MRG_UWaterloo Participation in the TREC 2018 Common Core Track

Maura R. Grossman and Gordon V. Cormack
University of Waterloo

The MRG_UWaterloo team from the University of Waterloo participated in the TREC 2018 Common Core Track. We used logistic regression to score and rank all documents from the Washington Post dataset, using pseudo-relevant and pseudo-nonrelevant training documents fetched from the Web using Google search.

For run **uwmrg**, the training set for each topic consisted of of the top ten links returned by a Google search for the words in the topic title and description. Each link was fetched and rendered as plain text using the command **lyx -dump**. Documents containing the the literal text **title:** and **description:** were excluded, as were documents containing **404 Not Found**. The former indicates a legacy copy of the topic statement from prior TREC efforts, while the latter indicates a defunct page.

In total, the training set contained 496 documents. For each topic we labeled **relevant** all the documents fetched using its title and description, and we labeled **not relevant** all the rest.

For run **uwmrgx**, we extracted the anchor text and query-based summary for each of the ten links provided in the Google-generated search engine result page. For each topic, these ten extracts were combined to form a single training document. Thus, the training set for each topic consisted of 50 documents, with one positive example and 49 negative examples.

We extracted each article in the Washington Post dataset and stripped the XML tags using **lyx -dump** to form a plain text rendering of each document. Normalized *tf-idf* feature vectors were created using code extracted from the TREC Total Recall Track Baseline Model Implementation (BMI).[1] The logistic regression implementation was Sofia-ML[2] with parameters **--learner_type logreg-pegasos --loop_type roc --lambda 0.0001 --iterations 200000**, also taken from BMI. For each topic, documents were sorted by score, and the top 10,000 were submitted to NIST.

Official TREC results are shown below.

	MAP	P@10	NDCG
uwmrg	0.2761	0.5000	0.5822
uwmrgx	0.2362	0.4360	0.5306

[1] http://cormack.uwaterloo.ca/trecvm/
[2] https://code.google.com/archive/p/sofia-ml/