A New Approach to Automatic Indexing using Journal Descriptors

Susanne M. Humphrey
National Library of Medicine
Bethesda, MD

INTRODUCTION

Work has been performed on a fully automated approach for the general categorization of documents based on a training set of document titles and abstracts and the journal-level indexing of the journals in which the documents are published. From this training set, associations are computed between the words in the titles and abstracts and the journal level indexing. A document outside the training set can be categorized using journal-level indexing terms when it has many words in common with those in the training set.

For example, words in titles and abstracts of MEDLINE citations to articles published in the journal Cancer can be said to be associated with MEDICAL ONCOLOGY, which is the journal descriptor for the this journal in NLM's SERLINE. If the association is very strong, i.e., if certain words in a document appear more often in documents in Medical Oncology journals, i.e., journals indexed MEDICAL ONCOLOGY in SERLINE, than in journals in other disciplines, we can then use this JD to index the individual document. If the words in this document are also strongly associated with journals having the JD PHARMACOLOGY, we can also consider this JD to be an indexing term for the individual document.

OBJECTIVE AND PREVIOUS APPROACHES

The goal of this research is to automatically index documents using a consistent, timely set of controlled indexing terms.

Previous approaches have been studied based on associating words in individual documents with their controlled indexing terms, requiring very large training sets of hundreds of thousands of documents. For example, the titles of 371,454 items from the AP newswire and their assigned keywords having some aspects of a controlled vocabulary were used in developing a probabilistic classifier (Lewis & Gale, 1994). Another study reported a production system based on statistical relations between words, phrases, and formulas in titles and abstracts from the abstract journal *Physical Briefs* and assigned indexing terms from a thesaurus for more than one million documents with about 125,000 documents added annually (Biebricher, Fuhr, Lustig, Schwantner & Knorz, 1988). A training set of about 4500 documents drawn from more than one million citations in INSPEC was used for statistically associating words from authors, titles, and abstracts with their indexing terms from the INSPEC thesaurus (Plaunt & Norgard, 1998). The trouble with these approaches is that they depend on the considerable intellectual effort by humans having indexed all those documents. To keep up with current knowledge, this indexing must be ongoing. Furthermore, the indexing in the training set would become inconsistent as indexing vocabularies and schemes change; due to volume, it is normally prohibitive to re-index document collections retrospectively.

Automatic categorization by discipline-like indexing terms (numbering in the hundreds in contrast to indexing thesauri normally with preferred terms numbering in the thousands or tens of thousands) has been investigated previously using Subject Field Codes (124 major fields, e.g., Business, Economics, Medicine, Meteorology, and 250 subfields) manually assigned by lexicographers to tens of thousands of word senses in LDOCE, Longman's Dictionary of Contemporary English (Liddy, Paik & Woelfel, 1993; Liddy & Paik, 1992). Words in a collection of test documents were tagged with the appropriate SFCs according to the LDOCE assignments, and statistical algorithms were applied to cluster the documents into meaningful groupings. The SFCs provided "an intermediate level representation of a text's contents" rather than a final representation for the clusters. For example, analysis of clusters showed they corresponded to topics such as "airlines" and "medical treatment", but the system was not designed to automatically provide controlled indexing terms for these clusters.

By contrast, the approach taken in our research depends on the intellectual effort of indexing at the journal level, usually a matter of a few thousand indexed items. Also, with journal-level indexing, there is some hope of maintaining the indexing over the years for currency with the domain and for consistency of indexing. The limited set of controlled indexing terms (143 unique JDs used in SERLINE) would not be expected to change radically over the years. Most of the terms are names of disciplines, i.e., Medical Oncology, Pharmacology, etc., as mentioned earlier. Maintaining journal-level descriptors and assigning them to journals are normal functions at NLM. Conceivably, these functions could reasonably be performed in other specialized domains, not limited to the scientific and technical.

METHODOLOGY

Our training set, which comes from another ongoing research project, is a sample taken from MEDLINE indexing input during 1993, consists of 3995 citations contained in 1466 different journals, having total JD counts as follows: 1016 journals have one JD, 370 have two JDs, 69 have three JDs, and 11 have four JDs. There are 123 unique JDs in this training set. To compute the JD profile for a word, we initially compute the number of occurrences of a word in journals with a particular JD. For example, in our training set the word ANTAGONIST or the plural variant occurs 48 times in documents in journals described by the JD PHARMACOLOGY. We divide this by the total number of occurrences of this word in the dataset, which is 137, giving us the ranking for this JD. We do this for each JD, resulting in the following display, ranking all the JDs for journals in which ANTAGONIST/S occurs:

```
WORD = ANTAGONIST
VARIANTS = ANTAGONIST ANTOGONISTS
TOTAL CITATION COUNT = 90
TOTAL OCCURRENCES = 137
TOTAL JD COUNT = 27
OCCURRENCES OF WORD VARIANTS PER JD / TOTAL OCCURRENCES, BY COUNT:
|PHARMACOLOGY| 48/137 = 0.350365
|BIOCHEMISTRY| 23/137 0.167883
|DRUG THERAPY| 17/137 0.124088
|NEUROSCIENCES| 13/137 = 0.094891
|ENDOCRINOLOGY| 11/137 = 0.080292
etc.
```

Alternatively, we also compute a profile based on the number of citations with at least one occurrence of the word in journals with a particular JD, divided by the total number of citations with at least one occurrence of the word. For example, ANTAGONIST/S occurs in 30 citations in journals described by the JD PHARMACOLOGY, and we divide this number by 90, which is the total number of citations in which this word occurs. Again, we do this for each JD, appending the following result to the above display:

```
CITATION COUNT FOR WORD VARIANTS PER JD / TOTAL CITATION COUNT, BY COUNT:
|PHARMACOLOGY| 30/90 = 0,333333
|BIOCHEMISTRY| 15/90 0.166667
|NEUROSCIENCES| 9/90 0.1
|DRUG THERAPY| 8/90 = 0.088889
|PHYSIOLOGY| 7/90 = 0.077778
etc.
```

We also can display journal titles in the testset with their JDs and counts for each word. The first number in each line is the number of occurrences of ANTAGONIST/S in documents in the journal. The second number is the number of citations in which ANTAGONIST/S occurs for the journal.

```
14 7 |J Pharmacol Exp Ther| |DRUG THERAPY| |PHARMACOLOGY|
10 7 |Eur J Pharmacol| |PHARMACOLOGY|
 6 5 |Br J Pharmacol| |PHARMACOLOGY|
 6 3 |Pharmacol Biochem Behav| |PSYCHOPHARMACOLOGY| |BEHAVIOR| |PHARMACOLOGY|
     |BIOCHEMISTRY|
 5 2 |J Neurochem| |NEUROSCIENCES| |CHEMISTRY|
etc.
```

To compute a ranked list of JDs for a document outside the training set, we average the rankings for each JD in the JD profiles of words that occur in the document. As a sample document, we used the complete text of a document from the full-text file on DIALOG for the New England Journal of Medicine (File 444), briefly cited as follows:

```
00114110
Copyright 1995 by the Massachusetts Medical Society
Dexamethasone, Granisetron, or Both for the Prevention of Nausea and
Vomiting during Chemotherapy for Cancer (Original Articles)
 The Italian Group for Antiemetic Research.
 The New England Journal of Medicine
 Jan 5, 1995; 332 (1), pp 1-5
 LINE COUNT: 00338    WORD COUNT: 04676
```

The JD for the New England Journal of Medicine in SERLINE is MEDICINE which is not particularly helpful for categorizing documents within the biomedical domain (to our knowledge, JDs are not directly available as search terms in any bibliographic retrieval system). However, our experimental system provided ranked JDs as indexing terms for this document as follows, showing appropriate outstanding descriptors MEDICAL ONCOLOGY and PHARMACOLOGY, based on JD profiles of words in the full text matching words in the training set and using their associations with JDs computed from the training set.

```
JD'S AND RANK BASED ON WORD/VARIANTS OCCURRENCES, BY RANK:
("MEDICAL ONCOLOGY" 0. 188727)
("PHARMACOLOGY" 0.086162)
("BIOCHEMISTRY" 0.057378)
("MEDICINE" 0.050353)
etc.
JDS AND RANK BASED ON CITATION COUNT FOR WORD/VARIANTS, BY RANK:
("MEDICAL ONCOLOGY" 0.181567)
("PHARMACOLOGY" 0.09302)
("BIOCHEMISTRY" 0.068717)
("MEDICINE" 0.063437)
etc.
```

While processing this full text document, we noticed the end references and the possibility of using JDs associated with them. The fifteen end references were distributed among six JDs as follows. If end references in a document can be parsed for isolating serial titles, it should be possible to automatically extract the JDs for these titles from SERLINE as possible indexing terms.

DRUG THERAPY
 Drug Safety 1993;9:410-28
TOXICOLOGY
 Drug Safety 1993;9:410-28
MEDICINE
 Lancet 1991;338:478-9
 N EngI J Med 1984;311:549-52
 Lancet 1991;338:483-7
MEDICAL ONCOLOGY
 Am J Clin Oncol 1988;11:594-6
 Am J Clin Oncol 1989;12:524-9
 Eur J Cancer Clin Oncol 1987;23:615-7
 Oncology 1988;45:346-9
 Am J Clin Oncol 1987;10:264-7
 J Clin Oncol 1990;8:1063-9
 Eur J Cancer 1990;26:311-4
 Eur J Cancer 1991;27:1137-40. Erratum, Eur J Cancer 1991;27:1717
 Eur J Cancer 1990;26:Suppl 1:S28-S32
 Support Care Cancer 1994;2:171-6.
NEOPLASMS, EXPERIMENTAL
 J Natl Cancer Inst 1991;83:1169-73
HEALTH SERVICES
 Support Care Cancer 1994;2:171-6.

PROBLEM AREAS AND FUTURE WORK

Our plans for future investigation include work in the following problem areas: developing criteria for word selection, e.g., stopwords, word frequency, grouping word variants, part of speech; developing and testing normalization algorithms to counteract uneven word and citation counts associated with JDs; developing algorithms to group rankings in a result, with the set of best JDs listed first, followed by JDs of intermediate quality; and for documents containing end references, extending this research to using JDs of serial titles if they can be readily parsed from these references. Future work also includes using a larger training set such as a complete month's input to MEDLINE and using our approach in various information retrieval applications to get a better sense for its utility. Such applications might include: search terms as alternative to detailed human indexing, text representation using natural language processing, referral to information sources in multi-source systems, referral to similar documents, and accessing formatted texts not routinely indexed (parts of monographs, grey literature, Web documents, etc.).

REFERENCES

Biebricher, P., Fuhr, N., Lustig, G., Schwantner, M., & Knorz, G. (1988). The automatic indexing system AIR/PHYS -- from research to application. In Y. Chiaramella (ed.), *ACM SIGIR 1lth International Conference on Research & Development in Information Retrieval* (pp. 333-342). New York: Association for Computing Machinery.

Lewis, D. D., & Gate, W. A. (1994). A sequential algorithm for training text classifiers. In W. B. Croft & C. J, van Rijsbergen (eds.), SIGIR '94, *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (pp. 3-12). Springer-Verlag: London.

Liddy, E. D., Paik, W., & Woelfel, J. K. (1993). Use of Subject Field Codes from a machine-readable dictionary for automatic classification of documents, In R. Fidel, B. H. Kwasnik, & P. J. Smith (eds.), *Advances in Classification Research, Vol. 3, Proceedings of the 3rd ASIS SIGICR Classification Research Workshop* (pp. 83-100). Silver

Spring, MD: American Society for Information Science (published by Learned Information, Medford, NJ).

Liddy, E. D., & Paik, W. (1992). Statistically-guided word sense disambiguation. In *Intelligent probabilistic approaches to natural language, Papers from the 1992 Fall Symposium, Technical Report FS-92-04* (pp. 99-107). Menlo Park, CA: AAAI Press.

Plaunt, C., & Norgard, B. A. (1998). An association based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science*. 49(10) pp. 888-902,