A Dataset of Contractual Events in Court Decisions

Guilherme Paulino-Passos[1], Ken Satoh[2] and Francesca Toni[1]

[1]Department of Computing, Imperial College London, UK
[2]Principles of Informatics Research Division, National Institute of Informatics, Tokyo, Japan

Abstract

The promise of automation of legal reasoning is developing technology that reduces human time required for legal tasks or that improves human performance on such tasks. In order to do so, different methods and systems based on logic programming were developed. However, in order to apply such methods on legal data, it is necessary to provide an interface between human users and the legal reasoning system, and the most natural interface in the legal domain is natural language, in particular, written text. In order to perform reasoning in written text using logic programming methods, it is then necessary to map expressions in text to atoms and predicates in the formal language, a task referred generally as information extraction. In this work, we propose a new dataset for the task of information extraction, in particular event extraction, in court decisions, focusing on contracts. Our dataset captures contractual relations and events that affect them in some way, such as negotiations preceding a (possible) contract, the execution of a contract, or its termination. We conducted text annotation with law students and graduates, resulting in a dataset with 207 documents, 3934 sentences, 4440 entities, and 1794 events. We describe here this resource, the annotation process, its evaluation with inter-annotator agreement metrics, and discuss challenges during the development of this resource and for the future.

Keywords

Contract Law, Information Extraction, Language Resource

1. Introduction

The promise of automation of legal reasoning is developing technology that reduces human time required for legal tasks or that improves human performance on such tasks. In order to do so, different methods and systems based on logic programming were developed since the beginning of the field of AI and Law [1, 2], as well as more recently [3, 4]. However, in order to apply such methods on legal data, it is necessary to provide an interface between human users and the legal reasoning system, and the most natural interface in the legal domain is natural language, in particular, written text. In order to perform reasoning in written text using logic programming methods, it is then necessary to map terms and expressions in text to entities and predicates in the formal language, a task referred generally as information extraction [5, 6]. Thus, the identification of legally relevant events and facts from textual descriptions is an important intermediate task for automatic legal reasoning from raw text.

LPLR2023: Logic Programming and Legal Reasoning, workshop at the International Conference on Logic Programming, Imperial College London, London, United Kingdom, July 9-10, 2023
g.passos18@imperial.ac.uk (G. Paulino-Passos); ksatoh@nii.ac.jp (K. Satoh); f.toni@imperial.ac.uk (F. Toni)
0000-0003-3089-1660 (G. Paulino-Passos); 0000-0001-8194-1459 (F. Toni)

More concretely, PROLEG [3] is a logic-programming-based system able to reason about contracts, inferring, for instance, their validity, based on events regarding such contracts. In order to apply PROLEG to cases in text, a framework called `ContractFrames` was developed [7]. This framework applies a rule-based method for extraction of relations and events to be used in PROLEG. However, this methodology was not evaluated on real data and expects a more limited form of sentences.

Here we propose a new dataset[1] on legal relations and events about contracts in court decisions. We envisage that this resource will allow for both performance evaluation directly for information extraction and as a source for supervised learning training for the automatic analysis of legal text. Thus we aim on moving the previous `ContractFrames` work to naturally occurring text. This dataset enables the subtasks of named entity recognition (NER) and event extraction.

Our new dataset is concerned with the task of event extraction in court decisions, focusing on contracts (see Figure 1 for an example of annotated paragraph). It captures contracts (such as sales contracts, labour contract, and rental agreements), represented by trigger expressions in the text from which one can infer their existence. It also captures events that modify those contracts in some way, such as negotiations preceding a (possible) contract, the execution of a contract, or its termination (in other words, what is more specifically called "contract events" in [7]). The entities that participate in the contract or perform the contract events are also annotated. In a logic programming representation such as PROLEG, the extracted entities correspond to terms (including the contract itself, which is reified), while the entity types, relations, and events correspond to predicates [7].

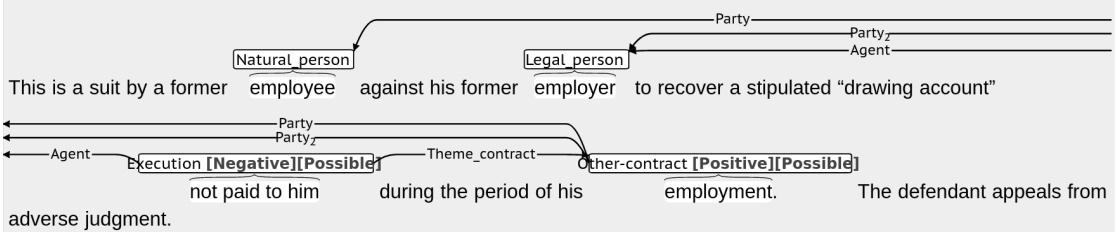
The paper is organised as follows. In Section 2, we present the methodology, chosen criteria, and guidelines for annotation. In Section 3 we discuss evaluation methodology and achieved outcomes of the annotation, including main challenges faced, and we illustrate disagreements between annotators. In Section 4 we discuss previous work on the annotation of events, and finally in Section 5 we discuss future work and conclusions.

2. Dataset Development

Choice and Overview of Corpus. We build our annotation on top of the U.S. Caselaw Dataset [8], a large dataset of high court decisions from different U.S. jurisdictions. In particular, our annotation is done on the Louisiana jurisdiction, and thus our cases consist of cases from Supreme Court of Louisiana and the Louisiana Court of Appeals. This decision is motivated by an aspect from the domain: Louisiana is a mixed jurisdiction, mixing aspects of civil law and common law, and has a special position of being a jurisdiction with a civil code in English. Although this is not crucial for the contract event extraction task, we believe work on this data would allow easier transfer to civil law jurisdictions. The documents were selected in batches, each randomly sampled from all Louisiana cases from Caselaw.

Annotation Methodology. For annotation, we had 6 law school students or graduates. They were selected from students from Europe and Japan, with fluency in English. They were asked

[1] Available at: https://doi.org/10.5281/zenodo.8098312




```
natural_person(employee1).
legal_person(employer1).
other_contract(employment1).
manifestation_fact(execution(employment1), negative, employee1,
  employer2, Date).
```

Figure 1: Example of annotation for an employment contract and below a possible PROLEG representation of the extracted information. Here, *employment* is labelled as a contract of type `Other-contract` and is deemed `Positive` (as in, existing, instead of negated) and a contract event of type `Execution` is labelled as `Negative`; both are labelled as `Possible`. Here polarity and possibility are with respect to factuality (see Table 1). The entities that are part of this contract and participate in the event (*employee* and *employer*) are also suitably annotated, with the `Party` relation to the contract, while the single entity that acts in the `Execution` event is related to it by the `Agent` relation.

to read the annotation guidelines, containing several examples of annotations, and annotate by themselves full, previously unseen opinions. We conducted the annotation using the brat annotation tool[2] [9]. We chose brat because it is a well-used open source web-based annotation tool, suitable for the annotation of entities, events and relations. We used the first rounds of annotation to iteratively improve our initial guidelines, based on the annotators' feedback. For most rounds, annotators were asked to annotate the same documents, initially to compare and achieve a shared understanding of the guidelines, but also to have a sizeable dataset for which one can measure appropriately the hardness of the task and reasonable disagreements.

Annotation Guidelines. We describe here the final guidelines. They are presented in the form of main principles and more detailed instructions for entities, relations, and events, and for the tagsets. Our main principles underpinning the guidelines are:

1. The annotation is guided by an interpretation which is sensible for a general legal specialist, regardless of jurisdiction.
2. Only annotate entities which participate in an event or occur in a sentence in which there is a `Contract` or a `ContractEvent`.
3. Any relations (including event participation) must only be between entities or events in the same sentence, or in neighbouring sentences.
4. When choosing the type of an entity, relation, or event, choose the most specific one that can be inferred from the text.

[2] http://brat.nlplab.org

5. Specifically for pronouns, only annotate one when it is a part of a relation or event.
6. For every pronoun, if you can find (in the same or neighbouring sentence) the entity it refers to, you must add the `same_reference` relation.
7. Only annotate events, entities and relations to which you can associate an actual expression in the text.
8. Annotate as many events, entities and relations as possible, following the above guidelines.

The first principle is a practical one, as it aims to liberate the annotator from worrying or spend time researching details from U.S. law. The second is motivated by the fact that our goal is finding contracts, not giving a full event annotation for every sentence in the dataset. The third principle aims to balance between unbound relation or argument annotation in the same document and providing some continuity and connection between sentences. The next instruction is important for the way the tagset is organised. Both entities and events are organised hierarchically, allowing some granularity for the annotation, as can be seen in Figure 3a. Event types were based on the Contract Workflow Ontology in the `ContractFrames` work [7]. If an annotator is unsure whether the term "defendant" in a case refers to a human being (thus, `Natural_person`) or perhaps a company (which would be a `Legal_person`), then the annotator should use the more general `Person`. This raises important issues for evaluation, that we discuss in Section 3.

As for principles 5 and 6, the inventory of relations between entities is very limited, since the focus of annotation is on events and their arguments. Most important is the `same_reference` relation, for co-reference. Since relations between entities in neighbouring sentences must be annotated, this is useful for tracking an entity. In addition, in our guidelines pronouns are to be annotated only if they are part of a relation or event, but every annotated pronoun must be linked to another entity by the `same_reference` relation, if there is co-referring entity in the same sentence or in a neighbouring sentence. Overlaps were allowed, and might be important when a single word or expression is the expression of both the contract and a contract event, or of a contract and one of its participating entities, such as in Figure 2.

Finally, the last two principles, 7 and 8, are instantiations of text-bound annotation principle and event-centred annotation, described in previous event annotation literature [10]. Principle 7 means that annotation requires to be bound to the text, in a way that the text itself provides clear evidence for the event being annotated. That is, every event annotation requires a trigger expression from which one can infer that, in that context, that expression is evidence for the event. This restricts annotation to sentences with clear indication of the event, forbidding annotators to annotate without textual evidence, even if an annotator could expect the existence of a contract in a specific scenario, based on their background knowledge. Principle 8 means that one should read the text searching for as many events as possible, within the bound of the previous rule, but possibly including multiple events per sentence, or events indirectly mentioned (but which can be associated to a textual expression).

There are elements which are not fully determined by the guidelines, since they might depend on what the annotator considers to be enough evidence of the contract or a description of the entity, and thus annotators were allowed some discretion. For instance, our guidelines do not specify whether to include or not modifiers of a noun as objects of a contract, such as in Figure 2, where in the example "house" is tagged as an entity, but an annotator could have annotated the entire "house at the countryside" span as the same entity. We compare the impacts of the

resulting disagreements on the next section.

An important aspect is of factuality annotation [11, 12]. Factuality annotation follows the tags from [12], seen in Table 1. Factuality is whether the mentioned event is presented by a source as a true event, that is, something that has happened, or as something different, such as a possible but uncertain event, a denied occurrence, or as a hypothetical scenario. Factuality may be analysed with respect to different sources: for instance, the author of the text may be reporting what another person said, and thus two different analyses are to consider the view from the author and to consider the view from the reported person. In our case, we only consider the author source, asking the annotator to evaluate whether the mentioned event indeed happened according to the report of the judge.

Table 1
Interpretation of factuality labels, adapted from [12]. Notice that some combinations are impossible.

	Positive	Negative	Underspecified
Certain	fact	negated fact	certain but unknown
Probable	probably	probably not	–
Possible	possibly	possibly not	–
Underspecified	–	–	unknown or uncommitted

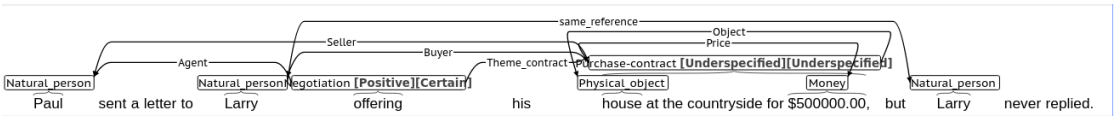


Figure 2: Example of an annotated sentence containing entity expressions overlapping a contract trigger expression. Notice how since the offer was made, but it is unclear whether it was accepted, the contract is listed as of underspecified factuality.

Lastly, for documents in which no contract or contract event was found, we asked annotators to annotate expressions which they found to be suggestive that the document will not contain contracts to annotate, using a special `Relevant_span` tag, as in Figure 4. It was not mandatory to annotate such a span for decisions without contracts. Annotators were instructed to use it only if they found an expression to make the existence of a contract in that document very unlikely. Notice that even an opinion which is not mainly concerned about contract law can still mention contracts, so should still be annotated. We envisage the `Relevant_span` tag as a sort of explanation for the lack of annotation in a document.

3. Dataset Analysis

The dataset has in total 207 unique documents, for a total of 3934 sentences and 300684 tokens. We report general statistics (totals and averages) of annotations in Table 2. Many of the documents have annotations from multiple annotators, and thus we show total countings including this multiplicity, total over documents averaging over annotators, and averages over annotators and documents.

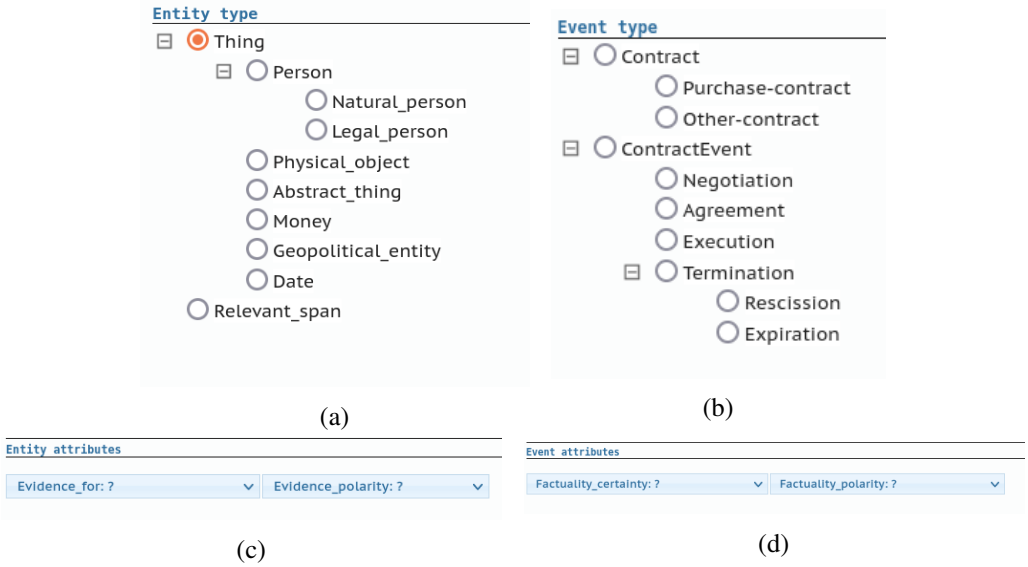


Figure 3: Annotation options.

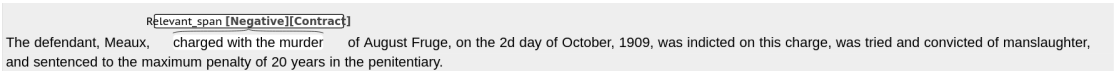


Figure 4: `Relevant_span` that indicates that the document is not about contract law.

Table 2
Statistics of the annotation. Counts include instances of sublabels, in line with the discussion in Section 3. Formally, A1 and A2 are obtained as follows. Let $n(a_i, d_j)$ be the number of annotations by annotator a_i on document d_j; let $S_j \subseteq \{1, \ldots, 7\}$ be the set of (indices of) annotators for document d_j, and let $c(d_j) = (\Sigma_{i \in S_j} n(a_i, d_j))/|S_j|$. Then A1$= \Sigma_{j \in \{1,\ldots,207\}} c(d_j)$ and A2=A1$/207$.

	`Thing`	`Contract`	`ContractEvent`	Events
Total	4440	1411	383	1794
Average across annotators, totalling over all documents (A1)	1668.7	527.25	149.75	677
Average across annotators and documents (A2)	8.06	2.55	0.72	3.27

Note that a substantial number of 94 documents out of the 207 have zero annotations (after analysis by annotators). Many of them have no real substantial text, and are mostly procedural considerations, although some of them do have content, but no annotator has found anything to be tagged. We have included them nonetheless for annotation in order to have the annotators make the judgement call on whether each of these documents is substantial or not. Since this carries information originated by the annotators, we include these documents with zero annotations in our dataset. Nevertheless, depending on the usage it might be advisable to exclude such documents.

Inter-annotator agreement measures. We measured inter-annotator agreement (IAA) in different ways, based on previous literature on event annotation [13, 14, 5]. For the purposes of the evaluation below, we are considering event triggers as entities, drawing no distinction between them and regular entities (that is, entities of type `Thing`, or one of its subtypes), unless otherwise specified. Furthermore, event arguments are the only relations evaluated in the metrics below.

Our metrics are based on the F1 score. Essentially this means that for a pair of annotators, we first consider annotations of one of then as the reference and calculate recall of the other annotation with respect to it, then roles are inverted, recall is again calculated, and finally we use the harmonic mean of those two values [15]. This way, F1 can be interpreted as how well one annotation predicts the other, where which one is the gold standard does not matter. Since the number of negative cases is very high, and the labels are not mutually exclusive, Cohen's kappa is not appropriate [5, 15]. An important aspect of the F1 score is that it can be used to handle sublabels more appropriately: if label L2 is a sublabel of label L1, annotator A has tagged a span as L1, and annotator B has tagged the same span as L2, then the sublabel information can be used to consider that, while annotator A has missed regarding B's annotation as the gold standard, annotator B annotated it correctly since L2 implies L1. We use this principle for all entity metrics.

Thus the main metrics we use are:

- for entities, evaluate
  - by text, character-by-character: for each character which is part of an entity for annotator A, check whether it is also part of an entity for annotator B;
  - by entity, strictly: for each entity, it is considered matched only by an entity with text spans have an exact match;
  - by entity, relaxed: for each entity, it is considered matched by an entity with any overlapping text;
- for relations, first find a match for each entity; if there are no matches for either entity, this is considered a miss; then, check whether for the matched entities there is a matching relation; entities are matched "optimistically": an entity is considered a match if it has at least some character overlap, it is the entity with maximum character overlap, and it is either a subtype or a supertype of the original type.

We evaluate those metrics by considering every pair of annotators, calculating a metric for this pair over the entire dataset, restricted to the documents which have been considered by both annotators, and then evaluating agreement between every possible pair according to the metric and averaging the results by pair of annotators [5].

Results. We present our agreement results in Table 3. As expected, the character-based evaluation is the stricter metric, since unmatched entities with longer spans have more weight than a matched entity with a shorter span (which are typically easier to agree on), while in the entity-based metrics, they count as a single mistake. For entities and contracts in general, agreement has been moderate, ranging from 40% to 60%, depending on the metric. However, annotations for `ContractEvent` show weaker agreement. This is partly due to some pairs of annotators having very few documents in common, biasing some results. This, coupled with the low number of `ContractEvents`, pushes down the average over annotators. Restricting to annotator pairs with more documents in common have agreement metrics between 27% and 46%.

Table 3
Inter-annotation agreement metrics, in average F1 score over annotators. Sublabels are considered and only event arguments are reported as relations, in line with the discussion in Section 3.

Entities	`Thing`	`Contract`	`ContractEvent`	All entities and events
characters	46.80%	46.56%	20.56%	46.08%
entity - strict	49.63%	56.29%	21.21%	49.79%
entity - relaxed	61.30%	64.32%	31.57%	60.84%

	any relation	`Theme`	`Object`	`Theme_contract`
Relations	45.10%	42.61%	75.19%	26.29%

Disagreement analysis. Some disagreements can be attributed to annotators missing information in the sentence. In Figure 5a, we can see as the first annotator did not specify the theme (in order words, the object) of the contract. Some contracts seem to have been missed by some annotators, on occasion.

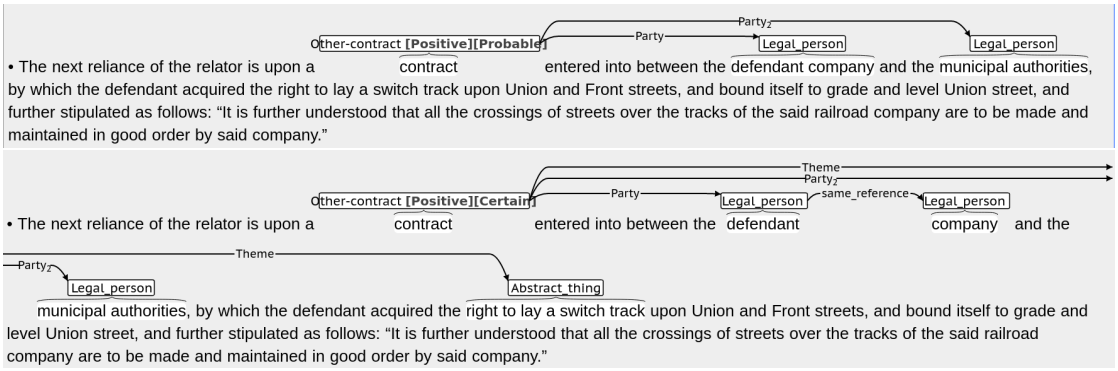
Guidelines require annotation of every entity in a sentence with an event, even if the entity itself does not participate in an event, but it was hard to make annotators follow this instruction consistently. This instruction is intended for improving agreement so that, even if annotators disagree on what are the entities participating in the event, at least one can verify that both agree on the existence of the same entities. However, frequently annotators miss such non-participating entities . This can be seen in Figure 5a: the second occurrence of the word "defendant" is not annotated by anyone. Although this was raised in meetings with the annotators, it is still a frequent source of disagreement in the data.

Another source of disagreement is the difference in what was included in the annotation for entities or event triggers. Annotators made different choices on when to include modifiers as parts of the entities. A simple but occurring example is in Figure 5b, where the determiner "the" was included by one annotator and omitted by the other. While this was discussed in iterations with annotators, this was not strictly defined in the guidelines, since some modifiers might be deemed necessary for specifying the entity, or as necessary for triggering an event, and thus this was open to annotator discretion. Figure 5c shows a less trivial example, involving coreferences.

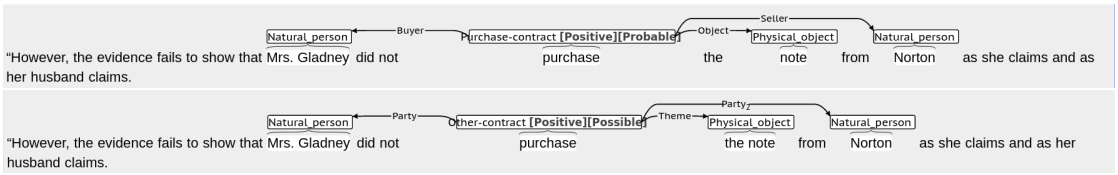
Finally, a form of disagreement particularly severe for `ContractEvents` is that sometimes annotators agreed on the mention of a `Contract`, but disagreed on whether the trigger itself also meant a `ContractEvent` or not. That is, whether the sentence describes a change in the dynamic of the contract, or if it is just an evidence for the contractual relationship. This occurs in Figure 5d.

4. Related Work

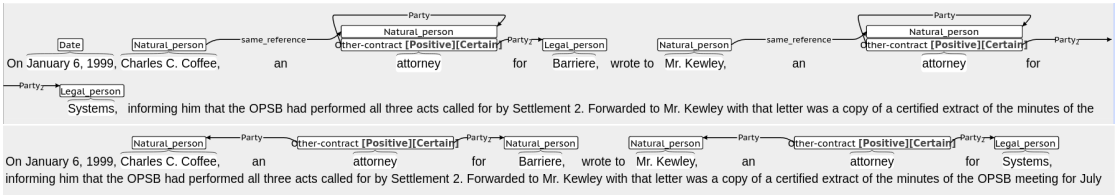
Events have been a recent topic of research in AI&Law, including a dataset of annotated events in court decisions [6] also presenting. Their work consists of 30 decisions of the European Court of Human Rights (ECHR) with time-related events annotated. Their work is motivated by the goal of automatically generating timelines from text. While motivated by `ContractFrames` [7], our work, in contrast, is based on decisions from the Supreme Court of Louisiana and the



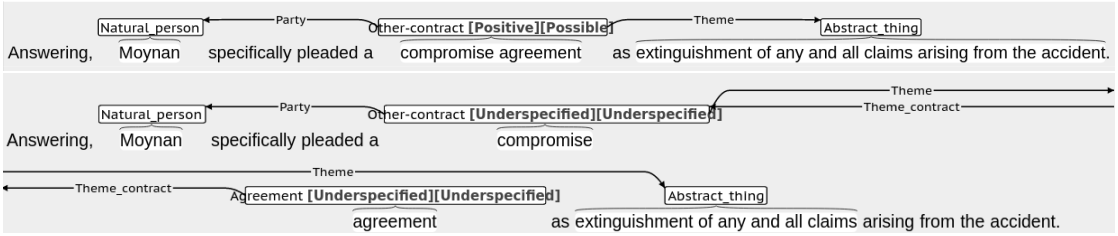
(a) Second annotator did not include a `Theme` for the occurring contract as the first annotator did.



(b) Disagreement regarding whether to include the determiner “the” in the entity. There is also a disagreement on whether the annotated contract is a `Purchase-contract` or an `Other-contract`, likely by a mistake by the second annotator.



(c) Both annotators have considered that the word “attorney” is a trigger for a `Contract`, since the sentence states that “Charles C. Coffee” is an attorney for “Barriere”, but disagreed on how to annotate the `Natural_person` entity. Notice that we allow annotations to overlap.



(d) One annotator understood the word “agreement” only as a mention of a (possible) contract, while the other as an `Agreement` event with underspecified factuality. Our understanding of the guidelines is that the second way is preferred, but this factuality makes it a subtle case.

Figure 5: Selection of disagreements by pairs of annotators.

Louisiana Court of Appeals, and we are focused on the goal of finding contractual relations and events that affect those relations, as distinguished from other facts or legal events. Even though the life cycle of events is relevant for annotation, the focus is not on capturing every event in the

text and their temporal order. This information is more sparse, and thus we are able to present a dataset with more documents.

Works on annotation and automatic extraction of events have received significant attention for more than a decade by the bioinformatics community [10], being also the topic of shared tasks in that field [14, 16]. Bioinformatics also faces the challenges of being a technical domain, being knowledge-intensive and requiring specialist annotation.

More generally, information extraction is an important topic in AI&Law, such as for ontology population [17]. A shared task from SemEval-2023 contains tasks on detecting entities relevant to the procedure, such as case number, petitioner, and cited statutes, on Indian legal decisions [18]. There are also open software available for some usual tasks such as parsing document structure, extracting named entities, and structured information such as citations [19, 20], although for some specific settings and jurisdictions.

5. Conclusion and Future Work

In this work we present a new dataset for contracts and events related to their workflows in case text. We contribute with the resource itself, and present the used methodology, the guidelines that resulted in the dataset, and our quantitative (inter-annotator agreement) and qualitative evaluations of the dataset, discussing reasons for the disagreements.

We leave for future work the analysis of performance of state-of-the-art NLP methods on this dataset. Recent state-of-the-art performances have been dominated by large contextualised language models such as BERT [21], available in the form of pre-trained models, along with variants [22, 23], as well as in-domain models, pre-trained in legal text [24, 25]. Such models have been deployed for the information extraction task, such as in the biomedical domain [26, 27, 28], but also in law [25] (although [29] reports an information extract in which BiLSTMs outperform BERT models). Thus an evaluation of the performance of BERT models on this dataset would be a natural next step.

Since this work was motivated by [7], an important future work is evaluating their rule-based system on this new dataset, and comparing that performance to a machine-learning-based system, with respect to the extraction of events. Another possible comparison is at the reasoning level, that is, to connect a system trained on this dataset to automatically parse events from text to a reasoning system such as PROLEG [3], allowing a comparison with [7] at the downstream reasoning task instead of at the event extraction task.

Acknowledgments

Guilherme Paulino-Passos would like to thank the National Institute of Informatics, Tokyo, Japan, for supporting his visit to Japan that made this work possible, as well as Capes (Brazil, Ph.D. Scholarship 88881.174481/2018-01). Francesca Toni also acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No.101020934, ADIX), as well as support from J.P. Morgan and the Royal Academy of Engineering, UK, under the Research Chairs and Senior Research Fellowships scheme. Ken Satoh acknowledges support by JSPS KAKENHI Grant

Number, JP22H00543 and JST, AIP Trilateral AI Research, Grant Number JPMJCR20G4. All authors would also like to thank the annotators who participated in this project: Arisa Ishikawa, Ayça Aysoy, Mihiro Ikeda, Ryohei Yoshida, Tom Lenters, Yasuaki Mori, Yuki Omata, Yuto Mori. We also thank anonymous reviewers for their feedback on earlier drafts of this work.

References

[1] M. J. Sergot, F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond, H. T. Cory, The British Nationality Act as a logic program, Commun. ACM 29 (1986) 370–386.

[2] T. J. M. Bench-Capon, G. O. Robinson, T. Routen, M. J. Sergot, Logic programming for large scale applications in law: A formalisation of supplementary benefit legislation, in: Proceedings of the First International Conference on Artificial Intelligence and Law, ICAIL '87, Boston, MA, USA, May 27-29, 1987, ACM, 1987, pp. 190–198. URL: https://doi.org/10.1145/41735.41757. doi:10.1145/41735.41757.

[3] K. Satoh, K. Asai, T. Kogawa, M. Kubota, M. Nakamura, Y. Nishigai, K. Shirakawa, C. Takano, PROLEG: an implementation of the presupposed ultimate fact theory of Japanese civil code by PROLOG technology, in: T. Onoda, D. Bekki, E. McCready (Eds.), New Frontiers in AI, JSAI-isAI Workshops, volume 6797 of *LNCS*, Springer, 2010, pp. 153–164. doi:10.1007/978-3-642-25655-4_14.

[4] J. P. Morris, Spreadsheets for Legal Reasoning: The Continued Promise of Declarative Logic Programming in Law, Master's thesis, Faculty of Law and Department of Computing Science, University of Alberta, 2020.

[5] P. Thompson, S. A. Iqbal, J. McNaught, S. Ananiadou, Construction of an annotated corpus to support biomedical information extraction, BMC Bioinformatics 10 (2009). doi:10.1186/1471-2105-10-349.

[6] E. Filtz, M. Navas-Loro, C. Santos, A. Polleres, S. Kirrane, Events matter: Extraction of events from court decisions, in: S. Villata, J. Harasta, P. Kremen (Eds.), Legal Knowledge and Information Systems - JURIX 2020, volume 334 of *Frontiers in AI and Applications*, IOS Press, 2020, pp. 33–42. doi:10.3233/FAIA200847.

[7] M. Navas-Loro, K. Satoh, V. Rodríguez-Doncel, Contractframes: Bridging the gap between natural language and logics in contract law, in: K. Kojima, M. Sakamoto, K. Mineshima, K. Satoh (Eds.), New Frontiers in AI, Springer International Publishing, 2019, pp. 101–114.

[8] Caselaw access project, 2018. URL: https://case.law.

[9] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, brat: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations Session at EACL 2012, Association for Computational Linguistics, Avignon, France, 2012.

[10] J. Kim, T. Ohta, J. Tsujii, Corpus annotation for mining biomedical events from literature, BMC Bioinform. 9 (2008). doi:10.1186/1471-2105-9-10.

[11] R. Saurí, J. Pustejovsky, Factbank: a corpus annotated with event factuality, Lang. Resour. Evaluation 43 (2009) 227–268. doi:10.1007/s10579-009-9089-9.

[12] R. Saurí, J. Pustejovsky, Are you sure that this happened? assessing the factuality degree of events in text, Comput. Linguistics 38 (2012) 261–299. doi:10.1162/COLI_a_00096.

[13] S. Kulick, A. Bies, J. Mott, Inter-annotator agreement for ere annotation, Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation (2014). doi:10.3115/v1/w14-2904.

[14] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, J. Tsujii, Overview of bionlp'09 shared task on event extraction, Proceedings of the Workshop on BioNLP Shared Task - BioNLP '09 (2009). doi:10.3115/1572340.1572342.

[15] G. Hripcsak, Agreement, the f-measure, and reliability in information retrieval, Journal of the American Medical Informatics Association 12 (2005) 296–298. doi:10.1197/jamia.m1733.

[16] J.-D. Kim, Y. Wang, N. Colic, S. H. Beak, Y. H. Kim, M. Song, Refactoring the genia event extraction shared task toward a general framework for ie-driven kb development, Proceedings of the 4th BioNLP Shared Task Workshop (2016). doi:10.18653/v1/w16-3003.

[17] L. Robaldo, L. di Caro, L. A. Alemany, M. Palminari, S. Villata, Deliverable D2.4: Ontology population: connecting legal text to ontology concepts and instances, Technical Report, Mining and Reasoning with Legal Texts (MIREL), 2018.

[18] A. Modi, P. Kalamkar, S. Karn, A. Tiwari, A. Joshi, S. K. Tanikella, S. K. Guha, S. Malhan, V. Raghavan, Semeval 2023 task 6: Legaleval - understanding legal texts, CoRR abs/2304.09548 (2023). URL: https://doi.org/10.48550/arXiv.2304.09548. doi:10.48550/arXiv.2304.09548. arXiv:2304.09548.

[19] M. Bommarito, D. Katz, E. M. Detterman, Lexnlp: Natural language processing and information extraction for legal and regulatory texts, InfoSciRN: Legal Informatics (Topic) (2018).

[20] ICLR&D, Blackstone library, ???? URL: https://research.iclr.co.uk/blackstone.

[21] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692.

[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: 8th International Conference on Learning Representations, ICLR, 2020.

[24] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, D. E. Ho, When does pretraining help?, Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (2021). doi:10.1145/3462757.3466088.

[25] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, Findings of the Association for Computational Linguistics: EMNLP (2020). doi:10.18653/v1/2020.findings-emnlp.261.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019). doi:10.1093/bioinformatics/btz682.

[27] P. Su, Y. Peng, K. Vijay-Shanker, Improving BERT model using contrastive learning

for biomedical relation extraction, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 1–10. doi:10.18653/v1/2021.bionlp-1.1.

[28] P.-T. Lai, Z. Lu, BERT-GT: cross-sentence n-ary relation extraction with BERT and Graph Transformer, Bioinformatics 36 (2021) 5678–5685. doi:10.1093/bioinformatics/btaa1087.

[29] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Neural Contract Element Extraction Revisited: Letters from Sesame Street, 2021. arXiv:2101.04355v2.