Historical Author Affiliations Assist Verification of Automatically Generated MEDLINE® Citations

Tehseen F. Sabir, Susan E. Hauser, Ph.D., George R. Thoma, Ph.D.
National Library of Medicine, NIH, DHHS, Bethesda, MD

High OCR error rates encountered in author affiliations increase the manual labor needed to verify MEDLINE citations automatically created from scanned journal articles. This is due to poor OCR recognition of the small text and italics frequently used in printed affiliations. Using author-affiliation relationships found in existing MEDLINE records, the SeekAffiliation (SA) program automatically finds potentially correct and complete affiliations, thereby reducing manual effort and increasing the efficiency of creating the citations.

The National Library of Medicine (NLM) developed the Medical Article Records System (MARS)[1] to automatically construct key fields of a MEDLINE bibliographic record from the OCR text of the scanned first page of journal articles. Prior to completing the record, the OCR text fields are manually verified and edited if necessary. The Seek Affiliation (SA) software program reduces operator time needed to verify the affiliation field, which tends to include a disproportionate number of OCR errors[2].

NLM's MEDLINE database contains over 14 million indexed citations to the biomedical journal literature. Most citations include a list of one or more authors and the affiliation of the first author. Many authors publish frequently while at the same institution. For each article processed by MARS, SA obtains from the MARS database the OCR text for the first author name, up to three additional author names, and the affiliation field. Using e-utilities from NLM's Entrez system[3], SA retrieves up to 100 MEDLINE citations whose first author name matches one of the OCR author names, giving priority to the OCR first author, and removing duplicate affiliations. SA's similarity scoring algorithm generates a score for each MEDLINE affiliation based on its similarity to the OCR affiliation. SA generates MARS database records for the 5 highest scoring affiliations that exceed a threshold determined via a ROC evaluation[4]. SA records are retrieved by the MARS verification program and the SA affiliations are presented to the operator as options to the OCR affiliation text. The operator may select an affiliation as is, or may edit it to create the correct affiliation.

SA options were introduced to MARS operators in December, 2005. Usage data for January and February, 2006, are:

- Total articles processed by MARS: 20258
- Total articles for which SA found a suggested affiliation: 4788 (24%)
- Total articles for which operators selected SA affiliations: 671(3.3%, 14%)

An SA affiliation was chosen for 671 articles, representing 3.3% of the total articles processed and 14% of the articles for which there were SA suggestions. Of the nine MARS verification operators four select SA affiliations at a higher rate than the overall average. These four operators select an SA affiliation for 6.1% of the total articles processed, or for 23.5% of the articles that have SA affiliations. The following case, where the operator chose an SA affiliation to replace the OCR affiliation text, illustrates how SA affiliations are useful:

OCR Aff Text: Klinik und Poliklinik des Kindes- und Jugendalters der Universiat zu K61n
(errors and incomplete)
SA Aff Selected: Klinik und Poliklinik f¨ur Psychiatrie und Psychotherapie des Kindes- und Jugendalters der Universit¨at zu K¨oln, Robert-Koch-Str. 10, 50931 K¨oln.
(full affiliation as it appears in an existing MEDLINE citation)

Because correcting the affiliation field is frequently labor intensive, we conclude that SA contributes to an overall improvement in speed and accuracy of the MARS system. As the MEDLINE database grows, the probability of finding correct affiliations in biomedical citations for a given author name will increase beyond the initial statistics given.

References

1. Thoma GR. Automating data entry into MEDLINE . Proc. 1999 Symp. on Document Image Understanding Technology, Apr 1999; College Park, MD: Institute for Advanced Computer Studies; 217-8.
2. Hauser SE, Sabir TF, Thoma GR. OCR correction using historical relationships from verified text in biomedical citations. Proc. 2003 Symp. on Document Image Understanding Technology, Apr 2003; College Park, MD: Institute for Advanced Computer Studies; 171-7.
3. U.S. National Institutes of Health, National Library of Medicine. Entrez Programming Utilities. http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.
4. Hauser SE, Schlaifer J, Sabir TF, Demner-Fushman D, Thoma GR. Correcting OCR text by association with historic datasets. Proc. SPIE Electronic Imaging, January 2003. SPIE Vol. 5010: 84-93.