Evaluation of Biomedical Citation Selector

Max Savery, Melanie Huston, MS, James Mork, MSc, Olga Printseva, PhD, DSc, Alastair Rae, PhD, Susan Schmidt, MLS, Dina Demner-Fushman, MD, PhD
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction

Finding and accessing scientific literature in PubMed partially relies on indexing articles for MEDLINE® using Medical Subject Headings (MeSH®). Currently, more than 5,000 journals in MEDLINE are fully indexed, where the majority of articles published by these journals have MeSH indexing assigned. However, for approximately 130 selectively indexed journals, articles are only indexed if, after manual review, they are determined to be relevant to biomedicine and the life sciences. In recent years, the number of these articles has increased substantially; for example, in 2018 about 125,000 were added to PubMed. To reduce the number of articles that must be manually reviewed, we have developed the Biomedical Citation Selector (BmCS), a high recall machine learning system that identifies the articles that require indexing[1]. Here, we manually evaluate BmCS's automated predictions to better understand its performance in the indexing workflow.

Methods

For this evaluation, we manually reviewed 4,763 articles predicted by BmCS to be out-of-scope for indexing at a 99.5% level of confidence. We chose only articles from chemistry journals because BmCS has been shown to perform more poorly on these articles than it does on those from other disciplines[1]. In a subsequent test, we selected 1,248 articles of any discipline, predicted to be in-scope at a 99.5% level of confidence. To manually review the articles, we designed a Microsoft Access form. The title and abstract of each article was reviewed by two of ten different reviewers, and a final label of in-scope or out-of-scope was assigned after considering both reviewers' input.

Results

The human reviewers agreed with 4,612 out-of-scope predictions generated by BmCS and disagreed with 151. However, 147 of the misclassified articles were publication types that should always be classified as in-scope, such as comments or errata. Because BmCS will handle these with rules in the future, we removed them from the evaluation. Table 1 reflects this adjustment. The reviewers agreed with all 1,248 articles predicted to be in-scope. Accuracy for the in-scope class was computed to be 100% and for the out-of-scope class, 99.9%. Fleiss' kappa, used to assess interrater agreement, was .981.

Table 1. Agreement, disagreement and accuracy between human reviewers and BmCS

	Agreement	Disagreement	Accuracy
Out-of-scope	4,612	4	99.9%
In-scope	1,248	0	100%

Conclusion

In order for BmCS to effectively contribute to the indexing pipeline, it must infrequently label in-scope articles as out-of-scope. This is what we observed: After adjusting the predictions as discussed above, only 0.1% of the system's out-of-scope predictions were incorrect. These results suggest that the system will effectively reject out-of-scope articles from the indexing pipeline. In addition, the perfect in-scope predictions indicate that we will be able to automatically send articles labeled as in-scope with 99.5% confidence to the human indexers, with relatively few false positives. In general, these results demonstrate that BmCS will reduce the number of articles from selectively indexed journals that require manual review by at least the amount we expect, 54%[1]. Further development of BmCS will focus on exploring the effect of more permissive prediction thresholds, as well as integration into the indexing production environment.

References

1. Rae A, Savery M, Mork J, Demner-Fushman D. A high recall classifier for selecting articles for MEDLINE indexing. Submitted to AMIA, 2019.