University of Waterloo Participation in the TREC 2007 Spam Track

Gordon V. Cormack

Abstract

This is the first year that we have submitted participant runs to TREC (Cormack is track coordinator). In parallel with running the track, we have investigated two new filtering methods, inspired by the excellent results that others have shown in the task.

- **Dynamic Markov Compression (DMC) modeling (wat2 prefix).** We invented the DMC model some 20 years ago for the purpose of data compression. Bratko showed at TREC 2005 that compression models could work well, and we found that DMC could work even better than the PPM and CWT models used by Bratko. On previous TREC tasks, DMC has been competitive with the best (Bratko et al, JMLR December 2006), and one of our runs (wat2 prefix) implements this method.
- **Logistic Regression on character 4-grams. (wat1 prefix).** We and others (Cormack et al., SIGIR 07; Sculley, SIGIR 2007) have observed that character 4-grams work much better than bags of words, escpecially if only the first 3k or so characters are considered.

  We use simple on-line graduate ascent implementation of logistic regression (Goodman and Yih, CEAS 2006) based on 4-grams. While our tests show that this approach is perhaps not quite as good as Sculley's on-line SVM, it is much simpler and faster, being implemented in its entirety in about 100 lines of C code. It also offers the potential advantage that its output is a probability (actually log-odds) which makes it easier to interpret and calibrate when considered as a component of a system to filter spam. SVM output, in contrast has only a geometric interpretation which cannot directly be mapped to a probability.

- **On-line fusion of DMC and Logistic regression. (wat3 prefix)** We used on-line logistic regression to fuse the results of our wat1 and wat2 methods giving a very fast algorithm that appears to give better results than any previously used at TREC. It remains to be seen whether this method beats Sculley's ROLSVM (other than in speed, where it prevails hands down.)
- **Threshold-based active learning (wat4 prefix).** We adapted our logistic regression filter (wat1 prefix) to train only on messages for which the absolute value of the score was less than some threshold. The TREC 2006 has used to calibrate the threshold so as to approximate the best asymptotic performance for various quota values.