Some Lessons Learned To Date from the TREC Legal Track (2006-2009)

Douglas W. Oard, Jason R. Baron and David D. Lewis

February 24, 2010

For four years now, the Text Retrieval Conference (TREC) Legal Track administered by the US National Institute of Standards (NIST) has untaken yearly studies evaluating the application of Information Retrieval (IR) methods to e-discovery in the context of U.S. civil litigation. In this short paper, we distill some of what has been learned. As we write this, analysis is not yet complete for some parts of the 2009 Legal Track, while the 2010 Legal Track is only now beginning. "Lessons learned" are, therefore, a work in progress. In an effort to be concise, we stick to bullet points – much more could be said about any of this. We are indebted to our colleagues for much of what we have learned, but we emphasize that these are our personal observations and that others might see things differently. Other perspectives on the Legal Track can be found in TREC proceedings papers by organizers and participants at http://trec.nist.gov/.

Terms of Reference

- Calling it the "Legal Track" may have been a mistake. Many people think of searching legislation and/or case law, rather than e-discovery, when they hear that term.
- Our sole focus to date has been on document review for responsiveness. Review for responsiveness is just one of several aspects of e-discovery to which IR techniques could be applied; others include early case assessment and privilege review.
- Review for responsiveness has applications beyond the civil litigation setting from which the term "e-discovery" arises, including close cousins such as "second requests" in the antitrust context, requests filed under state freedom of information laws or the federal Freedom of Information Act (FOIA), and some forms of criminal investigations.
- Some IR researchers focus principally on designing and studying human processes that are supported by automated tools, while others focus principally on designing and studying automated tools that could in principle support many different processes. The Legal Track has included both "interactive" tasks focused on the study of processes, and "batch" tasks focused on the study of algorithms and automated systems that embody them.

Some Background on TREC

- It is a common misconception that TREC – which includes tracks that address a wide range of challenges – exists to *conduct* evaluation. It is true that TREC conducts evaluations, but more importantly TREC exists to *facilitate* evaluation. TREC facilitates evaluation by developing evaluation data sets ("test collections"), by developing evaluation methods (ways of using those test collections to measure how well something works), by helping to build research communities that come together around important problems, and by generating baseline results to which the results of future work can be compared.

- What we broadly call "the scientific method" relies on the replicability of experiments. The evaluation resources developed in TREC must therefore be reusable. Reusable evaluation resources facilitate an evaluation-guided research paradigm in which (system-oriented) researchers iterate between using evaluation results as a basis for improving system design and testing those putative improvements by generating new evaluation results.
- Experience across a range of TREC tracks has shown that large, reusable test collections can be affordably created. The key idea is to sample and manually assess documents from the results returned by diverse systems. In combination with robust effectiveness measures, this can allow fair evaluation of systems which did not exist at the time of test collection creation. Indeed, one motivation for TREC to conduct annual evaluations is to attract results from a range of high performing systems, to support test collection construction.
- These annual evaluations produce baseline results as an important byproduct. The results truly are baselines, not "best effort" or upper bounds, for several reasons: (1) a strict schedule must be enforced to collectively guide sampling, so time available to each team is limited, (2) the process is conducted only once each year, so there is no opportunity to fix simple errors after seeing the results, (3) teams may (and often do) focus on measuring the effect of varying some aspect of their system rather than on optimizing overall system effectiveness. Indeed, industry participants often field research prototypes rather than operational systems, so TREC participants must agree not to use TREC results in advertising.

Some Lessons Learned

- The greatest success of the Legal Track has been bringing together experts in e-discovery and IR. The participating IR researchers in particular have gained an appreciation of the immense human effort that e-discovery represents in the modern legal system.
- An important byproduct of the first four years of Legal Track evaluations is a reusable test collection containing about 7 million scanned business documents from tobacco lawsuits, about 100 discovery requests, and a set of sampled assessments of responsiveness sufficient to support evaluation of future systems. In 2009, we started creating a second test collection, this one based on several hundred thousand email messages and attachments. The initial collection is not yet large enough for reliable reuse (we now have 7 production requests; the best estimates are that at least 40 production requests are needed). We do hope to ultimately create a reusable email test collection, possibly as early as 2010.
- The fidelity with which we modeled review for responsiveness improved substantially over this period in two ways. First, we developed sampling approaches and effectiveness measurements that could handle realistic (e.g., very large) response sets. Second, we learned to model a (relatively expensive) iterative human-in-the-loop process that reflects an actual use case with far higher fidelity than can be achieved using batch evaluation.
- Our thinking on evaluation of interactive search has evolved from an initial focus on retrieval effectiveness to a more nuanced view that interaction defines a continuum and that the central issue is therefore the balance between cost and effectiveness.

- Effectiveness estimates are, of course, impacted by details of the evaluation process. For example, measured effectiveness generally increased substantially when participants were allowed to appeal selected initial assessments. Our thinking on sampling and the design of evaluation measures evolved considerably as we developed ways of limiting inconsistencies that result when multiple human assessors of relevance are used, and we also learned to make explicit estimates of the accuracy of our measurements. We note that sensitivity to evaluation details is particularly strong when the proportion of responsive documents is low, a fact with practical consequences for anyone evaluating competing e-discovery solutions.
- For the vast majority of our production requests, fewer than half of all responsive documents were retrieved by a Boolean query negotiated by lawyers without interactive access to the collection. This was despite thoughtful keyword choices and the use of Boolean, truncation, and proximity operators in a formally correct fashion. While perhaps surprising, this result is consistent with every IR research study of which we are aware in which the number of unretrieved responsive documents from a large collection has been reliably estimated. We agree with the common view that the cause of this problem is the difficulty of anticipating the range of ways in which language is used to convey ideas of interest in a large collection.
- Some improvements over Boolean queries have been demonstrated using fully automated techniques based on statistical analysis of term usage–the TREC 2008 paper from the University of Waterloo is a good exemplar–but the demonstrated improvements are still relatively small. Choosing where to cut off the list of documents produced by a typical statistical ranked retrieval system has proven to be a surprisingly difficult issue.
- Intriguing results from the TREC 2008 interactive task suggest that the benefits of greater human effort may be more apparent in recall (the fraction of the responsive documents that are found) than in precision (the fraction of the found documents that are responsive). Every team in the 2008 interactive task achieved similar precision for the one production request that all teams tried, but they obtained vastly different levels of recall. Recall appeared to correspond very roughly to the human effort that each team invested. Extrapolating these results to e-discovery in general would not yet be appropriate, but we hope to gain further insight into this possible effect from the 2009 and 2010 interactive tasks. In the mean time, a set of papers published in the 2009 IEEE Conference on Systems, Man and Cybernetics by authors from H5 and elsewhere offer the best available description of this result.

Looking Forward

- Many TREC tracks enjoy a rich afterlife, spawning new venues where interested parties can work together. These include TREC-like frameworks originating outside NIST (often outside the US), as well as “data challenges” at academic workshops. Interest is growing in some way of certifying what constitutes “reasonable” processes for reviewing documents for responsiveness, and for other phases of e-discovery. TREC’s Legal Track is just one of several cooperative activities around e-discovery, and it will be natural for them to interact.
- Watch for a 2010 special issue on “E-Discovery” in *Artificial Intelligence and Law Journal*.