Challenges for Information Extraction from Dialogue in Criminal Law

Jenny Hong
Stanford University
jennyhong@cs.stanford.edu

Catalin Voss
Stanford University
catalin@cs.stanford.edu

Christopher D. Manning
Stanford University
manning@cs.stanford.edu

Abstract

Information extraction and question answering have the potential to introduce a new paradigm for how machine learning is applied to criminal law. Existing approaches generally use tabular data for predictive metrics. An alternative approach is needed for matters of equitable justice, where individuals are judged on a case-by-case basis, in a process involving verbal or written discussion and interpretation of case factors. Such discussions are individualized, but they nonetheless rely on underlying facts. Information extraction can play an important role in surfacing these facts, which are still important to understand. We analyze unsupervised, weakly supervised, and pre-trained models' ability to extract such factual information from the free-form dialogue of California parole hearings. With a few exceptions, most F1 scores are below 0.85. We use this opportunity to highlight some opportunities for further research for information extraction and question answering. We encourage new developments in NLP to enable analysis and review of legal cases to be done in a post-hoc, not predictive, manner.

1 Introduction

Our criminal justice system struggles to balance "the value of treating like cases alike, and the value of treating each case individually." (Bell et al., 2021) In criminal law, machine learning has been proposed as a tool to improve consistency in decision making, but to date, research efforts have primarily focused on *codified justice* – processes that make a determination given a limited set of case factors and using specifiable rules, such as a risk assessment used for a probation classification. However, various legal contexts balance a standard of codified justice with a standard of *equitable justice*, which requires decision-makers to apply moral principles to individuals' unique situations.

How can natural language processing aid equitable justice? Equitable justice centers human discretion and the uniqueness of each individual, but nonetheless is based on factual information. The facts of each case are typically discussed and interpreted through dialogue. Often, the dialogue produces transcripts, which are available as public records. Usually, the sheer length of transcribed conversational text all but prohibits any meaningful form of quantitative review, because of the immense effort involved in manually annotating case factors. NLP methods for information extraction over speech can assist in identifying the underlying facts of a case from hearing transcripts. The factors can then be used in statistical analyses of a decision-making process to (a) provide historical understanding over case records that are otherwise locked away in a filing cabinet, and (b) identify specific outlier cases for reconsideration of fair and equitable decision-making where human capacity for review is constrained. By applying information extraction post-hoc rather than filling in a data table or computing a risk score at the time of a hearing, the decision-maker retains full autonomy in conducting a legal process using their own discretion. In this role, information extraction supplements, but never fully supplants, the need for dialogue and transcripts. A broad set of stakeholders can then contribute to identifying the factors that may be relevant in comparing cases.[1]

We present a case study of the capabilities of information extraction methods for dialogue and identify areas for further research in the criminal law context. We have obtained a nearly complete dataset of 35,105 parole hearing transcripts from the State of California for individuals serving life sentences between 2007 and 2019. The California

[1] Bell et al. (2021) describes this approach in the context of the parole system in California. We provide a discussion of the ethical implications of our work in Section 9.

Proceedings of the 1st Workshop on NLP for Positive Impact, pages 71–81
Bangkok, Thailand (online), August 5, 2021. ©2021 Association for Computational Linguistics

parole hearing system serves as a useful case study because (1) California has one of the largest prison systems in the U.S., (2) the hearings are transcribed and available on the public record, (3) the hearings are relatively long (about 20,000 words) and illustrate the challenges of long dialogue, (4) human annotation of the hearings is expensive, and (5) the hearings are one continuous dialogue in a single sitting between a decision-maker and a parole candidate, with brief statements from the candidate's attorney. In comparison, criminal trials are much longer, present many forms of exhibits which are often not digitally available, and contain many additional complexities.

We have identified 11 case factors representative of the types of features (binary, multi-class, date, and numerical) that are relevant to the parole decision-making system and illustrate a range of challenges in information extraction. We evaluate three families of models on this task: (1) an unsupervised data programming paradigm (Ratner et al., 2016) extended to weak supervision, (2) pre-trained question answering models based on DistilBERT (Sanh et al., 2019) and Longformer (Beltagy et al., 2020), and (3) classification models based on BERT (Devlin et al., 2019) that are each fine-tuned to predict a single task.

Most models fall below an F1 score of 0.85 for most of the features. The different feature types challenge each of the models in different ways. Data programming remains a largely rule-based approach and works best when the keywords indicative of a label are clear, such as the penal code or a numerical education score. Pre-trained question answering models maintain or improve performance on most categories, except for boolean questions, which remains an area of active development. Surprisingly, all models perform poorly on extracting the risk assessment score, which relies on three simple keywords "low," "moderate," or "high."

Information extraction from long dialogues remains an open challenge, especially when the extraction tasks are not entity-based. We call on research in information extraction to move beyond entity-based tasks in order to tackle the range of tasks relevant for legal dialogue. We also emphasize the need for all methods to handle longer context windows. Long context windows are not merely a byproduct of underdeveloped retrieval methods; they are inherent to the level of personal detail required to apply equitable justice.

2 Related Work

2.1 Information Extraction and Question Answering

Information extraction spans a number of tasks, but neural approaches have concentrated on binary relation extraction. Many relation extraction tasks are performed on only the sentence level (Nguyen and Grishman, 2015; Adel et al., 2016; Levy et al., 2017; Karita et al., 2019; Luo et al., 2019), but techniques have emerged for cross-sentence or even document-level relation extraction (Yao et al., 2019). Compared to information extraction, question answering allows for a greater range of tasks, represented by the diversity of question formulations (Rajpurkar et al., 2016) and is an alternative approach to the task of creating parole hearing annotations.

For both information extraction and question answering, current top-performing models are pretrained large language models (Devlin et al., 2019; Radford et al., 2019) that have been fine-tuned on specific tasks, such as question answering.

Applications to dialogue focus on entity-based tasks like argument extraction (Swanson et al., 2015), named entity recognition (Chen and Choi, 2016; Choi and Chen, 2018; Bowden et al., 2018), relation extraction (Yu et al., 2020), and task-based extraction (Fang et al., 2018; Finch et al., 2020; Liang et al., 2020). Dialogue-like settings are relatively new for question answering. CoQA (Reddy et al., 2019) aims to answer questions over a written text in an abstractive way, but it is only conversational in that multiple questions can be asked of the same source text sequentially. FriendsQA (Yang and Choi, 2019) answers extractive questions about a multiparty dialogue. The questions are considered to be asked of the dialogue, by a third party outside the dialogue. Like FriendsQA, DREAM (Sun et al., 2019) also uses a dialogue as its source text, but its answers are multiple-choice.

2.2 Machine Learning for Criminal Law

Machine learning in law has mainly relied on tabular data, and mostly for prediction, e.g., policing (Ferguson, 2017; Barrett, 2017; Goel et al., 2016), pre-trial detention (Kleinberg et al., 2018a), sentencing (Elek et al., 2015). Retrospectively, past human (and algorithmic) decisions can be analyzed through the lens of algorithmic fairness, which seeks to understand the way machine learning models or human decisions systematically encode bias

```
Somebody actually took the time to count
up all your 115s and make a list of them
for me, and they covered the gambit,
but I am very surprised that you're
not a gang member.  We've got attempted
murder here in '01, deadly weapon in
'02, battery with a deadly weapon in '05,
pruno, '06, mutual combat, '06, deadly
weapon, '06, battery of peace officer,
'06.  And that seems to be sort of the
general way your life goes.  You picked
up a couple of these in 2013.
```

Figure 1: Example of a section of a hearing during which the deputy commissioner discusses the recent disciplinary history (recorded on Form "115") of the candidate. This occurs about halfway into a 50-page hearing. One extraction task is to identify the date of the most recent disciplinary writeup.

(Dwork et al., 2012; Barocas et al., 2017; Corbett-Davies et al., 2017; Corbett-Davies and Goel, 2018; Kleinberg et al., 2018b; Ho and Xiang, 2020).

Within natural language processing, computational linguistics has been used to scale up lexical analyses of various contexts, such as policing (Voigt et al., 2017) and judicial decisions (Danescu-Niculescu-Mizil et al., 2012). Lexical features can also be used in downstream analysis (Altenburger and Ho, 2019). Relational information extraction has been applied in the context of using named entities (e.g. attorneys, law firms, judges, districts, and parties of a case) as features for downstream risk analysis for intellectual property litigation (Surdeanu et al., 2011). However, both extractive and abstractive question answering are still largely unexplored in legal texts.

3 Data

Our text corpus consists of 35,105 parole hearing transcripts, averaging 18,499 words each, covering 15,852 unique individuals from 2007–2019 parsed from PDF documents. Each hearing is attended by a presiding and a deputy parole commissioner, the parole candidate, and typically an attorney for the candidate. Often, hearings also include a district attorney representative from the county of the commitment offense, who makes a statement, and a victim or their next-of-kin, who may make a statement. Some hearings are attended by visitors who do not participate in the dialogue. The majority of the conversation occurs between the parole candidate, their attorney, and the presiding commissioner.

3.1 Feature Selection

We selected 11 features from a set of case factors identified in discussion with legal scholars[2], former parole candidates, advocacy groups including appellate attorneys, representatives from the California Governor's office, and the Parole Board.

Four features are binary: `off_mur1` ("Do the controlling offenses include first-degree murder?"), `proggang` ("While in prison, did the parole candidate participate in gang-related programming?"), `da_opp` ("Did the district attorney attend the hearing and oppose parole?"), and `job_offer` ("Does the parole candidate have an offer letter for a job post-release?").

Two features are multi-class: `edu_level` ("What is the parole candidate's education level?"), which falls into one of five categories: "no high school or GED," "high school or GED or CHSPD," "some college courses," "college degree," or "other"; and `risk_assess` ("What is the risk score assigned by the psychological evaluation?"), which also has five categories: low, low/moderate, moderate, moderate/high, and high.

Three features are dates. Various dates are mentioned in the course of a parole hearing. Two that are usually stated at the start of the hearing are the MEPD (minimum eligible parole date) and the date that the parole candidate was received into the California Department of Corrections and Rehabilitation (CDCR). Discussing disciplinary writeups that occurred in prison is another key part of the hearing, and we use `last_writeup` to denote the year of the most recent such writeup.

Finally, two features are numerical. One is `yrserved`, the number of years the parole candidate has served in state prison. Another is `tabe`, a measure of educational attainment that corresponds roughly to grade levels (10.5 corresponds to finishing half of 10th grade, where 12.9, corresponding to high school completion, is the highest score).

The context window, or section of dialogue required to identify a feature, varies greatly. Figure 1 shows an example of a context window for the `last_writeup` task. In other hearings, the context window may be longer, e.g., the commissioner may decide to focus on the "mutual combat" in 2006 and speak about the single incident in depth before returning to the list of Forms 115.

[2] All 11 features are identified as more than marginally predictive in Bell (2019) and Young et al. (2015)'s studies of California parole hearings.

Feature	Num. Train	Num. Val.
off_mur1	16,201	1,867
proggang	563	48
da_opp	1,173	106
job_offer	1,173	106
edu_level	1,174	106
risk_assess	1,173	106
mepd	1,174	106
last_writeup	563	48
year_received	10,866	1,261
tabe	367	36
yrserved	982	94

Table 1: Training and validation split sizes for each feature.

3.2 Annotation

We collected annotations over a subset of transcripts from three sources. CDCR provided the controlling offense for 26,780 transcripts, which yields off_mur1. We scraped CDCR's "Inmate Locator" website to obtain year_received for each parole candidate. Bell (2019) provided human labels for 426 juvenile lifer parole hearings for a superset of the 11 factors.

We manually labeled 827 transcripts with 118 features with a team of 11 research assistants who were trained and supervised by a legal expert. Through the process of annotation, we narrowed down the 118 proposed fields through multiple rounds of annotations and inter-rater reliability evaluations. The first round of annotations included all 11 features. Subsequent rounds dropped tabe and proggang.

We split data into training and validation sets by sampling at the transcript level. We withheld an additional portion of the data in a separate test split that is not uncovered for the present work in progress. A subset of training transcripts was designated "development" and used for inspection during model development, in particular for developing human intuition for writing label functions.

Because not all features are covered by all label sources, the amount of labeled data varies by feature across the splits. Table 1 includes the number of examples in each group.

4 Human Performance

To compute a human performance baseline for the reliability with which the selected features can be extracted from transcripts, we use Cohen's κ coefficients. Because the overlap of annotators varies by feature, we compute a mean κ-statistic per feature, weighted by the number of documents that overlapped between the annotators . For the kth feature and two labelers i, j, $i \neq j$, let $\kappa_k(i,j) = \frac{p_0 - p_e}{1 - p_e}$, where p_0 is the relative observed agreement among labelers i and j and p_e is the probability of chance agreement under the observed data available for the labelers and let $N_k(i,j)$ be the number of documents for which i and j overlap on feature k. Table 2 reports the statistic

$$\hat{\kappa}_k = \frac{\sum_{i \neq j} N_k(i,j) \cdot \kappa_k(i,j)}{\sum_{i \neq j} N_k(i,j)}.$$

Feature	Human $\hat{\kappa}$ IRR
off_mur1	0.94
proggang	0.93
da_opp	0.99
job_offer	0.77
edu_level	0.92
risk_assess	0.80
mepd	0.61
last_writeup	0.69

Table 2: Inter-rater reliability $\hat{\kappa}$ score of human annotators for each feature

5 Extraction Models

5.1 Weakly Supervised Models

Labeling features for parole hearings is burdensome; each hearing takes about one hour to annotate per person. An alternative approach is to generate a noisy but larger dataset using data programming (Ratner et al., 2016). Data programming improves on purely rule-based methods by learning to automatically weight rules, also known as labeling functions, to produce a probabilistic label. When combined, multiple labeling functions $\lambda_1, \ldots, \lambda_n$ can comprise a high-quality estimate of a single label y. For example, for the task of classifying whether a candidate has a count of first-degree murder, λ_1 can be an indicator of whether the phrase "first degree" appears in the first ten conversational turns. Or, a labeling function might instead relying on neural sentiment analysis models. We wrote a set of labeling functions for each extraction task. We also wrote a retrieval heuristic that selects a number of conversational turns from the transcripts over which labeling functions are run.

We use two strategies to produce an estimate $\hat{y}$ from multiple labeling functions. **Snorkel** MeTaL proposes an unsupervised method (Ratner et al., 2018). Supervised methods can also be used, e.g. using linear or logistic regression to learn a weighting of the labeling functions to produce an estimate. In our case, we use logistic regression for the binary variables, where learning a prior makes sense, and prior-free constrained least squares regression for all other variables. We call this method weakly supervised labeling functions, or **WSLF**.

5.2 Pre-Trained Language Models

Data programming generalizes the knowledge of domain experts; pre-trained language models generalize the knowledge of a large English corpus.

We first use models fine-tuned for question answering, which allows us to use a single model for a wide range of features. We study two question answering models: DistilBERT (Sanh et al., 2019) fine-tuned on SQuAD (Rajpurkar et al., 2016) and Longformer (Beltagy et al., 2020) fine-tuned on SQuAD 2.0 (Lee et al., 2020). We call these two models **QA1** and **QA2**, respectively. Through QA1, we hope to understand the overall performance gain, if any, from pre-training. Through QA2, we hope to understand any advantages of using a model with a longer context window (4,096 tokens) that can handle unanswerable questions, which are common in this corpus.

Our second approach is to model each task as a classification task and to fine-tune a language model for each task. We first fine-tune the base BERT model (Devlin et al., 2019) on all parole hearing text, including unlabeled documents. We then train a classifier layer on the labels produced in data generation, because of how limited human labels are. We train a separate model for each task (as opposed to a single multi-head multi-task model), i.e. there is one model to predict the binary feataure `off_murl`, another one to predict the binary feature `proggang`, and so on. We call this approach task fine-tuned, or **Task-FT**.

6 Results

Table 3 reports the average F1 score across all classes. Binary and multi-class features have natural F1 score interpretations. Date features are quantized into years, and both numerical features have natural quantizations. The TABE score is already quantized to the nearest tenth of a point, and the years served rounded to the nearest year.

Because Snorkel, WSLF, and Task-FT models are trained for a given class, their results are given in the space of the label of the task, whether that is a binary label or a date, for example. However, both QA1 and QA2 models are extractive question answering models, i.e. the answers returned are taken from the text of the hearing. In some cases, the text needs additional processing to be transformed into a label. The transformation may be human intervention, such as in the case of `edu_level`, where the extractive answer "ninth grade" and needs to be translated into a categorical answer "no high school or GED." In other cases, such as with dates, the transformation can be partially or fully automated, such as by parsing answers like "March the 6th, 2019" into the MEPD year, 2019, using tools such as SUTime (Chang and Manning, 2012).

Overall, WSLF does well on most classification tasks, though it is beaten by QA2 on `risk_assess` and by the more powerful classifier Task-FT on `off_murl`. QA2 is strongest on dates and generally outperforms QA1. Task-FT performs best on a variety of tasks, but surprisingly, it does not always improve over WSLF and Snorkel, even though its training process uses the very labels produced by the data programming methods, but augmented with even more information, the underlying text itself.

7 Discussion

Our case study on extracting features from parole hearings illustrates many outstanding challenges in question answering, information extraction, and text classification. Addressing these challenges is key to using NLP for positive impact in criminal law. The tasks posed by the parole dataset do not fall neatly into relation extraction, which has been the focus of neural information extraction. For legal domain tasks, human labels are scarce and expensive, which raises the question of whether weak supervision may be a more efficient allocation of labels than direct supervision. Legal hearings are long and don't fit neatly into the context window size of a neural model, which raises questions about how neural question answering systems can address this task. We answer the questions in turn.

Can weakly supervised methods be successfully used to reduce the cost of data annotation? Data programming provides the opportunity to produce a large number of labels, but it still comes

Binary Features	**Snorkel**	**WSLF**	**QA1**	**QA2**	**Task-FT**	**Avg. # Words**
off_mur1	0.78	0.74	0.76*	0.78*	**0.80**	974
proggang	0.66	**0.87**	0.42*	0.53*	0.64	13,270
da_opp	**0.83**	**0.83**	0.73*	0.76*	**0.83**	5,219
job_offer	0.52	**0.63**	0.58*	0.53*	0.46	9,973
Multi-class Features	**Snorkel**	**WSLF**	**QA1**	**QA2**	**Task-FT**	**Avg. # Words**
edu_level	0.37	**0.41**	0.13*	0.30*	0.34	12,990
risk_assess	0.48	0.51	0.46	**0.53**	0.51	12,326
Dates	**Snorkel**	**WSLF**	**QA1**	**QA2**	**Task-FT**	**Avg. # Words**
mepd	0.74	0.83	0.79	0.79	**0.87**	2,405
last_writeup	0.27	0.03	0.35	**0.42**	0.24	4,811
year_received	0.47	0.01	0.73	**0.76**	0.15	1,700
Numerical	**Snorkel**	**WSLF**	**QA1**	**QA2**	**Task-FT**	**Avg. # Words**
tabe	0.87	0.88	0.87	0.90	**0.94**	972
yrserved	**0.28**	0.08	**0.28**	0.20	0.13	18,603

Table 3: F1 scores of information extraction models and the average number of words in the context windows that were the input text for each model. Scores with * in the QA columns required manual intervention to convert the extractive answer into a binary or multi-class label.

at the cost of requiring experts to translate domain knowledge into programs for each task. Rather than spending one hour labeling one document, an expert may spend dozens of hours designing labeling functions for a single task, e.g. "Does the parole candidate have a job offer?" Once designed, labeling functions are usually computationally light. In producing a final model, adding even weak supervision can improve performance, as seen by improvements of weakly supervised learning functions (WSLF) over the unsupervised Snorkel approach. But unsupervised and weakly supervised techniques mainly perform well only when the tasks can be framed as classification, or when the extractive procedure is relatively simple, such as finding a one-digit decimal TABE score. Reserving some human labels to supervise a WSLF approach outperforms the unsupervised Snorkel method.

Can neural question answering successfully address parole hearings? Neural question answering systems have the flexibility of handling a large range of question formulations and feature types. Compared to other models, this flexibility improves the performance on date features, but surprisingly, on only one additional task, risk_assess.

Boolean questions remain an outstanding challenge. Reading comprehension datasets like CoQA (Reddy et al., 2019) and BoolQ (Clark et al., 2019) include such questions but leave a substantial performance gap for future work. The reliance on manual conversion of some answers to binary or multi-class labels is problematic.

In general, including on date features, the most common failure mode for QA1 and QA2 is to return an incorrect answer of a correct type. For example, for yrserved, the models frequently returned any number they found in the context passage, such as the sentence (e.g. "15 years to life") or any other time range (e.g. "It was around two years I was part of that gang.")

How big a problem is document length? Long context windows continue to challenge all models present, especially neural models. Although developing retrieval models for dialogue can help narrow the context window for downstream question answering applications, an even bigger challenge is the fact that even with an ideal retrieval model, the "correct" context window can still be long. In conversation, speakers are free to go on tangents. More importantly, in the case of legal hearings, speakers elaborate on case factors, attending to detail (as they should), which can greatly prolong a hearing. For example, in discussions of the psychological risk score, both data generation methods and neural question answering systems fail to identify the sentence and keyword containing "low," "moderate," or "high." We suspect that this is because discussions of all risk factors are usually several thousand words long. The score can be mentioned at the very beginning or very end, but often it is tucked away somewhere in the middle.

8 Conclusion

Parole hearing transcripts go into a great amount of detail in discussing numerous case factors centered around a single named entity, an incarcerated individual who has reached their parole eligibility date. The lack of relational structure and long format of these hearings makes information extraction from transcripts very challenging using several very different approaches from modern NLP.

We estimate that an F1 score of 0.80–0.85 across a broad set of features would provide the ability to conduct meaningful downstream research on a hearing-driven decision-making process like parole. To flag individual cases for reconsideration, we believe that the bar likely lies even higher, since misclassifications often cause outliers. The performance of present models approaches the level at which we can provide useful automatic extraction tools to parole stakeholders for some features, especially certain binary ones. However, for other, seemingly simple medium- and high-cardinality tasks, much work remains.

We plan to conduct future experiments to provide more transparency to model performance. The opaque nature of NLP modeling perplexes our legal collaborators: "How can you identify whether a candidate has participated in gang-related rehabilitation programming but not pick out the risk assessment score from a choice of three words?"

The largest challenge moving forward remains natural language understanding in the face of document length. Of course, length is not the only problem and other artifacts of spoken dialogue cause challenges, including interruptions, corrections, and colloquial speech. Improved retrieval techniques or even summarization methods can help assess the extent to which document length remains a challenge and possibly mitigate its impact. However, there is no getting around the level of detail that is regarded as due process.

One solution is to incorporate the hierarchical nature of dialogue (Asher and Vieu, 2005). Within a discussion about risk assessment, a parole commissioner may ask about various sub-factors, such as mental illness, or behavior toward other individuals in prison. We suspect that the word "low," "moderate," or "high" can appear in any of those sub-topics without referring to the risk score. We hope to conduct further research to assess the need for and viability of a hierarchical model. Conversely, an extractive model sometimes picks up on risk-related words in the sub-topics, rather than returning to the higher level question of the risk scores.

Common sense knowledge will also play a role in solving this challenge. In one section of a hearing, the commissioner says, "And, uh, I note that you – you have both a high school diploma and GED, is that correct?" Over the course of the next eight thousand words, the parole candidate describes his life, from playing sports in high school, to having a child, to the chaos of teenage co-parenting, to night school, to getting married, and to moving cities to protect his children. Later on, the commissioner revisits the record and says, "You've taken some college classes," which the candidate himself failed to mention. In addition to understanding the topics and sub-topics in which education occurs, the `edu_level` task benefits from real-life knowledge about educational levels. The WSLF model performs well because of tailored labeling functions that encode information about "high school" and "college."

Finally, cross-sentence reference resolution remains important. In Figure 1, the question of the most recent Form 115 can be answered in a short context window. Yet, extracting the answer requires resolving the reference of "these" in " You picked up a couple of these in 2013."

While the amount of attention to personal detail in these hearings presents the biggest challenge to our extraction models, individualized attention is also precisely what defines *equitable justice*. We hope that the NLP community will take up this challenge.

9 Ethical Implications

Our work raises ethical questions about the use of NLP in criminal law. We argue that machine learning can have a positive impact in a decision-making process like parole when it is applied as a review tool. NLP can provide transparency into millions of pages of hearing dialogue that would otherwise remain inaccessible for any form of analysis. It is possible to use information aggregation as part of a toolkit that centers human discretionary judgment and uses technology to promote consistency, reconciling our desire for a human-led decision-making process with the reality that human discretion introduces inconsistencies and systemic biases. The analysis of the present work falls under the umbrella of the "Recon Approach" (Bell et al., 2021) and serves the purposes of conducting *reconnaissance* at the systemic level and creating an opportunity for *reconsideration* of individual cases.

The dual use objection. Perhaps the most prominent objection to the Recon Approach is analogous to the "dual use" argument for sentencing (Leins et al., 2020). While we have developed information aggregation tools for a review use case, what is there to stop someone turning that around and using these exact same features and for a codified justice use case?

In the California parole context, employing technology for a predictive, rule-based system requires legislative parole reform and an overhaul of California's approach to criminal data record keeping. As it is currently constructed, the Board of Parole Hearings operates with great discretion. Parole hearings are based only in part on data that is available before the hearing. For example, parole hearings often discuss mitigating pre-commitment factors such as the living circumstances of an individual at the time that the crime was committed, touching on topics such as childhood abuse, gang membership, or neighborhood crime. These data are often not even available in sentencing transcripts. Even for factors that are available in records before the hearing, such as a candidate's disciplinary conduct in prison, the data often only exists in archived handwritten reports that prison staff aggregate prior to the hearing. The data are read out in semi-structured form for the first time by the commissioner during the hearing. It is therefore not possible to extract a meaningful number of the features that are currently considered for a parole decision in California without first conducting a hearing. [3]

The risk assessment path. A second ethical question is whether features extracted from hearing dialogue can be used as the input to a risk assessment algorithm before a decision is reached. While constructing a such a risk assessment algorithm is possible in theory, we believe that such an algorithm would be hard to construct and virtually meaningless in the context of parole. Unlike applications to sentencing (Chen et al., 2019; Hu et al., 2018; Zhong et al., 2018), the outcome variable for parole is unclear. Lifer recidivism is extremely low (under 3% in California) and it has not risen even as the parole grant rate has increased from 3% to over 20% in the past two decades (Committee on Revision of the Penal Code, 2020).

Impact on mass incarceration. Finally, a third common question about our work is whether it is possible to use automatically extracted factors for increased review of parole grants, thus increasing the rate at which grants are overturned and contributing to the cycle of mass incarceration. The existing parole review process in California makes additional denials and reversals of grants unlikely. Immediately after a parole hearing, two parole commissioners make a recommendation to grant or deny parole. In the next 120 days, the decision is reviewed by the Parole Board. Afterward, the Governor has 30 days to review the decision before it becomes final. In practice, all parole grants are reviewed, but both the Parole Board and the Governor's review unit say that they lack the resources to review many denials. If the decision is a grant, the candidate is released from prison and the outcome is final. However, if the decision is a denial, nothing changes; the parole candidate remains in prison. So what happens if a prisoner is denied parole, but the decision was in fact inconsistent with the parole decision process? It means there is very limited opportunity to reconsider the case, possibly leaving a prisoner incarcerated much longer than necessary. If an analysis based on features extracted using NLP can identify outlier cases, this is actionable. The Governor may request a review, the Parole Board may advance the date of a hearing, or an appeals attorney may petition a court. On the other hand, there exists no basis on which we should assume that either the Governor or the Parole Board would overturn more hearings when provided with more data about the parole process.

[3] A related question is why proponents of codified justice or social scientists do not ask commissioners to tabulate factors in a hearing as the input to an algorithm, preempting the need for NLP. (Bell et al., 2021) provides a response to this: First, many parole stakeholders greatly value the "human factors" of the parole process; neither the legislature nor the Parole Board believe that an entirely tabular approach is appropriate. Second, by asking the agency that is conducting the hearings to tabulate such data, we postulate that CDCR would provide reliable annotations for all relevant factors. However, sometimes the agency under scrutiny of a review process is not incentivized to provide key data in structured form. For example, the Parole Board in California refused to provide race data for its parole candidates until it faced repeated litigation. Finally, in order to identify systemic inequities, the Recon Approach relies on a broad set of stakeholders to propose factors of inquiry, and knowledge of which factors are relevant may only become available after the fact, such as when legislation changes years after a hearing.

References

Heike Adel, Benjamin Roth, and Hinrich Schütze. 2016. Comparing convolutional neural networks to traditional models for slot filling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 828–838, San Diego, California. Association for Computational Linguistics.

Kristen M Altenburger and Daniel E Ho. 2019. Is yelp actually cleaning up the restaurant industry? a re-analysis on the relative usefulness of consumer reviews. In *The World Wide Web Conference*, pages 2543–2550.

Nicholas Asher and Laure Vieu. 2005. Subordinating and coordinating discourse relations. *Lingua*, 115(4):591–610.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NeurIPS tutorial*, 1:2.

Lindsey Barrett. 2017. Reasonably suspicious algorithms: predictive policing at the United States border. *N.Y.U. Review of Law & Social Change*, 41:327.

Kristen Bell. 2019. A stone of hope: Legal and empirical analysis of California juvenile lifer parole decisions. *Harvard Civil Rights-Civil Liberties Law Review*, 54:455.

Kristen Bell, Jenny Hong, Nick McKeown, and Catalin Voss. 2021. The Recon Approach: A new direction for machine learning in criminal law. *Berkeley Technology Law Journal*, 37.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Kevin Bowden, Jiaqi Wu, Shereen Oraby, Amita Misra, and Marilyn Walker. 2018. SlugNERDS: A named entity recognition tool for open domain dialogue systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA).

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China. Association for Computational Linguistics.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles. Association for Computational Linguistics.

Jinho D. Choi and Henry Y. Chen. 2018. SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Committee on Revision of the Penal Code. 2020. Parole release and penal code section 1170(d)(1) resentencing: Overview.

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023*.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 797–806.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Jennifer K Elek, Roger K Warren, and Pamela M Casey. 2015. *Using risk and needs assessment information at sentencing: Observations from ten jurisdictions*. National Center for State Courts.

Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew Guthrie Ferguson. 2017. Illuminating black data policing. *Ohio State Journal of Criminal Law*, 15:503.

Sarah E Finch, James D Finch, Ali Ahmadvand, Ingyu Choi, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D Choi. 2020. Emora: An inquisitive social chatbot who cares for you. In *3rd Proceedings of Alexa Prize*.

Sharad Goel, Justin M Rao, and Ravi Shroff. 2016. Personalized risk assessments in the criminal justice system. *American Economic Review*, 106(5):119–23.

Daniel E Ho and Alice Xiang. 2020. Affirmative algorithms: The legal grounds for fairness as awareness. *University of Chicago Law Review Online*.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. 2019. A comparative study on transformer vs RNN in speech applications. pages 449–456. IEEE.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018a. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018b. Algorithmic fairness. In *AEA papers and proceedings*, volume 108, pages 22–27.

Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. 2020. SQuAD2-CR: Semi-supervised annotation for cause and rationales for unanswerability in SQuAD 2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5425–5432, Marseille, France. European Language Resources Association.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, and Zhou Yu. 2020. Gunrock 2.0: A user adaptive social conversational system. In *3rd Proceedings of Alexa Prize*.

Fan Luo, Ajay Nagesh, Rebecca Sharp, and Mihai Surdeanu. 2019. Semi-supervised teacher-student architecture for relation extraction. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pages 29–37, Minneapolis, Minnesota. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Combining neural networks and log-linear models to improve relation extraction. arXiv:1511.05926.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, and Christopher Ré. 2018. Snorkel metal: Weak supervision for multi-task learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in Neural Information Processing Systems*, 29:3567.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108. Version 4.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Mihai Surdeanu, Ramesh Nallapati, George Gregory, Joshua Walker, and Christopher D Manning. 2011. Risk analysis for intellectual property litigation. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, pages 116–120.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Kathryne M Young, Debbie A Mukamal, and Thomas Favre-Bulle. 2015. Predicting parole grants: An analysis of suitability hearings for California's lifer inmates. *Federal Sentencing Reporter*, 28:268.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940. Association for Computational Linguistics.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.