University of Waterloo at TREC 2010: Legal Interactive

Mark D. Smucker[1], Charles L. A. Clarke[2], Gordon V. Cormack[2], and Olga Vechtomova[1]

[1]Department of Management Sciences, University of Waterloo
[2]David R. Cheriton School of Computer Science, University of Waterloo

Abstract

This year the University of Waterloo (UW) participated in the TREC Legal Interactive track and used the same process as last year except that this year we used three different human operators as opposed to only one as UW did last year. We participated in three topics: 301, 302, and 303. Relative to other participants, we performed well on one of the three topics. For two of the topics, low recall significantly hurt our F1 scores. Overall, we believe a contributing factor in our lower performance this year was with our interaction with the topic authorities, which resulted in our failing to understand the wide range of what constituted relevant material.

1 Introduction

This year we had two goals for our participation in the legal interactive track. Our first goal was to test the system and process used by the University of Waterloo (UW) in 2009 [1] with human operators other than one of the system's creators. The second goal was to bring up to speed other researchers at UW on the legal interactive track process. We attempted to achieve these goals by having three different researchers tackle one of three different topics. The three researchers are the co-authors of this paper minus Gordon Cormack, who is one of the creators of the UW system.

In 2009, UW participated in four topics and in all four of these topics, UW achieved the highest F1 [2]. The process used to achieve these results involves a human operator conducting an initial search for relevant documents using a standard search interface that produces 10 results with variable length summaries in response to a query. The operator can judge the summary as relevant, not relevant, or flag it as seen with no judgment. From a summary, the operator can click to view the whole document and parent email and attachments.

After searching for and judging documents in this fashion, the process then switches to an active learning process. A classifier is trained on the judged documents. Random documents from the collection are used as additional non-relevant documents if the training set is unbalanced. The classifier ranks the documents from most likely to least likely to be relevant. Using this ranking, the operator judges unjudged documents. After another period of judging, the classifier can be trained on the new, larger set of judgments and this process is repeated. At the conclusion of judging, the distribution of relevant documents is estimated to determine a cutoff that aims to optimize F1.

The operator's job is to train the classifier to be a model of what the topic authority (TA) would consider relevant and not relevant, and throughout the whole process, the operator engages the TA as appropriate to help maintain a focus on what is and is not relevant.

While the final ranking is automatically produced, the operator's choices from the initial search to the end of the active learning stage may have a significant impact on performance.

To perform a crude analysis of the sensitivity of the process to the operator in charge, this year we replaced UW's 2009 sole operator, Gordon Cormack, with three other IR researchers (the three other co-authors of this paper). Of note, the three IR researchers had no previous experience with the legal interactive track. While we did this experiment primarily to see how robust the UW process is to different human operators, we also did this experiment out of necessity — Gordon Cormack was one of the track's organizers this year.

2 Methods

The methods employed were the same as UW used in 2009 [1]. We submitted results for 3 topics: 301, 302, and 303. A different operator searched for relevant

	Estimated Recall			Estimated Precision			Estimated F1		
	watlint10		All Runs	watlint10		All Runs	watlint10		All Runs
Topic	Recall	Rank	Mean	Precision	Rank	Mean	F1	Rank	Mean
301	0.019	5/5	0.131	0.578	3/5	0.502	0.036	5/5	0.160
302	0.169	2/6	0.134	0.732	1/6	0.464	0.275	2/6	0.181
303	0.134	6/6	0.488	0.773	1/6	0.558	0.228	6/6	0.461

Table 1: Estimated recall, precision, and F1 for our run, watlint10. Also shown is the rank of the run among submitted runs with rank 1 being the best scoring run. The mean recall, precision, and F1 is shown for all of the participant runs submitted for each of the three topics.

	Search		Active Learning		
Topic	Non-relevant	Relevant	Non-relevant	Relevant	Total
301	223	256	2381	653	3513
302	503	141	783	57	1484
303	359	872	1165	1484	3880

Table 2: Number of relevance judgments made per topic using the manual search interface and the active learning interface.

documents and trained the classifier for each topic.

2.1 Results and Discussion

Table 1 shows the results for our submission, watlint10, compared to the other participants for topics 301, 302, and 303. Compared to other participants, we did best on topic 302 and effectively tied for best performance on this topic (F1 of 0.275 vs. 0.277).

Across the three topics, we tended to have good precision but low recall. In particular, we believe that our interaction with the topic authorities on topics 301 and 303 contributed to our low recall on these topics.

For topic 301, the operator exchanged a number of email messages with the TA, including discussions about specific documents. Nonetheless, the operator's understanding of what was relevant turned out to be much narrower than the TA's understanding, an error which we believe substantially contributed to the low recall values for that operator. Our recall on topic 302 was low, but the recall was above average compared to other participants. For topic 303, there was a general lack of interaction with the topic authority, which we believe contributed to our low recall on this topic.

Overall, our results in 2009 were dramatically better than this year. As described in the introduction, in 2009 we had a single operator while this year we used three new operators. Without a measure of how last year's operator would have performed this year, we can't conclusively isolate the cause of this year's lower performance.

Perhaps the most significant difference between 2009 and 2010 is that this year considerably fewer documents were judged for relevance. In 2009, 50,000 judgments were completed for 4 topics, which is an average of 12500 documents judged per topic. This year, 8,877 judgments were completed for 3 topics, which is an average of 2959 documents per topic. Table 2 shows detailed counts of judgments for each topic.

This difference in the amount of judging may be caused by both the topics this year compared to last year as well as the difference between human operators. For example, between the three operators this year, they individually judged 3513, 1484, and 3880 documents.

Unfortunately, our experiment was not setup to cleanly compare the effect of varying the human operator. Each operator worked on a different topic. Nevertheless, we did succeed at our second goal, which was for the co-authors to learn more about the legal interactive track. Going forward we now know that the operator of our system has to actively engage the topic authority to clearly understand the scope of a topic's relevance, which can be much broader than in non-legal domains.

3 Conclusion

This year the University of Waterloo (UW) used the same process for the legal interactive track as UW

used in 2009 except that instead of one human operator, UW used three different operators. While we found differences between the behavior of the operators in terms of the number of documents judged, the number of documents judged is not a good predictor of final performance. We performed well compared to other participants on one of three topics. On the other two topics, we believe that our interaction with the topic authorities contributed to our low recall on each of these topics.

4 Acknowledgments

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

[1] G. V. Cormack and M. Mojdeh. Machine learning for information retrieval: TREC 2009 web, relevance feedback and legal tracks. In *TREC 2009*. NIST.

[2] B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard. Overview of the TREC 2009 legal track. In *TREC 2009*. NIST.