"When to Stop": Waterloo (Cormack) Participation in the TREC 2016 Total Recall Track

Gordon V. Cormack & Maura R. Grossman

University of Waterloo

In the course of developing tools for the 2015 Total Recal Track, Track Co-Coordinators Gordon V. Cormack and Maura R. Grossman created an autonomous continuous active learning ("CAL") system, which was provided to participants as the baseline model implementation ("BMI") [http://plg.uwaterloo.ca/∼gvcormac/trecvm/]. BMI employs the technology-assisted review ("TAR") approach described by Cormack and Grossman [2]; the only difference is that BMI employs logistic regression implemented by Sofia ML [https://code.google.com/p/sofia-ml/], instead of SVM*light* [http://svmlight.joachims.org/]. BMI was reprised, unchanged from TREC 2015, except for the addition of a default "call-your-shot" stopping rule indicating the system's estimate of the point at which a reasonable compromise between recall and effort had been achieved.

The Waterloo (Cormack) team submitted runs using BMI for the "Athome" and "Sandbox" tasks. The only change that was made to BMI was to incorporate two different "call-your-shot" criteria that the authors had previously reported at SIGIR 2016 [1]:

- *The "Target" Method.* The Target Method involves first reviewing randomly selected documents until ten are judged relevant. Next, BMI method is run until these ten documents are discovered. It is known that this method achieves over 90% recall on average, and over 70% recall, with 95% probability
- *The "Knee" Method.* The Knee method involves a simple geometric algorithm to identify a "knee," or negative inflection point in the gain curve. The method stops when the slope folowing the knee diminishes to less the one-sixth of the slope before the knee.

As a baseline, we used the default TREC rule:

- *The "BMI" Method.* Stop with the number of documents reviewed exceeds $2\hat{R}+1000$, where $\hat{R}$ is the number of relevant documents judged relevant thus far.

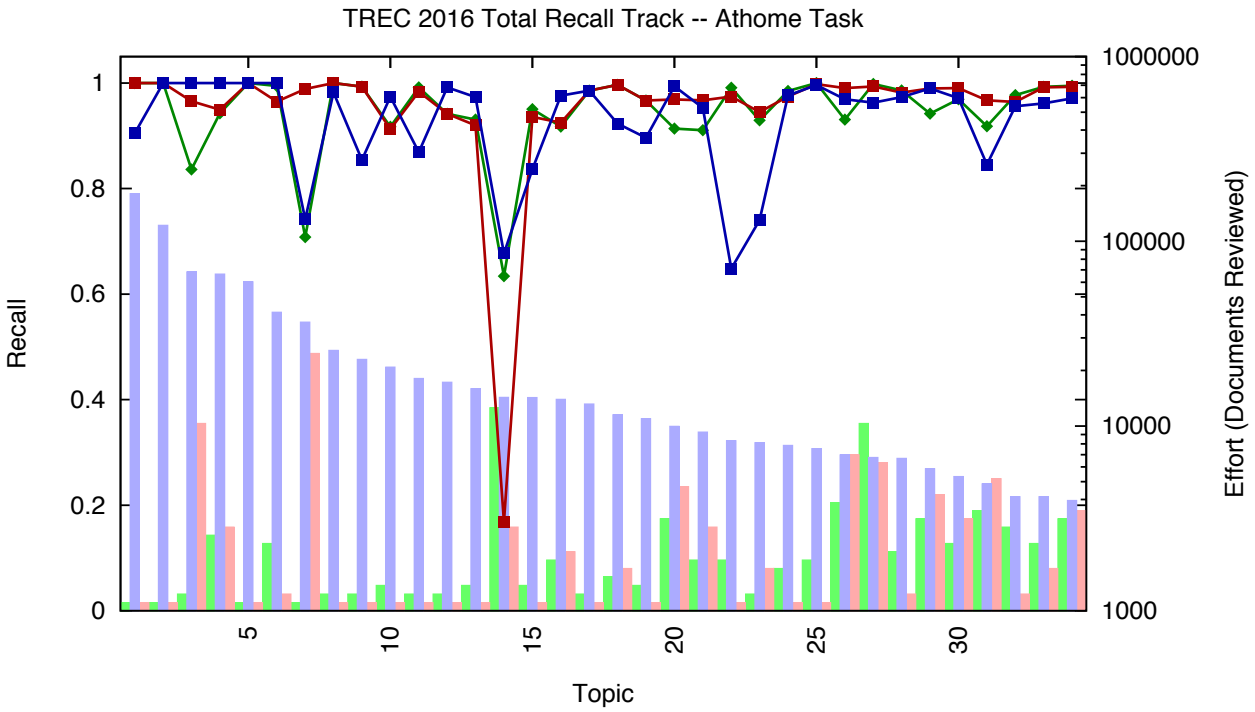
Results are summarized in Figure 1, on Page 2. Overall, BMI achieved very high recall on all but one topic. Neither of the methods we tested improved substantially on the default BMI Method. The Knee Method achieved slightly higher recall (0.950 vs. 0.945, on average over 34 topics) at the expense of slightly higher effort (3091 vs. 2452 documents reviewed, on average). The Target Method, while achieving a statistical guarantee of achieving high recall, achieved empirically lower recall than the other two methods (0.926), at the expense of an order of magnitude more review effort (25752 documents reviewed per topic). The increased effort was the result of random sampling to identify the initial ten relevant documents; a huge burden when there are few relevant documents to be found in the collection.

Overall, our results confirm the efficacy of all stopping criteria, notwithstanding one outlier topic. We examined this topic, and concluded that the Knee Method had missed entirely an expansive "mail-in" campaign consisting of thousands of essentially identical messages.

References

[1] G. V. Cormack and M. R. Grossman. Engineering quality and reliability in technology-assisted review. In *SIGIR 2016*.

[2] G. V. Cormack and M. R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv:1504.06868*, 2015.



Run	*Recall*	*Effort*
BMI	.945	2452
Knee	.950	3091
Target	.926	25752

Fig. 1: Summary results for three stopping criteria, for each topic in the TREC 2016 Total Recall "Athome4" task. Recall is shown as a line graph; corresponding effort is shown as a corresponding bar below (log scale). Average recall and effort are shown in the bottom-left inset. The BMI and Knee Methods achieve comparable recall $(0.945, 0.950)$ and effort $(2452, 3019)$. The Target Method achieves inferior recall (0.926) and much higher effort (25752).