Resolving Hierarchical Ambiguity in Indexing Recommendations

James G. Mork, MSc, Dina Demner-Fushman, MD, PhD,
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, DHHS, Bethesda, MD

Introduction

Finding the main points of a given text and mapping them to MeSH® facilitates MEDLINE® indexing, assignment of key terms by authors, cataloguing, etc. One of the problems in assigning terms using the MeSH hierarchies is to decide on the level of specificity that is required for each document. The decision is based on the context of the document, as well as the nature of the indexing task, whether a more general term, a more specific term, or both levels of terms should be assigned. For the task of assisting the MEDLINE indexer, the decision could be based on the indexing rule of assigning the most specific MeSH term appropriate to the article, except in cases where the indexer's expert judgement is to apply an allowed exception. In this work, we rely on a corpus-based approach to learn how to select an appropriate specificity of the term when our tool, The NLM Medical Text Indexer (MTI)[1] suggests both a general and a specific MeSH term from the same MeSH tree.

Methods

We used 763,227 citations that were indexed in 2015 and also had MTI recommendations (henceforth referred to as *Corpus*). MTI recommended 6,465,133 MeSH terms and the human indexers assigned 8,454,900 MeSH terms to the *Corpus*. We focused on citations for which MTI recommended both a more general MeSH term and a more specific MeSH term from the same MeSH tree – for example, *Vaccines* (D20.215.894) is more general than *Anthrax Vaccines* (D20.215.894.135.063). When MTI recommends both of these terms for a given citation, we call it *Hierarchical Ambiguity*. In the *Corpus*, there are 985,738 (15.25% of MTI total) *Hierarchical Ambiguity* pairings from MTI and 726,861 (8.60% of human total) from the human indexing, or almost half the rate of occurrence we see from MTI.

Results

The easy solution would have been to just not recommend the more general term per the rule of indexing the most specific term. The problem is that the exceptions noted earlier still constitute 8.60% of the human indexing. In our *Corpus*, we would lose 314,869 (31.94%) of the correct general term recommendations by always recommending only the most specific terms. Table 1 shows the distribution of indexer assignments where MTI assigned both terms.

Table 1. Human indexing as judgements for MTI recommended Hierarchical Ambiguity.

	Gen Wrong/Spec Right	Gen & Spec Wrong	Gen Right/Spec Wrong	Gen & Spec Right	Overall
Term Count	426,242	244,627	**186,503**	**128,366**	985,738
% of Suggestions	43.24%	24.82%	**18.92%**	**13.02%**	15.25%
			314,869 (31.94%)		

Ignoring MTI recommended pairs in which both the general and specific recommendations were wrong in the *Corpus*, we identified 6,399 *Hierarchical Ambiguity* pairs in the remainder of the *Corpus* where the general term was wrong on average 82.88% of the time accounting for 52.40% (223,361) of the *Gen Wrong/Spec Right* recommendations in Table 1. A simple rule to always remove the general term recommendation when one of these 6,399 *Hierarchical Ambiguity* pairs is identified in the MTI results, provided us with an improvement in Precision from 0.6292 to 0.6406 (**+1.81%**) while only dropping Recall from 0.6463 to 0.6419 (**-0.68%**) for our Test Collection.

Conclusion

Understanding where the problematic Hierarchical Ambiguities are in the MTI recommendations has provided us with a way of eliminating over 50% of the erroneous general term results with very little loss to Recall.

Acknowledgments

This work was supported by the intramural research program of the NIH, U. S. National Library of Medicine.

References

1. James G Mork, Antonio J Jimeno-Yepes, Alan R Aronson. The NLM Medical Text Indexer system for indexing biomedical literature. BioASQ Workshop, Valencia, Spain, September 27, 2013.