Conceptual Questions in Developing Expert-Annotated Data

Megan Ma
meganma@law.stanford.edu
Stanford Center for Legal Informatics,
Stanford Law School
Palo Alto, California, USA

Brandon Waldon
bwaldon@stanford.edu
Stanford University
Palo Alto, California, USA

Julian Nyarko
jnyarko@law.stanford.edu
Stanford Law School
Palo Alto, California, USA

ABSTRACT

In this paper, we argue that nuanced expert annotation often requires a significant rethinking of the traditional paradigms of data annotation. In a small pilot study, we find that even the most highly trained experts demonstrate significant heterogeneity in their evaluation of the document-level coherence of bespoke contracts. The outcomes of our study provide preliminary considerations of how paradigms of document annotation should fully utilize expert annotations in bespoke contexts.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **General and reference** → **Design**; **Empirical studies**; **Experimentation**; • **Theory of computation** → *Semantics and reasoning*; • **Applied computing** → **Document preparation**; **Annotation**; **Law**.

KEYWORDS

data annotation paradigms, large language models, contract review, domain expertise, legal NLP

ACM Reference Format:
Megan Ma, Brandon Waldon, and Julian Nyarko. 2023. Conceptual Questions in Developing Expert-Annotated Data. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594536.3595139

1 INTRODUCTION

It has been argued that specialized domains, such as the legal field, are rarely exposed to research in deep learning due to the high costs of expert annotations [2]. Coupled with the proprietary nature of legal documents, few datasets are broadly available for research. Accordingly, methodology around how expertise may be used to create legal, and specifically contract, datasets remain relatively unexplored. Coupled with the recent explosion of interest in legal applications of generative AI, existing practices around data annotation requires further assessment. This paper aims to reflect on the role and use of expertise in data annotation for Legal NLP. We put forth a small qualitative study to assess the annotation practices of an expert-annotated contract dataset.

Previous evaluation efforts in the domain of contract review can broadly be categorized into two groups: (1) information retrieval (e.g., CUAD [2], Lease Contract Review [5], etc.); and (2) document-level coherence (i.e., ContractNLI [3]). While most prior research has focused on the former (e.g., identifying the presence or absence of individual clauses), recent studies have proposed to treat contracts as "systems" in which individual clauses can interact in complex ways [9]. As such, the answer to a legal inquiry might often not be found in a single provision. Instead, it is important to understand the overall structure of the agreement.

In this paper, we argue that nuanced expert annotation often requires a significant rethinking of the traditional paradigms of data annotation. Current approaches are designed to allow experts to extract limited pieces of information from more bespoke texts such as contracts. However, this use of expertise is rather one-dimensional and does not account for the full scope of annotator proficiency. In effect, the dominant prescriptive paradigm offers a narrow scope of how expertise is translated and operationalized in the labeling process, and thereby, may be underutilizing expert knowledge. For example, traditional models assume that there is a correct answer to most annotation efforts. Heterogeneity in human annotations is viewed as a deficiency that should be removed. However, in our qualitative assessment, we find that even the most highly trained experts demonstrate significant heterogeneity in their evaluation of the document-level coherence of bespoke contracts. This, in turn, suggests that a full evaluation of document-level coherence necessitates a more permissive pooling of the information that liberally combines the signal received from the different annotators. Perhaps more importantly, the heterogeneity in expert annotations implies that models trained on pooled labels should be able to significantly outperform individual human judgments. With increasingly powerful models, including the most recent developments in large language models (LLMs), this appears more feasible than ever. Training is becoming less concerned with quantity of data available, and more significantly towards quality and contextual relevance. This development is particularly important in a field that is historically data-scarce and highly expensive to solicit expert input for.

The outcomes of our study provide preliminary considerations of how paradigms of document annotation should fully utilize expert annotations in bespoke contexts. We conclude that the substantive variability in contractual analysis offers reason for more diversity in expert annotation processes, and could enable opportunity for future large language models (LLM) to not only automate, but also improve the quality of contract review.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595139

2 THE STUDY

2.1 Setup

To conduct this study, we partnered with an international law firm, DLA Piper. 11 mid to senior-level lawyers with a minimum of 8 years' experience as a practicing attorney were asked to review a random sample of five contracts from an expert-annotated dataset – in this case, the CUAD dataset. Four of the attorneys were senior partners at the firm with over 17 years of experience. The heterogeneity across seniority allowed us to assess whether experience is concurrent with expertise, and whether this has an additional corollary impact on contract review. Furthermore, the attorneys had diverse expertise, ranging from technology transactions to employment litigation.

CUAD, short for Contractual Understanding Atticus Dataset [2], is a dataset comprised of over 500 highly bespoke contracts,[1] annotated to identify salient portions of contracts to increase the efficiency of review. The CUAD dataset consists of a broad range of contracts, from distribution and service to complex development and commercialization agreements. The attorneys were divided into two groups, with each group tasked to provide their comments on a non-overlapping set of five contracts. We stratified the agreements across contract types to ensure wide representation. Within each contract type, we chose one contract at random. This process allows us to better evaluate whether the contract review process differs across varying fields of law and whether it is largely consistent across attorneys regardless of specialization.

The tasks assigned to the attorney-participants required their feedback on (1) whether any clauses were in conflict with each other and (2) whether any provisions interacted with one another in a way such that the consequences of the commitments could only be understood if both clauses were taken into consideration. We also requested that the attorneys specify how they determined the clauses in the contract were contradicting and/or interacting. More importantly, we asked each group to review the contracts in their entirety. This would mirror the conditions of contract review and would take into account the potential, multilayered nature of contracts [9]. On average, the reviewed agreements were approximately thirty pages long.

We defined conflict as "Provision A conflicts with Provision B if it's impossible to satisfy both provisions simultaneously" and interaction effect as "Provision A interacts with Provision B if any changes to A will have a simultaneous effect on B." For instance, if a confidentiality clause in a given contract states that "no information with regards to Case X shall be disclosed," while another sub-provision in that contract allows for the "free access to all records between parties." There is an evident conflict between provisions. On the other hand, if a confidentiality clause in a given contract states that "no information with regards to Case X shall be disclosed, except as required by law," while another sub-provision in that contract states, "records shall be made available for inspection unless the records are exempt from disclosure," there is an interaction effect. In this case, Case X would typically be excluded from inspection with the exception of a mandated audit (e.g., IRS tax audit).

In designing the aforementioned tasks for our participating attorneys, we took inspiration from ContractNLI [3]. ContractNLI is distinct from existing prior research in contract review automation, as the authors design a more complex annotation task with the intention to evaluate against an entire document rather than at an individual clause level. Their seminal work introduced the significance of linguistic questions (e.g., natural language inference) in legal document analysis, as well as a sensitivity towards real-world cases for these tools. ContractNLI is interested in context inferred between clauses.

In contrast to the identification of discrete terms within a clause, ContractNLI focuses on how individual contractual clauses relate and operate with one another. In ContractNLI, annotations were multi-task, requiring both classification and evidence identification. Similarly, we asked the attorneys to not only identify relevant clauses, but also explain briefly why they were selected. Unlike ContractNLI, our tasks extended beyond a single type of contract (i.e., non-disclosure agreements) and focused on agreements that were already executed. Furthermore, rather than using synthetic data, we asked the attorneys to evaluate the contracts as is, to better reflect how contracts from the opposing party are received and analyzed (known as third party paper). Again, we wanted to ensure conditions of the study were a close parallel to contract review in practice.

2.2 Limitations

Due to the complexity of the annotation task and the high level of expertise of our annotators, our dataset is necessarily small. Indeed, we estimate that the total cost of our annotations is $220,000.[2] As such, this study is best understood as an initial, exploratory effort, rather than a comprehensive, quantitative assessment. In addition, we note that the contracts represented in our sample are among the most bespoke agreements, and do not represent agreements that a company might conclude routinely throughout its usual business operations.

2.3 Hypothesis

Prior literature has observed two paradigms, prescriptive and descriptive, that have competing motivations in terms of their outcomes for training models [7]. Prescriptive annotations discourage annotator subjectivity, tasking them to encode a single belief through specific guidelines. In contrast, descriptive annotations "encourage annotator subjectivity to create datasets as granular surveys of individual beliefs." [7] The methodology applied in the annotation of legal tasks is currently aligned with the prescriptive account, and for good reason: information extraction tasks for contractual review, such as entity extraction, should be consistent. However, when applied in the context of more complicated legal reasoning tasks, it may be worthwhile to explore whether a descriptive paradigm may be more effective, given the complexity and alleged diversity of contractual analysis.

[1] The authors collected contracts from the public, Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, maintained by the U.S. Securities and Exchange Commission (SEC). Contracts found in this database are required by the SEC to file them publicly and thereby, are deemed to be material agreements. This means that these contracts are not made in the ordinary course of business and are heavily negotiated.

[2] This is based on attorneys' hourly rate multiplied by the number of hours spent reviewing the contracts and the total number of attorneys.

The prescriptive paradigm is observable in CUAD, for example. In the CUAD paper, law students completed the annotations, but were subjected to contract review training prior to the exercise [2].[3] Each individual annotation was then verified against three additional annotators to ensure consistency and correctness. The law students also followed annotation standards set forth by the authors of the paper. In contrast, when developing a tool to assist with the contract redlining, it may be more useful to follow the descriptive paradigm with annotations from senior associates and partners. In this case, there is a higher value in gathering multiple perspectives, including identifying interpretational variation of contractual language.

Interestingly, ContractNLI introduced more complexity to legal datasets, having integrated questions of natural language inference (NLI) in the process. Though ContractNLI does not use expert annotations, their method appears to align more closely with the descriptive paradigm. The authors asked non-expert annotators to select the clause, and subsequently, the relevant NLI label[3]. While an example-oriented guideline was offered to the annotators, it behaved as a clarifying tool to help non-experts better understand the task rather than as a verification mechanism for consistency. Perhaps more importantly, labels from annotators were pooled. This suggests that training models for more analytical contract reasoning tasks could indeed benefit from a descriptive rather than prescriptive approach.

Consider, for example, the following anecdote shared by one of the partners from the study. This partner is a specialist in transactions, notably for contracts in an international commercial context. While the attorney has over 18 years of experience in the drafting and negotiation of contracts, they frequently seek input from their colleague in litigation. This is because the field of dispute resolution provides a lens into how certain contractual language could inadvertently give rise to risks of misinterpretation, and thereby, devolve into a misalignment of party interests. The partner described how their colleague often read the contract in a manner that far exceeded this partner's realm of interpretation. This offered a perspective to contract drafting that the partner, by their own account, could have never achieved alone. The partner would then take their colleagues' feedback and make the necessary adjustments. Similarly, we hypothesize that substantive variability exists across legal expertise, particularly at the mid-senior level. More importantly, this heterogeneity should not be treated as a weakness in training, but rather an opportunity to improve the quality of contract review by developing tools that could interpret multiple perspectives on a single issue.

2.4 Results and Observations

Given the limited data available, we focus on specific, noteworthy examples in the annotation process. We observed that evidentiary signs of consistency were rare. More often, a broader concern for language use in the contracts, and their associated implications, was raised by the attorney-participants. At a higher level, we found a general lack of agreement among participating attorneys around clauses that were deemed to be conflicting, and/or have an interaction effect. Among the 43 unique conflicts identified, on average, only 2 annotators agreed on the inconsistency. That is, of the contracts reviewed by the attorneys, only two contracts had identifiable convergence in their review of clauses that were deemed to conflict and/or interact. While attorneys had noted that clauses did conflict and/or interact in the same contract, the inconsistencies identified were typically not the same set of clauses.

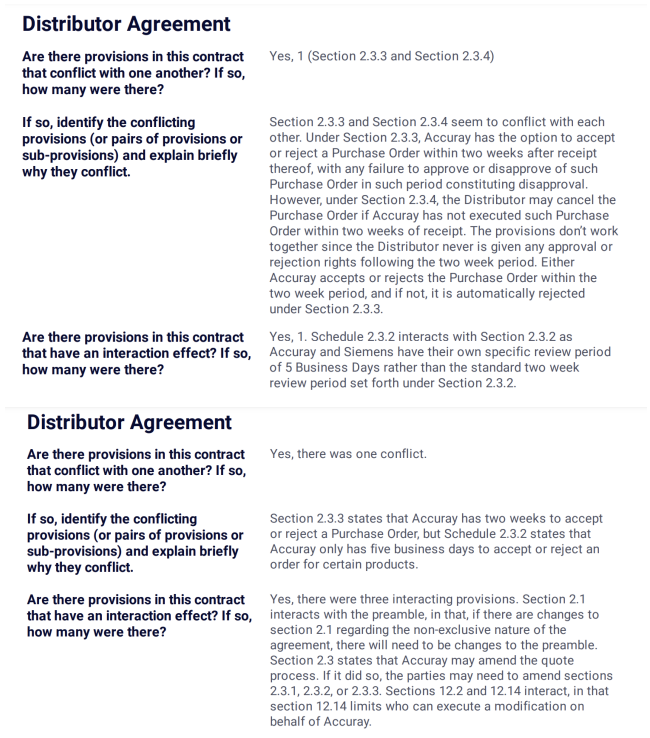

Figure 1: Sample comparative responses from the qualitative pilot.

Consider the example in Figure 1. In this example, both attorneys identified similar clauses of concern, though to a varying degree. One attorney had underscored an outright conflict between Sections 2.3.3 and Schedule 2.3.2. Yet, the other attorney merely raised it as having an interactive effect. Interestingly, both attorneys had similar years of experience, but with different specializations.

Uniformly among the partners that participated in the pilot, all raised concerns about the defined terms in the contracts. One partner noted,

> There is a lengthy definitions section. Any change to a definition will impact the clauses in which that definition is used. Also, wherever there are cross-references to other sections, those sections must be read together with the section where the cross-reference appears (and all definitions used therein) in order to determine the full meaning of the clause.

[3]This was described as roughly 70-100 hours of instruction from experienced lawyers through workshops and video lectures.

In several follow-up interviews with partners of the firm, they noted that they were apprehensive about their feedback and comments on the contracts. One partner described the oddity of the contractual phrasing, as many were written with "unusual language" and appeared to be rather one-sided. The language was regarded as imprecise, and "could be construed to be inconsistent and in conflict with themselves and/or other areas of the agreement" (see Figure 2).

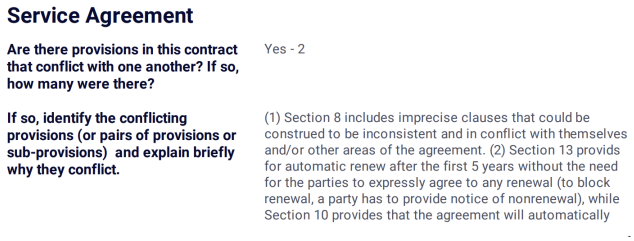

Figure 2: Sample response from the qualitative pilot.

Moreover, the underlying contractual positions between parties felt heavily imbalanced with little-to-no leverage for negotiation. This is peculiar provided that these contracts were not considered standard form contracts, whereby terms and conditions are typically asserted by one party to the other. Moreover, it was difficult for the partners to glean the relationship between parties and their relative business context. Consider the comment from another partner:

> It is relevant to point out that the review of agreements like this requires understanding of the business and context of the agreement. Reading a contract "cold" without an understanding of business and context is likely to result in fewer issues being flagged. Of the five contracts I reviewed, my personal knowledge encompassed three, while the other two were in unfamiliar contexts. I felt much more comfortable with comments on the three to which I could readily comprehend the context of the document.

Similarly, another partner also remarked that certain agreements felt incredibly niche and required highly technical knowledge that few attorneys, even ones with immense experience, could offer. Even with standards around contractual review, identifying clauses of concern could be challenging. The operative effects of the particular language would be largely unknown.

Again, these findings come as no surprise, given the population surveyed in this qualitative pilot is far too small to make generalizable claims. Nevertheless, they do inform that, at minimum – and even within a single firm – there is not necessarily a standard method of parsing contracts across attorneys. Rather, interpretation of the relationships between clauses and their prospective impact appear to differ across specialization and experience. This suggests that even if the contractual clauses and contract types may be identifiable, there remain serious limitations in the applicability of current models to support more nuanced contract review work.

3 REFLECTIONS FROM THE STUDY

The observations from the study demonstrate sufficient variability across legal expertise, particularly in exercises of contractual review beyond information retrieval. We observed that the most experienced attorneys hardly ever arrive at the same, or even, similar conclusion. Rather, they appear to be identifying pieces to the puzzle, which could prove to be useful in exercises of issue-spotting and/or improving the quality of their work products. The study, therefore, brings to light the significance of future work in the space of expert annotation. Coupled with the advent of generative AI, the use of descriptive annotation paradigms in expert domains require further assessment.

In a recent paper, "MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding," we have seen an increasing turn to reasoning as a necessary component to the labelling tasks[8]. More importantly, the authors address the underexplored use of models for reading comprehension tasks of highly complex legal texts. MAUD, the Merger Agreement Understanding Dataset, however, follows a similar methodology in the annotation process to CUAD. The training set was largely used to extract and identify key clauses. While there are signs of increased rigor in the verification of annotations, namely by experienced M&A lawyers, there remains a limited use of expertise in the training of the model. Again, the annotation process followed a prescriptive paradigm and aimed to discourage inter-annotator subjectivity.

For example, the clauses in question were drawn from the 2021 ABA Study that captured prevalence of provisions in private target M&A transactions[6]. This particular study considered market trends in clauses found in contracts, revealing the "golden standard" of contractual language in M&A agreements. As this study already relied on the collective expertise of M&A practitioners, further use of M&A expertise for the verification of annotations appears to be a conservative use case to leverage the data gathered from the 2021 ABA Study. In contrast, an interesting opportunity may be to utilize this data in a comparative manner, evaluating risk associated with departing from standard language for M&A agreements. This would enable attorneys to not only understand what the golden standard is, but why this may be the case.

4 FUTURE DEVELOPMENTS AND NEXT STEPS

Though these datasets have enabled the broader research community to build and train with legal data, we now face an increasing possibility of evolving models that do not require the labelling demands that prior models needed. At the wake of the next generation of LLMs, how should we be rethinking the labeling process in the expert annotation space? That is, what inputs should we be seeking in order to advance the research in Legal NLP? As we continue to hear about the various ways in which generative AI are finding integration, existing methods of labelling and curating expert data may become antiquated. Consequently, it is crucial that we reflect on how to best leverage the diversity present in legal expertise, as it may become increasingly necessary to distinguish annotation processes for models used in complex legal reasoning tasks.

As noted, many of the current expert-annotated datasets are largely used for information extraction [1, 2, 8][4] and apply a prescriptive paradigm. However, our study has shown that heterogeneity is found across expertise, suggesting, instead, the need to apply a descriptive annotation paradigm. More experiments with descriptive data annotation, such as the pooling of annotator labels in clusters, and/or training on the distributions of labels given by annotators, should be conducted to preserve the integrity of diverse feedback. We imagine then future opportunities to train LLMs on multi-annotator model architectures to enable a plurality of contractual insights. This supports not only how attorneys currently provide highly bespoke contractual review, but also is consistent with existing use cases for descriptive data annotation (e.g., subjective tasks like modelling abuse detection or hate speech). Accordingly, the next phase of our research partnership with DLA Piper will experiment with descriptive annotation processes to help train and fine-tune models that could support capturing how legal experts read contracts. More specifically, we will be gathering blacklines[5] of contracts from our attorney-participants, hoping to better capture how contractual language compares across expertise and specialization. We anticipate that these experiments could advance LLM research in building quality legal analysis tools and, eventually generative legal drafting tools, that are more closely aligned with legal practice.

5 CONCLUDING REMARKS

To reiterate, we recognize that the qualitative study is a rather narrow glimpse into the questions explored about expert-annotated datasets. Equally, we remain cognizant and do not claim that the aforementioned observations provide any definitive conclusions on the quality of these annotations. Instead, we hope to highlight the complexity and heterogeneity of feedback from legal experts, suggesting that there are fertile research grounds to explore translating expertise in annotation processes. For example, there exists symptoms of rising interest in the medical space to seek descriptive annotations to improve the quality of medical imaging [4]. In any event, the increasing prevalence, and prospects, of generative AI provide an opportunity to reassess existing methods of expert annotation, enabling a potential shift from contract review automation to contract refinement.

ACKNOWLEDGMENTS

This research is made possible thanks to our affiliate partnerships at the Stanford Center for Legal Informatics (CodeX) and to the funding support from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

REFERENCES

[1] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. LexGLUE: A benchmark dataset for legal language understanding in English. *arXiv preprint arXiv:2110.00976* (2021).

[2] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268* (2021).

[3] Yuta Koreeda and Christopher D Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799* (2021).

[4] Khiem H Le, Tuan V Tran, Hieu H Pham, Hieu T Nguyen, Tung T Le, and Ha Q Nguyen. 2022. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. *arXiv preprint arXiv:2203.10611* (2022).

[5] Spyretta Leivaditi, Julien Rossi, and Evangelos Kanoulas. 2020. A benchmark for lease contract review. *arXiv preprint arXiv:2010.10386* (2020).

[6] Jessica C. Pearlman. 2021. *2021 ABA PRIVATE TARGET MERGERS ACQUISITIONS DEAL POINTS STUDY.* Retrieved May 5, 2023 from https://www.klgates.com/2021-ABA-Private-Target-Mergers-Acquisitions-Deal-Points-Study-12-31-2021

[7] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475* (2021).

[8] Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding. *arXiv preprint arXiv:2301.00876* (2023).

[9] Spencer Williams. 2020. Contracts as systems. *Del. J. Corp. L.* 45 (2020), 219.

Received 6 May 2023

[4] These include the aforementioned datasets of CUAD and MAUD, but also in considerations of benchmark datasets such as LexGLUE.

[5] This may be defined as the benchmark language (golden standard) that a specific firm uses when drafting contracts.