OCR Correction Using Historical Relationships from Verified Text in Biomedical Citations

Susan Hauser, Tehseen Sabir, George Thoma

Lister Hill National Center for Biomedical Communications
National Library of Medicine
National Institutes of Health
Department of Health and Human Services
Bethesda, Maryland 20894

Abstract

The Lister Hill National Center for Biomedical Communications has developed a system that incorporates OCR and automated recognition and reformatting algorithms to extract bibliographic citation data from scanned biomedical journal articles to populate the NLM's MEDLINE® database. The multi-engine OCR server incorporated in the system performs well in general, but fares less well with text printed in the small or italic fonts often used to print institutional affiliations. Because of poor OCR and other reasons, the resulting affiliation field frequently requires a disproportionate amount of time to manually correct and verify. In contrast, author names are usually printed in large, normal fonts that are correctly recognized by the OCR system. We describe techniques to exploit the more successful OCR conversion of author names to help find the correct affiliations from MEDLINE data.

1 Background

The Medical Article Records System, MARS, was developed by the Lister Hill National Center for Biomedical Communications to semi-automatically generate electronic bibliographic records from paper-based journal articles for the National Library of Medicine's MEDLINE® database[1] [2]. The system incorporates a commercial OCR server to convert the scanned page and several in-house developed modules to automatically construct the record from the OCR output text [3] [4]. Human operators verify and correct the automatically extracted fields, and type in those fields that are not automatically generated.

For several reasons, the affiliation field requires a disproportionate amount of time to correct and verify [5] [6]: 1) the affiliation field is often printed in small and/or italic fonts which are poorly converted by the OCR server; 2) the printed affiliation field frequently contains multiple affiliations, one for each author, while only the affiliation of the first author is included in the bibliographic record; 3) affiliations from the United States are to end with "USA", which frequently is not in the printed affiliation; 4) the OCR server does not recognize diacritical characters, which are common in non-USA affiliations, and because the standard keyboard does not include diacritical characters, the verification operator must insert individual diacritics by selecting from special "keys" on the monitor screen.

Figure 1 is typical of what the operator sees while verifying the affiliation field. The top of the screen displays part of the scanned image, while the bottom of the screen displays the OCR text for the field being corrected, with low-confidence characters highlighted in red. In this case, the operator must delete three affiliations, and correct several words in the first affiliation, the only one to be retained by MEDLINE's conventions.

Figure 1 also illustrates the usual case, where author names are printed in a large, regular font. The OCR server correctly converts characters printed in large, non-italic fonts, and numerals printed in any font. The one zip code that does appear in the affiliation is correctly converted, even though it is printed in italics.

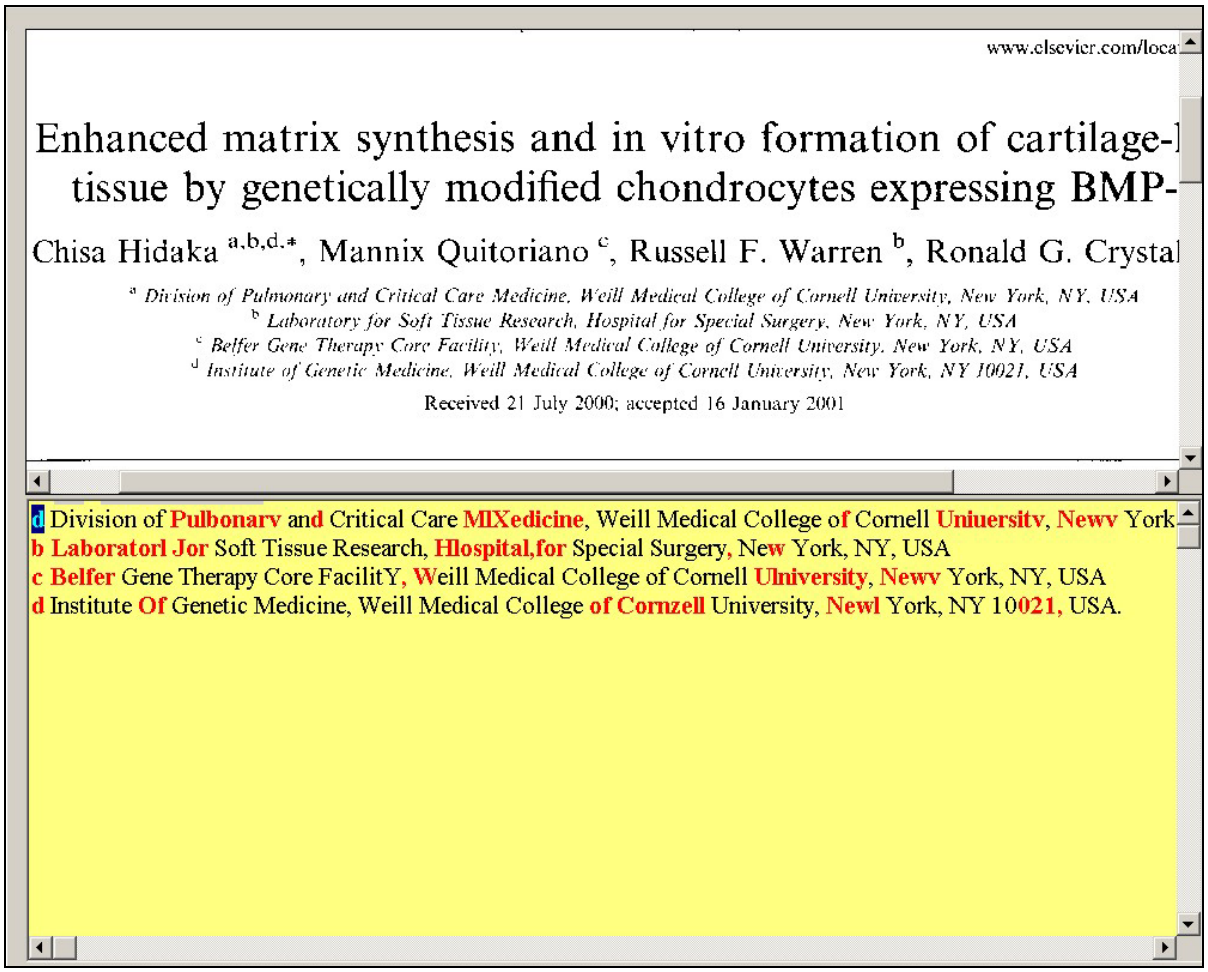

Figure 1. The verification operator's view of the OCR text from an affiliation field (bottom panel) and the corresponding image (top).

The MEDLINE database contains 12 million indexed citations for biomedical journal articles. Most of these citations include a list of one or more authors and the affiliation of the first author. Many authors publish repeatedly while at the same institution. Our objective is to use the historical author and affiliation relationships from this large dataset to find potentially correct, complete affiliations based on the author text and the affiliation text in the OCR output, and to present these affiliations to the verification operator in addition to the OCR text. The operator selects the affiliations as is, or edits them to create the correct affiliation field. Even if the affiliation presented is not structured exactly as the printed one, the operator may determine that it is easier, and more reliable, to edit the alternate affiliation than to correct the affiliation text from the OCR.

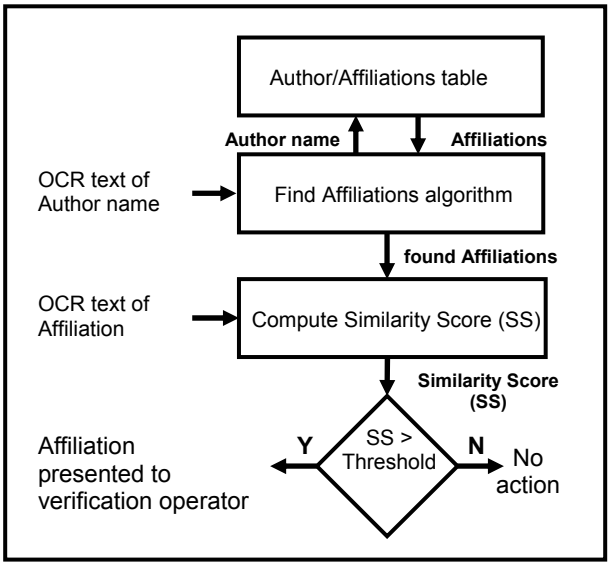

Figure 2. How the author/affiliation relationship is used.

Figure 2 shows how the module to find alternate affiliations is structured. The module uses the OCR text of the author name to query the author/affiliation table for potential affiliations. Each affiliation found in the table is compared to the OCR text of the affiliation to determine if it is the same institution as the OCR affiliation. If it is determined to be the same institution, it will be presented to the verification operator in addition to the OCR affiliation text.

2 Previous Work

A preliminary study was conducted with MARS output data, using a table of about 324,000 author/affiliation pairs extracted from MARS verified data, and a test set of about 20,000 first authors extracted from a separate set of MARS verified data [7]. 47% of the author names in the test set were found in the author/affiliation table. For 43% of these authors, at least one of the affiliations associated with the author name was the same institution as the institution in the OCR text. Thus we anticipate that a correct affiliation can be offered to the verification operator for at least 20% (= 47% x 43%) of the articles that are processed.

Based on these encouraging data, we constructed a more complete database of author/affiliation relationships from MEDLINE. This table contains over 700,000 unique author/affiliation pairs taken from MEDLINE entries from 2000 through 2002.

Of the 394,768 unique author names in the author/affiliation table, about 34% are paired with more than one affiliation. Thirty five author names are paired with over 100 affiliations. Thus, even when an author name is found in the author/affiliations table, it is still necessary to determine which, if any, of the associated affiliations is the same one as that from the printed article. This is done by comparing the affiliations found in the table to the OCR affiliation text and computing a similarity score. The scoring algorithm must take into account the possibility of errors in the OCR affiliation such as character substitutions, omissions or inclusions, the possibility of text in the OCR affiliation that is irrelevant to the final affiliation, such as "Dr. Smith is from…", and the possibility that the OCR text includes affiliations for authors other than the first author.

A partial-matching algorithm was developed that uses an edit distance threshold on a word by word basis, and then finds chains of such partially-matched words. The similarity score is calculated as the ratio of the total number of words in the two longest chains of partially-matched words to the number of words in the shorter of the two affiliations being compared. Tweaking the algorithm to exclude short words, a few stop words, all-digit sequences and email addresses yielded promising results when tested with a small set of ground truth data extracted from records processed by the MARS system [8]. "Good" results are when a threshold for the calculated similarity score reliably separates the affiliations from the table that are the same institution as the OCR affiliation from the affiliations from the table that are not.

The following two examples are cases where high similarity scores correctly indicate the same institutions, i.e. true positives.

Example 1:
OCR Text:

> The Wailter &i Eli-a Hall Ins.titfte of Medical Researchl, Post Office Box the Royal Melboulrlle Hospital 3050, Victoriat, Atustralia

Found Affiliation:

> The Walter and Eliza Hall Institute of Medical Research, Post Office Box the Royal Melbourne Hospital 3050, Victoria, Australia.

Verified Affiliation:

> The Walter & Eliza Hall Institute of Medical Research, Royal Melbourne Hospital, Victoria, Australia.

In example 1, the OCR text contains several errors plus some extraneous text. The found affiliation contains no incorrect text.

Example 2:
OCR Text:

> ' Laboratori de Quitnica Farmaceutica, Facultat de Farmacia, Universitat de Barcelona, Avda Diagonal sln, E-08028 Barcelona, Spain
> b Laboratoire de Chinlie Generale, Consertatoire National des Arts et Mttiers, 292, rue Saint-Martin, F-75141 Paris, France

Found Affiliation:

> Laboratori de Qu´imica Farmac`eutica, Facultat de Farm`acia, Universitat de Barcelona, Spain.

Verified Affiliation:

> Laboratori de Qu´imica Farmac`eutica, Facultat de Farm`acia, Universitat de Barcelona, Spain.

The OCR text in example 2 contains many errors, including missing diacritics, and an extra affiliation. The found affiliation is correct as is, including the diacritical marks.

There are also cases where high similarity scores result from comparing affiliations that are not the same. The following two examples are cases where high similarity scores are associated with two different institutions, i.e. false positives.

Example 3:
OCR Text:
*Departments of Pneumonology and t Oncology and Radiotherapy, Medical University of Gdansk, Poland

Found Affiliation:
Department of Oncology and Radiotherapy, Medical University of Gdansk, Poland.

Verified Affiliation:
Department of Pneumonology, Medical University of Gdansk, Poland.

In example 3, the first of the two departments listed in the OCR text is the correct affiliation, but the second department results in a higher score because those words are adjacent to the rest of the string.

Example 4:
OCR Text:
Bone Marroiw and Stem Cell Transplantation Center, Emory University, Atlanta. GA. USA

Found Affiliation:
Denver, CO 80262, USA.
boyer_e@hub.tch.harvard.edu

Verified Affiliation:
Bone Marrow and Stem Cell Transplantation Center, Emory University, Atlanta, GA, USA.

Example 4 illustrates the downside of excluding certain words from the partial-matching algorithm. After removing short words, email addresses and all-digit sequences from the found affiliation, the only word left is “Denver”, which is an edit distance of two from “Center”. Because the denominator of the final calculation is the shorter string, the similarity score is: 1 (word in the longest matching string of words) / 1 (word in the shorter affiliation) = 1.0, the highest possible score.

There are also cases where low similarity scores result from comparing affiliations that are the same institution. The following two examples are cases where low similarity scores are calculated for the same institutions, i.e. false negatives.

Example 5:
OCR Text:
Lehrsttuhlt flir Titcrzuicht und Allen,leiler LandlwirtschlaA

Found Affiliation:
Lehrstuhl fur Tierzucht und Algemeine Landwirtschaftslethe, Universit¨at M¨unchen, Germany.
Detlef.Pietrowski@gen.vet-med.uni-muenchen.de

Verified Affiliation:
Lehrstuhl f¨ur Tierzucht und Algemeine Landwirtschaftslethe, Universit¨at M¨unchen, Germany.

The OCR text of example 5 is so poor that only one of the longer words, “Lehrsttuhlt”, is within an edit distance of 3 (the edit distance threshold in effect for these tests) of one of the words in the found affiliation, “Lehrstuhl”. The similarity score is 1 (word in the longest matching string of words) / 5 (words in the shorter affiliation) = 0.2.

Example 6:
OCR Text:
Zur ch Universty Psych atric Hospital, Zurich Switzeriand

Found Affiliation:
Zurich University Psychiatric Hospital, Zurich, Switzerland.

Verified Affiliation:
Zurich University Psychiatric Hospital, Zurich, Switzerland.

The OCR server occasionally inserts extra spaces when converting italic text. Although the individual characters in the OCR of example 6 are mostly correct, two significant words are split by extraneous spaces, resulting in a similarity score of 3 (words in the longest matching string of words) / 6 (words in the shorter affiliation) = 0.5.

3 Current Work

Our current efforts are focused on improving the similarity scoring module to correctly identify more of the same institutions (reduce the number of false negatives) and correctly exclude more of the institutions that are in fact different (reduce the number of false positives).

Toward that end, our first task was to create a larger set of ground truth data to use for testing. Data records were copied from thirty-one journals that had been processed by the MARS system. Some of these are known to have poor quality OCR of the affiliation field in one or more articles. Of the 650 total articles in these journals, the author name for 436 articles was found in the author/affiliation table. A dataset was constructed to include the OCR text, the found affiliation text and the

verified text for each of the found affiliations for this set of 436 articles. The dataset of 2310 records includes one or more records per article depending on how many affiliations are associated with the author name.

A special program was written to display each found affiliation text and corresponding verified affiliation text from the dataset and record a human operator's decision of whether or not they are the same institution. The decisions were accumulated for five separate operators. For each pair, if the majority of the operators determined that they are the same institution, a '1' was appended to the record in the dataset. Otherwise, a '0' was appended to the record. The dataset now became a ground truth set suitable for use in testing modifications to the similarity scoring module. 514 (22.3%) of the OCR affiliation/found affiliation pairs in the test set are the same institution. The other 1796 pairs are not.

Our first exploration of the similarity scoring module was to revisit the "bag of words" method to matching. In this approach, rather than look for strings of partially-matched words, we only look for a partially-matched occurrence in the OCR affiliation text of each significant word in the found affiliation text. The similarity score is the number of such words divided by the number of significant words in the found affiliation. "Significant" words are those that satisfy the minimum word length requirement, are not in the stop word list, are not enclosed in parentheses and are not an email address. To reduce the possibility of a strong match between the words in the found affiliation and words in the OCR affiliation text that are from affiliations other than the first affiliation, the OCR affiliation text is further edited if it is longer than 1.5 times the length of the found affiliation: ending lines of the OCR affiliation are successively removed until the remaining text is no longer than 1.5 times the length of the found affiliation. As expected, the bag of words method does calculate a low similarity score for many non-same institutions. However, higher scores do not conclusively indicate that the text being compared is for the same institution.

The bag of words method was tested with four minimum word length and minimum edit distance requirements. The resulting true positives, false positives, true negatives and false negatives for a similarity score threshold of 0.85 are shown in Figure 3. The threshold is selected by plotting all of the test results and visually choosing the score at which there is a sharp increase in the number of true positives and a sharp decrease in the number of true negatives.

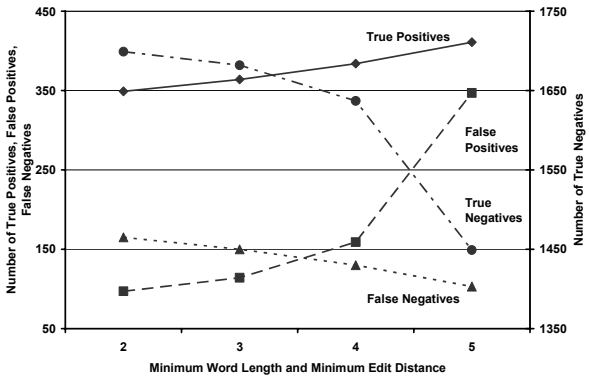

Figure 3. The effect of minimum word length and minimum edit distance on true and false positives and negatives using *bag of words* partial matching with a similarity score threshold of 0.85.

We also reexamined the original algorithm that calculated the similarity score as the ratio of the total number of words in the two longest chains of partially-matched words to the number of words in the shorter of the two affiliations being compared. This time we truncated long OCR affiliations as described for bag of words matching, and we did not remove all-digit words from consideration, reasoning that zip codes and street addresses could be useful components of a chain of partially matched words.

The chains of words method was tested with four minimum word length and minimum edit distance requirements. The resulting true positives, false positives, true negatives and false negatives for a similarity score threshold of 0.80 are shown in Figure 4. The method for selecting the threshold is the same as for Figure 3.

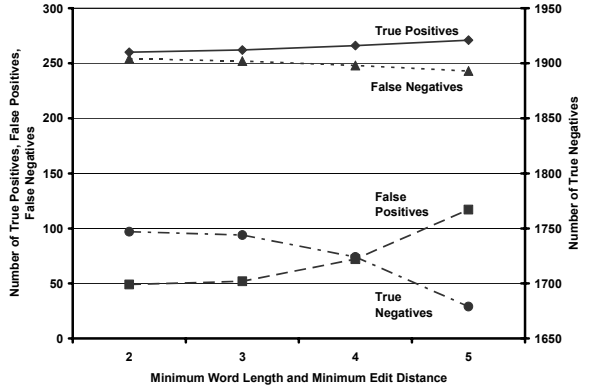

Figure 4. The effect of minimum word length and minimum edit distance on true and false positives and negatives using *chains of words* partial matching with a similarity score of 0.80.

The true positive and false positive values from Figures 3 and 4 are shown together in Figure 5.

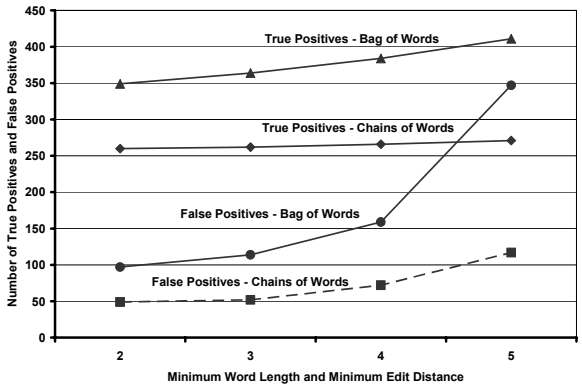

Figure 5. Comparing *bag of words* and *chains of words* method.

For both matching methods, true and false positives increase as minimum word length and minimum edit distance increase. However, true positives increase approximately linearly as a function of these requirements, while false positives increase as a power function. The tradeoff between true and false positives even extends to the choice of method: the bag of words method achieves higher rates of true positives at the cost of higher rates of false positives. However, at minimum word length and edit distance of less than 5, the ability of each method to separate the true and false positives is about the same, i.e., the number of true positives minus the number of false positives is about the same.

4 Future Strategies

There are several ideas to explore that have the potential for improving the similarity scoring algorithm. These include:

- Try other combinations of minimum word length and minimum edit distance, for example, a minimum word length of 3 and a minimum edit distance of 2.

- Explore a two-pass scheme in which the bag of words method, with one set of minimum requirements, is used to eliminate obvious non-matching affiliations, and then use the chains of words method, with another set of minimum requirements, to score the rest.

- Weight the contribution to the final score based on the word that is partially matched. For example, the words "Department" and "University" are less useful in identifying a particular institution than the words "Immunology" and "California." We have frequencies of word occurrence in MEDLINE affiliation fields [5] that can be used to build a table of weights. Additional information may be obtained through syntactical analysis of the found affiliation for words that are important for identifying that institution.

5 Conclusions

Simple partial matching schemes are adequate for eliminating obvious mismatches between the affiliation in the OCR text and affiliations found in the historical data. Likewise, they are adequate for finding matches in cases where the quality of the OCR text is good enough for partial matching of most of the significant words. Other techniques will be needed to reliably match affiliations for those cases where the OCR text is poor, or includes multiple affiliations. We continue to explore such techniques with confidence that we will be able to achieve reliable selection of affiliations to reduce the manual labor of the verification operator.

References

[1] Thoma GR. Automating data entry for an online biomedical database: a document image analysis application, *Proc. 5th International Conference on Document Analysis and Recognition (ICDAR'99)* (Bangalore, India, 1999) 370-3.

[2] Thoma GR. Automating the production of bibliographic records for MEDLINE, an R&D report of the Communications Engineering Branch, LHNCBC, NLM, Bethesda, Maryland (2001) 91 pp. (archive.nlm.nih.gov)

[3] Kim J, Le DX, Thoma GR. Automated labeling of bibliographic data extracted from biomedical online journals, *Proc. SPIE: Document Recognition and Retrieval X* **5010** (January 2003) 47-56.

[4] Thoma GR, Ford G. Automated data entry system: performance issues, *Proc. SPIE: Document Recognition and Retrieval IX* **4670 (**January 2002) 181-90.

[5] Ford G, Hauser SE, Le DX, Thoma GR. Pattern matching techniques for correcting low confidence OCR words in a known context, *Proc. SPIE: Document Recognition and Retrieval VIII* **4307** (January 2001) 241-9.

[6] Lasko TA, Hauser SE. Approximate string matching algorithms for limited-vocabulary OCR output correction, *Proc. SPIE: Document Recognition and Retrieval VIII* **4307** (January 2001) 232-40.

[7] Schlaifer J, Hauser SE. OCR affiliations: Feasibility considerations and numerical scoring for correction from past datasets. Internal project report of the Communications Engineering Branch, LHNCBC, NLM, Bethesda, Maryland (2001).

[8] Hauser SE, Schlaifer J, Sabir TF, Demner-Fushman D, Thoma GR. Correcting OCR text by association with historic datasets, *Proc. SPIE: Document Recognition and Retrieval X* **5010** (January 2003) 84-93.