Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat

Christopher Norman,[1,2] Mariska Leeflang,[2] Pierre Zweigenbaum,[1] Aurélie Névéol[1]
[1] LIMSI, CNRS, Université Paris Saclay, F-91405 Orsay
[1] Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
{christopher.norman, pierre.zweigenbaum, aurelie.neveol}@limsi.fr
m.m.leeflang@amc.uva.nl

Abstract

Systematic reviews in e.g. empirical medicine address research questions by comprehensively examining the entire published literature. Conventionally, manual literature surveys decide inclusion in two steps, first based on abstracts and title, then by full text, yet current methods to automate the process make no distinction between gold data from these two stages. In this work we compare the impact different schemes for choosing positive and negative examples from the different screening stages have on the training of automated systems. We train a ranker using logistic regression and evaluate it on a new gold standard dataset for clinical NLP, and on an existing gold standard dataset for drug class efficacy. The classification and ranking achieves an average AUC of 0.803 and 0.768 when relying on gold standard decisions based on title and abstracts of articles, and an AUC of 0.625 and 0.839 when relying on gold standard decisions based on full text. Our results suggest that it makes little difference which screening stage the gold standard decisions are drawn from, and that the decisions need not be based on the full text. The results further suggest that common-off-the-shelf algorithms can reduce the amount of work required to retrieve relevant literature.

Keywords: Evidence Based Medicine, Information Storage and Retrieval, Review Literature as Topic

1. Introduction

Systematic reviews seek to systematically gather all published evidence addressing a given research question and analyze the aggregate results. Systematic reviews constitute some of the strongest forms of scientific evidence, are an integral part of evidence based medicine, and serve a key role in informing and guiding public and institutional decision-making (Wright et al., 2007).

One limiting factor of systematic reviews is that they tend to be prohibitively costly to produce.[1] The number of references needed to be manually screened in order to satisfy the requirement that virtually all relevant articles have been identified can number in the tens of thousands. Often only some dozens of these references are selected for the final meta-analysis, and the selection process may require months of work for several reviewers (O'Mara-Eves et al., 2015).

The screening process starts with identifying an initial set of candidate references, typically by searching databases using boolean queries handcrafted by experts. From this initial set of references, reviewers first screen for inclusion based on titles and abstracts, and then based on the full text (O'Mara-Eves et al., 2015) as illustrated in figure 1. In this paper we will call the references excluded in the first screening stage No ('N'), references excluded in the second screening stage Maybe ('M'), and references included in the final analysis Yes ('Y').

This selection is divided into two stages because while final decisions can only be based on the full text of articles, many references can be rejected based only on title and abstract. Retrieving the full text articles, which often needs to be done manually, is generally only feasible for a fraction of the articles in large systematic reviews (Tsafnat et al., 2014). However, even though humans approach screening as a two-step process, automation methods to date have generally approached the problem as a one-step process to find the relevant articles.

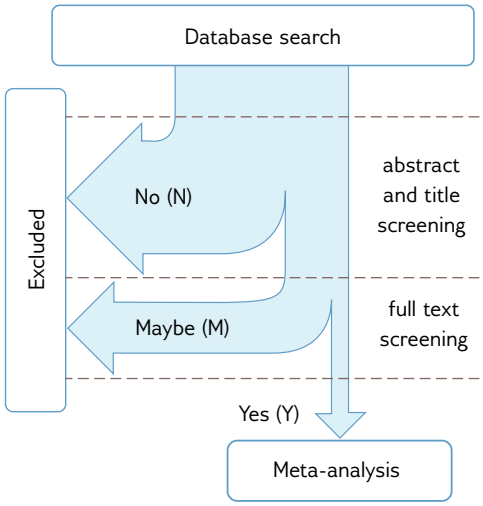

Figure 1: Overview of the data flow during the screening process in systematic reviews.

In this paper we ask if there is value in recognizing the distinction between each successive stage of the process. Our contribution is two-fold: First, we conduct experiments to inform methodology choices for automating the literature screening, and to find ways to improve the quality of constructing datasets used to train such retrieval methods. Second, we experiment on an existing reference dataset and introduce a new, complementary dataset.

[1] Although primary clinical research is often more expensive.

2. Related Work

Methods for automation have been attempted with varying degrees of success in technology assisted review in several topics in biomedicine (O'Mara-Eves et al., 2015). Technology assisted review has also been implemented in other fields with similarly stringent recall requirements, such as patent search (Stein et al., 2012), and electronic discovery (Grossman and Cormack, 2011). Automated document discovery is typically cast as a ranking or classification problem (O'Mara-Eves et al., 2015).

Common methods for automation include Support Vector Machines and variants of Naive Bayes, including Complement Naive Bayes (Matwin et al., 2010), and Multinomial Naive Bayes (Matwin and Sazonova, 2012). Other methods have been tried, including Voting Perceptrons (Cohen et al., 2006), Decision Trees (Bekhuis and Demner-Fushman, 2010), Evolutional SVM (Bekhuis and Demner-Fushman, 2010), WAODE (Bekhuis and Demner-Fushman, 2010), kNN (García Adeva et al., 2014), Rocchia (García Adeva et al., 2014), hypernym relations (Fiszman et al., 2010), ontologies (Sun et al., 2012), Generalized Linear Models (Shekelle et al., 2012), Gradient Boosting Machines (Shekelle et al., 2012), Random Indexing (Jonnalagadda and Petitti, 2014), and Random Forests (Khabsa et al., 2016). Few of the methods proposed have been evaluated on common datasets however, and it is therefore difficult to draw conclusions about relative performance (O'Mara-Eves et al., 2015).

Recently, Khabsa et al. (2016) proposed using random forests, and compared the performance of their system with the reported performance of earlier systems on Cohen's 15 reviews (see section 4.). Other methods have also been evaluated on the same dataset (Jonnalagadda and Petitti, 2014). For these reasons, and because the dataset is publicly available we will use this dataset as our baseline.

However, even though humans approach screening as a series of filters of increasingly fine granularity, all methods we have reviewed in previous literature approach the problem as a one stage process.

3. Objective

We construct an automatic screening system using a standard, off-the-shelf classifier. We describe our implementation and compare it with the state of the art to show that it functions as intended. We then apply our implementation on two datasets for systematic reviews, one of which is novel, in order to answer the following questions:

1. Can we separate the screening into two stages?
2. Do we need examples from all stages of screening (Y, M, N)?
3. Should the positive labels match the decisions in the first or second stage of the screening?

To our knowledge, these questions have not yet been considered by existing literature.

Note that the aim of this study is not to improve upon the state of the art, but to investigate how different labeling schemes affect datasets for literature screening.

Dataset	Topic	Y	M	N
Yearbook	ClinicalNLP (2017)	11	70	244 (177)
	ClinicalNLP (2016)	23	60	267 (191)
Cohen	CalciumChannelBlockers	100	180	938
	ACEInhibitors	41	142	2361
	BetaBlockers	42	260	1770
	Opiods	15	33	1867
	OralHypoglycemics	136	3	364
	Statins	85	88	3292
	SkeletalMuscleRelaxants	9	25	1609
	Antihistamines	16	76	218
	ProtonPumpInhibitors	51	187	1095
	Triptans	24	194	453
	NSAIDS	41	47	305
	ADHD	20	64	767
	AtypicalAntipsychotics	146	218	756
	UrinaryIncontinence	40	38	249
	Estrogens	80	0	288

Table 1: The distribution of class labels in each dataset. The Yearbook makes an additional separation of N into references that are off-topic and those that are on-topic but does not fit the research question of the review. The number of off-topic references is given in parentheses.

4. Datasets

To address our research questions, we use two datasets that label not only Y and N judgments, but explicitly mark the M subset.

The datasets each consist of references in the form of PubMed® identifiers (PMID) with corresponding inclusion labels (i.e. Y, M, or N) and topic labels. Article metadata, as well as titles and abstracts, are not included in either dataset, but can be downloaded from Medline® using the Entrez API.[2] The distribution of references from each review stage is reported in Table 1.

Like in the majority of previous literature, we assume that labeled training data is available, which is generally not true for new reviews.

Training data might however exist from past reviews on the same or similar topics. We call such cases where the training data is drawn from similar, but not exactly the same topic, *inter-topic* training.

It may also be possible to have reviewers label small batches of references, and use these as training data for the remainder of the process. Furthermore, systematic reviews sometimes need to be updated, in which case we can use the data from previous iterations for training. We call such cases where the training data is drawn from exactly the same topic *intra-topic* training.

4.1. The Yearbook Dataset

We construct this dataset by using the references that were considered on topic in the review on clinical NLP done by Névéol and Zweigenbaum (2016; 2017) for the IMIA Yearbook of Medical Informatics.

This review is updated annually, and the resulting dataset illustrates systematic reviews updates. In each iteration, previous data can be leveraged to train an *intra-topic* classifier. This dataset is made available in CSV and JSON format,[3] and is planned to be updated to incorporate future iterations of the review.

[2] https://www.ncbi.nlm.nih.gov/home/develop/api/

[3] Available from DOI: 10.5281/zenodo.1173076

Topic \ Measure	Intertopic			Intratopic			
	WSS@95	AUC		WSS@95		AUC	
			(Cohen)		(Khabsa)		(Khabsa)
CalciumChannelBlockers	.129	**.759**	.712	**.398**	.287 (RF)	.825	**.873** (SVM)
ACEInhibitors	.566	**.817**	.806	**.629**	.523 (CNB)	.917	**.951** (RF)
BetaBlockers	.400	**.837**	.801	**.511**	.367 (CNB)	.863	**.893** (RF)
Opiods	.301	**.885**	.856	**.590**	.554 (CNB)	.905	**.913** (RF)
OralHypoglycemics	.072	**.657**	.573	**.111**	.080 (CNB)	.568	**.781** (SVM)
Statins	.266	**.826**	.773	**.436**	.400 (RF)	.873	**.915** (RF)
SkeletalMuscleRelaxants	.241	.828	**.836**	**.429**	.371 (RF)	.740	**.794** (RF)
Antihistamines	.073	**.652**	.620	**.149**	.148 (CNB)	.650	**.722** (SVM)
ProtonPumpInhibitors	.377	**.823**	.793	**.307**	.288 (RF)	.826	**.880** (RF)
Triptans	.464	.819	**.823**	.303	**.312** (RF)	.792	**.909** (SVM)
NSAIDS	.671	**.912**	.899	**.537**	.528 (CNB)	.861	**.951** (SVM)
ADHD	.128	**.591**	.469	.616	**.668** (VP)	.908	**.951** (RF)
AtypicalAntipsychotics	.162	**.759**	.653	**.210**	.206 (CNB)	.779	**.835** (RF)
UrinaryIncontinence	.374	**.887**	.851	**.422**	.411 (RF)	.784	**.890** (SVM)
Estrogens	.176	**.693**	.588	.292	**.375** (CNB)	.689	**.887** (SVM)

Table 2: Results comparing our implementation to the state of the art. Intertopic results report the average over 5 runs. Intratopic results report the average over 10 runs (5×2 cross validation). Both cases use (Y||MN). Intertopic state of the art results are taken from Cohen (2008). Intratopic state of the art results are taken from Khabsa et al. (2016), who also report results on Complement Naive Bayes (CNB) by Matwin et al. (2010), Voting Perceptrons (VP) by Cohen et al. (2006), and Support Vector Machines (SVM) by Cohen (2008). Exact intertopic AUC scores are not explicitly reported by Cohen (2008) and have instead been extracted from Figure 1 in his paper.

4.2. The Cohen Dataset

In one of the early papers on screening automation, Cohen et al. (2006) constructed a dataset from 15 systematic reviews on drug efficacy. This dataset was later extended to 18 (Cohen et al., 2010), then to 24 reviews (Cohen et al., 2009). The smaller dataset comprising 15 reviews has been made available (Cohen et al., 2006).[4] Several methods, including Voting Perceptrons (Cohen et al., 2006), Complement Naive Bayes (Matwin and Sazonova, 2012), SVM (Cohen, 2006; Cohen et al., 2009; Cohen, 2008), Random Indexing (Jonnalagadda and Petitti, 2014), and Random Forests (Khabsa et al., 2016) have been tested on this dataset, and we can therefore use this dataset to compare our performance against previous work.

This dataset illustrates leveraging training data from similar topics. For each subtopic, data from the other subtopics may be leveraged to build an *inter-topic* classifier.

5. Document Ranking Method

We construct a ranker by extracting bag-of-n-grams ($n \leq 3$) over words in the titles and abstracts. We use both tf-idf scores and binary features, and both stemmed and unstemmed versions. The n-grams from the background, method, results, and conclusion of the abstract are also each considered in separation. We also extract article metadata, namely author-assigned keywords, journal names, and publication types. For Cohen we also extract MeSH terms, but omit these for Yearbook since MeSH terms are generally not yet available when reviews are updated.

We use a ranking approach only. In practice we ignore the decision boundary used by the logistic regression, and instead leave the decision as to where to stop the search entirely to the reviewer(s). Point measures, such as recall, can therefore only be computed as a function of the position in the ranked list.

We use the implementation of logistic regression in sklearn (Pedregosa et al., 2011) trained using stochastic gradient descent, i.e. the SGDClassifier trained using log loss. We train the ranker for a maximum of 100,000 iterations.

We generally follow the setup of Cohen et al. (2006), and Khabsa et al. (2016). For intra-topic cross validation we use 2-fold cross validation on each topic and repeat this 5 times. For intertopic training we report the average of 5 repetitions. In each experiment we report the average and standard deviation over all folds and repetitions. All hyperparameters remain constant throughout each experiment.

Unless otherwise stated, we use the default settings for all parameters. We train the ranker and calculate the AUC similarly to Cohen et al. (2009; 2008). Cross validation was done both inter-topic and intra-topic similarly to the later work of Cohen et al. (2009), and results are reported for each case. We also report the WSS@95 scores (Cohen et al., 2006) in order to compare our results against the naive bayes methods of Matwin et al. (2011). We handle class imbalance by (pseudo)randomly undersampling the majority class to have the same number of instances as the minority class. We however observe that this yields poor results when the number of examples in the majority class is low, and therefore include a minimum of 500 majority class examples.

[4] The old link has however expired. The data can now be found at `http://skynet.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html`

We increase the weights on the relevant references to 80 to emulate differing costs of misclassification. We also chose $\alpha = 10^{-4}$ as a reasonable value for the regularization term for the Cohen dataset, and $\alpha = 0.05$ for Yearbook. We selected these values through experimentation on one of the topics in Cohen (CalciumChannelBlockers), and the first iteration of the Yearbook dataset (2016).

5.1. Experimental Setup

We perform two types of experiments;
First, we run our implementation on the Cohen dataset and compare it with the reported performance of previous work. We do this in order to verify the correctness of our algorithm.
Second, we perform experiments where we enumerate different ways to treat Y, M, and N labels as positive and negative examples.
We test if it is feasible to emulate the way humans conduct systematic reviews by considering a two-stage approach where we first separate YM from N, and then Y from M.
We test whether treating the M subset as positive or negative labels impacts the performance by comparing the performance when separating YM from N with the performance when separating Y from MN.
And finally, we evaluate models where we treat the M subset as positive examples during training but negative during testing in order to test whether classification in earlier stages generalize to classification in later stages.
We report the work saved over sampling at 95% recall (WSS@95) (Cohen et al., 2006) and the area under the receiver operator characteristic curve (AUC) (Cohen, 2008) in order to bring our results in line with previous literature (Khabsa et al., 2016). The WSS@95 metric measures the theoretical work saved when using the model to retrieve 95% of the relevant articles.

6. Results

We present our comparison with the state of the art in Table 2. In Tables 3a–3c we present the results of our experiments using data with different compositions of examples in terms of Y, M, and N.

7. Discussion

In this section we discuss the results, in order to verify that our system works as intended, and to address the questions we set out in Section 3. Objective.

7.1. Performance of our System

Intuitively: based on the WSS@95 scores (Tables 3a, 3b), our method could save the reviewers from having to look at 46 (Antihistamines) to 1058 (BetaBlockers) references depending on the topic, or about 605 references on average. The results of our implementation are comparable to state of the art results across the board (Table 2). Our implementation exhibits equal or better results for intertopic training (Table 2). For intratopic training, our implementation exhibit worse results in terms of AUC, but better scores in terms of WSS@95. Our implementation seems to perform worse than the state of the art mainly on the topics where there are no or very few M (OralHypoGlycemics, Estrogens). It is also possible that the additional features used by Khabsa et al. (references cited) can explain some of the difference in results.

7.2. Can We Separate the Screening into Two Stages?

Separating the screening into two stages would entail first screening in terms of (YM||N) followed by (Y||M). However, from Tables 3a–3c it is clear that while (Y||MN) is feasible, (Y||M) is considerably more difficult than (Y||N) or (Y||MN) (Tables 3a–3c). The ranker is however doing a slightly better job on BetaBlockers and Triptans (Tables 3b and 3c).

In particular, when separating Y from M, the ranker is not performing much better than chance on many of the topics. This is to be expected, since M represent those references the human annotators required the full text to judge, and it would be unreasonable to expect the ranker to be able to judge these based only on title and abstract.

Consequently, we can certainly perform (YM||N) as an initial step, but (Y||M) would at the very least require ranking the full text articles.

7.3. Do We Need Examples from All Stages of Screening (Y, M, N)?

We observe similar results for (Y|M|N) and (Y||MN) on Cohen, i.e. we can train a ranker using positive examples that were included based on title and abstract (Y+M), even if these were to turn out to be non-relevant upon inspecting the full text (M). On the Yearbook dataset we observe better scores for (Y|M|N) than (Y||MN), likely due to the number of Y available for training (23) being much smaller. In Table 3b we can generally observe similar results for (Y|M|N) and (Y||MN), the exceptions being Triptans and NSAIDS where we observe better results for (Y||MN). We also observe similar results for (Y|M|N) and (Y||MN) on the Yearbook data. On some topics we observe better results for (Y|M|N), but the difference is small.

Furthermore, both (Y||N) and (M||N) seem to give reasonable results, although these results are not directly comparable to the results for (Y|M|N). We can also observe that (Y||N) is generally easier than (M||N). This could be due to Y containing fewer borderline cases.

Consequently, we do need positive examples drawn from Y or M, as well as negative examples drawn from N. It seems to make less difference whether we consider M to be positive or negative examples and we may be able to exclude either Y or M in training.

Interestingly it seems from Table 3a that it is more difficult to classify in terms of (YM||N) than (Y||MN) on Cohen, but the inverse is true on Yearbook. This might be explained by the small number of Y on Yearbook (11), and we can observe the same on the topics in Cohen with few Y (SkeletalMuscleRelaxants, ADHD). OralHypoglycemics have only 3 M and Estrogens no M at all, and we therefore exclude these topics from the results.

	(Y||MN)		(YM||N)		(Y|M|N)		(Y||M)		(Y||N)		(M||N)	
	WSS	AUC	WSS	AUC	WSS	AUC	WSS	AUC	WSS	AUC	WSS	AUC
Yearbook	.003	.625	.229	.803	.189	.808	.012	.481	.020	.738	.256	.785
Cohen	.449	.839	.265	.768	.472	.814	.163	.557	.423	.832	.239	.714

(a) Intra-topic results averaged over 10 runs (5 × 2 cross validation) for different dataset compositions. The averages were computed using weights proportional to the number of articles in each topic (Y+M+N, Y+M, Y+N, or M+N).

	(Y||MN)				(YM||N)				(Y|M|N)			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
Topic	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP (Yearbook)	.003	.000	.625	.005	.229	.011	.803	.001	.189	.008	.808	.002
CalciumChannelBlockers	.398	.098	.825	.024	.218	.056	.764	.030	.338	.073	.790	.012
ACEInhibitors	.629	.158	.917	.020	.277	.050	.800	.021	.598	.126	.879	.027
BetaBlockers	.511	.157	.863	.030	.187	.047	.730	.025	.476	.210	.831	.021
Opiods	.590	.193	.905	.052	.366	.096	.817	.033	.705	.063	.881	.035
OralHypoglycemics	.111	.048	.568	.026	.138	.068	.579	.036	.089	.020	.583	.026
Statins	.436	.176	.873	.021	.254	.094	.779	.025	.421	.101	.864	.015
SkeletalMuscleRelaxants	.429	.221	.740	.113	.264	.180	.826	.064	.445	.116	.746	.057
Antihistamines	.149	.089	.650	.089	.126	.038	.566	.026	.239	.092	.596	.013
ProtonPumpInhibitors	.307	.191	.826	.044	.167	.043	.731	.023	.378	.058	.770	.037
Triptans	.303	.237	.792	.075	.300	.039	.746	.030	.412	.067	.691	.026
NSAIDS	.537	.184	.861	.022	.402	.072	.755	.042	.458	.057	.727	.024
ADHD	.616	.148	.908	.026	.697	.096	.910	.017	.828	.057	.906	.011
AtypicalAntipsychotics	.210	.044	.779	.012	.123	.024	.714	.027	.284	.057	.803	.022
UrinaryIncontinence	.422	.144	.784	.032	.207	.089	.660	.040	.475	.072	.750	.038
Estrogens	.292	.089	.689	.026	.266	.093	.715	.040	.319	.056	.693	.026

(b) Intratopic results averaged over 10 runs (5 × 2 cross validation) for different dataset compositions.

	(Y||M)				(Y||N)				(M||N)			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
Topic	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP (Yearbook)	.012	.000	.481	.005	.020	.002	.738	.003	.256	.004	.785	.001
CalciumChannelBlockers	.141	.039	.590	.030	.421	.106	.852	.024	.208	.069	.743	.032
ACEInhibitors	.165	.083	.631	.059	.410	.370	.918	.032	.256	.063	.771	.020
BetaBlockers	.383	.096	.737	.021	.515	.135	.870	.034	.190	.031	.713	.018
Opiods	.131	.096	.526	.006	.592	.205	.906	.064	.249	.177	.762	.045
OralHypoglycemics	.058	.000	.387	.167	.105	.039	.579	.030	.754	.194	.826	.112
Statins	.125	.052	.560	.037	.439	.184	.879	.047	.240	.086	.708	.028
SkeletalMuscleRelaxants	.240	.143	.547	.017	.297	.149	.668	.078	.226	.163	.800	.067
Antihistamines	.204	.165	.554	.062	.161	.090	.700	.033	.128	.073	.583	.036
ProtonPumpInhibitors	.159	.052	.584	.022	.421	.168	.852	.026	.122	.046	.694	.032
Triptans	.199	.130	.695	.072	.437	.244	.880	.042	.272	.064	.746	.028
NSAIDS	.129	.050	.576	.056	.479	.185	.851	.017	.316	.094	.723	.027
ADHD	.193	.138	.588	.093	.707	.169	.938	.021	.639	.170	.916	.013
AtypicalAntipsychotics	.112	.023	.548	.017	.259	.114	.792	.030	.113	025	.629	.031
UrinaryIncontinence	.090	.038	.550	.024	.433	.159	.792	.033	.121	.103	.591	046
Estrogens	-	-	-	-	.233	.034	.686	.038	-	-	-	-

(c) Intratopic results averaged over 10 runs (5 × 2 cross validation) for different dataset compositions.

Table 3: (Y||MN) denotes results using Y as the positive class. (YM||N) denotes results using Y and M as the positive class. (Y|M|N) denotes results using Y and M as the positive class in training, and Y as the positive class in evaluation. (Y||M) denotes results using Y as the positive, and M as the negative class. (Y||N) denotes results using Y as the positive, and N as the negative class. (M||N) denotes results using M as the positive, and N as the negative class. Estrogens has no M, and is consequently excluded from the calculations of the results for (Y||M) and (M||N).

7.4. Can We Use M as Positive Examples for Training?

Cohen et al. previously discovered that while intratopic data is generally better than intertopic data (Cohen et al., 2006), the less targeted intertopic data can complement the intratopic data if the intratopic data is scarce (Cohen et al., 2006; Cohen et al., 2009). Our results suggest the same (Table 2), but also that we can generally use M as training examples to complement the Y. The intuition behind these ideas is similar: while it is generally important to have training data targeted for the particular problem, it is also important to have sufficient amounts of data, and less targeted training data can provide a supplement if only scarce amounts of data is available.

We can further compare the results for intratopic (Y|M|N) versus the results for intertopic (Y||MN) in Tables 3b and 2 to get a sense of whether complementing our training data by using M as positive examples works better than complementing our training data with less targeted data from similar topics.

We observe better results for intertopic (Y||MN) for OralHypoglycemics, SkeletalMuscleRelaxants, Antihistamines, Triptans, NSAIDS, and UrinaryIncontinence. This might in part be explained by OralHypoglycemics, SkeletalMuscleRelaxants and Antihistamines having few Y. We observe better results for intratopic (Y|M|N) on ACEInhibitors, ProtonPumpInhibitors, and ADHD. It is not clear why we observe this difference on these topics.

7.5. Strategies for Ranking Articles

From Tables 3a–3c it seems that there is no single approach that is clearly better for any kind of data. Which approach works best depends on the number of articles in each class, as well as the exact nature of articles in each stage. What parts of the data to e.g. use for training must therefore be decided based on the characteristics of the dataset, or by testing multiple approaches.

The results and conclusions of this study guided the strategic choices we made for the system submitted to the CLEF eHealth shared task *Technology Assisted Reviews in Empirical Medicine* (Norman et al., 2017; Kanoulas et al., 2017). We submitted four runs using different machine learning methods: 1) the (YM||N) approach described here 2) an (YM||N) approach using standard logistic regression (i.e. not trained using SGD), and 3) two variations of logistic regression with active learning, where the system starts using the (Y|M|N) approach and later switches to using the (Y||MN) approach once a sufficient number of Y have been discovered.

On the Cohen dataset approach 2 worked better than approach 1 for intratopic training and vice-versa, and we could reliably see improvements over either of these by using active learning. On the CLEF data however, approach 1 achieved much better results than either approach 2 or 3. We believe that this was at least partly due to the small number of relevant articles per topic in the CLEF dataset (Norman et al., 2017).

Our participation placed third to fifth in the evaluation overall, depending on metric used, and placed first among the systems not using active learning.

7.6. Limitations

This work relied on two datasets and a ranker developed in-house. It is not clear how the results generalize to other domains and datasets, or to other machine learning methods.

We observe fairly large variance for many of the runs (Tables 3b, 3c), and on many topics. This is particularly problematic for the WSS metric, but it also affects the AUC metric even averaged over ten repetitions. For instance, Estrogens has no M, and we should therefore expect the same results for (Y||MN) and (YM||N), yet we observe differences roughly equal to the standard deviation for the AUC. Previous literature generally do not report their variance, which complicates the comparison with previous results.

7.7. Future work

We are working on extending the system to use additional machine learning methods, including deep artificial neural networks, and to complement the system with information retrieval methods.

8. Conclusion

We find that in order to train rankers to automate the screening process we need to use 1) examples of excluded references (N), *and* 2) references included in either the first (M) or second stage of the screening (Y). In the systematic reviews, the M are those articles that were excluded after reading the full text, and so are in reality negative examples. However, our results suggest that these can still be used as positive examples for training. It may well be possible to construct an accurate ranker using only the M as the positive examples, without any real positive examples (i.e. Y) at all.

Our best results are achieved with (Y||MN) on the Cohen dataset, whereas our best results are achieved with (Y|M|N) on the Yearbook dataset. Given that the distribution of the labels is similar in both datasets it is likely that greater contribution of the M on the Yearbook dataset is due to its smaller size. For any new systematic review we only have whatever training data we label ourselves, and data scarcity is therefore one of the major issues we need to overcome. Even for systematic review updates the amount of positive training data available is typically modest since the number of included articles in any systematic review tends to be small (the Y column in Table 1).

Since the number of references that are provisionally included based on title and abstract (Y+M) can outnumber the final includes (Y) by almost ten to one (Table 1), using examples of M in addition to Y suggests a straightforward way to increase the amount of training data available (i.e. the Y|M|N approach), and thus potentially overcome the data scarcity problem, particularly if we do not have access to inter-topic training data. This does not seem to have been considered in previous work.

Our results also agree with the state of the art and suggest that common-off-the-shelf machine learning algorithms can accurately predict topical relevance of candidate articles for inclusion in systematic reviews.

In light of the results, we recommend that future datasets intended to be used either for training or for evaluation of

document screening should include a tripartite labeling reflecting the two filtering stages in manual systematic reviews. Strictly, only the distinction between YM and N is necessary for training, but we still likely want to only treat Y as positive during evaluation, since only these would be considered relevant for the purposes of the systematic review.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

9. Bibliographical References

Bekhuis, T. and Demner-Fushman, D. (2010). Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics*, 160(PART 1):146–150.

Cohen, A. M., Hersh, W. R., Peterson, K., and Yen, P. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. pages 206–219.

Cohen, A. M., Ambert, K., and McDonagh, M. (2009). Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update. *Journal of the American Medical Informatics Association*, 16(5):690–704.

Cohen, A. M., Ambert, K., and McDonagh, M. (2010). A Prospective Evaluation of an Automated Classification System to Support Evidence-based Medicine and Systematic Review. *AMIA Annual Symposium Proceedings*, 2010:121–125.

Cohen, A. M. (2006). An effective general purpose approach for automated biomedical document classification. *AMIA Annual Symposium proceedings*, pages 161–165.

Cohen, A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. *AMIA Annual Symposium proceedings*, pages 121–5.

Fiszman, M., Bray, B., Shin, D., Kilicoglu, H., Bennett, G., Bodenreider, O., and Rindflesch, T. (2010). Combining Relevance Assignment with Quality of the Evidence to Support Guideline Development. *Stud Health Technol Inform*, 160(1):709—-713.

García Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M., and Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, 41(4 PART 1):1498–1508.

Grossman, M. and Cormack, G. (2011). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17(3).

Jonnalagadda, S. R. and Petitti, D. (2014). A new iterative method to reduce workload in the systematic review process. *Int J Comput Biol Drug Des*, 6(0):5–17.

Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2017). Overview of the CLEF technologically assisted reviews in empirical medicine. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. CEUR-WS.org.

Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., and Ouzzani, M. (2016). Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482.

Matwin, S. and Sazonova, V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, 19(5):917–917.

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., and O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–53.

Matwin, S., Kouznetsov, A., Inkpen, D., and O'Blenis, P. (2011). Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure. *Journal of the American Medical Informatics Association*, 1(18):author reply 105.

Névéol, A. and Zweigenbaum, P. (2016). Clinical natural language processing in 2015: Leveraging the variety of texts of clinical interest. *IMIA Yearbook*, pages 234–239.

Névéol, A. and Zweigenbaum, P. (2017). Making sense of big textual data for health care: Findings from the section on clinical natural language processing. *Yearbook of medical informatics*, 26(01):228–233.

Norman, C., Leeflang, M., and Névéol, A. (2017). LIMSI@CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):5.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Shekelle, P. G., Dalal, S. R., and Shetty, K. D. (2012). A Pilot Study Using Machine Learning and Domain Knowledge To Facilitate Comparative Effectiveness Review Updating. *AHRQ*.

Stein, B., Hoppe, D., and Gollub, T. (2012). The impact of spelling errors on patent search. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 570–579. Association for Computational Linguistics.

Sun, Y. B., Yang, Y., Zhang, H., Zhang, W., and Wang, Q. (2012). Towards evidence-based ontology for supporting systematic literature review. *Evaluation and Assessment in Software Engineering*, 2012(1):171–175.

Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., and Coiera, E. (2014). Systematic review automation technologies. *Systematic reviews*, 3(1):74.

Wright, R. W., Brand, R. A., Dunn, W., and Spindler, K. P. (2007). How to write a systematic review. *Clinical orthopaedics and related research*, 455:23–29.