Constructing Test Collections using Multi-armed Bandits and Active Learning

Md Mustafizur Rahman[1], Mucahid Kutlu[2], Matthew Lease[1]
[1]School of Information, [2]Computer Engineering Department,
[1]The University of Texas at Austin, [2]TOBB University of Economics and Technology

Qatar National Research Fund
Member of Qatar Foundation
TEXAS The University of Texas at Austin
TOBB ETÜ University of Economics & Technology

Research Goal

Can we develop a test collection **without organizing** a shared task?

Shortcomings of a shared task

- Organizing a shared task is difficult, slow, and expensive
- Sometimes, it is impossible to garner enough participants, such as for less studied languages (e.g., Turkish) or search tasks (e.g., historical search).

Challenges

- How to allocate budget across topics?
- How to select documents for annotations?

Background

A test collection consists of
I. A collection of documents
II. A set of topics
III. A set of relevance judgments

Test collections are typically constructed
- by organizing a ***shared task,*** where multiple teams participate and submit their document rankings for the given document collection and the set of topics
- by applying the ***pooling***, where the top-ranked ***K*** documents from each submitted ranking system are selected for relevance judging

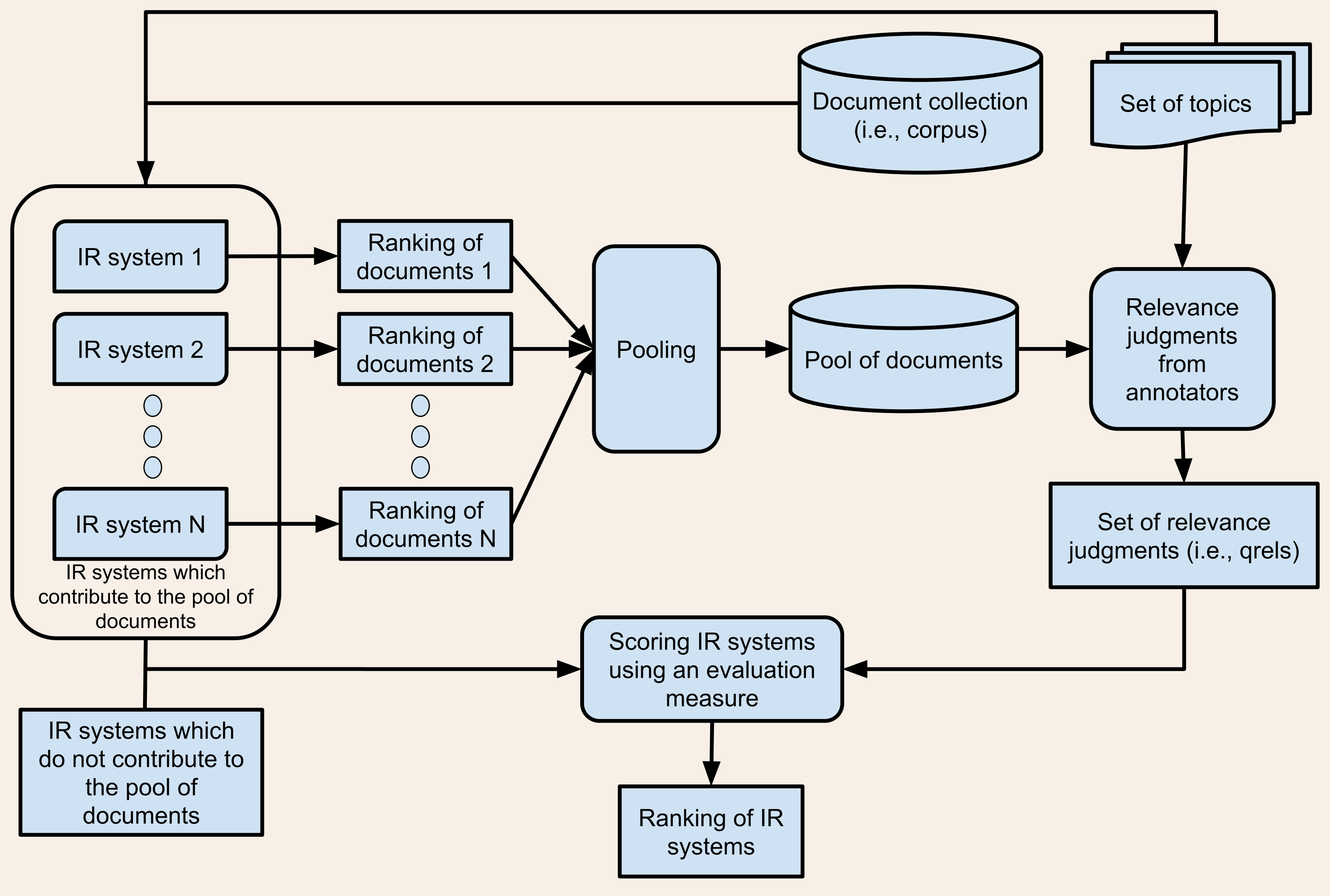

Figure 1: The typical steps involved in the construction of a test collection via pooling in a shared task

Methods

Topic Selection
- Different topics need different number of relevance judgments.
- Allocating a pre-defined budget across topics will incur more cost than actually needed.
- To find out as many as relevant documents, we frame the problem as an exploration-exploitation phenomenon where we
  - Either **exploit** an already selected topic
  - Or **explore** a new topic.
- We solve the exploration-exploitation phenomenon using ***Multi-armed Bandits*** (MAB) technique [1].

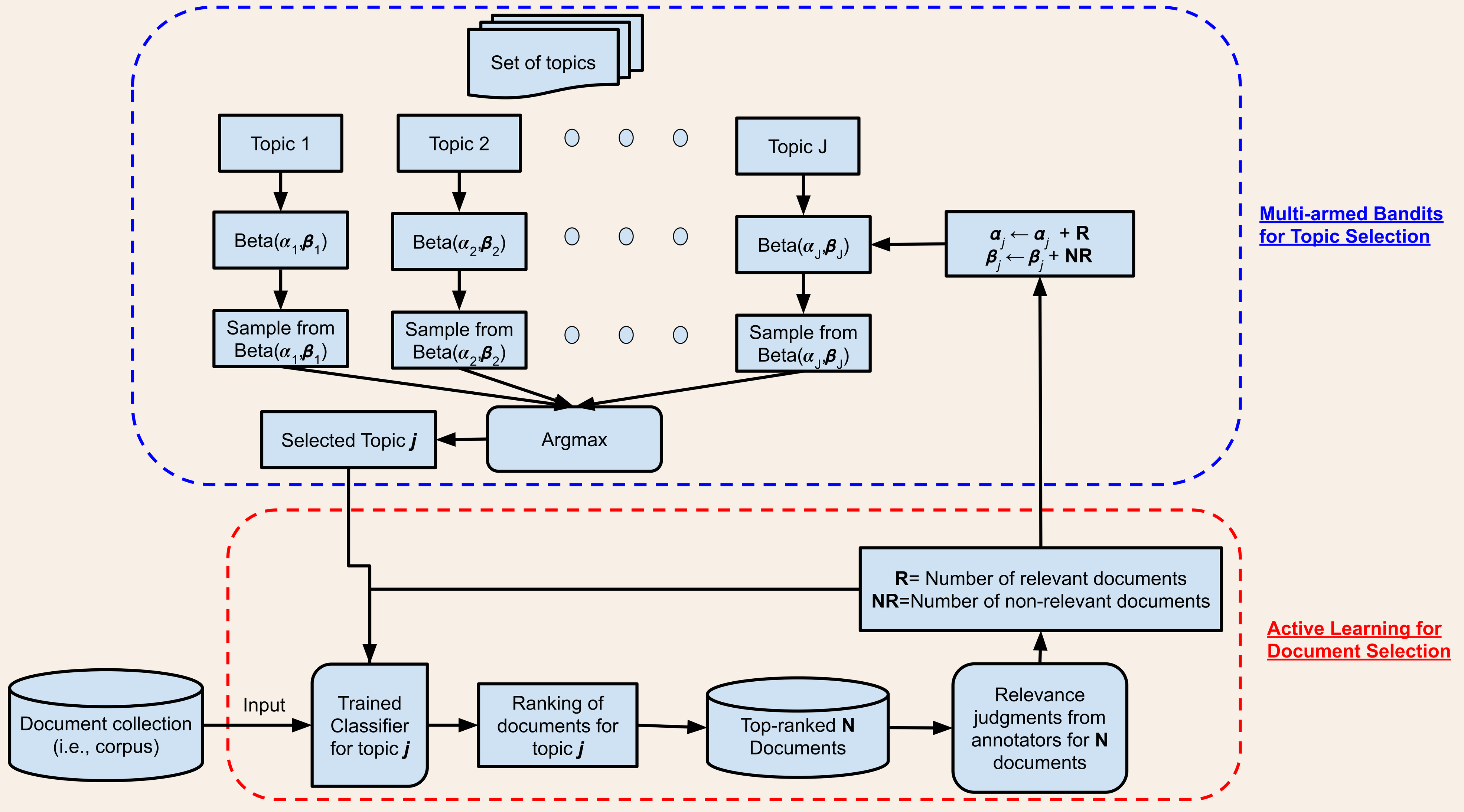

Figure 2: Two-phase Topic and Document Selection

Document Selection
- In the absence of a shared task, we do not have any IR system that can provide ranking of documents that can be used for selecting documents for annotations.
- We employ **active learning** (AL) [2], where a topic-specific classifier selects documents for annotations. The annotated documents in one iteration are employed to train the topic-specific classifier in the next iteration. More specifically, we apply ***Continuous Active Learning*** (CAL).

Results

Table 2: Avg. number of relevant documents found under varying budget per topic on TREC-8 for MTF [3], MM-NS [4], & MAB+CAL.

	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

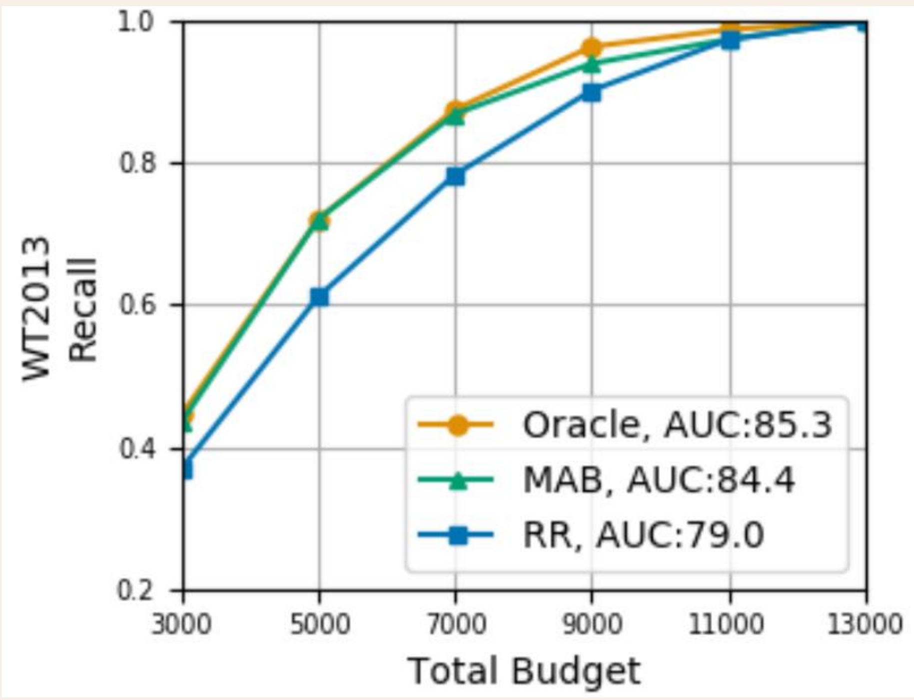

Figure 3: Recall of relevant documents achieved by Oracle, Multi-armed bandits (MAB), and Round-robin (RR) with CAL.

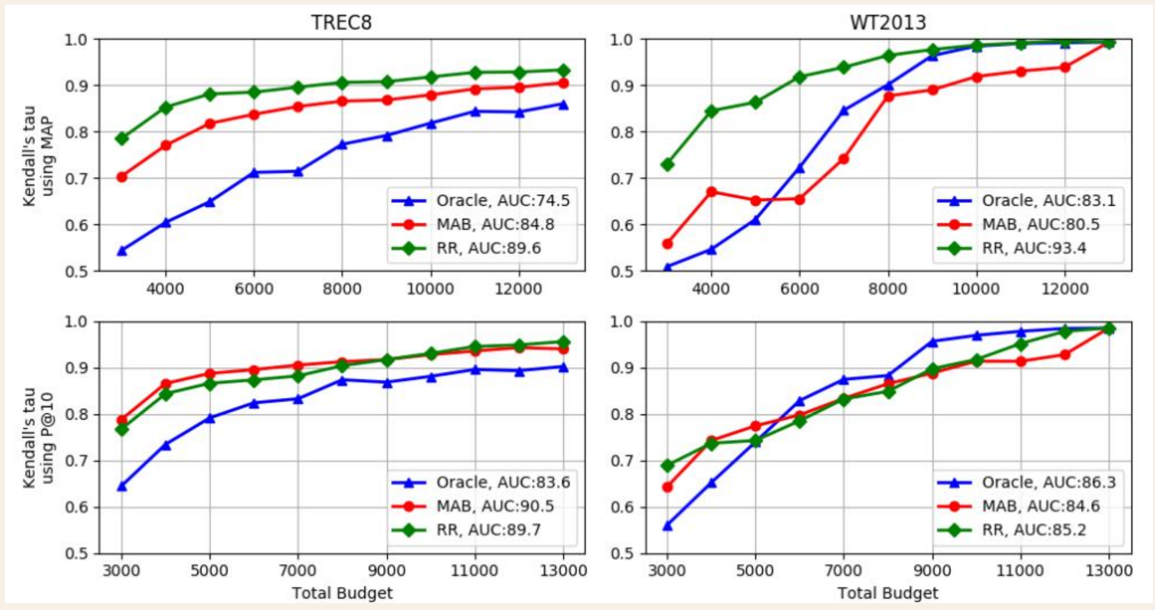

Figure 4: Result of Kendall's τ rank correlation score between the ranking produced using the official qrels and the ranking produced using qrels created by Oracle, MAB and RR with CAL.

Datasets

Table 1: Test collection statistics

Track	Collection	Topics	#Docs	#Judged	%Rel
WT'13[5]	ClueWeb12	201-250	52M	14,474	28.7%
TREC-8[28]	Disks45-CR[4]	401-450	528K	86,830	5.4%

Baselines

Topic Selection
- Oracle
- Round-robin (RR)
- Move-to-front (MTF) [3]
- MaxMean Non-Stationary (MM-NS) [4]

References

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

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

[3] 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

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