Waterloo (Cormack) Participation in the TREC 2015 Total Recall Track

Gordon V. Cormack, *University of Waterloo*
Maura R. Grossman, *Wachtell, Lipton, Rosen & Katz*[*]

January 24, 2016

In the course of developing tools for the 2015 Total Recall Track, co-coordinators Cormack and 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 essentially employs the approach described by Cormack and Grossman [http://arxiv.org/abs/1504.06868]; 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/].

The Waterloo (Cormack) team submitted runs using BMI for each of the five 2015 Total Recall test collections. The only change that was made to BMI was to add a provision to "call our shot" – that is, to indicate to the assessment server when we believed the run to be reasonably complete. Although the Track provided three milestones – "70recall," "80recall," and "reasonable" – we made no attempt to quantify the recall of our runs, and instead used the three milestones to indicate graduated levels of completeness, which one might interpret as "good," "better," and "best."

We investigated two methods for determining the completeness of our efforts:

- The *knee-finding method*: We employed a simple geometric algorithm to identify a "knee" or negative inflection point in the gain curve [http://www1.icsi.berkeley.edu/∼barath/papers/kneedle-simplex11.pdf]. We then computed the slope of the curve (*i.e.*, marginal precision) before and after the knee, and determined the review to be complete when the ratio of these slopes exceeded a given threshold: $\rho = 3.0$, $\rho = 6.0$, and $\rho = 10.0$, respectively, for our "70recall," "80recall," and "reasonable" stopping criteria. As we were concerned about the volatility of the slope estimates at low levels of effort, we configured our knee method to delay calling its shot until at least β documents had been retrieved: Knee100 always retrieved and reviewed at least $\beta = 100$ documents for review before employing the knee-finding algorithm; Knee1000 always retrieved and reviewed at least $\beta = 1000$ documents before employing knee-finding.

- The *2399 method*: In electronic discovery, much emphasis has been placed on the use of sampling to ensure adequate recall, and a sample size of 2,399 documents has been widely embraced (due to the fact that a sample of 2,399 allows one to estimate a proportion with a margin of error of ±2% and a confidence level of 95%). Such a sample is of little use in computing recall when the prevalence of relevant documents in the corpus is low, as we expected it to be for many of the 2015 Total Recall topics. We hypothesized that the effort to review an additional 2,399 documents would be better spent to review more documents so as to improve recall, rather than in a potentially futile effort to measure recall. To this end, we programmed our submission to call its shot once $N = \alpha r + 2399$ documents had been submitted for assessment, where r was the number of the N documents assessed relevant, and $\alpha = 1.0$, $\alpha = 1.1$, and $\alpha = 1.2$, respectively, for our "70recall," "80recall," and "reasonable" methods.

The results shown in Tables 1 through 3 show the results of the knee-finding and the 2399 methods for the 30 topics of the "athome1," "athome2," and "athome3" collections employed for the At Home task. The results shown in Tables 4 through 6 show the results of only the 2399 method for the 30 topics of the Sandbox task; resource constraints prevented us from testing the knee-finding method for this task.

Our results indicate that both methods were generally conservative when the highest values of the parameters α, β, and ρ were used, yielding very high recall levels. Knee-finding appears to have stopped prematurely for a few low-prevalence topics, and appears to have required unreasonably high levels of effort in some circumstances that appear to represent "difficult" topics, where initial precision was low and no clear knee formed. The 2399 method appeared much more stable, almost always achieving high recall. For low prevalence topics, it (by design) showed low precision; for example, when there were 23 relevant documents, the method would necessarily achieve no better than 1% precision. In many circumstances, however, it may nevertheless be worthwhile to review this many documents in order to ensure oneself that high recall has been achieved.

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

	Topic (R) – Athome1 Collection									
Topic:	100	101	102	103	104	105	106	107	108	109
# Relevant docs:	(4542)	(5836)	(1624)	(5725)	(227)	(3635)	(17135)	(2375)	(2375)	(506)
Knee $\beta = 100$ $\rho = 3$	0.9203	0.9854	0.8744	0.9836	0.8150	0.7593	0.0067	0.9571	0.8766	0.0059
	7703	*7704*	*2567*	*6982*	*498*	*5730*	*131*	*3479*	*3847*	*112*
Knee $\beta = 100$ $\rho = 6$	0.9679	0.9978	0.9310	0.9958	0.8855	0.8316	0.9974	0.9802	0.9735	0.0059
	11396	*9374*	*3847*	*8499*	*887*	*9374*	*24750*	*4698*	*7704*	*112*
Knee $\beta = 100$ $\rho = 10$	0.9813	0.9988	0.9581	0.9974	0.9163	0.8856	0.9977	0.9878	0.9764	0.0059
	15257	*11397*	*5730*	*10337*	*1233*	*15258*	*27255*	*6326*	*8499*	*131*
Knee $\beta = 1000$ $\rho = 3$	0.9207	0.9947	0.8966	0.9827	0.9119	0.7618	0.9929	0.9512	0.4577	0.9704
	7703	*8499*	*2842*	*6982*	*1106*	*5730*	*20403*	*3479*	*1233*	*2090*
Knee $\beta = 1000$ $\rho = 6$	0.9661	0.9976	0.9304	0.9951	0.9119	0.8393	0.9972	0.9827	0.4577	0.9763
	11396	*9374*	*3847*	*8499*	*1106*	*10337*	*24750*	*5189*	*1233*	*2317*
Knee $\beta = 1000$ $\rho = 10$	0.9830	0.9991	0.9483	0.9976	0.9119	0.8809	0.9978	0.9861	0.9764	0.9802
	16811	*11397*	*5189*	*10337*	*1106*	*15258*	*27255*	*5730*	*8499*	*3145*
2399 $\alpha = 1.0$	0.8989	0.9950	0.9360	0.9948	0.9692	0.7447	0.9935	0.9832	0.9251	0.9802
	6981	*8499*	*4252*	*8499*	*2842*	*5189*	*20403*	*5189*	*4698*	*3145*
2399 $\alpha = 1.1$	0.8989	0.9974	0.9360	0.9963	0.9692	0.7590	0.9965	0.9832	0.9389	0.9802
	6981	*9374*	*4252*	*9374*	*2842*	*5730*	*22473*	*5189*	*5189*	*3145*
2399 $\alpha = 1.2$	0.9194	0.9979	0.9360	0.9963	0.9692	0.7590	0.9973	0.9857	0.9389	0.9802
	7703	*10337*	*4252*	*9374*	*2842*	*5730*	*24750*	*5730*	*5189*	*3145*

Table 1: Athome1 Collection: Recall and Effort (italics) for various different stopping criteria.

	Topic (R) – Athome2 Collection									
Topic:	2052	2108	2129	2130	2134	2158	2225	2322	2333	2461
# Relevant docs:	(265)	(661)	(589)	(2299)	(252)	(1256)	(182)	(9517)	(4805)	(179)
Knee $\beta = 100$ $\rho = 3$	0.9245	0.8941	0.8625	0.8434	0.7659	0.0892	0.6209	0.9407	0.9523	0.3799
	497	*1373*	*1233*	*7704*	*708*	*131*	*152*	*16812*	*8499*	*131*
Knee $\beta = 100$ $\rho = 6$	0.9698	0.9622	0.9677	0.9361	0.9008	0.0892	0.8736	0.9767	0.9773	0.3799
	792	*2317*	*2317*	*13846*	*1527*	*131*	*390*	*24750*	*11397*	*131*
Knee $\beta = 100$ $\rho = 10$	0.9811	0.9758	0.9847	0.9622	0.9405	0.0892	0.9231	0.9841	0.9881	0.3799
	990	*3145*	*3479*	*20403*	*2317*	*131*	*631*	*30011*	*15258*	*131*
Knee $\beta = 1000$ $\rho = 3$	0.9849	0.8548	0.8353	0.8695	0.8611	0.9761	0.9560	0.9408	0.9517	0.9162
	1105	*1233*	*1106*	*8499*	*1106*	*1697*	*1106*	*16812*	*8499*	*1106*
Knee $\beta = 1000$ $\rho = 6$	0.9849	0.9637	0.9660	0.9356	0.8810	0.9873	0.9560	0.9766	0.9779	0.9162
	1105	*2317*	*2317*	*13846*	*1373*	*2317*	*1106*	*24750*	*11397*	*1106*
Knee $\beta = 1000$ $\rho = 10$	0.9849	0.9788	0.9796	0.9604	0.9405	0.9881	0.9560	0.9862	0.9875	0.9385
	1105	*3145*	*3145*	*20403*	*2317*	*2842*	*1106*	*33043*	*15258*	*1373*
2399 $\alpha = 1.0$	0.9925	0.9728	0.9830	0.6759	0.9524	0.9881	0.9835	0.8167	0.9193	0.9888
	2841	*2842*	*3145*	*4252*	*2842*	*3847*	*2842*	*10337*	*6982*	*2842*
2399 $\alpha = 1.1$	0.9925	0.9728	0.9830	0.6759	0.9524	0.9881	0.9835	0.8446	0.9386	0.9888
	2841	*2842*	*3145*	*4252*	*2842*	*3847*	*2842*	*11397*	*7704*	*2842*
2399 $\alpha = 1.2$	0.9925	0.9818	0.9830	0.7086	0.9524	0.9881	0.9835	0.8800	0.9534	0.9888
	2841	*3479*	*3145*	*4698*	*2842*	*4252*	*2842*	*12563*	*8499*	*2842*

Table 2: Athome2 Collection: Recall and Effort (italics) for various different stopping criteria.

	Topic (R) – Athome3 Collection									
Topic:	3089	3133	3226	3290	3357	3378	3423	3431	3481	3484
# Relevant docs:	(255)	(113)	(2094)	(26)	(629)	(66)	(76)	(1111)	(2036)	(23)
Knee $\beta = 100$ $\rho = 3$	0.4353	0.8496	0.6829	0.6923	0.9269	0.8636	0.4474	0.9820	0.9332	1.0000
	130	*131*	*1884*	*112*	*1106*	*112*	*112*	*1106*	*2842*	*112*
Knee $\beta = 100$ $\rho = 6$	0.4353	0.8584	0.9790	0.6923	0.9523	0.8636	0.4474	0.9847	0.9465	1.0000
	130	*152*	*4252*	*112*	*1527*	*131*	*112*	*1233*	*3847*	*112*
Knee $\beta = 100$ $\rho = 10$	0.4353	0.9912	0.9790	0.6923	0.9523	0.8636	0.4605	0.9847	0.9514	1.0000
	130	*344*	*4252*	*112*	*1527*	*131*	*152*	*1233*	*5189*	*112*
Knee $\beta = 1000$ $\rho = 3$	0.9961	0.9912	0.6882	1.0000	0.9316	1.0000	0.5263	0.9838	0.9224	1.0000
	1105	*1106*	*1884*	*1106*	*1106*	*1106*	*1106*	*1106*	*2567*	*1106*
Knee $\beta = 1000$ $\rho = 6$	0.9961	0.9912	0.9666	1.0000	0.9523	1.0000	0.5263	0.9892	0.9470	1.0000
	1105	*1106*	*3479*	*1106*	*1527*	*1106*	*1106*	*1233*	*3479*	*1106*
Knee $\beta = 1000$ $\rho = 10$	0.9961	0.9912	0.9733	1.0000	0.9634	1.0000	0.5263	0.9892	0.9504	1.0000
	1105	*1106*	*3847*	*1106*	*1884*	*1106*	*1106*	*1233*	*4252*	*1106*
2399 $\alpha = 1.0$	0.9961	1.0000	0.9823	1.0000	0.9841	1.0000	0.6184	0.9991	0.9499	1.0000
	2841	*2567*	*4698*	*2567*	*3145*	*2567*	*2567*	*3847*	*4698*	*2567*
2399 $\alpha = 1.1$	0.9961	1.0000	0.9823	1.0000	0.9841	1.0000	0.6184	0.9991	0.9499	1.0000
	2841	*2567*	*4698*	*2567*	*3145*	*2567*	*2567*	*3847*	*4698*	*2567*
2399 $\alpha = 1.2$	0.9961	1.0000	0.9852	1.0000	0.9841	1.0000	0.6184	0.9991	0.9504	1.0000
	2841	*2567*	*5189*	*2567*	*3145*	*2567*	*2567*	*3847*	*5189*	*2567*

Table 3: Athome3 Collection: Recall and Effort (italics) for various different stopping criteria.

	Topic (R) – Kaine Collection			
Topic:	Open	Restricted	Record	VA Tech
# Relevant docs:	(131698)	(14341)	(166118)	(20083)
2399 $\alpha = 1.0$	0.3839	0.6809	0.4550	0.8600
	53412	*12652*	*78361*	*20402*
2399 $\alpha = 1.1$	0.6280	0.7168	0.7940	0.9011
	94894	*13846*	*153054*	*22473*
2399 $\alpha = 1.2$	0.6605	0.7440	0.8801	0.9507
	104421	*15259*	*185276*	*27256*

Table 4: Kaine Collection: Recall and Effort (italics) for various different stopping criteria.

	Topic (R) – MIMIC II Collection [Part I]									
Topic:	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10
# Relevant docs:	(5881)	(3867)	(15101)	(7826)	(6123)	(5081)	(19182)	(11256)	(8706)	(8741)
2399 $\alpha = 1.0$	0.7432	0.8474	0.8125	0.6481	0.6335	0.5869	0.9819	0.7980	0.6986	0.7599
	6981	*5730*	*15258*	*7704*	*6326*	*5730*	*22473*	*11397*	*8499*	*9374*
2399 $\alpha = 1.1$	0.7779	0.8663	0.8643	0.6875	0.6681	0.5869	0.9926	0.8762	0.7864	0.8009
	7703	*6326*	*16812*	*8499*	*6982*	*5730*	*24750*	*13846*	*10337*	*10337*
2399 $\alpha = 1.2$	0.8109	0.8836	0.9548	0.7313	0.7013	0.6215	0.9949	0.9087	0.8256	0.8396
	8498	*6982*	*20403*	*9374*	*7704*	*6326*	*27255*	*15258*	*11397*	*11397*

Table 5: MIMIC II Collection,part I: Recall and Effort (italics) for various different stopping criteria.

Topic:	C11	C12	C13	C14	C15	C16	C17	C18	C19
# Relevant docs:	(180)	(2579)	(3465)	(2143)	(5143)	(8047)	(11117)	(16827)	(6828)
2399 $\alpha = 1.0$	0.9889	0.6506	0.5328	0.7354	0.9903	0.4710	0.6930	0.6749	0.6328
	2842	*4252*	*4252*	*4252*	*7704*	*6326*	*10337*	*13846*	*6982*
2399 $\alpha = 1.1$	0.9889	0.6506	0.5671	0.7354	0.9979	0.5103	0.7784	0.7696	0.6659
	2842	*4252*	*4698*	*4252*	*8499*	*6982*	*12563*	*16812*	*7704*
2399 $\alpha = 1.2$	0.9889	0.6758	0.5957	0.7485	0.9994	0.5469	0.8239	0.8585	0.6995
	2842	*4698*	*5189*	*4698*	*9374*	*7704*	*13846*	*20403*	*8499*

Table 6: MIMIC II Collection, part II: Recall and Effort (italics) for various different stopping criteria.