Melbourne at the TREC 2011 Legal Track

William Webber* and Phil Farrelly†

October 24, 2011

1 Approach

The Melbourne team was a collaboration of the University of Melbourne, RMIT University, and the Victorian Society for Computers and the Law. The approach taken was to train a support vector machine based upon textual features using active learning. Two sources of relevance annotations were used for different runs: the official annotations, provided by the topic authorities; and annotations provided by a member of the Melbourne team with e-discovery experience (though not legal training). We describe the SVM method used in Section 1.1, the run using official annotations in Section 1.2, and the run using the internal annotations in Section 1.3.

1.1 SVM classifier

Both methods used an SVM classifier with active learning. The features used were term occurrences with binary weights. The classifier implementation used was `SVM_PERF` [Joachims, 2006], optimized for error rate. Items for labelling were selected by ranking documents by increasing absolute predictive value (that is, placing the document the classifier is most uncertain about first), then sampling N from the top 1000, with probability inverse to rank. For all but the initial runs, probability of relevance was calculated using the SVM Platt method [Platt, 1999].

1.2 Official annotations

The runs using official annotations are tagged `cls`. These runs were developed for all three topics. The initial set of documents to obtain annotations for was obtained by extracted keywords from the topic descriptions and running them as keywords under the Lucene search engine, using default settings. These runs constituted the initial submissions, with probability of relevance assigned by scaling from 1.0 down to 0.0. The top 50 documents of these runs were submitted to request the initial assessments.

*University of Melbourne
†Potter Farrelly & Associates

Topic	Best	Median	Worst	cls	lrn
401	58.8%	28.0%	8.8%	14.3%	23.1%
402	58.8%	13.1%	2.3%	4.6%	n/a
403	72.0%	14.2%	3.1%	3.1%	n/a

Table 1: Best, median, and worst hypothetical F1 scores, and the scores achieved by the final Melbourne runs.

1.3 Internal annotations

The runs using internal annotations are tagged `lrn`. This run was only developed for Topic 401. The annotator interactively developed a small set of Boolean queries; these were sampled and annotated to form the initial learning data. Active learning then proceeded as for the official annotation run (Section 1.2), except that annotations were sought from the internal annotator, rather than the official run.

2 Results

The interim results of the Melbourne runs are compared with the officially reported interim best, median, and worst runs in Table 1. The Melbourne runs came below median for all topics, and the `cls` run scored worst for Topic 403.

References

Thorsten Joachims. Training linear SVMs in linear time. In Tina Eliassi-Rad, Lyle Ungar, Mark Craven, and Dimitrios Gunopulos, editors, *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226, Philadelphia, USA, August 2006.

John C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.