'probability matching' (that scores matched words according to a calculated confidence value and frequency of occurrence.)

Some data essential to a complete bibliographic record in MEDLINE cannot be automatically extracted, such as NIH Grant Numbers and Databank Accession Numbers. The major reason is that they appear in pages other than the scanned first page of the article. Such data is manually entered by a pair of "Edit" operators, a double-key process that ensures a high degree of accuracy. An EditDiff module then correlates these different entries and notes differences.

The output of the automated processes and the EditDiff module are then presented to the "Reconcile" operator who verifies the text and corrects errors. Following this text verification stage, the data is transmitted by the Upload module to the NLM's DCMS (Data Creation Maintenance System) which is later accessed by NLM indexers to add medical subject headings (MeSH terms) and keywords, thereby completing the bibliographic record for the MEDLINE database. The Admin workstation shown is used by the production supervisor to send a journal issue back to an earlier processing stage in case of errors.

The focus of this paper is the design of the Reconcile workstation.

Figure 1 MARS general schematic.



3 Reconcile workstation

The purpose of the text verification or Reconcile workstation is to enable an operator to check the accuracy of the bibliographic data extracted by the automated processes, as well as that entered manually by the Edit operator. Any errors are corrected at this stage before the citation is uploaded to the DCMS database. Desirable features of a text verification workstation include: (a) all text must be presented for verification, whether automatically or manually generated in upstream processes; (b) the original bitmapped document image must be displayed for reference; (c) characters detected with low confidence by the OCR must be highlighted so the operator may focus on them; (d) navigation between fields and

from one character to another must be rapid; (e) optionally, aids such as lexicon-based pattern matching systems may be provided to correct errors.

Two approaches to the design of this workstation were considered: the ‘conventional’ and the ‘Isomorphic.’ In this section we describe the conventional approach implemented in the MARS system, and in Section 4 the Isomorphic approach. The conventional approach to verifying the text output of any OCR system, or as in the case of MARS the output of a succession of automated processes, is to present the text in the same sequence as it appears on the printed page, and to highlight the low confidence characters (in color) in the text words. Then, as in our Reconcile workstation, the operator can "tab" quickly from one suspected character to the next and make the necessary corrections. This conventional approach has some drawbacks. For example, the operator must detect the suspected character surrounded by a mass of correct text. Also, the text must be corrected as encountered, thereby breaking the rhythm of identifying incorrect characters. However, this is the usual approach adopted in most, if not all, current text conversion systems and services.

The Reconcile operator is provided two GUI views: the general view and the split view. The general view in the Reconcile workstation’s GUI gives an overall picture of the bibliographic data captured (Figure 2). Prior to the operator verifying the contents of the bibliographic fields, the field windows are highlighted by background color: green for fields created by the automated processes, cyan for those entered manually, yellow for those created by combining the outputs of both automated and manual processes, and red if no text appears in the window. In the example shown, the windows for NIH Grant Numbers and Databank Accession Numbers appear in red, since these data were not entered by the Edit operator earlier in the workflow. (These fields are not captured automatically since they could appear anywhere in the article and not just on the scanned page.) Once the operator verifies the text, the field windows turn white, as shown for Pagination and Language in this example.

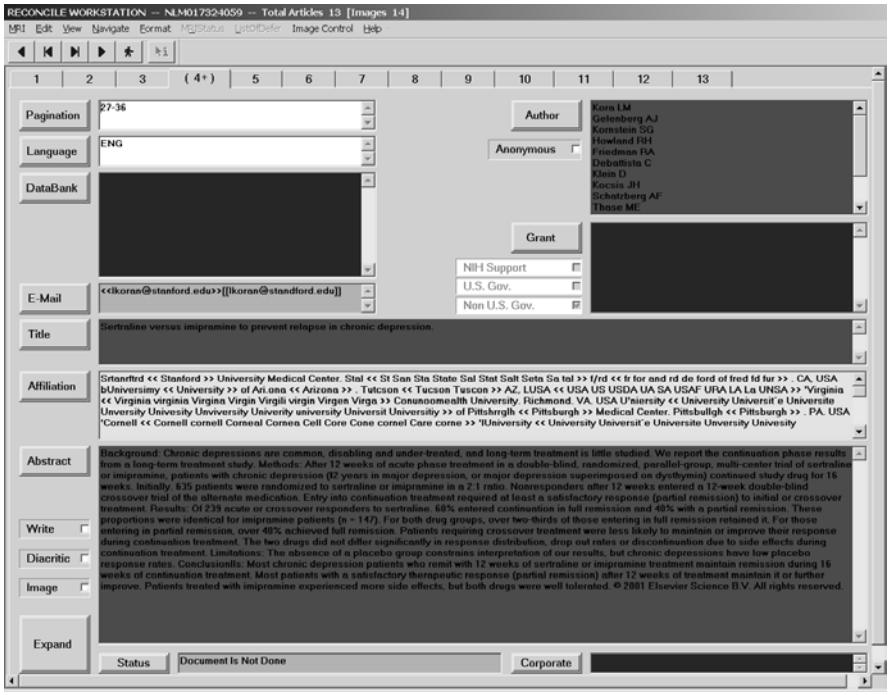


Figure 2 General view for all bibliographic fields in an article.

The split view (Figure 3) in the Reconcile workstation displays both the bitmapped image and the corresponding text to allow the operator to verify the text against the contents of the journal article. Low confidence characters are highlighted in red on the text to attract the operator's attention.

The Reconcile workstation provides the operator several additional functions: the operator may rename a field labeled incorrectly; activate a standalone OCR system to extract the field contents through the image or, alternatively, type in the text; if a page image is missing or duplicated, the operator may insert the missing page, or delete a duplicate; and if there are 'invalid' characters, the Reconcile software will convert them to the form required in MEDLINE.

Many of the functions in the Reconcile workstation are provided by a program called Character Verification[7], a module that allows the Reconcile operator to view the bitmapped document images and to verify the text in all the fields, both entered by the Edit operator, and that from the automated processes. It is based on two ActiveX controls, the Eastman Kodak Image ActiveX Control and Microsoft's Rich Text Editor ActiveX control, both embedded in the Reconcile software.

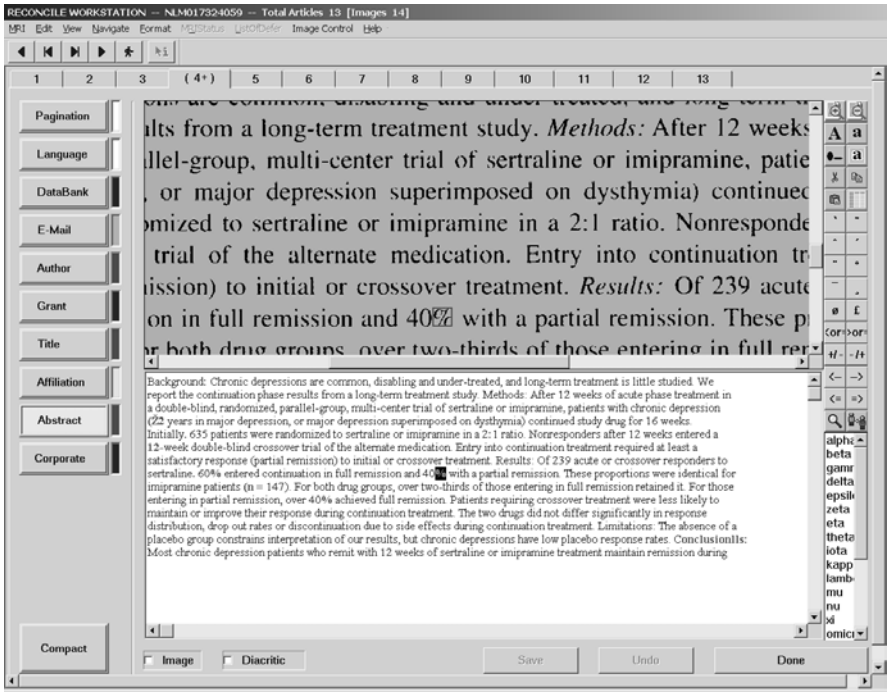


Figure 3 Split view shows both the bitmapped image and corresponding text. Low confidence characters are highlighted in the text window.

The Eastman Kodak Image control provides the functionality for the operator to view the image and perform manipulations such as rotation, and zooming in or out. It also provides a bounding box showing the area on the image that corresponds to the text that the operator is focusing on.

Character Verification also allows the operator to edit the field contents. The design of this editor is based on the Microsoft Rich Text Editor, and is derived from its heavily used functions such as copy, cut, paste, search and replace. The software can also relate the position of the text characters to the corresponding ones in the image, provide confidence levels, and allow the operator to enter diacritical marks, Greek letters, mathematical signs, change the first character of a selected word to upper case while leaving the rest in lower case, convert case, and complete words in the affiliation field from a partial output.

Character Verification and Reconcile's main program communicate via methods and events. Reconcile sets or gets the methods to instruct Character Verification. In turn, Character Verification fires events back to Reconcile to provide the information. It retrieves the OCR output for every article (bibliographic) field, including text contents (character codes), confidence levels and character coordinates, from the MARS database, and the images from a file server.

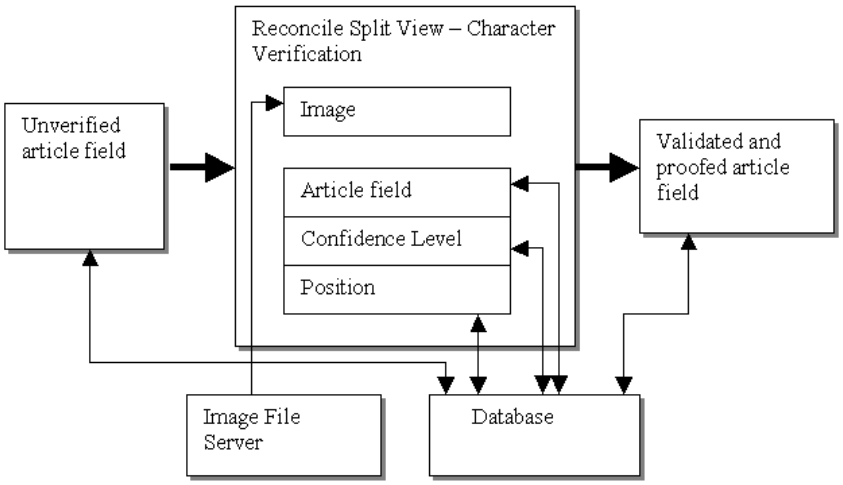


Figure 4 Character Verification displays label text, confidence and character position.

Character Verification also provides the results of the PatternMatch[5,6] program to the operator in a word list to choose from. An example is given in Figure 5 showing optional words in a dropdown box for the operator to select in an affiliation field.

4 Isomorphic method for text verification

To overcome the drawbacks in the conventional approach to text verification (noted in Section 3), we investigated an alternate method involving the simultaneous display of like characters. The hypothesis is that an operator would pick the odd one out quicker than reading through large numbers of words on the screen, would be required to do fewer keystrokes, and would experience lower eyestrain. These factors, we believe, would help achieve higher productivity. Proprietary systems, either commercial or research products, with similar functionality have been reported[8-10], but did not appear to allow for easy integration into our production system.

Our Isomorphic ("having the same form or appearance") method involves grouping like characters (drawn from a number of pages or journal issues at the same time) and displaying them in groups in a single window, as shown in Figure 6. Each character appears in its "edit box." Only low confidence **A - Z** and **0 – 9** characters, of the same type, would be displayed in groups. The example shows a set of characters in the edit boxes, mostly **e'**s, some of them a misreading of an **s** or an **E** as shown in the corresponding bitmapped images right above the edit boxes. Since context is important to detect poorly captured character shapes, the system must provide the display of the image fragment (a word or phrase) that provides the context in which

the (presumably) incorrect character appears. Such context will particularly help distinguish letters or numbers that appear similar, e.g., 1, I or 0,O.

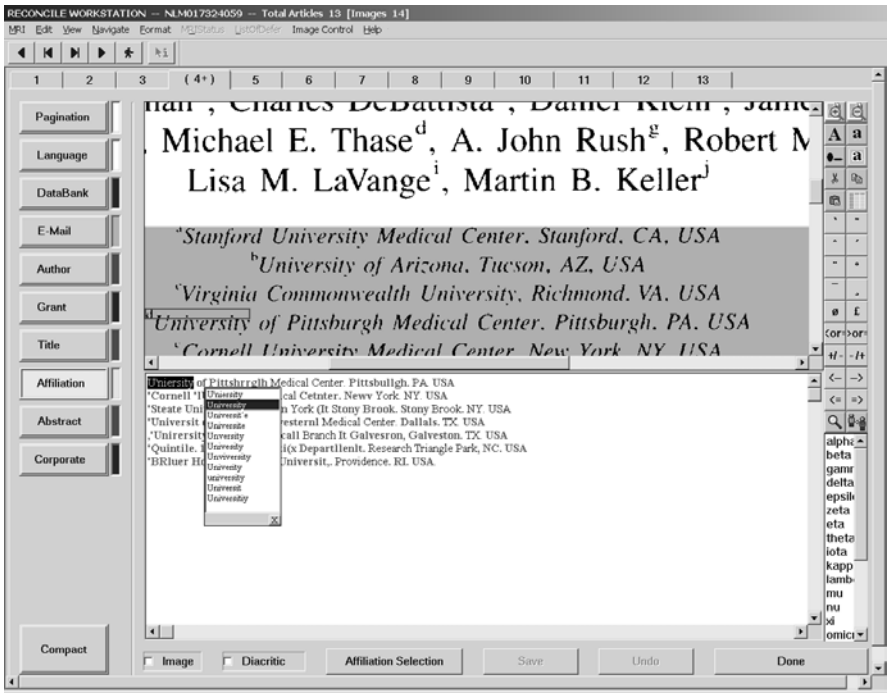


Figure 5 Operator can select alternative words in an affiliation field.

Our Isomorphic system is designed with the following functions:

An Automatic Context display of the TIFF image is available when any of the edit boxes is focused on.

By clicking on an edit box, or using the F1 thru F9 keys, the operator can correct the character. In Figure 6, the bitmap of each low confidence character appears above the edit text boxes. In case of ambiguity, the operator may type in a **'?'** that invokes the image fragment to provide the necessary context.

To continue, the operator can select **Next>>** and load the next batch of 9 characters, or select **<<Previous** for a second look at the previous 9. By selecting Next>>, all characters are set to high confidence.

All characters are displayed at low confidence, in case of an accidental clicking of the Next>> button.

Clicking OK will stop processing.

The Isomorphic system is implemented using Visual C++ with the Eastman Kodak Image libraries. Following the software development, we conducted a performance study using this system for reconciling, and measured the residual error rate and the time taken for correcting and verifying all the characters from a complete journal

issue at a time. Should further tests conclusively prove that the accuracy and time saved are an improvement over the current verification method, this module will be incorporated in the Reconcile workstation software.

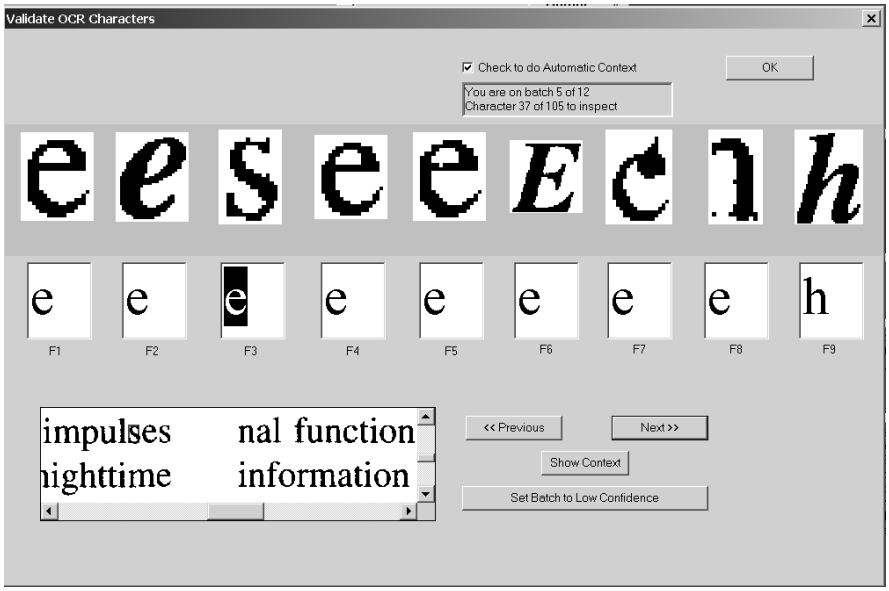


Figure 6 GUI for Isomorphic method.

5 Test results

Tests were conducted to determine the speed advantages of using the Isomorphic system over the conventional method. The Title and Abstract fields from 122 articles were selected for the tests. These fields were picked for the following reasons.

- These two fields account for nearly 99% of the OCR output data.
- The text in these fields is not reformatted.
- These fields have relatively good OCR data, unlike, say, the affiliation field that shows a high error rate on account of a preponderance of characters that are in italics and small sized fonts.

Two operators conducted the tests independently. The results presented in Figure 7 show that the Isomorphic approach is approximately three times as fast as the conventional method. For a normal daily work output of 600 bibliographic records, this corresponds to a saving of about 7 labor hours per day for the correction/verification of the Title and Abstract fields.

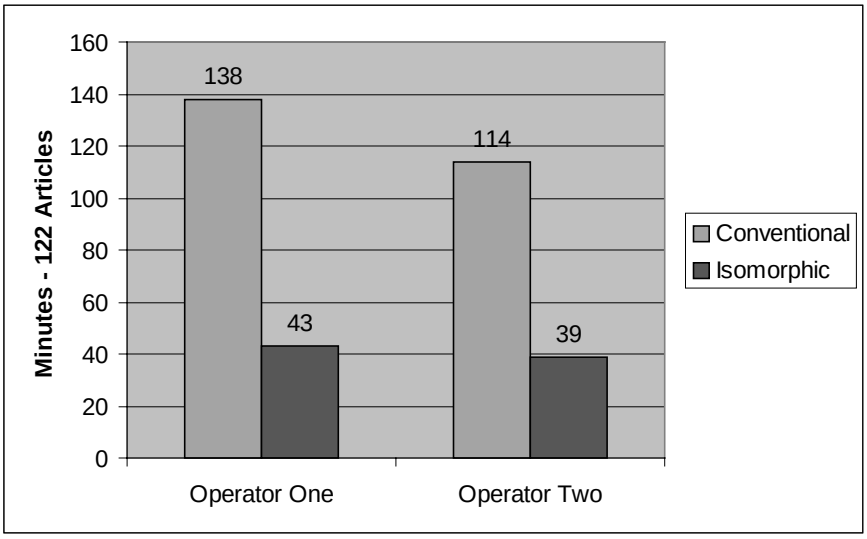


Figure 7 – Time taken to correct OCR errors using conventional and Isomorphic methods.

6 Summary

Text verification is a key stage in any OCR system or one that is a sequence of document image analysis and understanding processes. Alternative approaches, the conventional and the Isomorphic systems, have been described. Early performance test results show that text verification using the Isomorphic system, in which like characters are grouped and displayed simultaneously, may be done three times faster than with the conventional approach.

References

1. Automating the production of bibliographic records for MEDLINE. An R&D report of the Communications Engineering Branch, LHNCBC, NLM. Bethesda, Maryland. September 2001, 91pp. http://archive.nlm.nih.gov/~thoma/mars2001.pdf
2. Hauser SE, Le DX, Thoma GR. Automated zone correction in bitmapped document images. Proc. SPIE: Document Recognition and Retrieval VII, Vol. 3967, San Jose CA, January 2000, 248-58.
3. Kim J, Le DX, Thoma GR. Automated Labeling in Document Images. Proc. SPIE: Document Recognition and Retrieval VIII, Vol. 4307, San Jose CA, January 2001, 111-22.
4. Ford GM, Hauser SE, Thoma GR. Automatic reformatting of OCR text from biomedical journal articles. Proc.1999 Symposium on Document Image Understanding Technology, College Park, MD: University of Maryland Institute for Advances in Computer Studies; 321-25.

5. Ford G, Hauser SE, Le DX, Thoma GR. Pattern matching techniques for correcting low confidence OCR words in a known context. Proc. SPIE, Vol. 4307, Document Recognition and Retrieval VIII, January 2001, pp. 241-9.
6. Lasko TA, Hauser SE. Approximate string matching algorithms for limited-vocabulary OCR output correction. Proc. SPIE, Vol. 4307, Document Recognition and Retrieval VIII, January 2001, pp. 232-40.
7. Li Z. Character verification. Internal technical report, Communications Engineering Branch, August 23, 2001.
8. http://www.almaden.ibm.com/DARE/ui.html
9. http://www.clearlake.ibm.com/gov/ifp/sk_advantage.html
10. Moore A. The tricks to make OCR work better. Imaging Magazine. June 1994.