MULTIPLE APPROACHES TO FINE-GRAINED INDEXING OF THE BIOMEDICAL LITERATURE

AURELIE NEVEOL[1,2], SONYA E. SHOOSHAN[1], SUSANNE M. HUMPHREY[1],
THOMAS C. RINDFLESH[1], ALAN R. ARONSON[1]

[1]*National Library of Medicine, NIH*
Bethesda, MD 20894, USA

[2]*Equipe CISMeF, Rouen, France*

The number of articles in the MEDLINE database is expected to increase tremendously in the coming years. To ensure that all these documents are indexed with continuing high quality, it is necessary to develop tools and methods that help the indexers in their daily task. We present three methods addressing a novel aspect of automatic indexing of the biomedical literature, namely producing MeSH main heading/subheading pair recommendations. The methods, (dictionary-based, post- processing rules and Natural Language Processing rules) are described and evaluated on a genetics-related corpus. The best overall performance is obtained for the subheading genetics (70% precision and 17% recall with post-processing rules, 48% precision and 37% recall with the dictionary-based method). Future work will address extending this work to all MeSH subheadings and a more thorough study of method combination.

1. Introduction

1.1. *Indexing the biomedical literature*

To ensure efficient retrieval of the ever-increasing number of articles in the U.S. National Library of Medicine's (NLM's) MEDLINE® database, these documents must be systematically stored and indexed. In MEDLINE, the subject matter of articles is described with a list of descriptors selected from NLM's Medical Subject Headings (MeSH®). MeSH contains about 24,000 main headings covering specific concepts in the biomedical domain such as diseases, body parts, etc. It also contains 83 subheadings that denote broad areas in biomedicine such as immunology or genetics. Subheadings can be coordinated to a main heading in order to refer to a concept in a more specific way. NLM indexers select for each article an average of ten to twelve MeSH main headings (e.g., *Williams Syndrome*) or main heading/subheading pairs (e.g., *Williams Syndrome/genetics*). The indexing task is time consuming and requires skilled, trained individuals. In order to assist indexers in their daily practice, the NLM's Indexing Initiative [1] has investigated automatic indexing methods, which led to the development of the Medical Text Indexer (MTI) [2]. MTI is a software tool producing indexing recommendations in the form of a list of stand-alone main headings (i.e. not associated with subheadings) shown on request to the indexers while they work on a record in the MEDLINE Data Creation and Maintenance System (DCMS). Other work on the automatic assignment of MeSH descriptors to medical texts in English has also focused on stand-alone main headings [3-4]. While the indexing resulting from some of these automatic systems has been shown to approach human indexing performance as measured by retrieval [5], there is a need for automatic means to provide finer-grained indexing recommendations, namely main heading/subheading pairs in addition to stand-alone main headings.

In fact, there are both theoretical and practical reasons for this effort. From a theoretical point of view, the MeSH indexing manual [6] states that indexers must chose descriptors that reflect the content of an article by first selecting correct main headings and second by attaching the appropriate subheadings. Consequently, selecting an isolated main heading where a main heading/subheading pair should have been assigned is, strictly speaking, erroneous – or at best, incomplete. On the practical side, indexers do use both main headings and main heading/subheading pairs when indexing a document. Therefore, stand-alone main heading recommendations, while useful, will always need to be completed by attaching subheadings where appropriate.

The task of assigning MeSH descriptors to a document can be viewed as a multi-class classification problem where each document will be assigned several “classes” in the form of MeSH descriptors. When assigning MeSH main headings [4, 7] the scale of the classification problem is 23,883. Now, if one attempts to assign MeSH main heading/subheading pairs, the number of classes increases to 534,981. Many machine learning methods perform very well on binary classes but prove more difficult to apply successfully on larger scale problems. As regards MeSH main heading classification, the hierarchical relationships between the classes have been used to reduce the complexity of the problem [4, 7]. Previous work on producing automatic MeSH pair recommendations that relied on dictionary and rule-based methods seemed promising [10]. For these reasons, we are investigating similar methods here.

1.2. *Genetics literature*

Following the rapid developments of genetics research in the past twenty years, the volume of genetics-related literature has grown accordingly. While genetics literature represented about 6% of MEDLINE records for the year 1985[*], it represents over 19% of MEDLINE records for 2005[†].

In this context, it seems that providing fine-grained indexing recommendations for genetics literature is particularly important, as it will impact a significant portion of the biomedical literature. Therefore, we have elected to concentrate our effort in this subdomain for our preliminary work investigating automatic methods of providing MeSH pair indexing recommendations. This led us to focus on the subheadings genetics, immunology and metabolism which were found to be prevalent in the MeSH indexing of our genetics test corpus (see section 2.4).

1.3. *Objective and approach*

This paper presents the various methods we investigated to automatically identify MeSH main heading/subheading pairs from the text (title and abstract) of articles to be indexed for MEDLINE. The ultimate goal of this research is to add subheading-related features to DCMS when displaying recommendations to NLM indexers, in order to save time during the indexing process. A previous study of MTI usability showed that the possibility of selecting recommendations from a pick list saved look-up and typing time [8]. The ideal time-saving mechanism for subheading attachment would be to include relevant pairs in the current list of main headings available for selection. However, this solution is only viable if the precision of such recommendations is sufficiently high.

The possible obstacle that we foresee to including pair recommendations in the current pick list is that high precision for pair recommendations might be difficult to achieve without any human input throughout the process. Work in the area of computer-assisted translation [9] has shown the usefulness of interactive systems in the context of highly demanding cognitive tasks such as translation – or indexing. For this reason, we are considering the possibility of either dynamically showing related pair recommendations once the indexer selects a main heading for the record, or highlighting the most likely subheadings for the current record when indexers are viewing the list of allowable subheadings for a given main heading that they selected. The remainder of this paper will address the difficult task of producing the recommendations themselves.

[*] 19,348 citations retrieved by the query *genetics AND 1985 [dcom] AND MEDLINE [sb]* compared to 313,638 records retrieved by the query *1985 [dcom] AND MEDLINE [sb]* on 07/12/06.

[†] 114,530 citations retrieved by the query *genetics AND 2005 [dcom] AND MEDLINE [sb]* compared to 598,217 records retrieved by the query *2005 [dcom] AND MEDLINE [sb]* on 07/12/06.

2. Material and methods

In this section, we describe the three methods we investigated to identify main heading/subheading pairs from medical text. We also introduce the genetics corpus we used to evaluate the methods.

2.1. *Baseline dictionary-based method*

The first method we considered consists of identifying main headings and subheadings separately for a given document and then attempting to pair them.

Main headings are retrieved with the Medical Text Indexer [2] and subheadings are retrieved by looking up words from the title and abstract in a manually built dictionary in which each entry contains a subheading and a corresponding term or expression that is likely to represent the subheading in text. These terms are mainly derived from inflectional and derivational forms of the subheadings. They were obtained manually and tested on a general training corpus composed of a random 3% selection of MEDLINE 2004. Candidate terms were added to the dictionary if they benefited the method performance on the training corpus. For example, *gene*, *genes*, *genetic*, *genetics*, *genetical*, *genome* and *genomes* are terms corresponding to */genetics*. The dictionary contains 227 entries for all 83 subheadings, including 10 for */genetics*.

To obtain the pairs, the subheadings retrieved by the dictionary are coordinated with the main headings retrieved, if applicable. For each main heading, MeSH defines a set of subheadings called “applicable qualifiers” that can be coordinated with it (e.g. */genetics* is applicable to *Carcinoma, Renal Cell* but not *Odds Ratio*). In the dictionary method, all the legal pairs that can be assembled from the sets of main headings and subheadings retrieved are recommended. For example, two occurrences of the dictionary entry *genes* were found in the abstract of MEDLINE record 15319295, which means that */genetics* was identified for this record. Attempts were made to attach */genetics* to each of the twelve main headings recommended by MTI for this record, including *Carcinoma, Renal Cell* and *Odds Ratio*. The pair *Carcinoma, Renal Cell/genetics* was recommended because */genetics* is an allowable qualifier for *Carcinoma, Renal Cell*. However, */genetics* is not an allowable qualifier for *Odds Ratio*; therefore no other pair recommendation was made.

2.2. *Indexing rules*

The two methods detailed in this section are based on indexing practice, sometimes expressed in MeSH annotations. In previous work on the indexing of medical texts in French [10], indexing rules were derived from interviews with indexers. Similar rules were also available in the MedIndEx knowledge base

[11]. To build the sets of rules used here, we adapted existing rules [10-11] and manually created new rules. The rules were divided in two groups.

Post-processing rules

Post-processing (PP) rules build on a pre-existing set of indexing terms (i.e., the main heading recommendations from MTI), and enrich it by expanding on the underlying concepts denoted by the indexing terms within that set. Twenty-nine of these rules are currently implemented for */genetics* (as well as 11 for */immunology* and 8 for */metabolism*). Rules that were created in addition to the existing rules from MedIndEx and the French system (such as the example shown in figure 1) were evaluated using MEDLINE data. Specifically, we computed an estimated precision equal to the number of citations indexed with the trigger terms over the number of citations indexed with the trigger terms and the recommended pair[‡]. Only rules with an estimated precision over 0.6 were considered for inclusion in the rule sets.

According to the sample rule shown in Figure 1, a pair recommendation shall be triggered by existing MTI recommendations including the main heading *Mutation* as well as a *<DISEASE>* term[§]. Since *Mutation* is a genetics concept, an inference is made that */genetics* should be attached to the disease main heading. For example, both main headings *Mutation* and *Pancreatic Neoplasms* are recommended by MTI for the MEDLINE record 14726700. As *Pancreatic Neoplasms* is a disease term, the rule will be applied and the pair *Pancreatic Neoplasms/genetics* will be recommended.

> **If** the main heading *Mutation* and a *<DISEASE>* term appear in the indexing recommendations
> **then** the pair *<DISEASE>/genetics* should also be used.

Figure 1. Sample post-processing rule for the subheading genetics

[‡] For the sample rule shown in Figure 1, the estimated precision was 0.67. (On 09/06/06, the query *mutation [mh] AND (diseases category/genetics [mh] OR mental disorders/genetics [mh]*) retrieved 144,698 citations while *mutation [mh] AND (diseases category [mh] OR mental disorders[mh])* retrieved 216,749 citations)

[§] DISEASE refers to any phrase that points to a MeSH main heading belonging to the *diseases* or *mental disorders* categories.

Natural Language Processing rules

Natural Language Processing (NLP) rules use cues from the title or abstract of an article to infer pair recommendations. A sample NLP rule is shown in Figure 2. In the original French system, this type of rule was implemented by a set of transducers that exploited information on each term's semantic category (DISEASE, etc.) stored in an integrated electronic MeSH dictionary. Although very efficient, this method is also heavily language-dependent. For English, such advanced linguistic analysis of medical corpora is performed by NLM's SemRep [12], a tool that is able to identify interactions between medical entities based on domain knowledge from the Unified Medical Language System® (UMLS®).

> **If** a phrase such as "<*GENE*>[**] is associated with <*DISEASE*> " appears in text **then** the pair <*DISEASE*>*/genetics* should also be used.

Figure 2. Sample Natural Language Processing rule for the subheading genetics

Specifically, SemRep retrieves UMLS triplets composed of two concepts from the UMLS Metathesaurus® together with their respective UMLS Semantic Types (STs) and the relation between them, according to the UMLS Semantic Network. Hence, phrases corresponding to the pattern of the sample rule presented in Figure 2 would be extracted by SemRep as the triplet (gngm ASSOCIATED_WITH dsyn) where "gngm" denotes the ST "Gene or Genome", and "dsyn" denotes the ST "Disease or Syndrome". We can infer from this that there is an equivalence between the semantic triplet (gngm ASSOCIATED_WITH dsyn) and the MeSH pair <*DISEASE*>*/genetics* where "dsyn" and <*DISEASE*> refer to the same entity. In this way, the NLP rules were used to obtain a set of equivalencies between these UMLS triplets and MeSH pairs. Subsequently, a restrict-to-MeSH algorithm [13] was used to translate UMLS concepts to their MeSH equivalents. For example, the phrase "Association of a haplotype of matrix metalloproteinase (MMP)-1 and MMP-3 polymorphisms with renal cell carcinoma" occurring in the MEDLINE record 15319295 was annotated by SemRep with the triplet (gngm ASSOCIATED_WITH neop[††]) where the "Gene or Genome" was *MMP* and the "Neoplastic Process" ("neop") was *Renal Cell Carcinoma*. The latter UMLS concept can be restricted to its MeSH equivalent *Carcinoma, Renal Cell* and the

[**] GENE refers to any phrase that points to a MeSH main heading belonging to the GENE sub-hierarchy within the GENETIC STRUCTURES hierarchy.

[††] In the Semantic Types hierarchy, "neop" is a descendant of "dsyn". By inheritance, rules that apply to a given Semantic Type also apply to its descendants.

pair *Carcinoma, Renal Cell/genetics* is then recommended for the indexing. In the context of the genetics domain, we also use triplets retrieved by SemGen [14], a variant of SemRep specifically adapted to the identification of Gene-Gene and Gene-Disease interactions.

2.3. *Combination of methods*

In an attempt to assess the complementarity of the methods, we also evaluated the recommendations provided by any two methods. The combination consisted in examining all the recommendations obtained from two methods, and selecting only the concurring ones, if any. For example, the pairs *Ascomycota/genetics*, *Capsid Proteins/genetic and RNA Viruses/genetics* and *Totivirus/genetics* were recommended by the post-processing rules method for citation 15845253 while *Viruses/genetics*, *RNA Viruses/genetics* and *Totivirus/genetics* were recommended by the NLP rules for the same citation. Only the common pairs *RNA Viruses/genetics* and *Totivirus/genetics* are selected by combination of the two methods. In this case, the two pairs selected by combination were used to index the documents in MEDLINE. Two of the three discarded pairs (*Ascomycota/genetics* and *Viruses/genetics*) were not used by the indexers while the other one (*Capsid Proteins/genetics*) was.

2.4. *Test corpus*

All three methods (baseline dictionary-based, PP rules, NLP rules) were tested on a corpus composed of genetics-related articles selected from all citations indexed for MEDLINE in 2005. In order to avoid bias, the selection was not directly based on whether the articles were indexed with the subheading genetics. Instead we applied NLM's Journal Descriptor Indexing tool, which categorized the citations according to Journal Descriptors and also according to Semantic Types [15]. This categorization provided an indication of the biomedical disciplines discussed in the articles. For our genetics-related corpus, we selected citations that met either of these criteria:

- "Genetics" or "Genetics, Medical" were among the top six Journal Descriptors
- "genf" (Gene Function) or "gngm" (Gene or Genome) were among the top six Semantic Types

A total of 84,080 citations were collected and used to test the methods presented above. At least one of the subheadings genetics, immunology and metabolism appear in 53,903 of the corpus citations.

3. Results

3.1. *Independent methods*

Table 1 shows the performance of the methods of pair recommendation presented in section 2. For each method, we detail the results obtained for */genetics*, */immunology* and */metabolism*. We also indicate the overall figures (All) for the total number of recommendations obtained (Nb_rec), the total number of citations impacted (Nb_cit), the number of recommendations that were selected by MEDLINE indexers (Nb_rec+), the precision (PREC) and the recall (REC). Precision corresponds to the number of recommendations that were actually used by MEDLINE indexers over the total number of recommendations provided by the methods. Recall corresponds to the number of recommendations that were used by the indexers over the total number of pairs that were used by the indexers.

Table 1. Performance of MeSH pair recommendation

Method	**Nb_rec**	**Nb_cit**	**Nb_rec+**	**PREC**	**REC**
Dictionary (GE)	97,553	29,632	46,804	0.48	0.3663
Dictionary (IM)	6,691	1,629	2,326	0.35	0.1095
Dictionary (ME)	5,317	1,577	2,166	0.41	0.0200
Dictionary (All)	**109,561**	**31,476**	**51,296**	**0.47**	**0.1993**
PP (GE)	31,164	16,441	21,752	0.70	0.1703
PP (IM)	1,451	1,027	1,048	0.72	0.0493
PP (ME)	25,823	10,391	13,578	0.53	0.1253
PP (All)	**58,438**	**23,184**	**36,378**	**0.62**	**0.1413**
NLP (GE)	2,480	2,327	1,566	0.63	0.0123
NLP (IM)	97	91	26	0.27	0.0012
NLP (ME)	21	17	3	0.33	0.0000
NLP (All)	**2,598**	**2,435**	**1,605**	**0.62**	**0.0062**

3.2. *Combinations*

Table 2. Cross precision of MeSH pair recommendation methods

Method	**Dictionary**	**PP**	**NLP**
Dictionary	**0.47**	0.73	0.75
PP	0.73	**0.62**	0.87
NLP	0.75	0.87	**0.62**

Table 2 shows the precision and Table 3 shows the recall obtained when the methods are combined two by two (bold figures on the diagonal reflect the performance of the methods considered independently, as presented in Table 1).

Table 3. Cross recall of MeSH pair recommendation methods

Method	**Dictionary**	**PP**	**NLP**
Dictionary	**0.1993**	0.0498	0.0055
PP	0.0498	**0.1413**	0.0028
NLP	0.0055	0.0028	**0.0062**

4. Discussion

4.1. *General*

The performance of each method can vary considerably depending on the subheading it is applied to. Moreover, the global performance of all three methods seems higher for */genetics* than */metabolism* or */immunology*. This may be explained by the fact that genetics is a more circumscribed domain than metabolism and immunology. The best overall precision is obtained with the post-processing rules, and the best overall recall is obtained with the dictionary method. Similar observations could be made on a general training corpus, where the scope of the methods was mostly limited to the genetics-related articles.

4.2. *Error analysis*

To gain a better understanding of the results and how they might be improved, we have analyzed a number of recommendations that were made which were inconsistent with our reference (MEDLINE indexing) and therefore analyzed as errors. Table 4 presents a few characteristic cases. Most errors fall into these categories:

- Recommendation seems to be relevant
- Recommendation corresponds to a concept not substantively discussed
- Recommendation is incorrect

Especially with the NLP rules, there seem to be more cases where the recommendations address a relevant topic that is not discussed substantively in the article (e.g. PMID 15659801 in table 4). Sometimes, however, as shown in the example of PMID 15638374 in table 4, the concept denoted by the recommended pair seems relevant but not indexed. The added value of our tool could include reducing the number of similar omissions in the future.

Most “incorrect” recommendations come from the dictionary method which is the most simplistic. Another common source for errors is the case exemplified with PMID 15574482 in table 4 where a given post-processing rule can apply to several main headings, but only one of the candidates is relevant for subheading attachment. This situation was particularly prevalent with */metabolism* and resulted in a significantly lower precision for this subheading, compared to */immunology* and */genetics*.

Table 4. Analysis of sample erroneous pair recommendations

Recommendations	**Method**	**Error interpretation**
PMID 15574482 Seeds/GE Seedling/GE Oryza sativa/GE[‡‡]	PP: if MH *Plants, Genetically Modified* and a *<PLANT>* appear in the indexing, the pair *<PLANT>/genetics* should be used.	Three plants were discussed and the rule only applied to one, *Oryza sativa*, which was more specific (however, there is no direct ancestor-descendant relationship between the terms).
PMID 15638374 Phyllodes Tumor /GE	NLP: The text "The aim of the study was an evaluation of PCNA and Ki-67 expression in the stromal component of fibro-epithelial tumours."[§§] was interpreted by SemRep as "gngm LOCATION_OF neop" which translate into *Phyllodes Tumor/genetics*.	The recommended pair seems relevant for the article, although it doesn't appear in the MEDLINE indexing.
PMID 15659801 Liver Neoplasms /GE	Dictionary: The phrase "… gene expression in liver tumors … " contains the dictionary entry "gene", related to */genetics* which is an allowable qualifier for *Liver Neoplasms*, retrieved by MTI.	The concept is not substantively discussed in the article.

Error analysis can point to changes that should be made in the rules or formal concept description. Links between concepts in the case of PMID 15574482 in table 4 would make it possible to consider a filtering according to main heading specificity. For example if the fact that *Oryza sativa* is a more specific term than either *seeds* or *seedling* were available, one might consider

[‡‡] In this case, three pairs were recommended when applying the rule and only one (underlined) was correct.

[§§] The original phrase was edited to enhance legibility in the table enforcing a rule stating that subheadings should be only attached to the most specific term when several terms belonging to a same hierarchy are candidates for attachment.

4.3. *Complementarity of the methods*

The overlap in recommendations is not significant. As a result, using different methods will help cover more citations and increase the overall recall. However, the gain in precision obtained when combining several methods is offset by significant loss in recall. In fact, most of the recommendations resulting from the combination of methods concern the subheading genetics, especially where the NLP method is one of the combined methods. To overcome this problem we could consider the performance of post-processing rules and Natural Language Processing rules independently (e.g., there are 29 PP rules for */genetics*). Rules that achieve high precision individually may be used as such.

5. Conclusion and Future Work

We have presented three methods to provide MeSH main heading/subheading pair recommendations for indexing the biomedical literature. These methods were applied to a genetics-related corpus to provide recommendations including the subheadings genetics, immunology and metabolism. Although performance may vary considerably depending on the subheading and the method used, the results are encouraging and seem to indicate that some useful pair recommendations could be used in indexing in the near future.

In future work, we plan to expand the set of PP and NLP rules to cover all 83 MeSH subheadings. Investigating statistical methods to provide pair recommendations will be considered. For example, in the specific field of genetics, links between MEDLINE and other Entrez databases such as Gene could be exploited. Based on the results from the combination of methods, more elaborate combination techniques will be studied in order to lessen decrease in recall. Finer combinations at the rule level may be considered as well as other factors such as the influence of the specific genetics corpus we used. Finally, a qualitative evaluation of this work will be sought from the indexers at NLM.

Acknowledgments

This research was supported in part by an appointment of A. Névéol to the Lister Hill Center Fellows Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education, and in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The authors would like to thank Halil Kilicoglu for his help in the use of SemRep/SemGen and James G. Mork for his help in the use of MTI (Medical Text Indexer) during the experiments.

References

1. AR. Aronson, O. Bodenreider, HF. Chang, SM. Humphrey, JG. Mork, SJ. Nelson, TC. Rindflesch and WJ Wilbur. "The NLM Indexing Initiative". *Proc AMIA Symp.* 17-21 (2000).
2. AR. Aronson, JG. Mork, GW. Gay, SM. Humphrey, WJ. Rogers. "The NLM Indexing Initiative's Medical Text Indexer". *Proc. Medinfo*. 268-72 (2004).
3. P. Ruch, R. Baud, A. Geissbühler. "Learning-free Text Categorization". *LNAI.* **2780**, 199-204 (2003).
4. L. Cai and T. Hofmann. "Hierarchical document categorization with support vector machines". *Proc. CIKM*. 396-402 (2004).
5. W. Kim, AR. Aronson and WJ. Wilbur. "Automatic MeSH term assignment and quality assessment". *Proc AMIA Symp.*319-23 (2001).
6. http://www.nlm.nih.gov/mesh/indman/chapter_19.html (visited on 05/23/06)
7. M. Ruiz and P. Srinivasan. "Hierarchical neural networks for text categorization". *Proc. SIGIR*. 281–282 (1999).
8. C. Gay. "A MEDLINE Indexing Experiment Using Terms Suggested by MTI" *National Library of Medicine* Internal Report (2002).
9. P. Langlais, G. Lapalme and M. Loranger. "Transtype: Development-Evaluation Cycles to Boost Translator's Productivity" *Machine Translation* **15**, 77-98 (2002).
10. A. Névéol, A. Rogozan, SJ. Darmoni. "Automatic indexing of online health resources for a French quality controlled gateway." *Inf. Process. Manage.* **42**, 695-709 (2006).
11. SM. Humphrey. "Indexing biomedical documents: from thesaural to knowledge-based retrieval systems" *Artif. Intel. Med*. **4**, 343-371 (1992)
12. TC. Rindflesh and M. Fiszman. "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text"*J Biomed Inform.* **36(6)**, 462-77 (2003)
13. O. Bodenreider, SJ. Nelson, WT. Hole, and HF. Chang. "Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies." *Proc AMIA Symp*. 815-9 (1998).
14. TC. Rindflesch, B. Libbus, D. Hristovski, AR. Aronson, H. Kilicoglu. "Semantic relation asserting the etiology of genetic diseases." *Proc AMIA Symp*. 8-1 (2003).
15. SM. Humphrey. "Automatic indexing of documents from journal descriptors: a preliminary investigation." *J Am Soc Inf Sci Technol.* **50(8)**, 661-674 (1999).