High Recall Text Classification for Public Health Systematic Review

Paul McNamee[1] and **James Mayfield**[1] and **Samantha Y. Rowe**[2]

and **Alexander K. Rowe**[2] and **Hannah L. Jackson**[2] and **Megan Baker**[1]

[1] Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, USA
[2] U.S. Centers for Disease Control and Prevention, Atlanta, Georgia, USA
{mcnamee,mayfield}@jhu.edu, {say9,axr9,lhu2}@cdc.gov, megan.baker@jhuapl.edu

Abstract

Some information retrieval applications demand manageable levels of precision at high levels of recall. Examples include e-discovery, patent search, and systematic review. In this paper we present a real-world case study supporting a broad topic systematic review in the public health domain. We provide experimental results that demonstrate how retrieval performance on bibliographic citations can be materially improved. We attained an average precision of 0.57 and recall approaching 80% at a very reasonable screening depth. These results represent 18% and 23% relative gains over a baseline classifier. We also address pragmatic issues that arise when working on "noisy" real-world data, such as coping with citation records that often have empty fields.

Introduction

One approach for addressing the rising amount of published literature is reliance on systematic reviews. Systematic reviews are meta reviews that attempt to comprehensively survey extant literature and synthesize results to identify the best science from previously published studies. Their use is pervasive in the biomedical sciences and in public health policy, although the technique is applicable to any field. Creation of a systematic review is a substantial undertaking that requires a team of experts to follow a rigorous process that includes reviewing many thousands of potentially relevant studies for inclusion, followed by abstraction and analysis of those studies.

Crucial to the success of a systematic review is the thorough identification of studies that meet specified inclusion criteria. According to non-profit Cochrane (Lefebvre, Manheimer, and Glanville 2011):

> Systematic reviews of interventions require a thorough, objective and reproducible search of a range of sources to identify as many relevant studies as possible (within resource limits). This is a major factor in distinguishing systematic reviews from traditional narrative reviews and helps to minimize bias and therefore assist in achieving reliable estimates of effects.

Thus it is vital to identify a high fraction of pertinent documents from an examination of diverse sources.

Despite the increasing availability of text analysis tools for ranking, topical classification, and relation extraction, the *screening* (or *triage*) phase of systematic reviews is almost universally conducted by applying considerable human effort to craft Boolean queries and then laboriously evaluating search results that are ***unsorted*** by predicted relevance. This research explores supervised text classification with staggered annotation and iterative re-ranking to maximize human screening effort.

In the rest of this paper we briefly summarize related work and then discuss an ongoing systematic review, our data and experiments, and finally our conclusions.

Related Work

Prior work in automating screening has investigated several machine learning approaches. Cohen *et al.* published one of the earlier studies using voting perceptrons to rank citations (2006). Naïve Bayes, used for text categorization for over 50 years, has been used to classify articles in systematic reviews (Matwin et al. 2010; Frunza et al. 2011). Bekhuis and Demner-Fushman compared different classifiers for semi-automated screening (2012); they found that a support vector machine (SVM) approach attained the highest recall.

Other studies have also explored SVMs for systematic reviews: Mo *et al.* used SVMs, advocating addition of topic-model features (2015); Wallace *et al.* used ensembles of SVMs on three real-world datasets to reduce screening effort by between 40 to 50% (2010b); and, Yu *et al.* used SVMs to search PubMed for genetic associations for various diseases (2008).

A different recall-focused domain is legal e-discovery, the focus of the NIST Text REtrieval Conference (TREC) Legal Track (Oard et al. 2008). TREC also initiated a track exploring very high recall in 2015 (Roegiest et al. 2015).

Health Care Provider Performance Review

In low- and middle-income countries (LMICs) there are millions of deaths every year that could be prevented by an appropriate health intervention (*e.g.,* administration of medicines and vaccines). Improving the performance of health care providers (HCPs) such as health workers in hospitals, clinics, pharmacies, and communities is a key component for increasing coverage of health interventions. Many

Table 1: Sample Phase 1 citations.

Class	Docid	Article Title	Journal	Year	Comments
Pos	7901689	Field tests for rational drug use in twelve developing countries	Lancet	1993	No country is explicit
Pos	7219486	Reduction of mortality in rural Haiti through a primary-health-care program	N Engl J Med	1981	
Neg	16052396	Thromboembolism during hormone therapy in Japanese women	Seminars in Thrombosis and Hemostasis	2005	Japan is not a LMIC
Neg	6140346	Third World smoking–the new slave trade	Lancet	1984	No HCP performance

strategies exist to try to improve HCP performance, including training, supervision, and incentives. However, for health programs in the developing world to be effective, decision makers must know which strategies are most effective for a given context, and how much they cost. The *Health Care Provider Performance Review* (HCPPR) project[1] seeks to generate and disseminate evidence-based guidance for improving HCP performance in LMICs, which will ultimately lead to improved health for individuals and populations.

An earlier (*Phase I*) review examined studies from the 1970s to the mid-2000s. The current effort (*Phase II*) aims to identify the most salient studies from 2007 to 2015 to understand newer practices such as paying for performance and innovations in communication technologies.

The HCPPR inclusion criteria require a study to:

- take place in a low- or middle-income country;
- involve a quantitative evaluation of a strategy (broadly defined) to improve health worker performance; and
- use a robust evaluation design (*e.g.,* a randomized controlled trial).

A study involving any type of HCP, on any health topic, in any language, published or not, is eligible.

Data

Phase I

As a result of the *Phase I* review approximately 30k judgments were available from the literature prior to 2008. This data is extremely useful, both for building trained classifiers and for evaluating performance. Table 1 gives a few samples of positive (include) and negative (reject) citations from the *Phase I* data. We split the data chronologically into three partitions: *train* (70%), *dev* (15%), and *test* (15%). The *Dev* partition was occasionally used to inform parameter settings and assess the benefit of various features; the *Test* partition was used solely for evaluation. The data are skewed 30:1 in favor of rejection (see Table 2).

Phase II

The *Phase II* data comes from a heterogenous compilation of databases; we list the largest sources in Table 3. 34% of 14.5 million original citations are duplicates, and just over 2 million (21%) of the remaining 9.5 million do not have abstracts. Without abstracts classification is based on relatively short titles and limited metadata (*e.g.,* names of journals). We emphasize that at the start of our current project, no labels were available on the *Phase II* citations.

Table 2: *Phase I* data, partitioned for experiments.

	Train	Dev	Test	Total
Include	700	150	140	990
Reject	21,000	4,700	4,674	30,374
Total	21,700	4,850	4,814	31,364

Table 3: Principal *Phase II* data sources.

Citations	Source	URL
11,545,654	PubMed	www.ncbi.nlm.nih.gov
1,868,161	CINAHL	health.ebsco.com
518,543	EconLit	www.ebscohost.com
419,256	Cochrane	www.cochranelibrary.com
78,218	BLDS	blds.ids.cac.uk
35,605	ICTRP	www.who.int/ictrp/en

Experiments

We leveraged labeled data from the *Phase I* study to train a linear kernel support vector machine. We used the SVM^{light} tool for all of our experiments (Joachims 1999). SVMs are known for good performance in text classification, are resistant to overfitting, and can accept a large numbers of features. We use distance from the learned hyperplane to induce a ranking over the citations. The classes of features used are listed in Table 4. The majority of features are term occurrences weighted using TF/IDF.[2]

Citation Fields

On text classification tasks documents with greater amounts of text usually prove easier to classify, and we found that to be the case here. In Table 5 we report average precision using bags of words from: (a) title alone; (b) title+abstract; (c) title, abstract, and keywords (TAK); and (d), separate feature representations for bags of words from the title alone, and from TAK. The last method had the highest performance on our development set, so we chose this as our baseline model.

[1] www.hcpperformancereview.org

[2] TF/IDF weights combine term repetition (TF) and the discriminating power of a word (IDF). See (Salton and Buckley 1988).

Table 4: Features used for training classifiers.

Name	*Description*
title-bow	Bag of words from the title field.
tak-bow	Bag of words from title, abstract, and keyword fields.
lmic	Title (or TAK) mentions a low- or middle-income country, a demonym, or a phrase such as ‘third world’ or ‘developing country’.
hcp	Title (or TAK) mentions a health worker (*e.g.,* doctor, midwife, pharmacist, dentist, paramedic, *etc.*).
interv	Title (or TAK) references training, auditing, or another intervention to effect HCP performance.
xprod-title	Combined presence of *lmic*, *hcp*, and *intervention*. One of eight possibilities.
xprod-abs	Like *xprod-title* except over title and abstract.

Table 5: Average precision using bags of words composed from various fields.

Features	*Dev*	*Test*
Title alone	0.377	0.269
Title / Abstract	0.449 (+19%)	0.427 (+59%)
Title / Abs / KW	0.488 (+29%)	0.463 (+72%)
Title; Title / Abs / KW	0.524 (+39%)	0.484 (+80%)

Table 6: Gains in average precision.

Features	*Dev*	*Test*
Baseline	0.524	0.484
+*lmic*, +*hcp*, +*interv*	0.535 (+2.1%)	0.491 (+1.4%)
+*xprod-abs*	0.553 (+5.6%)	0.518 (+7.1%)
+*extra data*	0.614 (+17%)	0.570 (+18%)

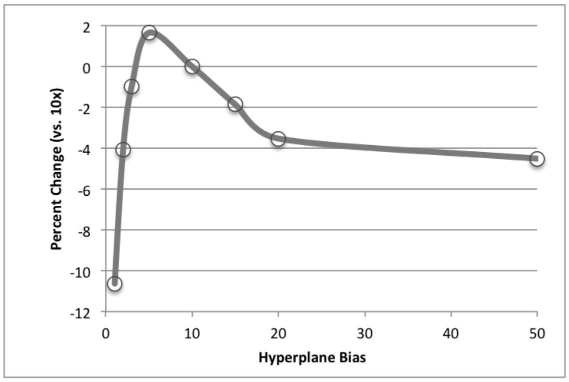

Figure 1: Percent change in average precision with a biased hyperplane compared to the standard setting. The out-of-the-box value ($j = 1$) experiences a 10% degradation.

Faceted Features

For the *Phase I* study human screeners created a complex, several hundred term Boolean query to identify candidate documents. The query corresponded to three facets of the inclusion criteria: that the study take place in a LMIC; that it be focused on health worker performance; and that it involve an intervention to improve health worker performance (*e.g.,* training, supervision). We assembled lists of suitable terms for each of these facets, and added three binary features corresponding to the presence of any such term in the citation record (see Table 4). For example, ‘Training Turkish Nurses to...’ would receive the +*interv*, +*lmic*, and +*hcp* features. But ‘Working Conditions of Turkish Nurses’ would not receive the +*interv* feature.

Additionally, since linear kernel SVMs do not automatically induce compound features, we generated “cross-product” features reflecting the $2^3 = 8$ possible combinations of these three atomic features from the title only, and, 8 similar features derived from both title and abstract.

The performance using each condition is given in Table 6. The cross-product features, which mirror human reasoning, are indeed helpful.

Phase II Annotations

Because some annotator time was available to produce judgments over novel *Phase II* data, we produced a ranking of the data and obtained judgments for citations at stratified levels in the ranking. This both helped us assess the performance of the system, and also gave valuable annotations to use as additional training data. We obtained 3,113 additional judgments (1,800 include; 1,313 reject), thereby nearly tripling the number of available positive examples. We then added all of these examples to the training data. The results listed on the last row of Table 6 show a sizeable improvement.

Biased Hyperplane

In our experiments we used SVM^{light} with default settings, except for adding the switch ‘-j 10’ based on limited testing on the *Dev* partition. This parameter biases the hyperplane away from the positive support vectors and towards the negative ones.[3] For this application, where the cost of missing a relevant item is high (*i.e.,* we need high recall[4]), this seems well motivated. In a post-hoc parameter sweep on the *Test* partition we found that a slightly lower value would have been optimal (see Figure 1).

Estimating Recall

To assist in planning the abstraction and analysis phases of the *Phase II* study, we needed to estimate precision at high recall levels in advance of having any judgments. Evaluating classifier performance can be done by randomly selecting a set of held out data, alternatively called a *certification*

[3] Wallace *et al.* (2010a) do something similar with the LibSVM package but found performance was not sensitive to the changes.

[4] In the literature *recall* is also known as *sensitivity*.

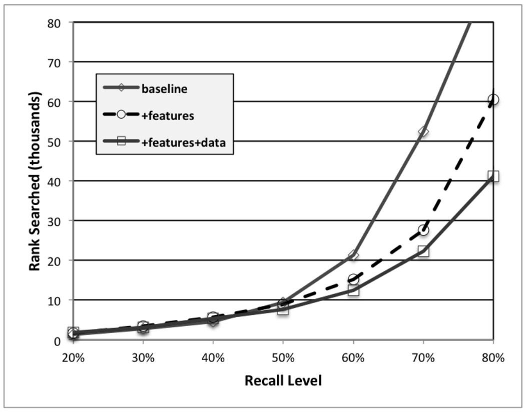

Figure 2: Estimated recall attained as a function of search depth (vertical axis). Three analyses were performed: (a) baseline; (b) after feature engineering; (c) and with additional training data. If 40k documents are screened, recall rises from 65% to 80%. Lower curves are better.

test set or *control set* (Bagdouri et al. 2013). We accordingly conducted experiments using 80% of our *Phase I* data for training, and reserving 20% for evaluation. We then mixed this 20% sample in with over 7 million deduplicated *Phase II* citations for which no labels are available, and ranked the entire set. Finally we identified the rank at which each known correct item was found. The results appear in Figure 2. By mapping recall levels to a number of documents in this way, a systematic review can accurately gauge how many documents must be pushed through to the abstraction and analysis phases to ensure a particular level of coverage.

Imperfect Data

Most approaches to supervised classification rely on test data being independent and identically distributed (iid) to the model's training data. However, the bibliographic data in our dataset is missing abstracts, the citation field with the largest amount of text, in about $\frac{1}{4}$ of cases. Ignoring the issue results in a 20% relative loss in average precision.

We trained separate models for the with and without abstract conditions, and ran both classifiers for each record. We investigated regression models to map abstract-less scores to their equivalent with-abstract score; however, the data were insufficiently correlated. By ablating abstracts for citations that had them we observed that a citations's rank can vary dramatically when using classifiers trained either with or without abstracts. Thus, we advocate that two ranked lists be passed on to the abstraction and analysis phases, one for entries with abstracts, the other for those without.

Conclusions

We described our use of automated classifiers to improve screening efficiency as part of an ongoing systematic review compiling evidence to inform public health policy decisions. In all systematic reviews high coverage is crucial for success. We demonstrated how to optimize performance at high recall levels when using linear SVMs for ranking. Specific techniques included feature engineering that exploits facets used in the human querying process; iterative retraining of models using sampled annotations; biasing hyperplane boundaries to reduce false negatives; distinguishing which field in a record each term is derived from; compensating for deficiencies in a linear kernel by creating cross-products of features; and processing documents with missing fields using separately trained classifiers.

References

Bagdouri, M.; Webber, W.; Lewis, D. D.; and Oard, D. W. 2013. Towards minimizing the annotation cost of certified text classification. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, 989–998. New York, NY, USA: ACM.

Bekhuis, T., and Demner-Fushman, D. 2012. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artif. Intell. Med.* 55(3):197–207.

Cohen, A. M.; Hersh, W. R.; Peterson, K.; and Yen, P.-Y. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13(2):206–219.

Frunza, O.; Inkpen, D.; Matwin, S.; Klement, W.; and O'Blenis, P. 2011. Exploiting the systematic review protocol for classification of medical abstracts. *Artif. Intell. Med.* 51(1):17–25.

Joachims, T. 1999. Making large-scale SVM learning practical. In Schölkopf, B.; Burges, C.; and Smola, A., eds., *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press. chapter 11, 169–184.

Lefebvre, C.; Manheimer, E.; and Glanville, J. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane Collaboration, 5.1.0 edition. chapter 6: Searching for Studies.

Matwin, S.; Kouznetsov, A.; Inkpen, D.; Frunza, O.; and OBlenis, P. 2010. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association* 17(4):446–453.

Mo, Y.; Kontonatsios, G.; and Ananiadou, S. 2015. Supporting systematic reviews using LDA-based document representations. *Systematic Reviews* 4(172):1–12.

Oard, D. W.; Hedin, B.; Tomlinson, S.; and Baron, J. R. 2008. Overview of the TREC 2008 Legal Track. In *Proceedings of the 17th Text REtrieval Conference (TREC)*.

Roegiest, A.; Cormack, G. V.; Grossman, M. R.; and Clarke, C. L. A. 2015. TREC 2015 Total Recall Track Overview. In *Proceedings of the Text REtrieval Conference*.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24(5):513–523.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2010a. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 173–182. ACM.

Wallace, B. C.; Trikalinos, T. A.; Lau, J.; Brodley, C.; and Schmid, C. H. 2010b. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics* 11(1):1.

Yu, W.; Clyne, M.; Dolan, S. M.; Yesupriya, A.; Wulf, A.; Liu, T.; Khoury, M. J.; and Gwinn, M. 2008. Gapscreener: an automatic tool for screening human genetic association literature in pubmed using the support vector machine technique. *BMC Bioinformatics* 9(1):205.