Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

Maura R. Grossman*
Wachtell, Lipton, Rosen & Katz
mrgrossman@wlrk.com

ABSTRACT

Using a novel evaluation toolkit that simulates a human reviewer in the loop, we compare the effectiveness of three machine-learning protocols for technology-assisted review as used in document review for discovery in legal proceedings. Our comparison addresses a central question in the deployment of technology-assisted review: Should training documents be selected at random, or should they be selected using one or more non-random methods, such as keyword search or active learning? On eight review tasks – four derived from the TREC 2009 Legal Track and four derived from actual legal matters – recall was measured as a function of human review effort. The results show that entirely non-random training methods, in which the initial training documents are selected using a simple keyword search, and subsequent training documents are selected by active learning, require substantially and significantly less human review effort ($P < 0.01$) to achieve any given level of recall, than passive learning, in which the machine-learning algorithm plays no role in the selection of training documents. Among passive-learning methods, significantly less human review effort ($P < 0.01$) is required when keywords are used instead of random sampling to select the initial training documents. Among active-learning methods, continuous active learning with relevance feedback yields generally superior results to simple active learning with uncertainty sampling, while avoiding the vexing issue of "stabilization" – determining when training is adequate, and therefore may stop.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Search process, relevance feedback.

Keywords: Technology-assisted review; predictive coding; electronic discovery; e-discovery.

*The views expressed herein are solely those of the author and should not be attributed to her firm or its clients.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2609601.

1. INTRODUCTION

The objective of technology-assisted review ("TAR") in electronic discovery ("e-discovery")[1] is to find as nearly all of the relevant documents in a collection as possible, with reasonable effort. While this study does not presume to interpret the common-law notion of what is reasonable, it serves to quantify the tradeoff between how nearly all of the relevant documents can be found (as measured by recall), and the human effort needed to find them (as measured by the number of documents that must be manually reviewed, which translates into time and cost).

In a typical review task, a "requesting party" prescribes relevance[2] by way of a "request for production," while the "responding party," its adversary, is required to conduct a review and produce the responsive, non-privileged documents identified as a result of a reasonable search. A study by Grossman and Cormack [8] shows that two TAR methods can be both more effective and more efficient than traditional e-discovery practice, which typically consists of keyword or Boolean search, followed by manual review of the search results. One of these methods, due to Cormack and Mojdeh [7], employs machine learning in a protocol we refer to as Continuous Active Learning ("CAL"). The other method, due to H5 [13], does not employ machine learning, and therefore is not considered in this study.

Often relying on Grossman and Cormack for support, many legal service providers have advanced TAR tools and methods that employ machine learning, but not the CAL protocol. These tools and methods, often referred to in the legal marketplace as "predictive coding," follow one of two protocols which we denote Simple Active Learning ("SAL") and Simple Passive Learning ("SPL"). Some tools that employ SAL have achieved superior results at TREC, but have never, in a controlled study, been compared to CAL. Tools that use SPL, while widely deployed, have not achieved superior results at TREC, and have not, in a controlled study, been compared to traditional methods, to SAL, or to CAL.

[1]See Grossman and Cormack [10] for a glossary of terms pertaining to TAR. See Oard and Webber [16] for an overview of information retrieval for e-discovery.

[2]The IR term "relevant" generally describes a document sought by an information-retrieval effort, while the legal term "responsive" describes a document that satisfies the criteria set forth in a request for production. In this study, the terms are used interchangeably; however, in the context of litigation, relevance may take on a broader meaning than responsiveness.

This study compares CAL, SAL, and SPL, and makes available a TAR evaluation toolkit[3] to facilitate further comparisons. The results show SPL to be the least effective TAR method, calling into question not only its utility, but also commonly held beliefs about TAR. The results also show that SAL, while substantially more effective than SPL, is generally less effective than CAL, and as effective as CAL only in a best-case scenario that is unlikely to be achieved in practice.

2. THE TAR PROCESS

The TAR process, in the abstract, proceeds as follows. Given a document collection and a request for production, a human operator uses one or more tools to identify documents to be shown to one or more human reviewers, who may or may not be the same individual as the operator. The reviewers examine these documents and label ("code") them each as responsive or not. More documents are identified using the tools, reviewed and coded by reviewers, and the process continues until "enough" of the responsive documents have been reviewed and coded. How many constitute "enough" is a legal question, which is informed by how much additional effort would likely be required to find more responsive documents, and how important those documents would likely be in resolving the legal dispute (*i.e.*, "proportionality considerations"). For our purposes, we consider the process to continue indefinitely, and track the number of responsive documents found (*i.e.*, recall) as a function of effort (*i.e.*, the number of documents reviewed and coded by reviewers). Using this information, the reader can determine retrospectively, for any definition of "enough," how much effort would have sufficed to find enough documents.

For this study, the operator is assumed to follow a strict protocol. All choices, including what tools are used, and when and how they are used, are prescribed by the protocol. In addition to satisfying the requirements for a controlled comparison, the use of a strict protocol may be appealing in the e-discovery context because the requesting party may distrust, and therefore wish to prohibit discretionary choices made by the operator on behalf of the responding party. The reviewers are assumed to code the documents they review in good faith, to the best of their abilities. In light of Grossman and Cormack [8, 9], and others [4, 18, 25, 26], it is unrealistic to assume the reviewers to be infallible – they will necessarily, but inadvertently, code some responsive documents as non-responsive, and vice versa.

The CAL protocol involves two interactive tools: a keyword search system and a learning algorithm. At the outset of the TAR process, the operator typically uses a keyword search to identify an initial set of documents to be reviewed and coded. These coded documents (often referred to as the "seed set") are used to train a learning algorithm, which scores each document in the collection by the likelihood that it is responsive. The top-scoring documents that have not yet been coded are then reviewed and coded by reviewers. The set of all documents coded thus far (the "training set") is used to train the learning algorithm, and the process of selecting the highest-scoring documents, reviewing and coding them, and adding them to the training set continues until "enough" of the responsive documents have been found.

[3]Available at `http://cormack.uwaterloo.ca/cormack/tar-toolkit`.

The SAL protocol, like CAL, begins with the creation of a seed set that is used to train a learning algorithm. The seed set may be selected using keywords, random section, or both, but, unlike CAL, the subsequent training documents to be reviewed and coded are selected using uncertainty sampling [15], a method that selects the documents about which the learning algorithm is least certain. These documents are added to the training set, and the process continues until the benefit of adding more training documents to the training set would be outweighed by the cost of reviewing and coding them (a point often referred to as "stabilization"). At this point, the learning algorithm is used for the last time to create either a set or a ranked list of likely relevant documents (the "review set"), which is subsequently reviewed and coded by reviewers.

The SPL protocol, unlike CAL or SAL, generally relies on the operator or random selection, and not the learning algorithm, to identify the training set. The process is typically iterative. Once a candidate training set is identified, the learning algorithm is then trained on these documents and used to create a candidate review set. If the review set is "inadequate," the operator creates a new candidate training set, generally by adding new documents that are found by the operator, or through random selection. The process continues until the review set is deemed "adequate," and is subsequently reviewed and coded by reviewers.

The TAR process addresses a novel problem in information retrieval, which we denote here as the "TAR Problem." The TAR Problem differs from well-studied problems in machine learning for text categorization [21] in that the TAR process typically begins with no knowledge of the dataset and continues until most of the relevant documents have been identified and reviewed. A classifier is used only incidentally for the purpose of identifying documents for review. Gain is the number of *relevant* documents presented to the human during training and review, while cost is the total number of *relevant and non-relevant* documents presented to the human during training and review.

3. SIMULATING REVIEW

To simulate the application of a TAR protocol to a review task, we require a realistic document collection and request for production, a keyword query ("seed query") to be used (if required by the protocol), and a simulated reviewer. To evaluate the result, we require a "gold standard" indicating the true responsiveness of all, or a statistical sample, of the documents in the collection.

Four review tasks, denoted Matters 201, 202, 203, and 207, were derived from Topics 201, 202, 203, and 207 of the TREC 2009 Legal Track Interactive Task – the same Topics that were used to evaluate Cormack and Mojdeh's CAL efforts at TREC [7, 12]. Four other review tasks, denoted Matters A, B, C, and D, were derived from actual reviews conducted in the course of legal proceedings. Statistics for the collections are provided in Table 1, and the requests for production are shown in Table 2.

The seed queries for the tasks derived from TREC, shown in Table 3, were composed by Open Text in the course of its participation in the TREC 2010 Legal Track Learning Task (which used the same topics as the TREC 2009 Legal Track Interactive Task), using a strategy that "attempted to quickly create [a] Boolean query for each topic" [24, page 5]. The seed queries for the tasks derived from legal proceedings,

Matter	Collection Size	# Rel. Docs.	Prevalence (%)
201	723,537	2,454	0.34
202	723,537	9,514	1.31
203	723,537	1,826	0.25
207	723,537	8,850	1.22
A	1,118,116	4,001	0.36
B	409,277	6,236	1.52
C	293,549	1,170	0.48
D	405,796	15,926	3.92

Table 1: Collection statistics.

Matter	Request for Production
201	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as *prepay transactions*.
202	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
203	All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
207	All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.
A	[Regulatory request]
B	[Regulatory request]
C	[Third-party subpoena]
D	[Regulatory request]

Table 2: Requests for production.

described in Table 3, were composed by or negotiated with the requesting party prior to the review process. The recall and precision of each of the seed queries (as measured with respect to the gold standard, discussed below) are shown in Table 3.

To simulate a reviewer, we use a "training standard" that consists of a relevance assessment for each document in the collection. If, during the course of simulating a particular review protocol, the reviewer is called upon to code a document, the assessment from the training standard – responsive or not responsive – is used for this purpose. The training standard does not represent ground truth; instead, it represents the coding decision that a fallible reviewer might render when presented with the document for review. For all of the simulated tasks, all of the positive training-standard assessments, and some of the negative assessments, were rendered by a reviewer during the course of a prior review. For the TREC-derived tasks, we used Cormack and

Matter	Seed Query	Recall	Prec.
201	`"pre-pay" OR "swap"`	0.436	0.038
202	`"FAS" OR "transaction" OR "swap" OR "trust" OR "Transferor" OR "Transferee"`	0.741	0.090
203	`"forecast" OR "earnings" OR "profit" OR "quarter" OR "balance sheet"`	0.872	0.034
207	`"football" OR "eric bass"`	0.492	0.167
A	`[7-term Boolean query]`	0.545	0.045
B	`[98-term Boolean query]`	0.991	0.019
C	`[46-term Boolean query]`	0.259	0.026
D	`[9-term Boolean query]`	0.945	0.325

Table 3: Keyword seed queries and their associated recall and precision.

Matter	Training Standard	
	Recall	Precision
201	0.843	0.911
202	0.844	0.903
203	0.860	0.610
207	0.896	0.967
A	1.000	0.307
B	0.942	0.974
C	1.000	0.429
D	0.961	1.000

Table 4: Recall and precision for the training standard used to simulate human review.

Mojdeh's TREC submissions;[4] for the legal-matter-derived tasks, we used the coding rendered by the first-pass reviewer in the course of the review. Documents that were never seen by the first-pass reviewer (because they were never identified as potentially responsive) were deemed to be coded as non-responsive. Overall, as measured with respect to the gold standard, the recall and precision of the training standard (shown in Table 4) indicate that the simulated reviewer achieves a high-quality – but far from perfect – result, by human review standards.

In contrast to the training standard, the gold standard represents ground truth. For the TREC-derived tasks, the gold standard consists of a stratified random sample, assessed by TREC using a two-stage adjudication process [12]. For the legal-matter-derived tasks, the gold standard consists of the documents produced to the requesting party, after a second-pass review and quality-assurance efforts. Each document in the gold standard is associated with an inclusion probability – the prior probability that it would have been included in the gold standard. Following TREC practice, the recall of any simulated review is estimated using the Horvitz-Thompson estimator [14], which weights each gold-standard document by the reciprocal of its inclusion probability.

Evaluation results are presented in two ways: as gain curves and as 75% recall-effort ("75% RE") values. A gain curve plots recall as a function of the number of documents reviewed. For any level of effort (as measured by the number

[4]Available at `http://trec.nist.gov/results.html`, subject to a usage agreement.

of documents reviewed), one can determine at a glance the recall that would be achieved, using a particular protocol, for that level of effort (see Figures 1 and 2 below). Conversely, for any recall level it is possible to determine what level of effort would be required to achieve that recall level. For the purpose of quantitative comparison, we tabulate 75% RE for all protocols (see Tables 5 and 6 below).

4. TAR PROTOCOLS

In this study, we used the same feature engineering and learning algorithm for every protocol, without any collection- or task-specific tuning. Following Cormack and Mojdeh [7], the first 30,000 bytes of the ASCII text representation of each document (including a text representation of the sender, recipient, cc or bcc recipients, subject, and date and time sent) was shingled as overlapping 4-byte segments. The number of distinct possible segments was reduced, by hashing, from $2^{32} = 4,294,967,296$ to $1,000,081$ (an arbitrarily chosen prime number near one million). Each feature consisted of a binary value: "1" if the feature was present in the first 30,000 bytes of the document; "0" if it was absent. For the learning algorithm, we used the Sofia-ML implementation of Pegasos SVM,[5] with the following parameters: "`--iterations 2000000 --dimensionality 1100000`."

For all protocols, we used a batch size of 1,000 documents. That is, the initial training set (the seed set) was 1,000 documents, and each iteration, whether CAL, SAL, or SPL, involved reviewing 1,000 documents and, if indicated by the protocol, adding them to the training set. Our primary experiments evaluated the specific formulations of CAL, SAL, and SPL described in Section 2; secondary experiments explored the effect of using keyword-selected versus randomly selected documents for the seed and training sets.

Our primary CAL implementation used, as the initial training set, 1,000 documents, randomly selected from the results of a search using the seed query. In each iteration, the training-set documents were coded according to the training standard, then used to train Sofia-ML, and hence to score the remaining documents in the collection. The 1,000 top-scoring documents were added to the training set, and the process was repeated 100 times.

Our primary SAL implementation used exactly the same 1,000-document keyword-selected seed set as CAL. Like CAL, in each iteration, the training-set documents were coded according to the training standard, then used to train Sofia-ML, and hence to score the remaining documents in the collection. These documents, ranked by score, constitute a candidate review set. Rather than implementing the decision as to whether stabilization had occurred, we recorded the candidate review set for future retrospective evaluation, and continued the training process. Unlike CAL, the 1,000 documents with the *least magnitude* scores were coded and added to the training set, and the process was repeated 100 times. In the end, the simulation yielded 100 different candidate review sets, corresponding to stabilization having occurred with a training-set size of 1,000, 2,000, ..., 100,000 documents. Each training-set size, when evaluated, yields a different gain curve, and a different 75% RE. Due to space considerations, we show gain curves only for the representative training-set sizes of 2,000, 5,000, and 8,000 documents. We report 75% RE for these three training-set sizes, as well as for the *ideal* training-set size, which in reality would not be known, since it requires the benefit of hindsight. The ideal training-set size is derived using the gold standard; 75% RE is calculated for every training-set size, and the lowest value is chosen.

Our primary SPL implementation used random selection throughout, as advocated by some SPL proponents. The initial training set (which we denote the "seed set," notwithstanding the fact that many SPL proponents use the same term to refer to the final training set) consisted of 1,000 randomly selected documents, and each iteration added 1,000 more randomly selected documents. As for SAL, we recorded the candidate review set after each iteration, and report gain curves for the representative training-set sizes of 2,000, 5,000, and 8,000 documents, as well as 75% RE for these training-set sizes, and for the *ideal* training-set size, as defined above.

Variants of these protocols, for which we report 75% RE, include using randomly selected documents as a seed set for CAL and SAL, using a keyword-selected seed set for SPL, and using an entirely keyword-selected training set for SPL.

5. PRIMARY RESULTS

As illustrated in Figure 1, the CAL protocol achieves higher recall than SPL, for less effort, for all of the representative training-set sizes. All eight graphs show the same basic result: After the first 1,000 documents (*i.e.*, the seed set), the CAL curve shows a high slope that is sustained until the majority of relevant documents have been identified. At about 70% recall, the slope begins to fall off noticeably, and effectively plateaus between 80% and 100% recall. The SPL curve exhibits a low slope for the training phase, followed by a high slope, falloff, and then a plateau for the review phase. In general, the slope immediately following training is comparable to that of CAL, but the falloff and plateau occur at substantially lower recall levels. While the initial slope of the curve for the SPL review phase is similar for all training-set sizes, the falloff and plateau occur at higher recall levels for larger training sets. This advantage of larger training sets is offset by the greater effort required to review the training set: In general, the curves for different training sets cross, indicating that a larger training set is advantageous when high recall is desired.

75% recall effort, shown in Table 5, illustrates the superiority of CAL over SPL, even when SPL is afforded the benefit of hindsight to choose the ideal training-set size. A simple sign test shows with statistical significance ($P < 0.01$) that CAL is superior to SPL according to 75% RE, and also according to recall effort for any other level of recall.

Figure 2 shows that the CAL protocol generally achieves higher recall than SAL. However, the SAL gain curves, unlike the SPL gain curves, often touch the CAL curves at one specific inflection point. The strong inflection of the SAL curve at this point is explained by the nature of uncertainty sampling: Once stabilization occurs, the review set will include few documents with intermediate scores, because they will have previously been selected for training. Instead, the review set will include primarily high-scoring and low-scoring documents. The high-scoring documents account for the high slope before the inflection point; the low-scoring documents account for the low slope after the inflection point; the absence of documents with intermediate scores accounts for the sharp transition. The net effect

[5]Available at `http://code.google.com/p/sofia-ml`.

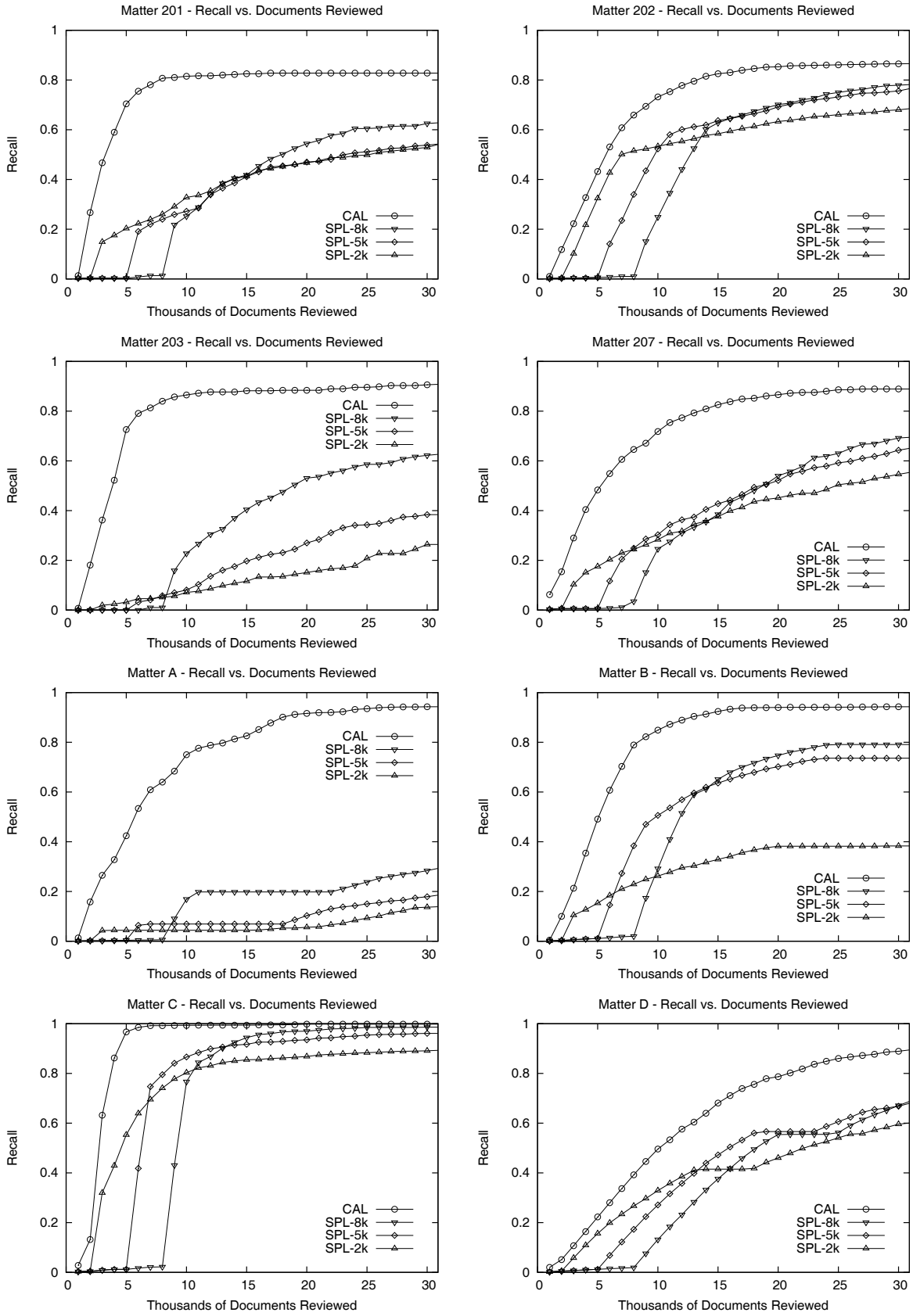

Figure 1: Continuous Active Learning versus Simple Passive Learning using three different training-set sizes of randomly selected documents.

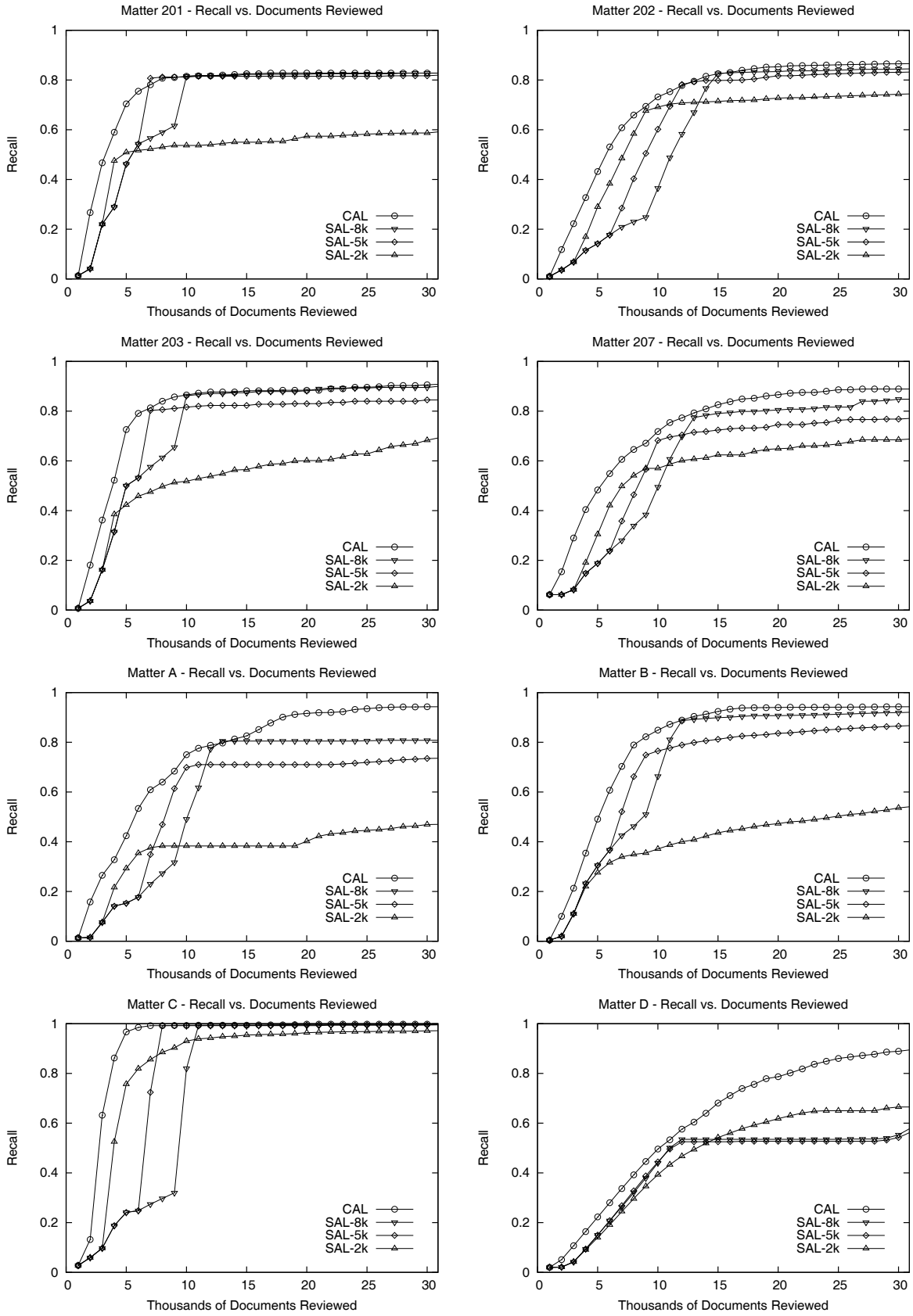

Figure 2: Continuous Active Learning versus Simple Active Learning using three different training-set sizes of uncertainty-sampled documents.

Matter	CAL	SAL				SPL			
		Training Set Size				Training Set Size			
		2K	*5K*	*8K*	*Ideal*	*2K*	*5K*	*8K*	*Ideal*
201	**6**	237	7	10	7	284	331	164	56
202	**11**	34	12	14	12	47	29	26	26
203	**6**	43	7	10	**6**	521	331	154	99
207	**11**	55	23	13	13	103	50	36	35
A	**11**	210	42	12	12	502	326	204	85
B	**8**	119	10	11	10	142	41	21	20
C	**4**	5	8	10	5	9	8	10	7
D	**18**	60	54	53	**18**	55	38	37	37

Table 5: 75% Recall Effort for Primary Results (measured in terms of thousands of documents reviewed). Bold numbers reflect the least possible effort to achieve the target recall of 75%.

Matter	CAL	CAL-seedran	SAL	SAL-seedran	SPL	SPL-seedkey	SPL-allkey
			Training Set Size		Training Set Size		
			Ideal	*Ideal*	*Ideal*	*Ideal*	*Ideal*
201	**6**	**6**	7	8	56	36	43
202	**11**	12	12	12	26	23	20
203	**6**	637	**6**	614	99	26	16
207	**11**	12	13	13	35	26	16
A	**11**	15	12	15	85	79	66
B	**8**	10	10	10	20	19	39
C	**4**	**4**	5	**4**	7	6	9
D	**18**	19	**18**	19	37	28	34

Table 6: 75% Recall Effort for Primary and Supplemental Results (measured in terms of thousands of documents reviewed). Bold numbers reflect the least possible effort to achieve the target recall of 75%.

is that SAL achieves effort as low as CAL only for a specific recall value, which is easy to see in hindsight, but difficult to predict at the time of stabilization.

Table 5 illustrates the sensitivity of the SAL and SPL results to training-set size, and hence the difficulty of choosing the precise training-set size to achieve 75% recall with minimal effort.

6. SUPPLEMENTAL RESULTS

To assess the role of keyword versus random selection at various stages of the training process, we evaluated the following variants of the primary protocols: (i) CAL-seedran, in which the seed set was selected at random from the entire collection; (ii) SAL-seedran, in which the seed set was selected at random from the entire collection; (iii) SPL-seedkey, in which the initial 1,000 training documents were the same keyword-selected seed set used for CAL and SAL in the primary protocols; and (iv) SPL-allkey, in which all training examples were selected at random from the results of the keyword seed query. 75% recall effort (with ideal training-set sizes, where applicable) for these variants, as well as the primary protocols, is shown in Table 6.

A comparison of the results for CAL and CAL-seedran shows that a random seed set generally yields the same or slightly inferior results to a keyword-selected seed set. In one case – Matter 203 – the random seed set fails spectacularly. The collection for this task has very low prevalence (0.25%), and the seed set of 1,000 random documents contained only two responsive documents, which were insufficient to "kick-start" the active-learning process. A comparison of the results for SAL and SAL-seedran shows the same general effect, including the degraded performance caused by random seeding for Matter 203.

A comparison of the results for SPL and SPL-seedkey shows that, as for CAL and SAL, the use of keyword selection for the *initial* training set generally yields superior results to random selection. A comparison of the results for SPL and SPL-allkey shows that, with two exceptions, keyword selection for the *entire* training set is superior to random selection. However, a comparison of the results for SPL-seedkey and SPL-allkey shows neither to be consistently superior; in four cases, using keywords for only the initial training set was superior, and in four cases, using keywords for the entire training set was superior.

In summary, the use of a seed set selected using a simple keyword search, composed prior to the review, contributes to the effectiveness of all of the TAR protocols investigated in this study.

7. DISCUSSION

7.1 Random vs. Non-Random Training

The results presented here do not support the commonly advanced position that seed sets, or entire training sets, must be randomly selected [19, 28] [*contra* 11]. Our primary implementation of SPL, in which all training documents were randomly selected, yielded dramatically inferior results to our primary implementations of CAL and SAL, in which none of the training documents were randomly selected. While it is perhaps no surprise to the information retrieval community that active learning generally outperforms random training [22], this result has not previously

been demonstrated for the TAR Problem, and is neither well known nor well accepted within the legal community.

Perhaps more surprising is the fact that a simple keyword search, composed without prior knowledge of the collection, almost always yields a more effective seed set than random selection, whether for CAL, SAL, or SPL. Even when keyword search is used to select *all* training documents, the result is generally superior to that achieved when random selection is used. That said, even if passive learning is enhanced using a keyword-selected seed or training set, it is still dramatically inferior to active learning. It is possible, in theory, that a party could devise keywords that would render passive learning competitive with active learning, but until a formal protocol for constructing such a search can be established, it is impossible to subject the approach to a controlled scientific evaluation. Pending the establishment and scientific validation of such a protocol, reliance on keywords and passive learning remains a questionable practice. On the other hand, the results reported here indicate that it is quite easy for either party (or for the parties together) to construct a keyword search that yields an effective seed set for active learning.

The principal argument in favor of random selection appears to be the concern that non-randomly selected training examples are "less than representative of the entire population of relevant documents" [19, pages 260-261], and therefore might "bias" the learning method, resulting in the exclusion of certain classes of relevant documents. It is easy to imagine that such an effect might occur with SPL; however, it is more difficult to imagine how such a bias could persist through the CAL process.

There are situations in which a finite random sample used as a training set could exclude an identifiable population of relevant documents. By way of example, consider a collection consisting of 1,000,000 emails and 100,000 spreadsheets, of which 10,000 emails and 1,000 spreadsheets were relevant. A random training set consisting of 1,100 documents would contain about 1,000 emails, of which about 10 were relevant, and about 100 spreadsheets, of which, as likely as not, none would be relevant. A machine-learning method might well infer that spreadsheets generally were not relevant, thereby exhibiting a blind spot. Random training tends to be biased in favor of commonly occurring types of relevant documents, at the expense of rare types. Non-random training can counter this bias by uncovering relevant examples of rare types of documents that would be unlikely to appear in a random sample.

7.2 Continuous vs. Simple Active Learning

The differences between the CAL and SAL results arise, we believe, from differences in the design objectives underlying their training methods. The underlying objective of CAL is to find and review as many of the responsive documents as possible, as quickly as possible. The underlying objective of SAL, on the other hand, is to induce the best classifier possible, considering the level of training effort. Generally, the classifier is applied to the collection to produce a review set, which is then subject to manual review.[6] The use of SAL raises the critical issues of (i) what is meant by the "best" classifier, and (ii) how to determine the point at which the best classifier has been achieved (commonly referred to as "stabilization" in the context of TAR). In this study, we arbitrarily define "best" to minimize the total training and review effort necessary to achieve 75% recall, and sidestep the stabilization issue by affording SAL the luxury of an oracle that determines immediately, perfectly, and without cost, when stabilization occurs. In practice, defining and detecting stabilization for SAL (and also for SPL) is "[p]erhaps the most critical question attendant to the use of technology-assisted review for the production of documents" [19, page 263]. In practice, recall and precision of candidate review sets are typically estimated using sampling, and stabilization is deemed to occur when an aggregate measure, such as F_1, appears to be maximized [17]. The choice of a suitable criterion for stabilization, and the cost and uncertainty of sampling to determine when that criterion has been met [3], are fundamental challenges inherent in the use of SAL and SPL that are not addressed in this study; instead, SAL and SPL have been given the benefit of the doubt.

With CAL, each successive classifier is used only to identify – from among those documents not yet reviewed – the next batch of documents for review. How well it would have classified documents that have already been reviewed, how well it would have classified documents beyond the batch selected for review, or how well it would have classified an independent, identically distributed sample of documents, is irrelevant to this purpose. Once it has served this narrow purpose, the classifier is discarded and a new one is created. Because the TAR process continues until as many as possible of the relevant documents are found, the nature of the of documents to which successive classifiers are applied drifts dramatically, as the easy-to-find relevant documents are exhausted and the harder-to-find ones are sought.

For SAL, where training is stopped well before the review is complete, we observed informally that uncertainty sampling was superior to relevance feedback, consistent with previously reported results in machine learning for text categorization [15]. For CAL, our results indicate relevance feedback to be superior.

7.3 When to Terminate the Review

Regardless of the TAR protocol used, the question remains: When to terminate the review? The answer hinges on the proportionality considerations outlined in (U.S.) Federal Rules of Civil Procedure 26(b)(2)(C) and 26(g)(1)(B)(iii), which, respectively, limit discovery if "the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues," and require that discovery be "neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action."

Whether the termination point is determined at stabilization (as for SAL and SPL), or deferred (as for CAL), eventually a legal decision must be made that a reasonable review has been conducted, and that the burden or expense of continuing the review would outweigh the benefit of any additional documents that might be found. The density of responsive documents discovered by CAL appears to fall off monotonically, thus informing the legal decision maker how

[6] In some circumstances – which have not been considered in this study – the review set may be produced to the requesting party without any subsequent review.

much effort would be necessary to find more documents; moreover, Cormack and Mojdeh [7] note that the scores of the responsive documents tend to a normal distribution, and that by fitting such a distribution to the scores, it is possible to estimate recall without resorting to sampling. That said, we leave to future research the issue of how best to determine when to stop.

7.4 Imperfect Training

It has been argued that the accuracy of the human review of the training set is critical, and that a "senior partner" [1, page 184], or even a bi-party committee, should review the training documents [2, page 7]. While the existing scientific literature indicates this concern to be overstated [6, 20, 27], our results further confirm that superior results can be achieved using a single, fallible reviewer. That said, a limitation of our evaluation toolkit is that our simulated reviewer always codes a given document the same way; a real reviewer would be influenced by factors such as the prevalence of responsive documents among those reviewed [23], the order in which the documents were reviewed, and any number of other human factors. We conjecture that these factors would tend to benefit CAL over the other protocols because: (i) the prevalence of responsive documents among those reviewed would be higher, especially at the outset of the review; (ii) similar documents would tend to be reviewed together by virtue of having similar scores; and (iii) the reviewer would gain early insight into the nature of responsive documents without having to wade through a haystack of random or marginal documents looking for an unfamiliar needle. Knowledge of the legally significant documents early in the review process is valuable in its own right. We leave it to future research to confirm or refute our conjecture.

7.5 Limitations

The prevalence of responsive documents in the eight review tasks varies from 0.25% to 3.92%, which is typical for the legal matters with which we have been involved. Others assert that these are examples of "low-prevalence" or "low-richness" collections, for which TAR is unsuitable [19]. We suggest that such assertions may presuppose an SPL protocol [11], which is not as effective on low-prevalence datasets. It may be that SPL methods can achieve better results on higher-prevalence collections (*i.e.*, 10% or more responsive documents). However, no such collections were included in this study because, for the few matters with which we have been involved where the prevalence exceeded 10%, the necessary training and gold-standard assessments were not available. We conjecture that the comparative advantage of CAL over SPL would be decreased, but not eliminated, for high-prevalence collections.

Our evaluation toolkit embodies a number of design choices, the effects of which remain to be explored. Our choices for feature engineering and learning algorithm are state of the art for text classification [5, chapter 11], and we have no indication that another choice would yield materially different results. We reprised most of the experiments in this study using logistic regression, instead of SVM, achieving similar results. A naïve Bayes classifier, on the other hand, achieved generally inferior results overall, but the same relative effectiveness among the protocols. A full exploration of feature engineering and classifier choices remains the subject of future research.

Finally, our use of a batch size of 1,000 was occasioned by efficiency considerations. In each of 100 iterations, we augmented the training set by 1,000 documents, trained the classifier, and scored every document in the collection. Each simulation required several hours of computation; the study required several weeks. For the CAL protocol only, we reran the simulations using a batch size of 100 – entailing ten times as much computation (*i.e.*, several days per simulation) – and achieved slightly better results. The effect of even smaller batch sizes on the effectiveness of TAR protocols remains an open question.

8. CONCLUSION

While the mechanisms and efficacy of active machine learning are well known to the information retrieval community, the legal community has been slow to adopt such technologies, which could help address the growing volume of electronically stored information in (U.S.) legal proceedings. Much of the resistance, we submit, is due to lack of awareness of differences among TAR methods and protocols, and over generalization from one TAR method (typically, a variant of SPL) to all TAR.

Because SPL can be ineffective and inefficient, particularly with the low-prevalence collections that are common in e-discovery, disappointment with such tools may lead lawyers to be reluctant to embrace the use of *all* TAR. Moreover, a number of myths and misconceptions about TAR appear to be closely associated with SPL; notably, that seed and training sets must be randomly selected to avoid "biasing" the learning algorithm.

This study lends no support to the proposition that seed or training sets must be random; to the contrary, keyword seeding, uncertainty sampling, and, in particular, relevance feedback – all non-random methods – improve significantly ($P < 0.01$) upon random sampling.

While active-learning protocols employing uncertainty sampling are clearly more effective than passive-learning protocols, they tend to focus the reviewer's attention on marginal rather than legally significant documents. In addition, uncertainty sampling shares a fundamental weakness with passive learning: the need to define and detect when stabilization has occurred, so as to know when to stop training. In the legal context, this decision is fraught with risk, as premature stabilization could result in insufficient recall and undermine an attorney's certification of having conducted a reasonable search under (U.S.) Federal Rule of Civil Procedure 26(g)(1)(B).

This study highlights an alternative approach – continuous active learning with relevance feedback – that demonstrates superior performance, while avoiding certain problems associated with uncertainty sampling and passive learning. CAL also offers the reviewer the opportunity to quickly identify legally significant documents that can guide litigation strategy, and can readily adapt when new documents are added to the collection, or new issues or interpretations of relevance arise.

There is no reason to presume that the CAL results described here represent the best that can be achieved. Any number of feature engineering methods, learning algorithms, training protocols, and search strategies might yield substantive improvements in the future. The effect of review order and other human factors on training accuracy, and thus overall review effectiveness, may also be substantial.

Nevertheless, the experimental protocol, evaluation toolkit, and results presented here provide a foundation for further studies to investigate these and other possible approaches to improve the state of the art in TAR for e-discovery.

9. ACKNOWLEDGEMENT

Cormack's research is supported by a Discovery grant and a Research Tools and Instruments grant from the Natural Sciences and Engineering Research Council of Canada.

10. REFERENCES

[1] *Da Silva Moore* v. *Publicis Groupe*, 287 F.R.D. 182, S.D.N.Y., 2012.

[2] Case Management Order: Protocol Relating to the Production of Electronically Stored Information ("ESI"), *In Re: Actos (Pioglitazone) Products Liability Litigation*, MDL No. 6:11-md-2299, W.D. La., July 27, 2012.

[3] M. Bagdouri, W. Webber, D. D. Lewis, and D. W. Oard. Towards minimizing the annotation cost of certified text classification. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 989–998, 2013.

[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter? In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674, 2008.

[5] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press, 2010.

[6] J. Cheng, A. Jones, C. Privault, and J.-M. Renders. Soft labeling for multi-pass document review. *ICAIL 2013 DESI V Workshop*, 2013.

[7] G. V. Cormack and M. Mojdeh. Machine learning for information retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks. *The Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.

[8] M. R. Grossman and G. V. Cormack. 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):1–48, 2011.

[9] M. R. Grossman and G. V. Cormack. Inconsistent responsiveness determination in document review: Difference of opinion or human error? *Pace Law Review*, 32(2):267–288, 2012.

[10] M. R. Grossman and G. V. Cormack. The Grossman-Cormack glossary of technology-assisted review with foreword by John M. Facciola, U.S. Magistrate Judge. *Federal Courts Law Review*, 7(1):1–34, 2013.

[11] M. R. Grossman and G. V. Cormack. Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review." *Federal Courts Law Review*, 1, to appear 2014.

[12] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the TREC 2009 Legal Track. *The Eighteenth Text REtrieval Conference (TREC 2009)*, 2009.

[13] C. Hogan, J. Reinhart, D. Brassil, M. Gerber, S. Rugani, and T. Jade. H5 at TREC 2008 Legal Interactive: User modeling, assessment & measurement. *The Seventeenth Text REtrieval Conference (TREC 2008)*, 2008.

[14] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[15] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[16] D. W. Oard and W. Webber. Information retrieval for e-discovery. *Information Retrieval*, 6(1):1–140, 2012.

[17] Y. Ravid. *System for Enhancing Expert-Based Computerized Analysis of a Set of Digital Documents and Methods Useful in Conjunction Therewith*. United States Patent 8527523, 2013.

[18] H. L. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

[19] K. Schieneman and T. Gricks. The implications of Rule 26(g) on the use of technology-assisted review. *Federal Courts Law Review*, 7(1):239–274, 2013.

[20] J. C. Scholtes, T. van Cann, and M. Mack. The impact of incorrect training sets and rolling collections on technology-assisted review. *ICAIL 2013 DESI V Workshop*, 2013.

[21] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[22] B. Settles. *Active learning literature survey*. University of Wisconsin, Madison, 2010.

[23] M. D. Smucker and C. P. Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602, 2010.

[24] S. Tomlinson. Learning Task experiments in the TREC 2010 Legal Track. *The Nineteenth Text REtrieval Conference (TREC 2010)*, 2010.

[25] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

[26] W. Webber, D. W. Oard, F. Scholer, and B. Hedin. Assessor error in stratified evaluation. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 623–632, 2010.

[27] W. Webber and J. Pickens. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 929–932, 2013.

[28] C. Yablon and N. Landsman-Roos. Predictive coding: Emerging questions and concerns. *South Carolina Law Review*, 64(3):633–765, 2013.