Design & Engineering of Intelligent Information Systems: PI9

Instructed by *Dr. Eric Nyberg*

Due on November 2, 2015

Vivian Robison vrobison

Going back to PI5, the last time we used precision metrics, we find:
MRR: 0.22760290556900725
MAP: 0.36346705103990634
This composite ranker with learned rates raises these metric scores to:
MRR: 0.29782082324455206
MAP: 0.4164951150125029

Performance according to these metrics has clearly improved. The scores are hampered by the number of questions with no relevant passages, where P@1 is 0 no matter what. A manual perusal of a subsection of questions suggests P@1 has improved quite a bit, which makes sense since we used that metric for parameter optimization. The learned weights for combining scores does a better job of weighing the ngram overlap score with the boolean question subject presence score than my original guessed-at weights. I picked 5-fold validation as a fairly common number of folds that wouldn't make each fold too small to be meaningful.

Training over the entire dataset, rather than just a subset, also helps raise the scores, but training the final weights on the same data we use at runtime is bad practice–if this wasn't what was intended by "Train the model on the entire dataset with the best hyperparameters you get in step 4" that needed to be made clearer.

Table 1: Average Precision at 1 over 5 folds

fold	av P@1
1	0.0
2	0.0625
3	0.0625
4	0.1875
5	0.0
all	0.2978