The University of Amsterdam (ILPS) at TREC 2015 Total Recall Track

David van Dijk[†‡], Zhaochun Ren[‡], Evangelos Kanoulas[‡], and Maarten de Rijke[‡]

[†] *Create-IT, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands*
[‡] *University of Amsterdam, Amsterdam, The Netherlands*

Abstract

We describe the participation of the University of Amsterdams ILPS group in the Total Recall track at TREC 2015. Based on the provided Baseline Model Implemention ("BMI") we set out to provide two more baselines we can compare to in future work. The two methods are bootstrapped by a synthetic document based on the query, use TF/IDF features, and sample with dynamic batch sizes which depend on the percentage of predicted relevant documents. We sample at least 1 percent of the corpus and stop sampling if a batch contains no relevant documents. The methods differ in the classifier used, i.e. Logistic Regression and Random Forest.

1 Introduction

The Total Recall track was introduced at TREC this year. In this track, participants implement automatic or semi-automatic methods to identify as many relevant documents as possible, with as little review effort as possible, from document collections containing as many as 1 million documents.

After downloading the collection and information need, participants must identify documents from the collection and submit them to the online relevance assessor which return the relevance labels.

The objective is to submit as many documents containing relevant information as possible, while submitting as few documents as possible, to the automated relevance assessor.

The track provided "Play-at-Home" and "Sandbox" evaluation. For "Play-at-Home" evaluation, participants ran the system on their own hardware, and participant could choose to run "automatic" or "manual", where the latter involved manual intervention. For the "Sandbox" evaluation a virtual machine needed to be submitted, containing a fully automated solution.

The Information and Language Processing Systems (ILPS) group of the University of Amsterdam participated in the "Play-at-Home" and "Sandbox" evaluation, without the use of manual intervention. In this paper, we explain the runs we submitted and their results.

Section 2 describes the methods we submitted, Section 3 lists the runs and their results, Section 4 contains our conclusion.

2 Methods

Before our final submissions we experimented with several methods on the provided test data. We tried different combinations of bootstrapping and sampling methods, features and classifiers. Beating the baseline [**?**] turned out challenging. None of the methods provided a significant improvement over the baseline. Therefore we decided to postpone this task, and for now submit two basic methods, that are slight variations on the baseline, in order to see the influence of dynamic sampling, a heuristical stopping criterion and different classifiers.

Our methods work as follows:

(1) Create synthetic document from query , code it as relevant.
(2) Temporarily code a randomly sampled document as not relevant.
(3) Add coded documents to initial training set.
(4) Train classifier on initial training set, classify documents in corpus.
(5) Review documents in descending order until we find at least one relevant and one not relevant.
(6) Initialise the training set with the reviewed documents.
(7) Set batch size B to 100.
(8) Train classifier on training set, classify documents in corpus.
(9) Select B highest scoring documents for review.
(10) Review the documents, coding them as relevant or not relevant.
(11) Add the reviewed documents to the training set.
(12) Set B to the part of relevant docs found in the batch, times 0.1% of the total amount of documents in the corpus.
(13) If B is zero, and there has been sampled less then 1% of the corpus, set B to 1% of the corpus minus the amount of documents already sampled.
(14) Repeat steps 8 through 13 until the amount sampled is over 1% of the corpus and B is zero.

For preprocessing some basic filtering and Porter stemming is applied. The two methods are bootstrapped by a synthetic document based on the query. We use TF/IDF features, and sample with dynamic batch sizes which depend on the percentage of predicted relevant documents. We sample at least 1 percent of the corpus and stop sampling if a batch contains no relevant documents. The methods differ in the classifier used, i.e. Logistic Regression and Random Forest. Scikit classifiers were used with default parameters.

3 Runs & Results

We report on the preliminary results for the **Athome** experiments (datasets **Athome1**, **Athome2** and **Athome3**, 10 queries each), as well as for one of the Sandbox tests, which was done on the MIMIC II clinical dataset[1](19 queries). We compare the results of our methods (*Baseline1* and *Baseline2*) against the provided baseline (*BMI*).

Athome1 (290K). On the Athome1 dataset, overall the *BMI* and *Baseline1* gave similar results and *Baseline2* performed less. Only on athome109 *Baseline1* outperformed *BMI*. *Baseline2* was in between the two from the start until around 0.5 recall, then dropped to third place again. Fig. 1 shows the results on topic athome101, reflecting the general performance. The picture shows the effort over recall levels (step size 0.05) until the maximum recall achieved by either *Baseline1* or *Baseline2*. Table 1 lists the topics for the dataset, the number of relevant documents per topic, followed by the maximum recall achieved

[1] https://physionet.org/mimic2/mimic2_clinical_overview.shtml

by *Baseline1*. At the bottom of the table the average maximum recall is provided (0.92730), which gives us some intuition on how well our stop criterion performed. An average maximum recall above 0.9 seems reasonable.

Athome2 (450K). Though the general picture remains, *Baseline1* performs a little less then *BMI* overall on the Athome2 dataset. *Baseline2* is still in third place. See fig. 2 for an example. Table 2 shows the number of relevant documents per topic are smaller in general then on the Athome1 dataset. The average maximum recall on *Baseline1* is a little higher as in the Athome1 experiments; 0.93733.

Athome3 (900K). The results from the Athome3 dataset gives similar results as seen so far. On one topic *Baseline1* outperforms *BMI*, see fig. 3. Table 3 reveals the relevant documents per topic are, on average, even smaller than on the Athome2 dataset. The *Baseline1*'s average maximum recall went up to 0.97049, which seems quite good.

MIMIC II. The size of the MIMIC II dataset is unknown to us at the moment of writing, as it was used in a sandbox test. Again we see similar results as before for several topics. But for some topics we see the performance of *Baseline2* getting closer to *BMI*. Another observation is that *Baseline1* stops way too soon for some topics. Fig. 4 provides an example of both observations. When we look at Table 4 we see that for most topics the stopping criterion is not performing well for *Baseline1* on this dataset. Only for 3 topics the maximum recall is above 0.9, the other 13 topics are below 0.73, 5 of those are close to zero. The *Baseline1*'s average maximum recall is at a 0.48989, not an acceptable level. Table 4 also shows the relevant documents per topic are quite large on average compared to the other datasets.

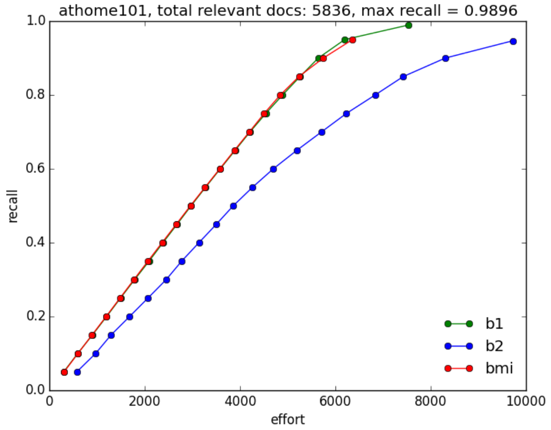

Figure 1: Athome101

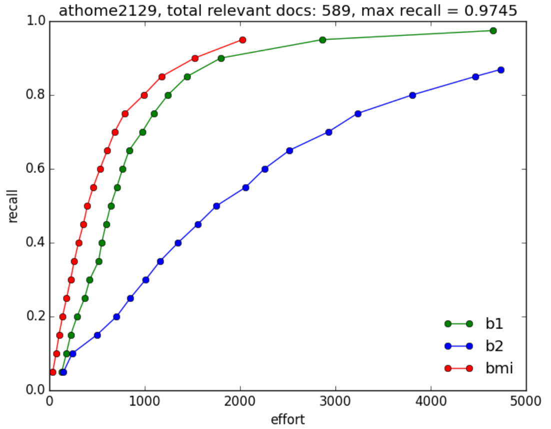

Figure 2: Athome2129

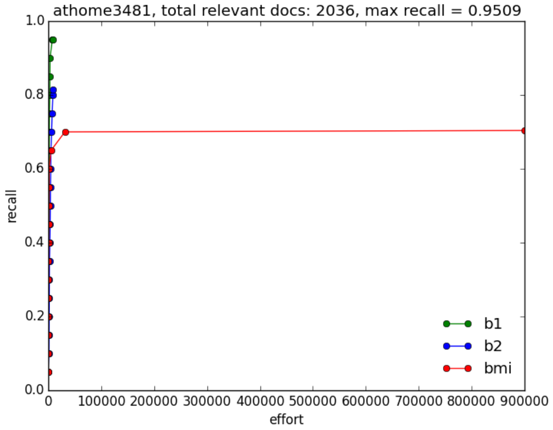

Figure 3: Athome3481

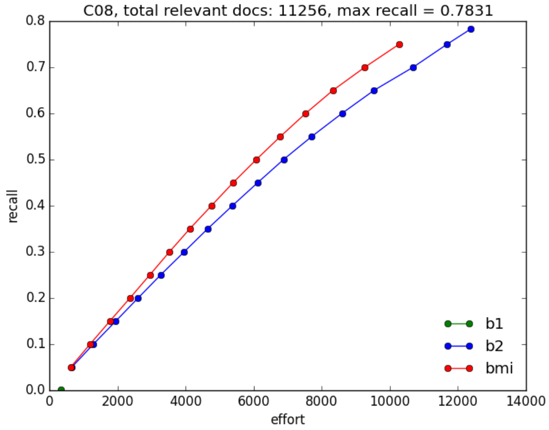

Figure 4: C08

ID	#Rel.	Max. recall
athome100	4,542	0.9036
athome101	5,836	0.9896
athome102	1,624	0.9206
athome103	5,725	0.9857
athome104	227	0.8546
athome105	3,635	0.7508
athome106	17,135	0.9879
athome107	2,375	0.9689
athome108	2,375	0.9390
athome109	506	0.9723
Avg. max. recall		0.92730

Table 1: Athome1 (290K), Baseline1

ID	#Rel.	Max. recall
athome2052	265	0.9962
athome2108	661	0.9803
athome2129	589	0.9745
athome2130	2,299	0.7747
athome2134	252	0.8651
athome2158	1,256	0.9881
athome2225	182	0.9561
athome2322	9,517	0.9032
athome2333	4,805	0.9463
athome2461	179	0.9888
Avg. max. recall		0.93733

Table 2: Athome2 (450K), Baseline1

ID	#Rel.	Max. recall
athome3089	255	0.9961
athome3133	113	1.000
athome3226	2,094	1.000
athome3290	26	0.9895
athome3357	629	0.9857
athome3378	66	0.9546
athome3423	76	0.8290
athome3431	1,111	0.9991
athome3481	2,036	0.9509
athome3484	23	1.000
Avg. max. recall		0.97049

Table 3: Athome3 (900K), Baseline1

ID	**#Rel.**	**Max. recall**
C1	5,811	0.6910
C2	3,867	0.7210
C3	15,101	0.0031
C4	7,826	0.6360
C5	6,123	0.5692
C6	5,081	0.4718
C7	19,182	0.9585
C8	11,256	0.0022
C9	8,706	0.0030
C10	8,741	0.6608
C11	180	0.9500
C12	2,579	0.0016
C13	3,465	0.0026
C14	2,143	0.5684
C15	5,143	0.9117
C16	8,047	0.4977
C17	11,117	0.6980
C18	16,827	0.4561
C19	6,828	0.5053
Avg. max. recall		0.48989

Table 4: MIMIC II, Baseline1

4 Conclusion & Discussion

This year we participated in the first run of the TREC 2015 Total Recall track. We submitted two variations on the baseline and submitted them for both **Athome** and **Sandbox** evaluation. The methods both differed from the baseline in the sampling method, i.e. we set the next batch size based on the percentage of correctly predicted documents in the current batch. The baseline's sampling method let's the batch size grow monotonically, based on a growing confidence in the classifier. We also applied a heuristical stopping criterion. The methods differ among each other by the classifier used; *Baseline1* used Logistic Regression, *Baseline2* a Random Forest classifier.

In the experiments, the baseline's sampling method outperformed our sampling method. The batch size turned out to have a substantial influence on performance.

Logistic Regression outperformed the Random Forest classifier overall. The influence of the classifier on performance varied among datasets.

The stopping criterion worked reasonably well on the Athome dataset, but it did not perform well on the MIMIC II dataset. As we do not have access to the latter dataset we have not been able to analyse this result yet. Our stopping criterion depends on our sampling method. As the sampling method did hurt performance, this dependency is unwanted and we aim to experiment with other stopping criteria on the task.

5 Acknowledgements

Acknowledgments. This research was supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, CI-14-25, SH-322-15, Amsterdam Data Science, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project nr 027.012.105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.