Fine-Grained Indexing of the Biomedical Literature: MeSH Subheading Attachment for a MEDLINE Indexing Tool

Aurélie Névéol, Sonya E. Shooshan, James G. Mork, Alan R. Aronson
U.S. National Library of Medicine, Bethesda, Maryland
National Institutes of Health, Lister Hill Center for Biomedical Communications
{neveola,sonya,mork,alan}@nlm.nih.gov

ABSTRACT

***Objective**: This paper reports on the latest results of an Indexing Initiative effort addressing the automatic attachment of subheadings to MeSH main headings recommended by the NLM's Medical Text Indexer. **Material and Methods**: Several linguistic and statistical approaches are used to retrieve and attach the subheadings. Continuing collaboration with NLM indexers also provided insight on how automatic methods can better enhance indexing practice. **Results**: The methods were evaluated on corpus of 50,000 MEDLINE citations. For main heading/subheading pair recommendations, the best precision is obtained with a post-processing rule method (58%) while the best recall is obtained by pooling all methods (64%). For stand-alone subheading recommendations, the best performance is obtained with the PubMed Related Citations algorithm. **Conclusion**: Significant progress has been made in terms of subheading coverage. After further evaluation, some of this work may be integrated in the MEDLINE indexing workflow.*

INTRODUCTION

Research Context

In light of the significant increase in the indexing load anticipated by the U.S. National Library of Medicine (NLM) in order to keep the MEDLINE® database up to date[1] in the next decade, tools must be developed in order to assist indexers in their daily task. In this paper, we discuss on-going research at NLM in the framework of the Indexing Initiative [1-2], a project which addresses the specific issue of automatic indexing methods. Current efforts aim at integrating a subheading attachment feature to the Medical Text Indexer (MTI) [2], a tool that provides NLM indexers with Medical Subject Headings (MeSH®) indexing recommendations.

MEDLINE indexing

Articles from prominent journals in the biomedical domain selected for inclusion in MEDLINE are indexed with MeSH descriptors to facilitate later retrieval through search engines such as Entrez used by PubMed. The MeSH thesaurus contains about 24,000 main headings representing medical concepts (e.g. *Alzheimer Disease*, *Kidney* or *Hypoglycemic Agents*) and 83 subheadings (e.g. *genetics*, *surgery* or *therapeutic use*) that may be coordinated to the main headings in order to refer to a more specific aspect of a concept. For example, an article specifically discussing anti-diabetic medication should be indexed with the descriptors *Diabetes Mellitus/drug therapy* and *Hypoglycemic Agents/therapeutic use* while only the main heading *Diabetes Mellitus* will be appropriate for a more general article addressing several issues related to diabetes. It must be stressed that for each main heading MeSH defines a set of "allowable qualifiers" that may be attached to it. As a result, certain pairs may not be formed, such as *Hypoglycemic Agents/genetics* as *genetics* is not an allowable qualifier for *Hypoglycemic Agents.*

Since 2002, the Medical Text Indexer has been producing automatic indexing recommendations for articles to be entered in MEDLINE. The recommendations are displayed in the Data Creation and Maintenance System (DCMS) interface so that indexers may view and/or use them when working on a MEDLINE record.

Tools designed to assist humans in a task that they would otherwise perform independently must be oriented towards reaching two distinct goals: 1/ high performance for the task at hand and 2/ adequate conveyance of results to the users of the tools. In fact, a recent evaluation of an indexing help system sought to establish that using the system did not impede the curation process [3]. Our objective is to provide NLM indexers with comprehensive MeSH recommendations that meet both criteria. We report on a recent evaluation of several subheading attachment methods and describe our interaction with the indexers while developing these methods.

MATERIAL AND METHODS

Test Corpus

The indexing methods described here were evaluated on a test corpus composed of 50,000 citations randomly selected from the MEDLINE 2006 baseline. The 2006 version of MeSH was used. To

[1] 1 million journal articles per year by 2015 *vs.* 635,469 in 2006

avoid bias in the evaluation of the methods, separate MEDLINE corpora were used for training.

"Jigsaw puzzle" methods

The "jigsaw puzzle" methods work by extracting MeSH main headings and subheadings relevant to an article separately, and then trying to attach the subheadings to appropriate main headings. In practice, main headings are paired with subheadings that MeSH defines as "allowable". For example, if the MeSH main headings *Alzheimer Disease* and *Kidney* are retrieved and the subheading *genetics* is also retrieved (see below for details on how terms may be retrieved), the pair *Alzheimer Disease /genetics* will be formed because genetics is an allowable qualifier for *Alzheimer Disease*. However, *genetics* is not allowable for *Kidney;* therefore, the two terms cannot be paired.

A dictionary method (DIC) introduced in [4] uses MTI-retrieved main headings. Subheadings are then extracted based on the presence of certain dictionary words or expressions in the title or abstract of the article. For example, the subheading *genetics* will be retrieved if words such as "gene", "genes", "genetic", "genetical", "DNA", "RNA", etc. are found. At first, the dictionary was composed of words that could be related to the subheadings based on the indexing manual[2] description of subheading use. It was then expanded thanks to a statistical fingerprinting of the subheadings over the entire MEDLINE collection, using a technique similar to [5]. For each subheading, the citations that used the subheading at least once in the indexing were collected to form a subheading characteristic corpus (SH). After stop words were removed, a score *S* was computed for each word *w* in the corpus SH_i as follows:

$$S_{w,SHi} = \frac{occ(w)_{SH_i}}{occ(w)_{MEDLINE}} * \frac{occ(w)_{SH_i}}{\sum\limits_{\forall w \in SH_i} occ(w)_{SH_i}} \qquad (1)$$

The score of a word is based on its frequency (number of occurrences) in the subheading corpus *vs.* the MEDLINE collection and its frequency in the subheading corpus *vs.* the frequency of all content words in this corpus.

The top 50 words according to this ranking were considered for addition in the dictionary. They were added if they improved the performance of the dictionary method on two training corpora[3]. Bigram statistics obtained from the subheading corpora were also used. As of March 2007, dictionary entries were augmented in this way for the 26 most frequent subheadings.

Alternatively, an MTI method works by inferring relevant subheadings based on the main headings themselves. For example, if a G13 category main heading (*Genetic Phenomena*) were retrieved by MTI, we infer that the subheading *genetics* might be relevant for indexing the article. It would then be attached to the main headings also retrieved by MTI, when applicable. There is at least one such rule for 82 of the subheadings. This study is the first to evaluate the use of subheadings retrieved with these rules.

Rule-based methods

Post-processing (PP) rules infer pair recommendations from a pre-existing set of indexing terms - in our case, MTI main heading recommendations. A sample rule is: "**If** the main heading *Mutation* and a *<DISEASE>* term[4] appear in the indexing recommendations, **then** the pair *<DISEASE>/genetics* should also be used." These rules were developed in the same spirit as the subheadings inferred in the MTI method above – in fact, *Mutation* is a G13 category term. However, they are much more specific as they define which type of main heading the subheading should be attached to. Furthermore, before a new rule is added to the set, it is evaluated on the training corpora used for the dictionary method.

Natural Language Processing (NLP) rules use cues from the title or abstract of an article to infer pair recommendations. More specifically, interactions between medical entities are retrieved from the text in the form of Unified Medical Language System® (UMLS®) triplets using SemRep [6]. UMLS triplets are 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. The knowledge expressed in these triplets is then translated into MeSH pairs using rules and a restrict-to-MeSH algorithm [7]. A sample rule would be that the triplet (enzyme AFFECTS disease or syndrome) translates into MeSH by attaching the subheading *enzymology* to the corresponding disease term[4]. However, some rules are more complicated and must be tailored to several term categories. For example, the triplet (therapeutic or preventive procedure TREATS disease or syndrome) translates into MeSH by attaching the subheading *surgery* if the procedure is surgical (term category E04) or the subheading *radiotherapy* if the procedure involves radiation (term category E02.815), etc. The PP and NLP rules are described in more detail in [4].

[2] http://www.nlm.nih.gov/mesh/indman/chapter_19.html (3/12/07)

[3] A corpus composed of ~17,000 citations randomly extracted from MEDLINE 2004 and a corpus of 100,000 citations randomly extracted from MEDLINE 2006 (distinct from our test corpus)

[4] i.e. a C or F03 category term

Statistical method

The PubMed Related Citations (PRC) method that we used was first introduced in [8]. It uses a k-nearest neighbors approach to find citations in the MEDLINE database that are similar to the new article to index. MeSH pair recommendations are then inferred from the existing indexing of the ten nearest neighbors.

Indexers' feedback

In order to obtain feedback from NLM indexers, pair recommendations were produced using the methods described above for three journal issues[5] to be entered in MEDLINE. The journals were chosen to fit the early focus of the project on the genetics domain. The recommendations were shown (on paper, at this stage) to the indexers only if they were provided by at least two methods. At first, the recommendations were presented using the full name of the subheadings (e.g. *therapeutic use* is the full name for the subheading abbreviated as *TU*) to mimic MEDLINE records. After first viewing the recommendations, indexers remarked that 1/ the full name for subheadings overcrowded the results, 2/ a list of subheadings that generally applied to a citation would be desirable and 3/ recurring mistakes could be avoided by filtering the results using a list of "stop main headings" for which they did not wish to see any subheading recommendations.

As a follow-up to these observations, we reprocessed the same articles using abbreviations for subheadings and applying filtering with the stop list of 92 main headings. In addition, we also produced a list of stand-alone subheadings statistically relevant to the citations using the PRC method. The indexers noted a significant improvement in the results. The next step of our work will consist in involving more indexers to collect feedback on a larger and more varied corpus.

RESULTS

Table 1 presents the performance obtained for each indexing method on the test corpus. In the second column ("scope") we indicate the number of subheadings for which the method is currently able to provide recommendations. The third column shows the precision (P), which corresponds to the number of pairs recommended by the method that were also selected by NLM indexers over the total number of pairs recommended. The fourth column shows the recall (R), which corresponds to the number of pairs recommended by the method[6] that were also selected by NLM indexers over the total number of pairs that were selected by NLM indexers. Finally, the last column shows the F-measure (F), which combines precision and recall with equal weight. The seventh line of the table shows the performance of the pair recommendations obtained from at least two methods after filtering was applied. Finally, the last line of the table shows the pool performance of the pair recommendations obtained from at least one method. The best performance according to each metric is bolded.

Indexing method	**Scope out of 83**	**P**	**R**	**F**
Dictionary	83	26	31	28
MTI	82	24	13	17
Post-Processing rules	19	**58**	5	9
NLP rules	20	38	2	4
PubMed Related Citations	83	35	54	**42**
At least 2 methods + filtering	83	44	29	35
Pool	83	26	**64**	37

Table 1 – Performance of pair recommendations

Table 2 presents the performance of stand-alone subheading recommendations with a selection of the methods – those expected to yield the best recall.

Indexing method	**P**	**R**	**F**
MTI	18	15	8
PubMed Related Citations	24	86	38
Dictionary	31	46	36

Table 2 – Performance of stand-alone subheading recommendations

	Subheading Counts
Allowable Subheadings (MTI)	59.40
Allowable Subheadings (MEDLINE)	54.33
Subheadings used by NLM indexers	3.51
Subheadings recommended by MTI	1.40 (0.53 used)
Subheadings recommended by DIC	5.61 (1.61 used)
Subheadings recommended by PRC	12.46 (3.01 used)

Table 3 – Average number of allowable, recommended and used subheadings per citation in the test corpus

To illustrate the impact of stand-alone subheading recommendations, Table 3 shows the average number of subheadings recommended per citation by the methods as well as the average number of subheadings that are applicable to MTI-retrieved main headings or MEDLINE reference main headings. Table 4 presents the indexing recommendations obtained with all the methods for a sample corpus citation (DIC refers to the dictionary

[5] Hum Hered. 2006;62(2), Genet Test. 2006 Fall;10(3) and Vet Immunol Immunopathol. 2006 Nov 15;114(1-2)

[6] As explained in [2], we expect the recommendations produced by these methods to be used interactively by the indexers. Therefore, only the pairs that involve a main heading selected by the indexers are considered when computing the metrics.

method, MTI to the MTI method, NLP to the Natural Language Processing rules, PP to the Post-Processing rules and PRC to the PubMed Related Citations). In this case, no recommendations were removed after the filtering step.

PMID - 16384987

Influence of treatment parameters on selectivity of verteporfin therapy.

PURPOSE: To improve selectivity of verteporfin therapy (PDT) in neovascular age-related macular degeneration (AMD) using modified treatment parameters. METHODS: Nineteen consecutive patients with predominantly classic choroidal neovascularization (CNV) in AMD were treated with 6 mg/m2 verteporfin given as bolus infusion. Patients received PDT with a fluence of either 25 or 50 J/cm2. Choroidal perfusion changes were evaluated by indocyanine green angiography (ICGA) at baseline, day 1, week 1, week 4, and month 3. Secondary outcomes were CNV closure rate and therapy-induced leakage documented by fluorescein angiography (FA). The safety of the treatment was assessed with ETDRS visual acuity. RESULTS: Complete CNV closure was achieved in all patients at day 1. Choroidal hypoperfusion was minimal in eyes treated with a reduced fluence of 25 J/cm2. Most patients treated with 50 J/cm2 showed significant choriocapillary nonperfusion at week 1, lasting as long as 3 months. A transient PDT-induced increase in leakage area in FA at day 1 was found to be more extensive in the 50-J/cm2 group. CONCLUSIONS: Bolus administration of verteporfin combined with a reduced light dose achieved improved selectivity of photodynamic effects, avoiding collateral alteration of the physiologic choroid while obtaining complete CNV closure. An increased selectivity with decreased effect on the surrounding choroid should be of advantage in verteporfin monotherapy as well as in combination strategies.

MEDLINE reference indexing	**Pair recommendations**	**Methods**
Capillary Permeability	Choroid/blood supply	DIC|PRC
Choroid/blood supply	Choroidal Neovascularization/complications	DIC|PRC
Choroidal Neovascularization/*drug therapy/etiology	Choroidal Neovascularization/drug therapy	DIC|MTI|PP|PRC
Fluorescein Angiography	Choroidal Neovascularization/therapy	DIC|MTI
Humans	Macular Degeneration/complications	DIC|PRC
Indocyanine Green/diagnostic use	Macular Degeneration/drug therapy	DIC|MTI|NLP|PP|PRC
Macular Degeneration/complications/*drug therapy	Macular Degeneration/therapy	DIC|MTI
*Photochemotherapy	Photochemotherapy/adverse effects	DIC|PRC
Photosensitizing Agents/*therapeutic use	Photosensitizing Agents/therapeutic use	DIC|PRC
Porphyrins/*therapeutic use	Visual Acuity/physiology	DIC|PRC
Tomography, Optical Coherence		
Treatment Outcome	*Additional recommendations not shown:*	
Visual Acuity	*16 DIC-only recommendations (none correct)*	
	11 PRC-only recommendations (including 3 additional correct)	
	Stand-alone subheading recommendations (PRC)	
	AE BS CO DI DU DT ET MT PA TU PH	

Table 4 – Pair recommendations obtained for a sample citation in the Test Corpus (correct recommendations in the right column are underlined)

DISCUSSION

Methods performance

The sample citation shown in Table 4 is quite representative of the results obtained over the test corpus. The DIC and PRC methods usually yield numerous recommendations, while MTI is more moderate and PP or NLP are sometimes sparse. The subheadings (here, *diagnostic use* and *etiology*) used by the indexers that do not appear in the pair recommendations are included in the stand-alone selection.

The performance obtained for the various methods is consistent with our aim in developing them: the highest precision is obtained with the rule-based methods (NLP and PP) while the best recall is obtained with the statistical method (PRC). The other two methods (DIC and MTI) have intermediate precision and recall. By applying the "at least 2 methods and filtering" rule as shown in Table 4, at least one pair recommendation was made for 76% of the citations in the test corpus. 84% of the recommendations filtered out using the stop-list are incorrect.

The selection of stand-alone subheadings to apply to a particular citation is achieved with 86% recall. Although precision is only 18%, it reduces the list of applicable subheadings for a citation by about 75% (from 54 down to 12), which the indexers find useful as it may save time in deciding which subheading to use.

The NLP results for *genetics* (/GE), *immunology* (/IM) and *metabolism* (/ME) vary from what we obtained on the genetics-related corpus [4]. There seems to be a significant recall increase for /IM and /ME and precision drop for /GE. It is possible that outside the genetics domain, the pairs predicted are no longer necessarily addressing substantively discussed concepts. However, recent updates in SemRep focusing on gene-disease interactions may also have had an impact. For these same subheadings, the DIC results show a slight drop in precision and significant recall increase due to the additions to the dictionary described above.

In general, we observe a significant variability across methods for a given subheading and across subheadings for a given method. For this reason, combining the different approaches is desirable.

Usability

The precision obtained by combining the methods (44%) is comparable to the inter-indexer agreement reported in [9]. Indexers say that they value the automatic recommendations if they can help save typing time or if they can trigger the idea of using a correct indexing descriptor. In this respect, recommendations that are close (even though not strictly identical) to what an indexer would really select are also useful. However, the down side of almost-correct recommendations is that they might confuse junior indexers who may not have sufficient training to distinguish between almost-correct and correct recommendations

Project progress and future work

Compared to the work reported in [4], we have significantly extended the scope of the project by covering the 26 most frequent subheadings[7] more thoroughly, instead of just three genetics-related subheadings. Moreover, statistical methods have been investigated to complement the dictionary and rule-based methods. In future work, we intend to resume this effort to address all 83 subheadings with all of our methods. In particular, on-going work addresses the automatic extension of the PP rule set using Inductive Logic Programming. We also believe that significant performance improvement may be achieved by optimizing the combination of the methods.

CONCLUSION

In this paper, we presented several methods addressing the automatic attachment of subheadings to MeSH main headings retrieved by MTI. In the past few months, significant progress has been made in this project. NLM indexers' participation helped improve both the performance and subheading coverage. Further indexer validation may lead to integrating a subheading attachment feature in DCMS.

Acknowledgements

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine and by an appointment of A. Névéol to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education. The authors would like to thank James Marcetich and Joe Thomas of NLM's Indexing Section for their interest and feedback on this work.

References

1. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC and Wilbur WJ. The NLM Indexing Initiative. *Proc AMIA Symp* 2000:17-21.
2. Aronson AR, Mork JG, Gay CW, Humphrey SM and Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Proc. Medinfo* 2004: 268-72.
3. Karamanis N, Lewin I, Seal R, Drysdale R and Briscoe T. Integrating Natural Language Processing with FlyBase Curation. *Proc. PSB* 2007:245-256.
4. Névéol A, Shooshan SE, Humphrey SM, Rindflesch TC and Aronson AR. Multiple approaches to fine-grained indexing of the biomedical literature. *Proc. PSB* 2007(12):292-303.
5. Liu Y, Brandon M, Navathe S, Dingledine R and Ciliax BJ. Text mining functional keywords associated with genes. *Proc. Medinfo* 2004:292-296.
6. Rindflesh TC and Fiszman M. 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).
7. Bodenreider O, Nelson SJ, Hole WT and Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proc AMIA Symp* 1998:815-9.
8. Kim W, Aronson AR and Wilbur WJ. Automatic MeSH term assignment and quality assessment. *Proc AMIA Symp* 2001:319-23.
9. Funk ME, Reid CA and McGoogan LS. Indexing consistency in MEDLINE. Bull. Med. Libr. Assoc. 1983:2 (71): 176-183.

[7] Based on MEDLINE as of December 2006