CLEARY GOTTLIEB

BACKSTOP

DISCOVERY INTELLIGENCE TECHNOLOGY

This paper presents the results of the collaborative entry of Backstop LLP and Cleary Gottlieb Steen & Hamilton LLP in the Legal Track of the 2009 Text Retrieval Conference (TREC) sponsored by the National Institute for Standards and Technology (NIST). The Legal Track served as a truncated replication of a document review of almost one million documents. Backstop software, assisted by attorney document review of less than one-tenth of one percent of the overall document set, classified the documents and achieved a combined accuracy rate ("F1 score") of approximately 80%.

Background

The team

Backstop LLP is a provider of automatic document-classification software. Using artificial intelligence, Backstop software learns from attorney classifications of documents for large-scale document reviews in legal cases. The software creates a model to replicate the attorneys' collective knowledge and classifies documents accordingly.

Cleary Gottlieb Steen and Hamilton LLP is a leading international law firm with 12 offices located in major financial centers around the world, employing approximately 1,100 lawyers from more than 50 countries and diverse backgrounds who are admitted to practice in numerous jurisdictions around the world. Its clients include multinational corporations, international financial institutions, sovereign governments and their agencies, as well as domestic corporations and financial institutions in the countries where its offices are located. Cleary received Chambers & Partners' inaugural International Law Firm of the Year award.

Cleary prides itself on its application of cutting-edge technology to provide clients with efficient, economical and accurate document review and production.

National Institute for Standards and Technology (NIST)

Founded in 1901, NIST describes itself as "a non-regulatory federal agency within the U.S. Department of Commerce. NIST's mission is to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life."

Text Retrieval Conference (TREC) Legal Track

TREC was started in 1992, co-sponsored by NIST and the U.S. Department of Defense, with the purpose to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. The Legal

Track of TREC is an annual evaluation of text-retrieval methodologies and techniques in the context of legal discovery.

The Task

Ground rules

The TREC Legal Track Interactive Task for 2009 involved a simulated Request for Production of documents (RFP) in a hypothetical lawsuit in the fallout of the collapse and bankruptcy of Enron. Each participating team was provided with a body of almost one million potentially responsive documents, and tasked to use its technology to identify, with as much accuracy as possible, documents responsive to the various topics in the RFP.

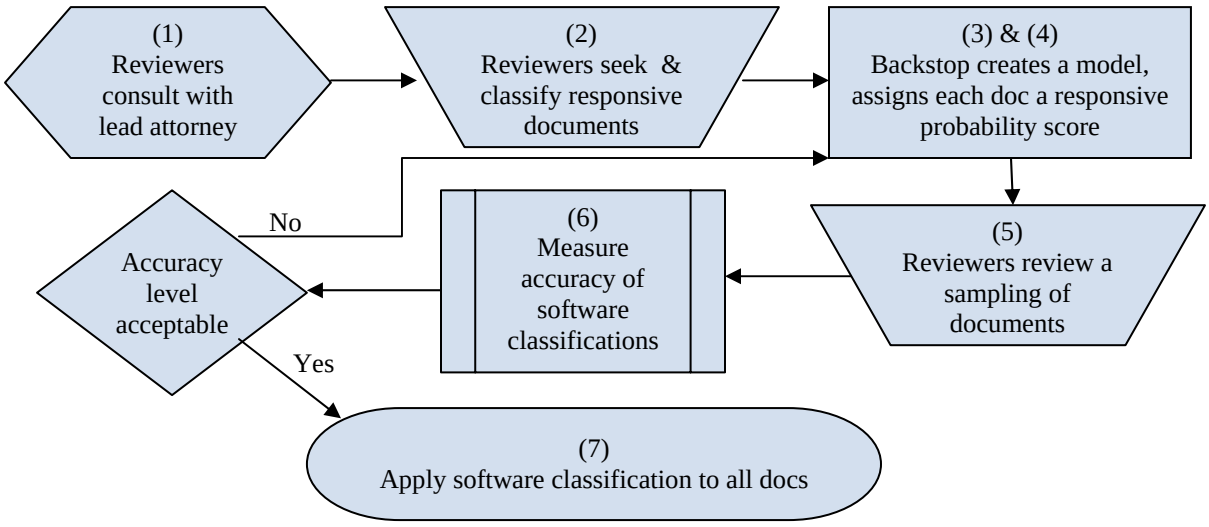
In particular, teams received a simulated complaint, supplementary materials, RFP, and some one million electronic documents. The complaint embodied a notional lawsuit filed against Enron. The RFP included seven specifications, or topics. For each topic, TREC provided a topic authority. The topic authority played the role of a law-firm partner responsible for responding to the RFP, with comprehensive knowledge regarding the facts and issues surrounding the complaint, and authoritative say over the relevance vel non of documents to the RFP specifications. Teams were allowed up to ten hours of consultation with the topic authority, replicating interaction that would occur on an actual document review. The electronic documents—taken from actual Enron files—represented the entire population of potentially responsive documents.

Backstop – Cleary team methodology

We participated in TREC 2009 to raise the level of discourse and awareness regarding Information Retrieval (IR) approaches to litigation document review. While the problems of measuring accuracy are well known in IR circles, they are particularly troublesome in the context of a nuanced litigation analysis. Also, unlike pure research or academic endeavors that focus on different methods of optimizing accuracy, the legal industry is further constrained by practical considerations of cost, time, and transparency. There is a dearth of scholarship exploring this constellation of concerns with any scientific rigor, and we wanted to demonstrate that there are measurable correlations between the amount of review effort invested, the number of refinement iterations made, and the accuracy of the final submission, even across topics of different levels of complexity. Additionally, we wanted to present a framework for discussing work flow models sounding in proportionality[1] that could be used to seed discussion in the legal community regarding best practice approaches to review for different types of matters, such as responding to third party subpoenas, HSR requests, and the like.

In general, the use of Backstop software entails the following basic methodology:

[1] See Fed. R. Civ. P. 26(b)(2)(C).



(1) Reviewers consult with supervising associate or partner (the "topic authority," for purposes of TREC).

(2) Reviewers seek documents responsive to topic using keyword searches and information gleaned from the supervising associate or partner (or TREC topic authority), with a goal of rapidly generating a critical mass of documents meeting the topic criteria, rather than constructing a model from a randomly generated review corpus.

(3) Backstop creates a classification model for each tag using reviewer-applied codes.

(4) Model assesses entire population of documents, assigning probability of responsiveness between 0 and 1.

(5) Reviewers review a sampling of the results provided by the model, in two categories:

- Documents not previously reviewed by a reviewer. The review of such documents
  - ➢ generates additional lines of inquiry for the topic authority
  - ➢ identifies additional possible search terms or responsive sub-topics
  - ➢ assesses model accuracy
  - ➢ iteratively trains the model

  To avoid the artificial suppression of recall (a measure of accuracy, defined in greater detail below) by biasing training documents with search term hits, some of the not-previously-reviewed documents are selected at random, while others are selected according to the responsiveness probability score described in step (4).

- Documents previously reviewed by a reviewer as to which the model yields a different classification from that applied by the initial reviewer. Such review

- identifies possible reviewer errors (particularly useful where the reviewer generated the initial training set largely through the use of search terms without reviewing each document)

- assesses model accuracy

- iteratively trains the model

(6) Measure the accuracy of the model as against the attorney reviewers' classifications. Using the information gained in the preceding steps, the reviewers renew the cycle. This progressively increases the quality and quantity of training data, as the reviewers' understanding of the case matures, and the accuracy of the model as it trains on the growing quantity and quality of reviewer-applied classifications.

(7) When the software has achieved a pre-determined, minimum level of accuracy for a given topic—say, 90% recall and 70% precision (both measures of accuracy are defined in greater detail below)—the software classifies all of the remaining documents (namely, those which have not been reviewed by an attorney).

For purposes of TREC, the Backstop – Cleary team followed a similar methodology, subject to a number of resource constraints:

> First, rather than a team of reviewers as would be customary on a large document review, we had only one reviewer per topic.
>
> Second and related, rather than reviewing a large set of documents comprising a non-negligible proportion of the documents, we were able to review only a few thousand documents largely comprising search term hits.
>
> Rather than successively training and refining the computer model over multiple iterations, we were able to run only few if any iterations. Thus, we were not able to continue the software-training process until the desired level of accuracy had been achieved.

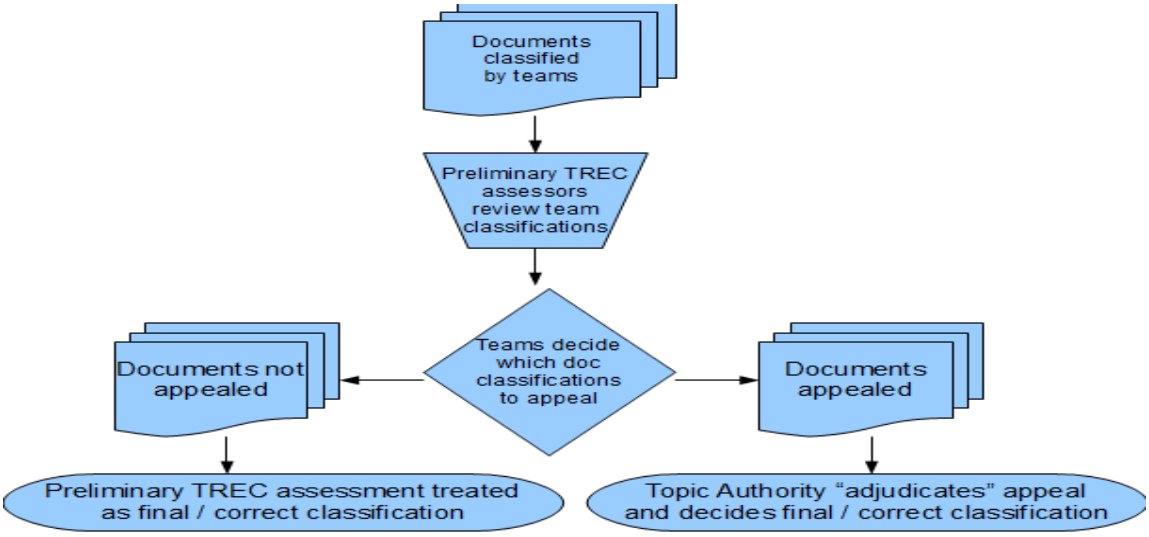
TREC evaluation of results

Using the resources described above, and its own technology, each team identified and submitted documents responsive to the various RFP topics.[2] A team of document reviewers—comprising law students, paralegals, and attorneys—then reviewed a sample of each team's submissions and classified them as correct or erroneous. Teams were afforded the opportunity to appeal these preliminary document classifications to the respective topic authorities, who then made a final

[2] Teams were not required to submit results for all seven topics. Our team initially took on four of the seven topics, but we were able to bring resources to bear on only one as the amount of excess capacity available to devote to TREC decreased over the course of the endeavor. Results for the other three topics were based on a one-off manual review of a small number of documents with no opportunity for iterative model training and little or no interaction with the topic authority.

determination for each appealed document. Based on this adjudication, TREC then assessed the accuracy of each team's classifications according to the three measures described below.

This diagram illustrates the TREC evaluation workflow:



Further details concerning the simulated complaint, RFP, documents, and evaluation methodology are available at the NIST TREC 2009 web site or from Backstop LLP.

Measuring accuracy

Accuracy measurements

Accuracy can be measured in many ways. The measurements used by TREC to evaluate accuracy are recall, precision, and F1 score:

> *Recall* is the percentage of actually responsive documents for a given tag that are correctly classified as such. In other words, of all the responsive documents that should have been produced for a given tag, what percentage actually were produced? (In the context of privilege, the question is the opposite—of all the privileged documents that should have been withheld, what percentage actually were withheld?) Seen another way, recall measures errors of omission. Recall is sometimes regarded as the most important measure of accuracy, because of the substantial sanctions that can result from failure to produce responsive documents.
>
> *Precision* is the percentage of documents classified as responsive that actually are responsive to a given tag. In other words, of all the documents produced for a given tag, what percentage were actually responsive? (In the context of privilege, as with recall, the question is reversed—of all the documents withheld as privileged, what percentage actually should have been withheld?) Seen another way, precision measures errors of inclusion. Precision is sometimes regarded as

less significant a measure of accuracy than recall, because producing too much information is far less likely to draw a sanction than failing to produce responsive documents.

F1 score is the harmonic mean of recall and precision. It is therefore a combined measure of recall and precision, providing a measure of overall accuracy. The F1 score weights recall and precision equally. In light of the greater significance of recall, some believe that a better measure of overall accuracy is the F2 score, which weights recall twice as heavily as precision.

In addition to the measures described above, an additional measure of accuracy may be easier to understand and useful in some contexts, although TREC did not use it as an accuracy measure because it can be misleading:

Agreement rate is the rate at which the software classifications agree with those of the attorney reviewers. In other words, of all the decisions that must be made with respect to responsiveness vel non to a particular topic, what percentage of the time did the software make the correct call?

TREC's measurement of accuracy

As noted above, TREC assessors did not review all of the nearly one million documents classified by the software. Rather, for each topic, TREC assessors reviewed a small fraction (between approximately 5,000 and 10,000) of those documents. That yielded what we term here "actual" recall, precision, and F1 score.

In order to project what those scores would be for the entire document set, TREC then divided all one million documents into different categories, according to which teams had classified each document as responsive or non-responsive. For example, if four teams participated in a given topic, there would be sixteen possible categories: four categories for documents classified as responsive by only one team (*i.e.*, documents classified as responsive only by Team 1, only by Team 2, etc.), six categories for documents classified as responsive by exactly two teams (Teams 1 & 2, Teams 2 & 3, etc.), four categories for documents classified as responsive by exactly three teams (Teams 1, 2 & 3, Teams 1, 2 & 4, Teams 1, 3 & 4, or Teams 2, 3, & 4), one category for documents classified as responsive by all four teams, and one category for documents classified as responsive by no teams. TREC then assessed the "actual" performance of each team on documents in each of the categories and weighted that performance according to the proportion of all documents in each category, yielding what we term here "projected" recall, precision, and F1 scores.

Backstop – Cleary team results

The Backstop – Cleary team focused its efforts on one topic (RFP specification). According to the TREC evaluation methodology, Backstop achieved the following projected accuracy measures:

Recall	Precision	F1 score
77%	83%	80%

In other words, after manually reviewing fewer than eight thousand documents—less than one-tenth of one percent of the overall data set—our team correctly identified over three-quarters of all of the responsive documents. Of the documents identified by our team as responsive, over 80% were in fact responsive. This suggests that a "good enough" result may potentially be obtained with even a very modest expenditure of effort, depending on the needs of the review. If greater accuracy were required, additional attorney review could be used to further train the model.

The Backstop – Cleary team, using data provided by TREC, calculated its actual performance scores (as defined above) as follows:

Recall	Precision	F1 score	F2 score	Agreement rate
84%	94%	89%	86%	97%

This level of accuracy compares favorably with levels generally observed in human reviewers.[3]

Conclusion

Software-assisted review and automated document classification can achieve a level of accuracy that compares favorably with levels generally observed in human reviewers, using orders of magnitude fewer hours of attorney resources. In particular, after reviewing a small fraction of the overall data set, levels of accuracy can be achieved that may be acceptable for compliance with a production burden (*e.g.* Hart-Scott-Rodino Second Request, subpoena, CID). With minimal effort, automated document classification can quickly identify a substantial body of responsive documents for early case assessment, to find important documents quickly.

[3] *See, e.g.*, Blair, David C. and Marion, M. E., *An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System*, Communications of the ACM 28(3), 289-99 (1985) (finding attorney search-term-assisted recall of approximately 20% in case involving 40,000 potentially responsive documents). Although this level of accuracy compares favorably to human review, it is lower than commonly observed in real-world cases conducted using the software. We believe the most probable explanation for the subpar performance relates to a quirk in treatment of web-browser URL's in the document set, which limited time and resources available did not permit us to address.

As noted above, our team also dabbled in three other topics, without devoting substantial resources to them. The results calculated from our submissions for those topics confirm that significant results can sometimes be achieved with minimal effort. For example, for one topic, we devoted twelve hours of document review and reviewed fewer than 1,000 documents (one one-hundredth of one percent of the document set). We achieved projected recall of 20% (33% actual), projected precision of 69% (79% actual), projected F1 score of 31.5% (46% actual), actual F2 score of 37%, and actual agreement rate of 96%. Even where not adequate for final compliance with a production burden, such results may be useful for early case assessment, to find important documents quickly, or for "jump-starting" a document review.