The Melbourne Team at the TREC 2010 Legal Track

William Webber[1], Falk Scholer[2], Mingfang Wu[2], Xiuzhen Zhang[2], Douglas W. Oard[3], Phil Farrelly[4], Sandra Potter[4], Steven Dick[5], and Phill Bertolus[6]

[1]University of Melbourne
[2]RMIT University
[3]University of Maryland
[4]Potter Farrelly Consulting
[5]Mallesons Stephen Jaques
[6]Wombat Technology

July 1, 2011

1 Overview

The Melbourne team was a collaboration between academic and industry groups. The team participated in both the learning and the interactive tasks of this year's Legal Track. The baseline run for the learning track employed true-relevance feedback, achieving respectable outcomes; the experimental runs added additional features and employed an SVM classifier, with poor results. The techniques developed for the learning task were then deployed in the interactive task. The classifier again achieved poor predictive quality, although final results place our production (non-significantly) first. We describe the learning task efforts in Section 2, and the interactive task in Section 3.

2 Learning Task

The team submitted three runs to the learning task. The first of these, `rmitindA`, built a query from topic keywords and performed retrieval using the BM25 similarity metric, along with true-relevance feedback (TRF) on the relevant seed documents. The remaining two runs trained an SVM classifier, using the TRF score as one amongst of seven features. One run, `rmitmlfT`, used all seven of the features; the other, `rmitmlsT`, used for each topic the set of features (except all features, or the TRF feature alone) which achieved the highest mean AUC on a ten-fold cross-validation.

Our philosophy this year was "we're not here to win; we're here to learn"; our results show that we have achieved at least the first of these objectives. The baseline TRF run performed respectably, coming at or above the median hypothetical F1 score

Id	Name	Description
1	isAttach	Is the document an attachment (rather than an email)?
2	hasContent	Is the document (email body) non-empty?
3	custodianRel	What proportion of seed documents from the custodian are relevant?
4	relWithin	How many seed documents within $t = 7$ days are relevant?
5	externalProp	What proportion of participants in an email are from non-Enron addresses?
6	numRecip	How many recipients does the email have?
7	trfScore	What score does the document receive under Okapi-based True Relevance Feedback?

Table 1: Features used in the classifier runs.

for all but one topic. Our machine classification runs, though, performed poorly, often well below the TRF run, which is surprising, given that the TRF score was a feature.

We also tried using Mechanical Turk for relevance assessment. The intention was to train a run with Turker-assessed documents added to the seed set. The Turkers performed our tasks very poorly, however, with 87% of them failing a simple trap question, making their output unusable.

2.1 Method

We describe the baseline true-relevance feedback run first, then the experimental machine classification runs built on top of it. We also describe our unsuccessful attempt at using Mechanical Turk to produce seed documents for a run.

True relevance feedback

Indexing and searching for the TRF run were carried out using the Lemur toolkit.[1] Stopwords were removed, and words were stemmed using the Porter stemmer. The Okapi BM25 model was used for matching and ranking documents. Keywords from the topic were used as the query. For feedback, all relevant documents from a topic's seed set were taken as positive examples, and the top 100 terms were added to the query. Unranked documents were assigned a minimal fixed score, and appended to the ranked search result. Result similarity scores were then normalised linearly to the range $[0, 1]$, to estimate probability of relevance.

Machine classifier

The other two learning runs used an SVM classifier, employing the SVM$^{\text{perf}}$ [Joachims, 2006] implementation. Seven features, summarized in Table 1, were implemented for

[1] http://www.lemurproject.org

Topic	Mnemonic	Subset features	Subset AUC	Full AUC	TRF AUC
200	Houses	3,4,7	0.72	0.71	0.63
201	Prepay transactions	1,2,3,4,5,7	0.91	0.92	0.62
202	FAS 140/125	3,5,7	0.91	0.91	0.79
203	Financial forecasts	3,5,7	0.83	0.79	0.79
204	Document shredding	3,5,7	0.82	0.79	0.83
205	Energy forecasts	3,5,7	0.90	0.90	0.85
206	Analyst reports	3,5,7	0.97	0.96	–
207	Fantasy football	3,5,7	0.95	0.92	–

Table 2: Features selected for the `rmitmlst` run for each topic, with their AUC scores under ten-fold cross-validation, plus the AUC score of using the TRF feature by itself. The classifier did not converge for the TRF score feature alone on Topics 206 and 207.

the classifier. Other, possibly more powerful features were planned, but were not implemented due to lack of time. The count-based features 4 and 6, and the real-valued feature 7, were transformed by adding one and taking the natural logarithm. All feature values were normalized to the $[-1, 1]$ range, based on the maximum and minimum unnormalized scores observed on the training examples.

Ten-fold cross-validation was performed on the seed documents to determine the best feature subsets for each topic. Cross-validation results, expressed as area under the ROC curve (AUC), are given in Table 2. The `rmitmlfT` run used all seven features, while the `rmitmlsT` run used the feature subset giving the best AUC values, excluding the set of all features and the set containing only feature 7, the TRF score. The features selected for `rmitmlsT` are reported in column 3 of Table 2. As the table shows, adding features supplementary to the TRF score generally improved cross-validation accuracy on the seed documents, often by a wide margin; however, the team's results on the official relevance judgments failed to replicate this improvement.

Mechanical Turk

We intended to use Mechanical Turk (MT) to create one of our runs. A sample of documents were submitted to MT for relevance assessment; the plan was to use the assessments gathered to train a separate or enhance an existing machine classifier run. Failing that, the MT assessments could at least be used to calibrate the probability of relevance values assigned to submitted documents. The results gathered from MT were, however, too low in quality even for the latter usage.

Our human intelligence tasks (HITs) for MT were designed as follows. A topic statement was presented to the Turker, accompanied by six documents (emails or text-version attachments). The Turker was required to assess the relevance of each document to the topic. For each document, the Turker had the choice of Highly Relevant, Relevant, Irrelevant, and Not in English. In fact, almost all of the documents in the

Topic	Ours			All participants		
	rmitindA	rmitmlfT	rmitmlsT	Best	Median	Worst
200	15.3%	2.3%	1.8%	25.8%	12.1%	1.8%
201	13.0%	6.8%	7.1%	53.5%	13.6%	1.5%
202	37.8%	41.7%	38.2%	70.6%	35.3%	4.5%
203	32.2%	11.3%	19.6%	39.4%	24.2%	3.2%
204	17.1%	5.3%	8.7%	26.6%	10.3%	5.3%
205	52.1%	34.5%	35.1%	52.1%	46.7%	18.0%
206	5.9%	18.7%	7.1%	37.0%	13.6%	3.4%
207	24.3%	11.2%	16.4%	90.3%	18.9%	6.7%

Table 3: Hypothetical F1 scores of our learning task runs, compared to all participant systems. A run's hypothetical F1 score is the F1 score the run would have achieved had it picked the optimal cutoff depth.

Enron corpus are in English. To set up a trap question, however, one of the six documents in every set was an email in German, taken from a technical mailing list. If the Turker failed to identify this document as not being in English, the HIT was rejected. Unfortunately, the rejection rate turned out to be very high: 87% of HITs failed the trap question, worse even than would occur by purely random clicking. Such a high failure rate questions the quality of the few HITs that passed the trap question, too. As a result, we judged the Mechanical Turk assessments to be unusable.

Our experience of Mechanical Turk is an interesting one. How can such a high failure rate occur on such an easy trap question? Was the payment offered too low (2 cents per hit)? Was the task too easy to automate (only buttons needed to be pressed)? Or did the task require too much attention from the Turkers?

2.2 Results

The hypothetical F1 scores, based upon an optimal cutoff depth, for the Melbourne runs are given in Table 3, along with the best, median, and worst hypothetical F1 scores across all participants. It is evident, first, that the TRF run performed slightly above the median for most topics; and second, that the machine classification runs generally performed worse than the TRF ones, even though they used TRF scores as a feature. We investigate the poor performance of the classifier in Section 2.3

Learning task runs had to include a probability of relevance for each document. Relevance estimates for the TRF run were formed by linearly scaling retrieval similarity scores to the $[0, 1]$ range; the resulting probabilities of relevance greatly overestimate yield. The classifier runs used the method proposed by Platt [1999] to derive probabilities from classifier predictions using cross-validation. The resulting sum of probabilities, however, was much too high. The sampling of seed documents appears to have been strongly biased in favour of relevant documents. If the sampling design had been available, it should have been possible to correct the estimates accordingly. In the absence of sampling information, the solution adopted was to scale probabilities by

Topic	Cross-validation		Ranking		Seed–qrel agreement	
	Seeds	Qrels	Seeds	Qrels	κ	Accuracy
200	0.71	0.82	0.73	0.66	0.30	0.71
201	0.92	0.77	0.91	0.69	0.06	0.50
202	0.91	0.96	0.91	0.94	0.34	0.92
203	0.79	0.89	0.81	0.68	0.19	0.60
204	0.79	0.79	0.83	0.60	0.31	0.71
205	0.90	0.90	0.90	0.86	0.17	0.84
206	0.96	0.96	0.98	0.77	0.02	0.19
207	0.92	0.92	0.93	0.58	0.14	0.78

Table 4: Tenfold cross-validation AUC values for the full feature set classifier on the seed documents and on the official qrels (left); AUC results for ranking the assessed seed and qrel documents using a model trained on the seed documents (middle); and agreement between seed and qrel assessors for commonly assessed documents, measured by Cohen's κ and accuracy (agreed assessments as proportion of all) (right).

the number of relevant documents found in last year's interactive task, for which reason we designate the classifier runs as technology-assisted, rather than fully automated.

2.3 Result analysis

The most striking result from our submission is the poor performance of the machine classifier runs. Cross-validation on the seed documents suggested that the addition of these features to the TRF score would in most cases lead to an improvement in accuracy (Table 2). The AUC scores for some topics were above 0.9, suggesting very accurate classification. Furthermore, training and cross-validating using the official assessments, instead of the seed documents, produces comparable AUC scores, as the first two columns of Table 4 attest. If, however, we take the model developed on the seed documents, and use it to rank first the seeds and then the official assessments, we see a sharp fall in scores on official compared to seed assessments, as shown in the third and fourth columns of Table 4. This fall suggests that the seed documents may be inaccurate examples for the official assessments.

Evidence of seed bias is found in the low level of agreement between the seed and official assessors, on documents assessed by both. The second last column of Table 4 provides the Cohen's κ scores for this agreement (a κ of 1 means perfect agreement, whereas a κ of 0 means only random agreement), while the last column gives the proportion of assessments the two groups agreed on. Agreement ranges from fair to random to (in the case of Topic 206) almost adversarial. Such low agreement between seed data and official assessments suggests that this is a challenging data set on which to train a classifier, and indeed our classifier achieves its best results (Table 3) on those topics (namely Topics 202 and 205) for which seed accuracy is the highest.

Seed bias cannot, however, be the only cause of poor classifier performance, since the classifier performed lamentably on some topics that have tolerable seed–qrel agree-

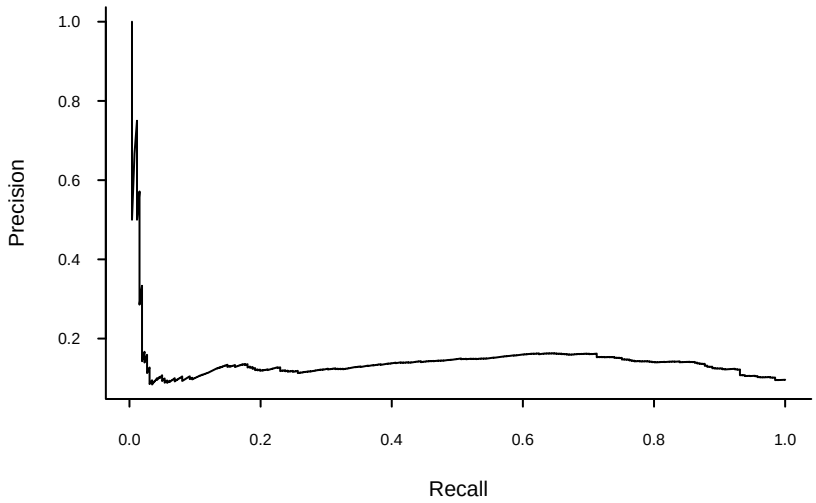

Figure 1: Recall-precision curve for the full-feature SVM run on Topic 200, using the official qrels, and considering only those documents contained in the official qrels.

ment, most notably Topics 200 and 207. Indeed, the classifier achieved an official AUC of 0.28 on Topic 207, meaning that it performed worse than a random ordering of documents.[2] Moreover, the true-relevance feedback run itself learns from the seed documents, and was able to do so without a general calamitous drop in performance.

Part of the classifier's failure to use seed documents to predict official assessments may lie in the features or classifier used. Consider the recall-precision curve over assessed documents for Topic 200, shown in Figure 1. The classifier pushes a few relevant documents to the top of the list, but then precision falls precipitously, only to start rising again further down the ranking. Such a result is suggestive of over-fitting: a few relevant documents mislead the classifier into promoting a largely irrelevant class.

A likely culprit for such over-fitting is Feature 3 of Table 1, the proportion of a custodian's seed documents which are relevant. Indeed, the upper ranks of the run displayed in Figure 1 are dominated by the emails of a particular custodian, a few of which are relevant, the rest not; meanwhile, a quarter of the way down the ranking, a larger cluster of relevant emails from another custodian are located. Perhaps the former custodian has only relevant instances in the seed set, and the latter only irrelevant ones. Such an error could be caused by misleading assessments by the seed assessor; but it could equally be caused by the biased way in which the seed documents were selected, from runs returned by previous participants, not at random from the collection as a whole. The resulting over-fitting would not be caught by standard ten-fold cross-validation, which assumes that the seed examples are accurate and unbiased. The textual features used in true-relevance feedback may be more robust to errors of this sort.

[2]The official AUC score is estimated over unassessed documents, whereas the AUC score in column 4 of Table 4 only considers assessed documents.

3 Interactive task

The Melbourne team also participated in this year's interactive task, submitting a run for Topic 302. The run was developed using a combination of human-directed Boolean keyword search and machine classification. Classifier effectiveness in locating relevant documents was low, however, perhaps in part because relevant documents were few. Appeals were selected by a three-fold review of conflicting assessments; this review offers insights into assessor agreement.

3.1 Method

A commercial e-discovery tool was used to perform keyword searches, browse the corpus, and tag documents as relevant or irrelevant. The assessed documents were then fed as seed documents to the classifier, using all the features listed in Table 1. The output of the classifier was used to select new documents for assessment, while human-directed searches continued in parallel.

The original intention was to assess documents the classifier found ambiguous; under the SVM model, documents sitting close to the separating hyperplane. In practice, though, the classifier had low precision even amongst the top-ranked documents, and greatly overpredicted the proportion of relevant documents. We judged it unlikely that documents nominally close to the hyperplane (with a prediction score near 0.0) were truly ambiguous.[3] Instead, the top few hundred unassessed documents were selected from the classifier for assessment at each iteration.

3.2 Results

Table 5 gives the precision of the classifier's top-n documents at each iteration of the run development, as measured by the team's own relevance assessments. Even though we were selecting the documents with the highest relevance prediction from the classifier, precision was low, ranging from 5% to 16%. The variation in precision at each iteration, though borderline significant ($\chi^2 = 12.2$, $p = 0.06$), is likely due to using different assessors. The low precision of the top-n results is despite the fact that the classifier was giving a positive relevance prediction to roughly a quarter of the corpus.

Human-directed searches, primarily by iteratively refined Boolean queries, were carried out in parallel to the classifier runs. Precision cannot meaningfully be compared between human-directed and classifier searches, as the methods and aims differ. Nevertheless, the human-directed searches were not producing many new relevant documents. (The jump in the total number of relevant documents between September 8th and September 14th is due in part to reassessment of earlier results.)

A more controlled comparison between human-directed and classifier search was carried out as part of the September 13th iteration. Three query documents, addressing three different aspects of the topic, were created by extracting extended sections of highly relevant text from assessed relevant documents. These query documents were then submitted as queries to the BM25 system, without relevance feedback or the use

[3] These problems may have been due to our use of AUC, instead of loss, as the SVM objective function.

		Proprietary deduplication					
Date	Method	Num	Rev	Rel	TA	!Rel	Precision
Aug-31	Top 100	153	135	7	7	121	5.5%
Sep-02	Top 100	87	86	14	1	73	16.1%
Sep-03	Top 200	258	253	15	4	235	6.0%
Sep-06	Top 100	150	150	17	1	134	11.3%
Sep-08	Top 100	87	87	9	0	79	10.2%
Sep-13	Top 200	188	188	15	0	175	7.9%
Sep-14	Top 200	250	235	18	1	217	7.5%
Incident	Top 100	132	120	8	0	113	6.6%
Prevent	Top 100	93	93	4	0	89	4.3%
Other	Top 100	104	88	3	0	85	3.4%
All Sep-06			28511	157	27	28356	
All Sep-08			28696	176	28	28521	
All Sep-14			29296	256	30	29027	

Table 5: Results of classifier runs at different dates. The undeduplicated top n emails and top n attachments were taken from the classifier at each iteration. "Num" is the number selected, after deduplication, and "Rev" the number reviewed. "Rel" is the number found relevant, "TA" the number referred to the TA, and "!Rel" the number found irrelevant (a document could, erroneously, be assigned to more than one reviewed category). The precision at each iteration is given in the final column. The rows labelled "Incident", "Prevent", and "Other" give results for three query documents lodged on September 13; see the text for further explanation. The "All" rows give totals for their respective dates, including results from human-directed searches.

of machine classification. The top n results for each of these artificial query documents are shown in the eighth through tenth rows of Table 5. The manually created queries were even less successful in locating relevant documents than the machine classifier.

We originally intended to rank documents using the classifier, then truncate the ranking at a cutoff based on cross-validation or manual sampling. In the event, however, the classifier proved to be too inaccurate. Even at the top of the ranking, only 10% of documents were relevant; therefore, accepting unreviewed documents on the classifier's recommendation would damage our run's precision. Therefore, our final submission consisted solely of documents that had been manually reviewed and assessed as relevant by us.

3.3 Interim assessments

Our final submission consisted of 326 officially deduplicated documents, across 265 distinct messages. As our classifier was still yielding around 8% relevant documents in the top 100 at each iteration, we were confident that there were more relevant doc-

Returned	Assessed		
	Relevant	Not relevant	Total
Returned	113	152	265
Not returned	235	11,542	11,777
Total	348	11,694	12,042

Table 6: Agreement between the submitted Melbourne run and the official, pre-appeal assessments on the documents included in the assessment sample. The 238 documents found unassessable by the assessors are excluded.

Stage	Precision	Recall	F1
Pre-appeal	0.22	0.080	0.120
Post-appeal	0.45	0.200	0.277
↪ Rank	4	1	1

Table 7: Estimated, pre-appeal (row 1) and post-appeal (row 2) effectiveness scores for the Melbourne interactive run, and post-appeal rank of Melbourne run amongst six participants for the topic (row 3).

uments in the collection, but we believed that the number was not large. The official, post-appeal assessments found 326 relevant documents in the assessment sample; the official estimate for the population has not been provided, but (inferring from our scores and the known data) is somewhere around 1,000.

The document-level agreement on the assessment sample between our returned run and the interim, pre-appeal assessments is shown in Table 6. Agreement is only fair ($\kappa = 0.35$), although our failing to return a document should not be taken as an explicit judgment that the document was not relevant. The interim results confirmed our finding that there are few relevant documents for the topic large. At the message level, the sampled assessment estimates that there are 740 relevant messages in the collection, and that our run located 59 of them. The estimated effectiveness scores from the interim assessments are shown in the first row of Table 7.

3.4 Appeals

There were altogether 387 document-level assessments that disagreed with our run: 235 documents we did not return were assessed relevant, and 152 documents we returned were assessed irrelevant. To determine which assessments to appeal, we performed a multi-assessor, blind re-assessment of the conflicting documents, on the basis of the topic authority's guidelines to the assessors. Seven of our team members took part in this review. Each document was randomly assigned to three different reviewers. Order of review was randomized, and reviewers were not told what the original return status or assessment of the document were.

Interim assessment	Proportion reviewed as relevant				
	0	$1/3$	$1/2$	$2/3$	1
Relevant	145	38	4	34	14
Not Relevant	72	42	0	26	12

Table 8: Agreement between reviewers and assessors on the documents whose assessment disagreed with our run. Each cell counts the number of the 387 reviewed documents falling into that category. In the columns, we count the proportion of the three reviewers who regarded the document as relevant; this excludes reviews of "unassessable", making the proportion of $1/2$ possible. Documents officially assessed as "unassessable" were likewise excluded from the review process.

Interim assessment	Final assessment			Total
	-1	0	1	
0	3	8	27	38
1	14	58	36	108
Total	17	66	63	146

Table 9: Adjudication outcome for appealed documents. Counts are for documents appealed by the Melbourne team. For instance, the first row, fourth column shows that 27 of our appeals against interim assessments of "irrelevant" (0) were upheld by the topic authority. The -1 assessment means "unassessable"; we appealed no assessments of this type.

The agreement between the threefold reviewers and the original assessors for each conflicting document is shown in Table 8. The reviewers disagreed with relevant assessments more often than they did with irrelevant ones. The review result was taken as the majority assessment of the three assessors for a document; if this result disagreed with the official assessors, the document was appealed. Document-level assessments were not appealed, however, if the result would not change message-level assessments; for instance, we did not appeal relevant assessments for message bodies if we agreed that an attachment to the message was relevant.

Altogether, some 146 of the 387 document-level assessments that conflicted with our run were appealed, making up 106 of the 281 conflicting message-level assessments. The adjudication results of these appeals is reported in Table 9. A surprising proportion (12%) of documents initially found assessable by the interim assessors were marked unassessable by the topic authority. Excluding these documents, 77% of our appeals against irrelevant judgments were upheld, as were 62% of our appeals against relevant judgments.

Our final, post-appeal scores is shown in the second row of Table 7. Success in appeals appreciably boosted our scores, though we lack the data to say what boost other teams received. We finished first out of six participants in recall and F1, but fourth in

Reviewer	I_2	I_3	A_1	A_2	A_3	A_4
I_1	0.42	0.51	0.05	0.37	0.09	0.32
I_2		0.13	0.21	0.31	0.21	0.41
I_3			0.13	0.52	0.36	0.42
A_1				0.22	0.22	0.16
A_2					0.45	0.49
A_3						0.47

Table 10: Inter-reviewer agreement (Cohen's κ) between the seven reviewers for conflicting assessments. Reviewers labelled I are from industry, A from academia. Reviews of messages where we disagree with a relevant assessment on the body, but agree that there is a relevant attachment, are excluded.

precision. The relative showings are surprising: we only submitted documents that we had manually reviewed, which should boost precision at the cost of recall. These results suggest that other teams were even more conservative in their productions. In any case, the wide errors bounds on all scores (as reported in the official results) make it difficult to state which participants had conclusively superior productions.

Part of the motivation for determining appeals by a randomized, multi-assessor review was to measure inter-assessor agreement. Table 10 shows the κ between each reviewer. Contested-relevance message bodies with agreed-relevance attachments are excluded, since these bodies are frequently near-empty, and reviewer agreement on rejecting a relevant assessment is misleadingly high. Three reviewers were from industry (though none had legal training), while four were from academia. Inter-reviewer agreement is variable, ranging from near-random at 0.05, to moderate at 0.52; there is no clear pattern of homogeneity within industry or academic reviewer groups. These agreement figures need to be interpreted with some care: on the one hand, the reviewers all took part in run development, and so might be expected to have a more coherent conception of relevance; on the other, though, the documents assessed are those whose relevance is contested, and may therefore be more difficult to determine.

3.5 Analysis

It appears that Topic 302 was a topic for which there were, in e-discovery terms, relatively few relevant documents. Even so, the performance of the machine classifier is disappointing. Only 10% of the top-ranked unassessed documents were actually relevant, and this proportion neither increased as the classifier was trained, nor decreased as the pool of relevant documents was depleted. The poor accuracy of the classifier meant that it could not be used to automatically classify relevant documents, but was only usable as a means of suggesting documents for manual review.

While the poor precision of the classifier's top ranks might be attributable to the sparsity of relevant documents, the waywardness of the classifier's relevance predictions is not. As mentioned previously, the classifier consistently gave a positive predic-

tion to a quarter of the documents in the corpus, meaning that it regarded them as more likely to be relevant than not. It is possible that this error is due to the nature of the training instances. These instances were far from randomly chosen. Instead, they consisted on the one hand of a large number of irrelevant documents (see the bottom rows of Table 5), belonging mostly to a few clearly irrelevant classes (for instance, historical calendar appointments regenerated as emails when Enron migrated mail servers); and on the other hand of a much smaller number of documents, chosen for review because of strong apparent evidence of relevance, though not in fact with a high proportion actually relevant.

4 Summary

A collaborative team of academic and industry participants, based in Melbourne, took part in both the interactive and the learning tasks of this year's Legal Track. The core retrieval method was a true-relevance feedback (TRF) run using the BM25 retrieval model. The TRF result was then supplemented with six other features, and an SVM classifier trained.

Three runs were submitted to the learning task. The first, using only the TRF scores, performed respectably. The other two, using machine classification based on the TRF score plus all or a subset of the other features, performed much worse, despite promising AUC scores under cross-validation. An analysis suggests, first, that the seed documents are unreliable indicators of officially assessed relevance, and second, that the features chosen for the classifier were not robust to seed bias.

A single run was submitted for Topic 302 of the interactive task. The process was a blend of machine classification and of human-directed search using a commercial e-discovery tool. The classification method was the same as for the learning task, namely true-relevance feedback supplemented with other features and used to train an SVM classifier. Even at the top of its ranking, the classifier returned a low proportion of relevant documents, perhaps due in part to there being few relevant documents in the collection; additionally, the classifier grossly overpredicted the prevalence of relevance. A multi-assessor review process was employed to determine appeals, showing variable agreement between reviewers. Nevertheless, our interactive run achieved (non-significantly) top scores in recall and F1.

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.