Appendix

A Ethical Implications

Our work raises ethical questions about the use of technology in criminal justice review procedures. We have engaged with our IRB, the parole board, the governor's office, parole attorneys, and formerly incarcerated parole candidates, to discuss how we frame our work around these issues. Our goal has been to enable human review that was previously impossible in a decision making process that allows for significant discretion. To be clear, we do not advocate for replacing human decision makers or reviewers, nor is the purpose of our work to develop a risk assessment tool for parole decisions. Rather, our work seeks to provide another avenue to all stakeholders in the parole process – the board, the governor, appellate attorneys, victims, those incarcerated, and the public – to review denial decisions that currently go unreviewed. We believe that our system provides no benefit to reviewing parole grants: the bar set by the board today is high and both the board and Governor's office already review every single grant in detail. However, our exploration of 30,734 transcripts provided by the board has yielded concerning examples of process anomalies that warrant further investigation. We believe that technology can play a role in facilitating the review of the massive amounts of historical legal data that currently goes unreviewed.

B Evaluation Details on Parole Hearings

Since there are no ground truth labels on which chunks include legal anomalies for our dataset, we collected annotations over a validation dataset that was held out at training time. We recruited and trained undergraduate and doctoral-level law students to identify instances of anomalous language in a held-out evaluation set of 315 parole hearings, as part of a larger structured labeling effort. After coding a hearing transcript for structural features, annotators were instructed to mark sentences in the transcript as language anomalies and provide an explanation. During the annotation task, annotators were required to read the entire transcript, so we estimate that the annotations produced have high true recall. Each annotation was mapped to one or more chunks of 256 tokens, containing the sentences highlighted as part of the annotation.

We then refined the annotations: the student labels were evaluated by a California parole attorney, who rated each annotation chunk on a scale from 1 to 3. A rating of **3** indicated that the chunk contained significantly anomalous language and would warrant additional review of the entire parole hearing in which it occurred. A rating of **2** indicated that the chunk contained anomalous language, but of a degree that may or may not require additional review of the hearing. Finally, a rating of **1** indicated the chunk did not actual contain anomalous language.

We consider only the annotations rated as **2** or **3** to be the "true anomalies" for the validation dataset. This process of checking non-expert human labels by a parole attorney also yields an approximate measure of human precision: the portion of student annotations that were rated a **2** or **3** by the expert.

B.1 Anomaly Detection as Information Retrieval

Since our goal is to assist human expert reviewers with a limited time-budget, we model the anomalies for each document in an information retrieval manner. Of primary interest, we evaluate the tradeoff between $\mathbb{E}[k]$, the number of chunks of text a human reviewer must read, and the true anomaly recall, as we vary the threshold of our model.

To estimate the precision of our annotations, we asked our parole expert to review a set of anomaly predictions made by our model using the same 1-3 rating scheme. We compute the mean reciprocal rank (MRR) (Voorhees et al., 1999). This can also be thought of as the average precision aggregated over the documents, where each document's score is evaluated over only the chunks that the expert reads before finding the first anomaly. This directly measures the time spent by a reader, rather than measuring the total size of what the system returns. Since we cannot ask our legal expert to perform this evaluation at all possible thresholds, we fix a threshold to achieve a desired recall-k tradeoff, and then determine the model's MRR at this threshold. We can compare this MRR to the human annotator precision given by the expert's review of human annotations.

References

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.