Constructing Test Collections using Multi-armed Bandits and Active Learning

Md Mustafizur Rahman
The University of Texas at Austin
nahid@utexas.edu

Mucahid Kutlu
TOBB Economy and Tech. University
m.kutlu@etu.edu.tr

Matthew Lease
The University of Texas at Austin
ml@utexas.edu

ABSTRACT

While test collections provide the cornerstone of system-based evaluation in information retrieval, human relevance judging has become prohibitively expensive as collections have grown ever larger. Consequently, intelligently deciding which documents to judge has become increasingly important. We propose a two-phase approach to intelligent judging across topics which does not require document rankings from a shared task. In the first phase, we dynamically select the next topic to judge via a multi-armed bandit method. In the second phase, we employ active learning to select which document to judge next for that topic. Experiments on three TREC collections (varying scarcity of relevant documents) achieve $\tau \approx 0.90$ correlation for P@10 ranking and find 90% of the relevant documents at 48% of the original budget. To support reproducibility and follow-on work, we have shared our code online[1].

KEYWORDS

Information Retrieval; Evaluation; Active Learning; Multi-Armed Bandits

ACM Reference Format:
Md Mustafizur Rahman, Mucahid Kutlu, and Matthew Lease. 2019. Constructing Test Collections using Multi-armed Bandits and Active Learning. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313675

1 INTRODUCTION

Cranfield-based evaluation of information retrieval (IR) systems [4, 23] relies on the construction of a *test collection*: a document collection, a set of search topics, and human judgments of document relevance for those search topics. However, large-scale IR evaluation is becoming increasingly challenging and economically infeasible because there are too many documents in the collection to be judged. Consequently, there is a growing need for improved methods to create test collections at minimal cost.

Developing a reliable, low-cost IR test collection requires intelligent budget allocation across search topics and careful selection of documents for human annotation. However, existing approaches [28, 32] either allocate the same budget across all topics or estimate a budget per topic rather than allocating the total budget dynamically across topics based on their judgment needs. For example, in TREC, each topic is traditionally allotted a budget that equals the size of its *top-k* pool, constructed by pooling the highest ranked k documents from the submitted runs for that topic.

However, developing a test collection by running a shared task has several drawbacks. First of all, running a shared task is difficult, slow, and expensive. In some cases, it may be nearly impossible to garner enough participants, such as for less studied languages (e.g., Turkish) or search tasks (e.g., historical search). In addition, Li and Kanoulas [17] recently noted that some document selection approaches based on participant rankings [9, 18] may overly bias selection toward the best-performing runs. Furthermore, Voorhees [28] reports empirical evidence that some dynamic document selection methods [18] can also produce less reusable test collections.

Consequently, if our goal is simply to build a new test collection at minimal cost, it would be preferable to be able to do this without having to run a shared task [26]. However, this poses its own set of challenges. In the absence of document rankings from shared task participants, we must find another means to select which documents should be judged for relevance to search topics. In addition, while the size of the *top-k* pools for different topics are known to vary widely, since we have no document rankings with which to apply pooling, we must find another means of allocating the judging budget across search topics.

We propose a two-phase approach to constructing test collections without needing to organize a shared task. Because it is known that evaluation becomes more reliable as more relevant documents are found [24], our main goal is to find as many relevant documents as possible for a given budget. In the first phase (topic selection), we want to select whichever topic is most likely to supply relevant documents. In the second phase (document selection), we want to select one or more documents for the given topic that are the most likely to be relevant. We implement the first phase via a multi-armed bandit (MAB) [21] method. We are not familiar with any prior work exploring MAB for intelligent topic selection. During the second phase, we investigate several alternative active learning [25] (AL) strategies to select which documents to judge for each search topic. Finally, we use the collected relevance judgments to update both the active learning classifier for document selection and the bandit statistics for topic selection, thereby iteratively improving topic and document selection in successive rounds.

Across three TREC collections with varying scarcity of relevant documents, our best method achieves $\tau \approx 0.90$ Kendall correlation on average with the original system ranking for P@10, using less than 50% of the budget allotted in the original pool. Results for MAP-based evaluation are weaker but still encouraging, suggesting the approach is promising but future work is still needed.

[1] https://github.com/mdmustafizurrahman/MAB_AL_TestCollection/

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313675

2 RELATED WORK

A considerable amount of research [1, 2, 8, 9, 17–20, 30, 31] has been conducted to select documents for human annotation. These approaches can be mainly categorized into two groups: i) static [1, 2, 19, 30, 31] and ii) dynamic [8, 9, 17, 18, 20] selection. In a static selection process, the relevance judgment process is initiated only after the document selection process is completed. In contrast, in a dynamic selection process, document selection and relevance judgment process occur in a feedback loop, where the relevance judgments that are collected so far affect the selection process of the subsequent documents. Our proposed approach is closely related to the dynamic selection approaches. However, our approach does not require organizing a shared-task while prior work relies on the runs submitted by participants of a shared-task.

2.1 Document Selection Strategies

Various dynamic document selection methods have been proposed to intelligently select which of the documents retrieved in a shared task should be judged. The *Move-to-front* (MTF) pooling [9] is one of the pioneer examples of the dynamic document selection process. For a given topic, MTF maintains a priority list of submitted runs. At any round, MTF selects the top-ranked unjudged document from the current highest priority run for annotation. Then based on the relevance judgment of the selected document, MTF updates the priority of the selected run and moves to the next round.

Carterette et al. [3] propose to construct minimal test collections to judge ranking systems with confidence. They select documents to be judged iteratively such that the document that might differentiate systems' performances the most is selected at each iteration. However, their method completely depends on submissions from multiple systems, while the main goal of our work is eliminating the massive cost of running shared-tasks.

Rajput et al. [20] develop a framework to construct test collections using an iterative reinforcement method between nuggets and documents. Their iterative process starts with a few manually created nuggets which are utilized to select documents for the relevance judgment. The relevant documents found in the relevance judgment are used to update the weights of the existing nuggets and extract new nuggets automatically which are then used in the subsequent iteration of document selection. However, should automatic nugget extraction fails (e.g. Web Track 2009), their document selection process solely depends on the submitted runs of the shared task, whereas we consider a no shared task context.

Losada et al. [18] develop a document selection process by utilizing several multi-armed bandits [21] approaches. Their best performing approach, the *MaxMean* (MM) approach, assigns a weight to a run proportional to the ratio of the number of relevant documents found, and the total number of documents judged so far for that run; then it selects the run with the highest weight. Losada et al. empirically prove their superiority over MTF in terms of the number of relevant documents found.

However, Li and Kanoulas [17] point out that dynamic document selection based on a shared task context creates a bias towards the good runs (e.g., runs with more relevant documents). Their proposed approach also utilizes a shared task by inducing a probability distribution from the participating systems, and a probability distribution over the ranks of the documents; then it actively samples documents from the joint distribution to construct a test collection.

Recently, Voorhees [28] further supports Li and Kanoulas's statement by developing the TREC 2017 Common Core track collection via utilizing *MaxMean*(MM) proposed by Losada et al. [18]. The author finds out that only a single run contributes the large percentage of relevant documents in the collection [28]. In contrast, our test collection construction approach in the absence of a shared task is free from this bias towards good runs.

The use of AL in our document selection phase closely follows work by Cormack and Grossman [7] (and we adopt their nomenclature). However, their study is situated in a rather different domain, "e-discovery", focusing on set-based retrieval rather than ranked retrieval. Moreover, the judging cost is measured differently in this domain: no document can be selected automatically since all of the documents must be reviewed for privilege following discovery.

2.2 Topic Selection Strategies

Previous studies [11, 12, 14, 24, 28, 32] treat topic selection as a budget allocation problem. These studies either allocate an immutable budget for a topic at the outset of the relevance judgment using various information e.g., the number of relevant documents at the *top-k* pool [28, 32], or recommend to utilize many topics with a fewer number of relevance judgments [24].

Zobel [32] proposes to collect more judgments for topics with a higher estimated number of relevant documents. The estimation is computed using the relevance judgment collected for each topic in its *top-k* pool, and according to Zobel, it follows the power law distribution with the rank of the document. Voorhees follows the same *top-k* pool with a different formulation ([28], Page 411) to estimate the budget for a topic. However, none of the approaches treat the budget allocation as a topic selection problem as we have done here, where the selection of a subsequent topic for relevance judgment depends on the relevance judgments collected so far. Besides, both of the approaches [28, 32] utilize a shared task which is absent in our task setting.

Prior work [11, 12, 14] also investigates the possibility of selecting a subset of topics. Guiver et al. [11] experimentally show that if we can find a "right" subset of topics, we can achieve a ranking of systems that is very similar to the ranking when all of the topics are used for evaluation. Even though they do not propose a method for finding the "right" subset of topics, their work motivated many researchers to work on the topic selection problem [12, 14]. Studies on topic selection treat each selected topic equally (i.e. collect the same amount of judgments per each), while no judgments are selected for the topics that are not selected. On the contrary, we collect a varying number of judgments per topic.

3 PROPOSED APPROACH

Our approach consists of two phases which alternate between topic selection and document selection. In topic selection phase, we estimate the prevalence of relevant documents for each topic, and select the topic with the highest estimated prevalence. In document selection phase, we use a classifier to predict the relevance of all unjudged documents for the selected topic, and use those predictions

to select which documents to judge next for that topic. We then use the collected relevance judgments to update both the given classifier and the topic statistics. This iterative process of topic and document selection continues until the budget is exhausted.

Algorithm 1 describes our two-phase approach in detail.

Algorithm 1: Two-phase Topic and Document Selection

Input : Unjudged documents U • batch size N • Budget b • Document selection policy p
Output: Relevance judgments $R^{1:J}$ for topics $1:J$

$R^{1:J} \leftarrow \emptyset$
for *topic* $j \leftarrow 1$ **to** J **do**
  Select seed document set $S^j \in U$ for topic j
  $R^j \leftarrow$ judge_relevance (j, S^j) ▷ Collect judgments
  $U \leftarrow U - S^j$ ▷ Update set of unjudged documents
  $b \leftarrow b - |S^j|$ ▷ Update remaining budget
  $C^j \leftarrow$ train_classifier (S^j, R^j)
  $\alpha_j \leftarrow 1, \beta_j \leftarrow 1$ ▷ Initialize Beta distribution for bandit

while remaining budget b >= batch size N **do**
  for *topic* $j \leftarrow 1$ **to** J **do**
    $P_j \leftarrow$ sample probability from $Beta(\alpha_j, \beta_j)$
  Selected topic $j \leftarrow \underset{j}{\text{argmax}}\, P_j$
  $\widehat{R^j} \leftarrow predict\_relevance(C^j, U)$ ▷ for all unjudged
  S^j = select_documents (U, $\widehat{R^j}$, batch size N, policy p)
  $R^j \leftarrow$ judge_relevance (j, S^j) ▷ Collect judgments
  $U \leftarrow U - S^j$ ▷ Update set of unjudged documents
  $b \leftarrow b - N$ ▷ Update remaining budget
  $C^j \leftarrow$ update_classifier (C^j, S^j, R^j)

  for *document* $i \leftarrow$ **to** N **do** ▷ Update bandits
    if $R^j(i)$ *is relevant* **then**
      $\alpha_j \leftarrow \alpha_j + 1$
    else
      $\beta_j \leftarrow \beta_j + 1$

3.1 Phase 1: Topic Selection

Each topic may need different number of judgments based on the number of relevant documents found in the collection. Thus, allocating a static pre-defined budget for all topics may incur more/less cost than required for topics. Therefore, we prefer to allocate the budget across topics dynamically based on the needs of each topic. Given the fact that we do not have any prior knowledge about the number of judgments needed for a topic, an exploration-exploitation situation arises naturally, which can be approached using the multi-armed bandits (MAB) techniques [21].

In the standard MAB problem, we have to repeatedly make a choice among J bandits (or slot machines), each of which has a hidden probability of winning and losing. Depending on the slot machine j we select at round t, we receive a reward (e.g. 1 for winning, or 0 for losing). In the long run, we want to maximize our final reward. This multi-armed bandits problem is a natural fit for our topic selection problem. Each topic j can be referred to as a slot machine and each of these topics has a hidden probability of supplying a relevant document. At each round t, based on the selected topic j, we either receive a relevant or non-relevant document. At any round t, we can estimate which topic is more likely to provide more relevant documents. The exploration-exploitation dilemma arises when deciding between whether we should keep selecting that topic or we should explore other topics; we do not know which decision will help us maximize the number of relevant documents.

To allocate the budget across topics dynamically, we apply a Bayesian approach [10] where each topic's hidden probability distribution over relevant and non-relevant documents is endowed with a prior distribution. With no prior knowledge about the topic's prevalence, we start with a uniform prior for each topic. Note that a uniform distribution is a special case of a $Beta(\alpha, \beta)$ distribution, when $\alpha = 1$ and $\beta = 1$. In our Algorithm 1, α and β denote the number of relevant and non-relevant documents found so far for a topic j, respectively. At any round t, we start by selecting a topic j from the $Beta(\alpha, \beta)$ distribution. Next we select N documents to judge using a document selection process (Section 3.2) for topic j. Finally, the binary outcome, O, (either relevant or non-relevant) for each selected document is used to update the hidden probability distribution of topic j. With a $Beta(\alpha, \beta)$ prior and a binary outcome, the posterior O is also a $Beta(\alpha + O, \beta + 1 - O)$ distribution.

3.2 Phase 2: Document Selection

In this phase, we select N documents to judge for a given topic j. However, in the absence of a shared-task, we do not have any ranking information of documents. Therefore, if we select documents randomly, it is very unlikely to find any relevant document. Thus, we employ active learning [25], a learning paradigm where a classifier decides which document should be annotated next in order to maximize the learning curve of the classifier.

3.2.1 Task Definition and Learning Model. To train a topic-specific active classifier, we must collect the topic-specific training data. Let us assume that a training pair for a topic j is denoted by $\left\langle x^i, y^i_j \right\rangle$, where x^i denotes the feature representation of document, i and y^i_j denote the binary relevance judgment for <document i, topic j>. We adopt logistic regression as a classifier to infer the probability of relevance $P(y^i_j|x^i)$ of each document x^i for topic j:

$$p(y^i_j|x^i) = h_\theta(x_i) = \frac{1}{1 + \exp(-\vec{\theta}^T x_i)} \tag{1}$$

with $\vec{\theta} \in \mathbb{R}^D$ denotes model parameters.

3.2.2 Document Selection Criteria. Utilizing the posteriori probability of documents for a given topic j, we can select documents to be judged in two different ways: *Simple Active Learning* (SAL) and *Continuous Active Learning* (CAL).

SAL [15] selects a document for relevance judgment when the classifier is most uncertain about the label of that document. We employ an entropy-based uncertainty function [25] for this:

$$Uncertainty(x) = -\sum_{y \in Y} P(y|x) \log P(y|x) \tag{2}$$

Table 1: Test collection statistics. As collections have grown larger, judging budgets have also shrunk, leading to increased prevalence of relevant documents in later tracks.

Track	Collection	Topics	#Docs	#Judged	%Rel
WT'14[6]	ClueWeb12[4]	251-300	52M	14,432	39.2%
WT'13[5]	ClueWeb12	201-250	52M	14,474	28.7%
TREC-8[29]	Disks45-CR[5]	401-450	528K	86,830	5.4%

where y is either relevant or non-relevant. With binary relevance, SAL selects:

$$x^{\star} = \underset{i}{\operatorname{argmin}} \ |p(relevant|x^i) - 0.5| \tag{3}$$

In contrast, CAL selects a document based on how likely the document is to be relevant according to the prediction of the classifier.

$$x^{\star} = \underset{i}{\operatorname{argmax}} \ p(relevant|x^i) \tag{4}$$

3.2.3 Seed Document Selection. In order to learn an initial classifier for each topic, a minimal *seed set* of relevance judgments for each topic is required. We assume a *single* off-the-shelf or in-house IR system (e.g., Apache Lucene[2] or Indri[3]) is used to produce a document ranking for each topic. The assessor is then asked to proceed down the document ranking for each topic until at least m number of relevant and m number of non-relevant documents have been found, or some maximum effort is reached without success, in which case, the topic is discarded. In this paper, we assume $m = 5$. Note that this single IR system is used only once at the outset to guide seed set annotation, and one might even do without this if one is willing to resort to boolean search or random selection to identify the seed documents [7].

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. We conduct our experimental evaluation on three TREC test collections (See **Table 1**). We assume binary relevance and collapse NIST graded relevance judgments to binary.

We use only pooled documents of each collection because unpooled documents are not judged and typically assumed to be not-relevant. However, relevant documents may exist outside the pool, and assuming them to be non-relevant could hurt classifier training and prediction. One could judge the unpooled documents selected by our method, though this could yield inconsistency between the original judgments and new judgments. This issue could be usefully revisited in future work. In the remainder of this paper, we refer to the set of pooled judged documents as the "qrels."

Document Selection. In the document selection phase, we have to represent the documents in a feature vector for the classifier of the active learning approach. We first pre-process the documents using `IndriBuildIndex`[6]. The pre-processing consists of text normalization, stopword [16] removal, and Krovetz stemming [13]. Finally, each pre-processed document is represented using a 15K dimensional TF-IDF [22] vector. Each iteration of AL selects one document to be judged next. In order to select the seed documents of AL, we randomly select one of the runs for each test collection and assume it as our off-the-shelf IR system.

Baselines. We have one baseline method for document selection and another baseline method for topic selection.

- **Simple Passive Learning (SPL):** This baseline method for document selection uses uniform random selection.
- **Round-Robin (RR):** RR simply cycles through topics and then repeats, thus allocating the same budget for each topic. No information is used regarding topic prevalence ratio or collected relevance judgments.

Upperbound. We develop **Oracle** as the upperbound for our topic selection method. Whereas MAB must learn the prevalence ratio of topics during dynamic topic selection, the Oracle knows the exact prevalence ratio of each topic at each round t. To do this, we initialize the α and β parameters of each topic with the topic's total number of relevant and non-relevant documents in the original qrels. At each round t, we update the parameters, decreasing α and β based on the relevance judgments collected during that round.

Table 2: Avg. number of relevant documents found under varying budget per topic on TREC-8. For MoveToFront (MTF), MaxMean Non-Stationary (MM-NS), and MAB+CAL topic selection, MAB+CAL consistently performs best.

	Average number of judgments per topic						
Method	**100**	**300**	**500**	**700**	**900**	**1100**	**all**
MTF	34.06	58.48	71.78	79.22	84.5	87.58	94.04
MM-NS	36.96	64.62	77.3	82.5	86.34	89.2	94.04
MAB+CAL	**46.3**	**78.4**	**86.5**	**90.3**	**91.3**	**93.5**	94.04

4.2 Results and Discussion

Effectiveness of topic selection. How well can MAB identify relevant documents? **Figure 1** compares the recall of MAB with Oracle and RR for WT2014 and WT2013 collections. The overall Area Under Curve (AUC) effectiveness across all budget points is reported in Figure 1. The x-axis of each plot represents the total allotted budget. Ultimately, each strategy achieves complete recall because all pooled documents are judged. As expected, RR is the weakest performing topic selection method across all six plots of Figure 1, suggesting that we should select topics intelligently to maximize recall of relevant documents. Also as expected, Oracle performs best in every case, and the recall curve of MAB consistently lies between Oracle and RR.

Dynamic vs. random document selection. Figure 1 also reports the performance of the three document selection strategies. We observe that CAL outperforms the SPL baseline in both test collections and in all topic selection strategies. However, we cannot infer the same conclusion about SAL because SAL outperforms SPL only in 3/6 of cases based on AUC scores. CAL selects the documents that are most likely to be relevant, yielding a higher recall than others. On the other hand, SAL selects documents whose relevance is uncertain, and thus tends to select more non-relevant documents, ultimately decreasing its recall.

The allotted budget for each topic. In **Figure 2**, we report the budget allotted to each topic using CAL for document selection, and MAB and RR for topic selection on TREC8, WT2013, and WT2014

[2] https://lucene.apache.org/
[3] https://www.lemurproject.org/indri.php
[6] www.lemurproject.org/indri.php

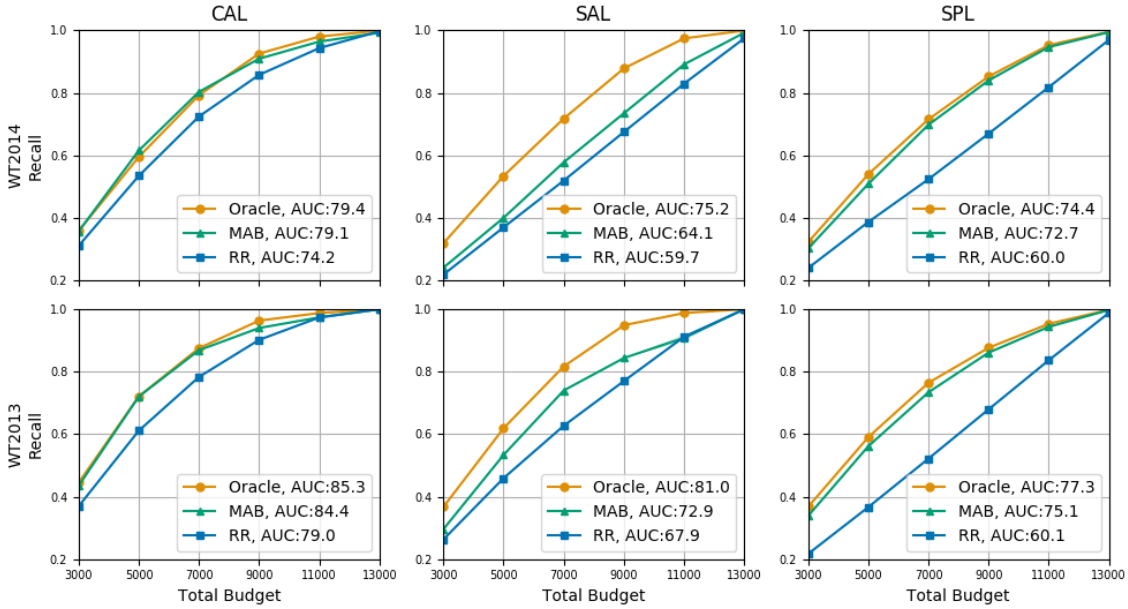

Figure 1: Recall of relevant documents achieved by Oracle, Multi-armed bandits (MAB), and Round-robin (RR) topic selection methods as a function of varying evaluation budget on WT2014 and WT2013 collections. The x-axis shows total cost over all topics. Plots are grouped vertically for CAL, SAL, and SPL document selection methods.

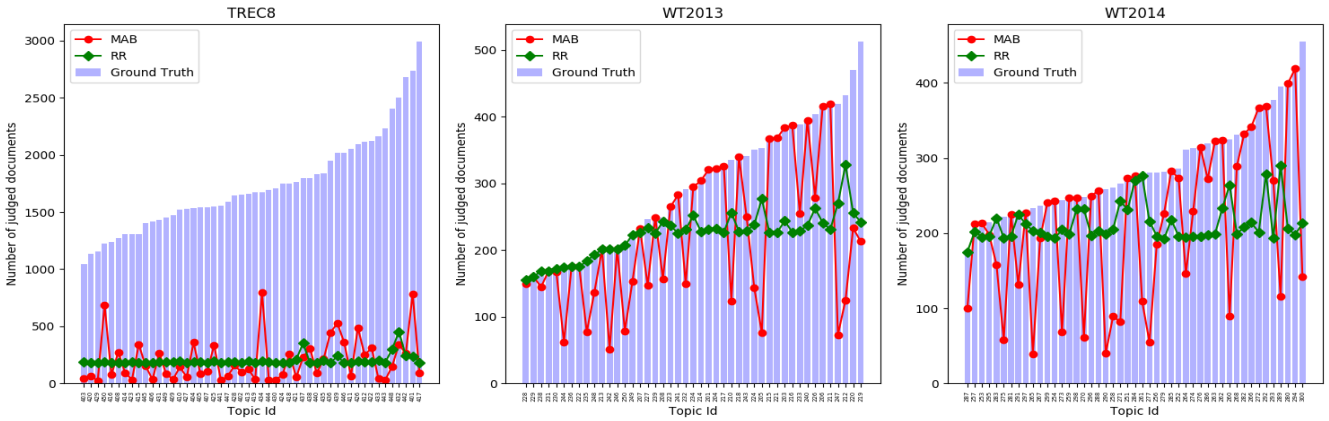

Figure 2: The budget (# of judged documents) across topics using MAB and RR with CAL document selection on TREC8, WT2013 and WT2014 test collections with total budget of 10,000. The bar plots show per-topic budget in the original qrels.

when the overall budget is 10,000. We can see that MAB-based topic selection varies the judging budget used across topics. For RR, the number of judgments per topic varies only due to the varying number of seed documents used to start AL. Ignoring the cost of seed documents, all topics would have the same number of judgments per topic in RR. Thus as expected, RR has a lower variance in budget allocation per topic than MAB.

Comparison against pool-based methods. We compare our approach, MAB with CAL (MAB+CAL) against two state-of-the-art pool-based dynamic document selection methods: MoveToFront (MTF) [9] and the MaxMean Non-Stationary (MM-NS) [18]. **Table 2** presents the average number of relevant documents found per topic of TREC-8 collection. We copy the MM and MTF results reported in [18]. Note that for MTF and MM-NS, we have to manually specify a fixed budget per topic. In contrast, MAB+CAL intelligently learns the budget allocation per topic dynamically.

Table 2 shows that MAB+CAL finds a higher average number of relevant documents per topic than MTF and MM-NS in all cases, *despite* the fact that both MTF and MM-NS utilize the document ranking information from the submitted runs in a shared task.

Rank Correlation. One of the ways to evaluate the reliability of a low-cost evaluation method is to compare the resultant ranking of IR systems with the ranking produced by all qrels. Thus, we build qrel sets using CAL for document selection, and RR, MAB, and

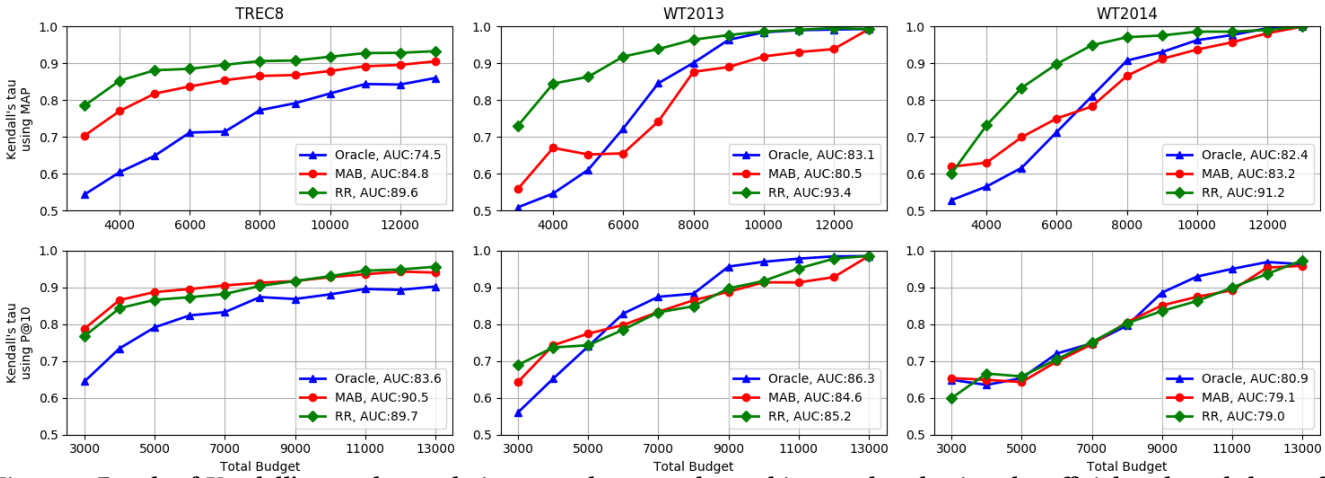

Figure 3: Result of Kendall's τ rank correlation score between the ranking produced using the official qrels and the ranking produced using qrels created by Oracle, MAB and RR with CAL over varying budgets on TREC8, WT2014 and WT2013 collections. Rankings are produced using MAP (top row) and P@10 (bottom row).

Oracle for topic selection varying the total budget from 3K to 13K, and calculate Kendall's τ rank correlation between ground truth ranking and the resultant ranking for each case. We use MAP and P@10 as the evaluation metric. The results are shown in **Figure 3** for TREC8, WT2013 and WT2014 test collections.

For the bottom row of Figure 3 (Kendall's τ rank correlation for P@10 evaluation), we see that with low judging budget, MAB provides better τ correlation than RR for test collections with lower prevalence ratio per topic (i.e., TREC8). Specifically, MAB achieves $\tau = 0.9$ (a traditionally-accepted threshold for acceptable correlation [27]) when the total budget is 6500, which is only 7.4% of the original allotted budget for TREC8. However, results are less clear for test collections with a higher prevalence ratio. For example, MAB outperforms RR for WT2014 but RR outperforms MAB for WT2013 in terms of AUC.

When we compute Kendall's τ for MAP evaluation (Figure 3, Top Row), both Oracle and MAB actually perform worse than RR across test collections. Since MAP is a recall-based metric, it considers the full judgment pool for each topic, whereas P@10 takes account only the first 10 documents. This suggests the current MAB approach may work well for shallow metrics but not for deep ones.

Recall that RR allocates the same budget across all topics, whereas MAB-based topic selection yields a varying number of documents across topics (Figure 2). This is because MAB seeks to optimize the number of relevant documents [9, 18]. Oracle optimizes this same objective function as MAB but even better, knowing the exact prevalence ratio of each topic, which MAB must learn. Results thus suggest that our MAB-based approaches spend the budget excessively on topics with high prevalence ratio, causing judging few documents for the topics with low prevalence ratio. Sometimes, MAB outperforms Oracle seemingly because its imperfect estimate of prevalence leads it to explore more and exploit less. Thus, while MAB approaches succeed in finding more relevant documents, we observe a metric divergence between what is being optimized (recall) and what we actually care about: reliable evaluation, as measured by rank correlation. Moreover, results underscore the importance of evaluating test collection creation methods via rank correlation, and not only by recall of relevant documents [18].

Furthermore, when we compare the results across collections for MAP score in Figure 3, we find that the performance difference between RR and MAB-based approaches are much higher in collections with a high prevalence ratio (i.e., WT2013 and WT2014) than a low prevalence ratio (i.e., TREC8). This signifies the fact that in terms of MAP, MAB-based approach produces a more reliable evaluation for test collections with a low prevalence ratio than test collections with a high prevalence ratio. This is another promising result of our MAB-based approach which encourages further empirical evaluation.

5 CONCLUSION AND FUTURE WORK

This work investigates the feasibility of developing a minimal cost test collection without organizing a shared task. In the absence of shared task document rankings to prioritize relevance judging, we utilize active learning [25] instead. Furthermore, we intelligently and dynamically allocate judging effort across topics via multi-armed bandits [21]. Results on the three different TREC test collections indicate that by utilizing an average of only 48% of the total budget (Table 1) allocated in these test collections, our best approach, MAB+CAL, not only finds 90% of the relevant documents (Figure 1) but also achieves a Kendall's τ rank correlation value of 0.90 with the original system ranking when the correlation is computed using P@10 (Figure 3, Bottom Row). However, further work is needed to effectively support more recall-oriented IR evaluation metrics such as MAP. Current results suggest that our MAB-based approaches spend the budget excessively on topics with high prevalence ratio, judging too few documents for topics with low prevalence ratio. Future work might usefully explore more sophisticated bandit approaches (e.g. contextual bandits).

Acknowledgements. This work was supported by NPRP grant # NPRP 7-1313-1-245 from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] Javed A Aslam, Virgil Pavlu, and Emine Yilmaz. 2006. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 541–548.

[2] Javed A Aslam and Emine Yilmaz. 2007. Inferring document relevance from incomplete information. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 633–642.

[3] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 268–275.

[4] Cyril Cleverdon. 1967. The Cranfield tests on index language devices. In *Aslib proceedings*, Vol. 19. MCB UP Ltd, 173–194.

[5] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. 2013. TREC 2013 Web Track Overview. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*.

[6] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. *TREC 2014 web track overview*. Technical Report. DTIC Document.

[7] Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 153–162.

[8] Gordon V Cormack and Maura R Grossman. 2018. Beyond pooling. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1169–1172.

[9] Gordon V Cormack, Christopher R Palmer, and Charles LA Clarke. 1998. Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 282–289.

[10] Ole-Christoffer Granmo. 2008. A Bayesian Learning Automaton for Solving Two-Armed Bernoulli Bandit Problems. In *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications (ICMLA '08)*. IEEE Computer Society, Washington, DC, USA, 23–30. https://doi.org/10.1109/ICMLA.2008.67

[11] John Guiver, Stefano Mizzaro, and Stephen Robertson. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* 27, 4 (2009), 21.

[12] Mehdi Hosseini, Ingemar J Cox, Natasa Milic-Frayling, Milad Shokouhi, and Emine Yilmaz. 2012. An uncertainty-aware query selection model for evaluation of IR systems. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 901–910.

[13] Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 191–202.

[14] Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2018. Intelligent topic selection for low-cost information retrieval evaluation: A New perspective on deep vs. shallow judging. *Information Processing & Management* 54, 1 (2018), 37–59.

[15] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 3–12.

[16] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. SMART stopword list. *Journal of Machine Learning Research* (2004).

[17] Dan Li and Evangelos Kanoulas. 2017. Active Sampling for Large-scale Information Retrieval Evaluation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 49–58.

[18] David E Losada, Javier Parapar, and Álvaro Barreiro. 2016. Feeling lucky?: multi-armed bandits for ordering judgements in pooling-based evaluation. In *proceedings of the 31st annual ACM symposium on applied computing*. ACM, 1027–1034.

[19] V Pavlu and J Aslam. 2007. *A practical sampling strategy for efficient retrieval evaluation*. Technical Report. College of Computer and Information Science, Northeastern University.

[20] Shahzad Rajput, Matthew Ekstrand-Abueg, Virgil Pavlu, and Javed A Aslam. 2012. Constructing test collections by inferring document relevance via extracted relevant information. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 145–154.

[21] Herbert Robbins. 1985. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*. Springer, 169–177.

[22] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[23] Mark Sanderson et al. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[24] Mark Sanderson and Justin Zobel. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 162–169.

[25] Burr Settles. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–114.

[26] Ian M Soboroff. 2013. Building Test Sollections (without running a community evaluation). In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 1132–1132. https://isoboroff.github.io/Test-Colls-Tutorial/Tutorial-slides/.

[27] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management* 36, 5 (2000), 697–716.

[28] Ellen M. Voorhees. 2018. On Building Fair and Reusable Test Collections Using Bandit Techniques. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 407–416. https://doi.org/10.1145/3269206.3271766

[29] Ellen M. Voorhees and Donna Harman. 2000. Overview of the Eighth Text REtrieval Conference (TREC-8). 1–24.

[30] Emine Yilmaz and Javed A Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, 102–111.

[31] Emine Yilmaz and Javed A Aslam. 2008. Estimating average precision when judgments are incomplete. *Knowledge and Information Systems* 16, 2 (2008), 173–211.

[32] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 307–314.