Subject Matter Experts

What Role Should They Play in TAR 2.0 Training?

John Tredennick

One of the givens of traditional technology-assisted review (“TAR”) is the notion that a subject matter expert (“SME”) is required to train the algorithm. During a recent EDRM webinar, for example, I listened to an interesting discussion about whether you could use more than one expert to train the algorithm, presumably to speed up the process. One panelist stated confidently that using four or five SMEs for training would be unworkable. (I guess they would be hard to manage.) But she wondered whether two or three experts might be OK.

In quick response, another speaker cautioned that consistency was the key to effective training against a reference set (a staple of traditional TAR). He cautioned that having one expert review the training documents was critical to the process.

I found myself wanting to jump into the conversation. (I couldn’t, of course, because webinars don’t work that way.) Putting aside whether experts are consistent even in their own tagging, are we really sure that TAR training is an “experts-only” process, as is suggested by many proponents of what I call TAR 1.0?

What about having review teams assist in the training process? What if we use experts to do what they do best—find good documents to help get the ranking started? Then send the highest-ranked documents to the review team so they can start right away? Let the expert continue to find good documents through witness interviews and search techniques or even some form of random sampling. That approach would allow the review team to get going right away, rather than wait for the expert to finish the document-training process.

My thoughts were not just pipe dreams—I had seen the results of our research examining just this question. Dr. Jeremy Pickens, our Senior Research Scientist, had done experiments using the TREC data from 2010 to see whether expert training provided better rankings than could a review team. The question is important because a lot of review managers find it difficult to keep their teams waiting while a senior person finds time to look at 3,000 or more documents.
necessary for TAR training.\(^1\) On top of that, the same expert is required to come back and train any time new uploads are introduced to the collection.

Dr. Pickens’ research was done using Insight Predict, our proprietary engine for predictive ranking. He did it in conjunction with our research on the benefits of continuous ranking, which I wrote about in a separate post. The goal for that work was to see if a continuous learning process might provide better ranking results and, ultimately, further reduce the number of documents necessary for review. Our conclusion was yes, that continuous ranking could save on review costs and cut the time needed for review.

Welcome to TAR 2.0, where we challenge many of the established notions of the traditional TAR 1.0 process.

**Where Do Experts Fit in TAR 2.0?**

If you accept the cost-saving benefits of continuous ranking, you are all but forced to ask about the role of experts. Most experts I know (often senior lawyers) don’t want to review training documents, even though they may acknowledge the value of this work in cutting review costs. They chafe at clicking through random and often irrelevant documents and put off the work whenever possible.

Often, this holds up the review process and frustrates review managers, who are under pressure to get moving as quickly as possible. New uploads are held hostage until the reluctant expert can come back to the table to review the additional seeds. Indeed, some see the need for experts as one of the bigger negatives about the TAR process.

Continuous ranking using experts would be a non-starter. Asking senior lawyers to review 3,000 or more training documents is one thing. Asking them to continue the process through 10,000, 50,000 or even more documents could lead to early retirement—yours, not theirs. I can hear it now: “I didn’t go to law school for that kind of work. Push it down to the associates or those contract reviewers we hired. That’s their job.”

So, our goal was to find out how important experts are to the training process, particularly in a TAR 2.0 world. Are their judgments essential to ensure optimal ranking or can review team judgments be just as effective? Ultimately, we wondered if experts could work hand in hand with the review team, doing tasks

\(^1\) Typical processes call for an expert to train about 2,000 documents before the algorithm “stabilizes.” They also require the expert to review 500 or more documents to create a control set for testing the algorithm and a similar amount for testing the ranking results once training is complete. Insight Predict does not use a control set (the system ranks all the documents with each ranking). However, it would require a systematic sample to create a yield curve.
better suited to their expertise, and achieve better and faster training results—at less cost than using the expert exclusively for the training.

Our results were interesting, to say the least.

Research Population

We used data from the 2010 TREC program\(^2\) for our analysis. The TREC data is built on a large volume of the ubiquitous Enron documents, which we used for our ranking analysis. We used judgments about those documents (i.e. relevant to the inquiry or not) provided by a team of contract reviewers hired by TREC for that purpose.

In many cases, we also had judgments on those same documents made by the topic authorities on each of the topics for our study. This was because the TREC participants were allowed to challenge the judgments of the contract reviewers. Once challenged, the document tag would be submitted to the appropriate topic authority for further review. These were the people who had come up with the topics in the first place and presumably knew how the documents should be tagged. We treated them as SMEs for our research.

So, we had data from the review teams and, often, from the topic authorities themselves. In some cases, the topic authority affirmed the reviewer’s decision. In other cases, they were reversed. This gave us a chance to compare the quality of the document ranking based on the review team decisions and those of the SMEs.\(^3\)

Methodology

We worked with the four TREC topics from the legal track. These were selected essentially at random. There was nothing about the documents or the results that caused us to select one topic over the other. In each case, we used the same methodology I will describe here.

For each topic, we started by randomly selecting a subset of the overall documents that had been judged. Those became the training documents, sometimes called seeds. The remaining documents were used as evaluation (testing) documents. After we developed a ranking based on the training

---

\(^2\) The Text Retrieval Conference is sponsored by the National Institute for Standards and Technology. (http://trec.nist.gov/)

\(^3\) We aren’t claiming that this perfectly modeled a review situation but it provided a reasonable basis for our experiments. In point of fact, the SME did not re-review all of the judgments made by the review team. Rather, the SME considered those judgments where a vendor appealed a review team assessment. In addition, the SMEs may have made errors in their adjudication or otherwise acted inconsistently. Of course that can happen in a real review as well. We just worked with what we had.
documents, we could test the efficacy of that ranking against the actual review tags in the larger evaluation set.  

As mentioned earlier, we had parallel training sets, one from the reviewers and one from the SMEs. Our random selection of documents for training included documents on which both the SME and a basic reviewer agreed, along with documents on which the parties disagreed. Again, the selection was random so we did not control how much agreement or disagreement there was in the training set.

**Experts vs. Review Teams: Which Produced the Better Ranking?**

We used Insight Predict to create two separate rankings. One was based on training using judgments from the experts. The other was based on training using judgments from the review team. Our idea was to see which training set resulted in a better ranking of the documents.

We tested both rankings against the actual document judgments, plotting our results in standard yield curves. In that regard, we used the judgments of the topic authorities to the extent they differed from those of the review team. Since they were the authorities on the topics, we used their judgments in evaluating the different rankings. We did not try to inject our own judgments to resolve the disagreement.

**Using the Experts to QC Reviewer Judgments**

As a further experiment, we created a third set of training documents to use in our ranking process. Specifically, we wanted to see what impact an expert might have on a review team’s rankings if the expert were to review and “correct” a percentage of the review team’s judgments. We were curious whether it might improve the overall rankings and how that effort might compare to rankings done by an expert or review team without the benefit of a QC process.

We started by submitting the review team's judgments to Predict. We then asked Predict to rank the documents in this fashion:

1. The lowest-ranked positive judgments (reviewer tagged it relevant while Predict ranked it highly non-relevant); and

---

4 Note that we do not consider this the ideal workflow. A completely random seed set, with no iteration and no judgmental/automated seeding, this test does not (and is not intended to) create the best yield curve. Our goal here was to put all three tests on level footing, which this methodology does.
2. The highest-ranked negative judgments (reviewer tagged it non-relevant while Predict ranked it highly relevant).

The goal here was to select the biggest outliers for consideration. These were documents where our Predict ranking system most strongly differed from the reviewer’s judgment, no matter how the underlying documents were tagged.

We simulated having an expert look at the top 10 percent of these training documents. In cases where the expert agreed with the reviewer’s judgments, we left the tagging as is. In cases where the expert had overturned the reviewer’s judgment based on a challenge, we reversed the tag. When this process was finished, we ran the ranking again based on the changed values and plotted those values as a separate line in our yield curve.

**Plotting the Differences: Expert vs. Reviewer Yield Curves**

A yield curve presents the results of a ranking process and is a handy way to visualize the difference between two processes. The X axis shows the percentage of documents that are reviewed. The Y axis shows the percentage of relevant documents found at each point in the review.

Here were the results of our four experiments.
The lines above show how quickly you would find relevant documents during your review. As a base line, I created a gray diagonal line to show the progress of a linear review (which essentially moves through the documents in random order). Without a better basis for ordering of the documents, the recall rates for a linear review typically match the percentage of documents actually reviewed—hence the straight line. By the time you have seen 80% of the documents, you probably have seen 80% of the relevant documents.

The blue, green and red lines are meant to show the success of the rankings for the review team, expert and the use of an expert to QC a portion of the review team’s judgments. Notice that all of the lines are above and to the left of the linear review curve. This means that you could dramatically improve the speed at which you found relevant documents over a linear review process with any of these ranking methods. Put another way, it means that a ranked review approach would present more relevant documents at any point in the review (until the end). That is not surprising because TAR is typically more effective at surfacing relevant documents than linear review.
In this first example, the review team seemed to perform at a less effective rate than the expert reviewer at lower recall rates (the blue curve is below and to the right of the other curves). The review team ranking would, for example, require the review of a slightly higher percentage of documents to achieve an 80% recall rate than the expert ranking.\(^5\) Beyond 80%, however, the lines converge and the review team seems to do as good a job as the expert.

When the review team was assisted by the expert, through a QC process, the results were much improved. The rankings generated by the expert-only review were almost identical to the rankings produced by the review team with QC assistance from the expert. I will show later that this approach would save you both time and money, because the review team can move more quickly than a single reviewer and typically bills at a much lower rate.

\textbf{Issue Two}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{trec_issue2}
\caption{TREC Issue 2}
\end{figure}

In this example, the yield curves are almost identical, with the rankings by the review team being slightly better than those of an expert alone. Oddly, the expert

\(^5\) In this case, you would have to review 19\% of the documents to achieve 80\% recall for the ranking based only on the review team’s training and only 14\% based on training by an expert.
QC rankings drop a bit around the 80% recall line and stay below until about 85%. Nonetheless, this experiment shows that all three methods are viable and will return about the same results.

**Issue Three**

In this case the ranking lines are identical until about the 80% recall level. At that point, the expert QC ranking process drops a bit and does not catch up to the expert and review team rankings until about 90% recall. Significantly, at 80% recall, all the curves are about the same. Notice that this recall threshold would only require a review of 30% of the documents, which would suggest a 70% cut in review costs and time.
Issue four offers a somewhat surprising result and may be an outlier. In this case, the expert ranking seems substantially inferior to the review team or expert QC rankings. The divergence starts at about the 55% recall rate and continues until about 95% recall. This chart suggests that the review team alone would have done better than the expert alone. However, the expert QC method would have matched the review team’s rankings as well.

What Does This All Mean?

That’s the million-dollar question. Let’s start with what it doesn’t mean. These were tests using data we had from the TREC program. We don’t have sufficient data to prove anything definitively but the results sure are interesting. It would be nice to have additional data involving expert and review team judgments to extend the analysis.

In addition, these yield curves came from our product, Insight Predict. We use a proprietary algorithm that could work differently from other TAR products. It may
be that experts are the only ones suitable to train some of the other processes. Or not.

That said, these yield curves suggest strongly that the traditional notion that only an expert can train a TAR system may not be correct. On average in these experiments, the review teams did as well or better than the experts at judging training documents. We believe it provides a basis for further experimentation and discussion.

Why Does this Matter?

There are several reasons this analysis matters. They revolve around time and money.

First, in many cases, the expert isn’t available to do the initial training, at least not on your schedule. If the review team has to wait for the expert to get through 3,000 or so training documents, the delay in the review can present a problem. Litigation deadlines seem to get tighter and tighter. Getting the review going more quickly can be critical in some instances.

Second, having review teams participate in training can cut review costs. Typically, the SME charges at a much higher billing rate than a reviewer. If the expert has to review 3,000 training documents at a higher billable rate, total costs for the review increase accordingly. Here is a simple chart illustrating the point.

<table>
<thead>
<tr>
<th>Expert Only</th>
<th>Hourly Rate</th>
<th>Review Rate</th>
<th>Documents</th>
<th>Total Cost</th>
<th>Time Spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Expert</td>
<td>$550</td>
<td>60</td>
<td>3,000</td>
<td>$27,500</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expert and Review Team</th>
<th>Hourly Rate</th>
<th>Review Rate</th>
<th>Documents</th>
<th>Total Cost</th>
<th>Net Time Spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 Reviewers</td>
<td>$60</td>
<td>60</td>
<td>3,000</td>
<td>$3,000</td>
<td>5.0</td>
</tr>
<tr>
<td>1 Expert</td>
<td>$550</td>
<td>60</td>
<td>300</td>
<td>$2,750</td>
<td>5.0</td>
</tr>
</tbody>
</table>

| Savings               |             |             |           | $21,750    | 40.0          |

Using the assumptions I have presented, having an expert do all of the training would take 50 hours and cost almost $27,500. In contrast, having a review team do most of the training while the expert does a 10% QC, will reduce the cost by
85%, to $5,750. The time spent on the combined review process changes from 50 hours (6+ days) to 10 combined hours, a bit more than a day.\(^6\)

You can use different assumptions for this chart but the point is the same. Having the review team involved in the process saves time and money. Our testing suggests that this happens with no material loss to the ranking process.

This all becomes mandatory when you move to continuous ranking. The process is based on using the review team rather than an expert for review. Any other approach would not make sense from an economic perspective or be a good or desirable use of the expert's time.

So what should the expert do in a TAR 2.0 environment? We suggest that experts do what they are trained to do (and have been doing since our profession began). Use the initial time to interview witnesses and find important documents. Feed those documents to the ranking system to get the review started. Then use the time to QC the review teams and to search for additional good documents. Our research so far suggests that the process makes good sense from both a logical and efficiency standpoint.

About the Author

John Tredennick is the founder and CEO of Catalyst, which provides secure, multi-language, document repositories for litigation and other complex regulatory matters. Prior to founding Catalyst, John spent over 20 years as a trial lawyer and litigation partner at Holland & Hart, a national law firm based in the Rocky Mountains.

John was an early pioneer in the field of legal technology. He was editor-in-chief of the multi-author, two-book series, Winning With Computers: Trial Practice in the Twenty-First Century (ABA Press 1990, 1991). Both were ABA best sellers focusing on using computers in litigation technology. At the same time, he wrote, How to Prepare for Take and Use a Deposition at Trial (James Publishing 1990), which he and his co-author continued to supplement for several years. He also wrote, Lawyer's Guide to Spreadsheets (Glasser Publishing 2000), and, Lawyer's Guide to Microsoft Excel 2007 (ABA Press 2009).

John has been widely honored for his achievements. In 2013, he was named by the American Lawyer as one of the top six “E-Discovery Trailblazers” in their

---

\(^6\) I used “net time spent” for the second part of this chart to illustrate the real impact of the time saved. While the review takes a total of 55 hours (50 for the team and 5 for the expert), the team works concurrently. Thus, the team finishes in just 5 hours, leaving the expert another 5 hours to finish his QC. The training gets done in a day (or so) rather than a week.
special issue on the “Top Fifty Big Law Innovators” in the past fifty years. In 2012, he was named to the FastCase 50, which recognizes the smartest, most courageous innovators, techies, visionaries and leaders in the law. London’s CityTech magazine named him one of the "Top 100 Global Technology Leaders." In 2009, he was named the Ernst & Young Entrepreneur of the Year for Technology in the Rocky Mountain Region. Also in 2009, he was named the Top Technology Entrepreneur by the Colorado Software and Internet Association.

John is the former chair of the ABA’s Law Practice Management Section. For many years, he was editor-in-chief of the ABA’s Law Practice Management magazine, a monthly publication focusing on legal technology and law office management. More recently, he founded and edited Law Practice Today, a monthly ABA webzine that focuses on legal technology and management. Over two decades, John has written scores of articles on legal technology and spoken on legal technology to audiences on four of the five continents. In his spare time, you will find him competing on the national equestrian show jumping circuit.