Information Extraction from Legal Documents: A Study in the Context of Common Law Court Judgements

Meladel Mistica[♦][*] Geordie Z. Zhang[♥][*] Hui Chia[♣][*]
Kabir Manandhar Shrestha[♥] Rohit Kumar Gupta[♥] Saket Khandelwal[♥]
Jeannie Marie Paterson[♣] Timothy Baldwin[♦] Daniel Beck[♦]
[♦] School of Computing and Information Systems
[♥] Melbourne Data Analytics Platform [♣] Melbourne Law School
The University of Melbourne, Australia
{misticam, geordie.zhang, chia.h}@unimelb.edu.au
{kmanandharsh, rohitkumarg, saketk}@student.unimelb.edu.au
{jeanniep, tbaldwin, d.beck}@unimelb.edu.au

Abstract

'Common Law' judicial systems follow the doctrine of precedent, which means the legal principles articulated in court judgements are binding in subsequent cases in lower courts. For this reason, lawyers must search prior judgements for the legal principles that are relevant to their case. The difficulty for those within the legal profession is that the information that they are looking for may be contained within a few paragraphs or sentences, but those few paragraphs may be buried within a hundred-page document. In this study, we create a schema based on the relevant information that legal professionals seek within judgements and perform text classification based on it, with the aim of not only assisting lawyers in researching cases, but eventually enabling large-scale analysis of legal judgements to find trends in court outcomes over time.

1 Introduction

The law is reason free from passion[1] — but you'll have to dig through hundreds of pages to find it.

In common law countries such as Australia, a core legal principle is the doctrine of precedent — every court judgement contains legal rulings that are binding upon subsequent cases in lower courts, though how legal rulings apply in subsequent cases is dependent on the facts of the case. When preparing to give a legal opinion or argue a case, lawyers spend many long hours reading lengthy judgements to identify therein the precedents that are salient to the case at hand. This time-consuming manual process has formed a barrier to large-scale analysis of legal judgements. Even though thousands of court judgements are published in Australia every year,[2] lawyers are only able to analyse small numbers of judgements, potentially missing broader trends hidden in the vast numbers of judgements that are published by the courts.

There is a growing body of research at the intersection of Law and Natural Language Processing, including prediction of court opinion about a case (Chalkidis et al., 2019a; Aletras et al., 2016), classification of legal text by legal topics or issues (Soh et al., 2019; Chalkidis et al., 2019b), and legal entity recognition (Cardellino et al., 2017). However, our ultimate goal is to assist lawyers in identifying sections of judgements relevant to their case at hand, as well as bulk analysis of cases to identify relationships between factual patterns and decision outcomes. For this reason, we model our initial study on the sentence-by-sentence identification of argumentation zones within academic and scientific texts (Teufel et al., 2009; Guo et al., 2010). However, these zoning papers do not account for the complex document structure of legal judgements, which have the potential to be structured as multiple sub-documents within the one court decision (see Section 3).

The overall goal of the project is to automate the extraction of information from legal judgements, to assist lawyers to more easily and quickly identify the type of information that they are looking for from a large number of judgements. The project also aims to enable the large-scale analysis of judgements by legal researchers in order to identify trends or patterns that may be occurring within judgments, for example identifying patterns

[*] Meladel Mistica, Geordie Z. Zhang, and Hui Chia contributed equally to this paper.

[1] Aristotle, Politica (Politics By Aristotle), written 350 B.C.E, translated by Benjamin Jowett

[2] For example, the Federal Court of Australia alone publishes around 1700–2500 judgements per year.

of facts that lead to particular results. This kind of analysis is relevant in predicting the outcome of complex cases and may also inform law reform. This part of the study reports on the initial phase of experimenting with the granularity of the annotation labels in developing our schema, as well as our initial experiments in automatically identifying these labels.

2 Background

Legal research as a broad term can include any form of research that is undertaken for the purpose of advancing legal advice, litigation or law reform, and can include research activities such as community surveys, comparative studies of legislation and the study of court judgments.

This project focuses solely on the activity of studying court judgments, as it is a crucial component of legal research in common law countries. For lawyers and legal researchers, court judgments are a key source of data for the purpose of legal research, though legal research in general can encompass other sources of data, such as legislation, international treaties, government reports etc.

When lawyers or legal researchers read a court judgment, what they are looking for is observations, opinions or decisions that the judge has made about how the law should be interpreted and applied in the particular context of the case before it. For example, what are the rules to resolve conflict between competing values, or what are the rules for resolving ambiguities of the meaning of a word in legislation? These observations, opinions and decisions by judges can be conceptualised as "law data" – data that legal researchers collect in order to understand how laws are being applied by courts to specific factual patterns and to predict how it may be applied in future scenarios.

Collecting data about how laws are interpreted is important at both the individual and the societal level. At the individual level, much of a lawyer's work is advising clients on what they need to do to comply with the law. Lawyers will research past court judgments to collect data about how the law has been interpreted in similar factual situations, in order to make an informed opinion about how the law is likely to be applied to the case at hand. At the societal level, legal researchers in academia, regulatory agencies and government collect data on how laws are being interpreted and applied to specific facts, in order to assess whether laws are delivering the desired social outcomes.

The field of legal research has conventionally relied mostly on qualitative data, and if there is quantitative data it is usually at a small scale. The reason for this is because "law data" is expressed in court judgments that are generally very long and complex free-form text. The only method for collecting "law data" has been through the manual reading of legal judgments by people with legal expertise. This is a very time-consuming process and therefore legal research has generally had to rely on small quantities of data.

The contribution that NLP can make to the legal field is to enable the automatic extraction of "law data" from court judgements, to increase the number of court judgments that legal researchers can analyse. The challenge for this project has been the novelty of the task of extracting complex data from court judgments. There is no established schema for extracting information from court judgments. The schema proposed in this study is the result of a multi-disciplinary approach to merging the categories of data that are useful to legal researchers and lawyers, with the categories of information that can be accurately labelled using text classification.

3 Corpus Development

We developed our initial proof-of-concept corpus from court judgements from the High Court of Australia,[3] which is the highest court in the Australian judicial system hierarchy. A court case may be decided by a single judge or a group of judges. In the case of a single judge, the court judgement is single-authored with one voice. When there are multiple judges, they can write a single judgement as a group, particularly if they are in agreement, or they can give separate reasons. In the latter case, the court judgement will then consist of multiple sets of reasons, structured as sub-components from the different judges, which together make up the entire judgement for that court case.

To legal domain experts, there are general patterns or sequences by which different types of information tend to appear within a judgement. However there is a high degree of variation between court judgements according to the writing style of the judge. For instance, one common document pattern begins with the explanation of the facts of the case, followed by the reasoning on how the rele-

[3] https://www.hcourt.gov.au/publications/judgements

Label	Description
Fact	Specific facts of that case, e.g. *The applicant entered Australia as an unauthorised maritime arrival on 5 September 2011.*
Reasoning	Legal principles considered, e.g. *The question that arises is whether the Tribunal failed to consider that the applicant faced a real probability of irreparable harm.*
Conclusion	Outcome of the case, e.g. *The Tribunal committed a jurisdictional error, the appeal should be allowed.*

Figure 1: Description of the Label Set

vant legal principles were applied, and then ending with their conclusion. But this is not always the case. Some judges will state their conclusions at the beginning, and then provide a detailed examination of the facts and legal reasoning. Where there are multiple sets of reasons within a single judgement, each set of reasons will have its own structure particular to that judge's writing style.

We limit our corpus to immigration law cases, and randomly selected 55 of these High Court judgements. These 55 documents contain over 9.5K sentences in total. Each of them was annotated at the sentence level with either Fact, Reasoning or Conclusion, which capture different aspects of the case as shown in Figure 1. In this initial corpus, Reasoning made up half of the labelled sentences. Of the remaining sentences, three quarters were labelled Fact, and one quarter Conclusion. The Fact and Conclusion segments of the case are usually what lawyers are most interested in. These portions of the document (judgement) contain unique details pertaining to the case, while the Reasoning category is a combination of original insights of this case and a recapitulation of previous relevant judgements.

Annotation For the annotations, we had 1 primary annotator (Annotator A), a qualified lawyer and legal researcher, who marked up all of the sampled High Court judgements. Annotator A had a label distribution of Fact: 38%, Reasoning: 50%, and Conclusion: 12%. We also had 2 secondary annotators (Annotators B and C): the first is a practising immigration lawyer, and the second has some legal training, but is not a fully qualified lawyer. We randomly selected 3 documents (judgements) for the secondary annotators to mark up. This made up 5% of the number of sentences of the whole corpus. Of those sentences, there were no three-way disagreements between the annotators. The Cohen's kappa (κ) between all three annotators shows very good 2-way agreement between all pairs of annotators. The inter-annotator agreement between A–B and B–C were 0.70, and between A–C was 0.73. A large majority of the 2-way disagreements involved Reasoning, with 81.5% of the disagreements being Reasoning-vs-Fact and Reasoning-vs-Conclusion, split roughly 50:50.[4]

4 Experiments

In order to assess the feasibility of using our corpus in a supervised setting, we perform experiments using a range of different models for sentence-level classification. The goal is to have a reasonable understanding of how difficult the task is, both in terms of our initial schema and training data size.

Data Processing Although the task is modelled at the sentence level, the corpus was split at the document-level for training, validation, and testing. This set-up emulates the real-world setting, where new documents are classified as a whole. We use a 80%:10%:10% split for training, development and testing (corresponding to 44:5:6 documents and 3000:1200:800 sentences, respectively). Since there is a smaller number of Conclusion sentences in court judgements, we perform undersampling over the training data only, by randomly deleting samples from the other majority classes to balance the number of training instances across the three labels. Note that this was performed for the training set only, and the development and testing sets were left untouched.

Methods As two baselines, we use: (1) a majority-class classifier, based on the training data;

[4] We note that the dataset will not be made publicly available because the project team does not have the right to publish this data. Whilst court judgments are in the public domain, there are copyright restrictions on republication. Republication of court judgments in an altered form, which our labelled dataset would be, is not allowed.

Model	P	R	Macro F1	Micro F1
RoBERTa	.64	.67	.65	.71
BERT	.64	.70	.65	.70
XLNet	.65	.70	.66	.72
MajorityClass	.20	.33	.25	.59
NBSVM	.55	.56	.55	.63

Table 1: Initial Performance Evaluation

and (2) the NBSVM model proposed by Wang and Manning (2012), which combines a naive Bayes model with a support vector machine, using a bag-of-words text representation. We compare this with a set of pre-trained language models, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019). We employ similar structures for these models: 12 layers of transformer blocks, a hidden layer size of 768d, and 12 attention heads. All models are trained by adding a single hidden layer with softmax output.[5]

Initial Results We evaluate our models using Precision, Recall, and Macro-averaged and Micro-averaged F1, showing the results in Table 1. The NBSVM model outperforms the majority class baseline by 0.30 in Macro F1. Using a pre-trained model further improves the performance, with XLNet increasing Macro F1 by 0.11 over the NBSVM baseline, and achieving the best results. While this is expected, since these models have been pre-trained over large amounts of textual data, it is still remarkable given how domain-specific court judgements are.

Incorporating Context While our initial results are promising, at 0.66 Macro F1 they still result in many errors. This undermines the potential of our approach to be deployed in real-world scenarios. In the remaining experiments, we explore a few approaches to improve performance, focusing on XLNet since it was our best model in the initial experiments.

One hypothesis is that the label of a sentence is affected by its context in the document. This is directly reflected in the annotation procedure, since annotators have access to the full document when labelling sentences. In order to test this hypothesis, we prepend each sentence with its two previous sentences in the document, and feed the sequence of three sentences into XLNet as input.

Model	Class	P	R	F1
XLNet	CONCLUSION	.42	.71	.53
	FACT	.72	.62	.67
	REASONING	.81	.77	.79
$\text{XLNet}_{\text{context}}$	CONCLUSION	.58	.80	.67
	FACT	.85	.83	.84
	REASONING	.82	.74	.78

Table 2: Results for XLNet without & with Sentential Context (Prepending the Previous Two Sentences)

Model	Class	P	R	F1
$\text{XLNet}_{\text{context}}$	CONCLUSION	.71	.57	.63
	FACT	.85	.85	.85
	REASONING	.83	.87	.85

Table 3: Results for $\text{XLNet}_{\text{context}}$ with Sentential Context but without Undersampling

We show the results of this approach in Table 2, comparing with the XLNet model used in the initial experiments without sentential context. We also break down the results across the three individual classes, to get a better understanding of any differences in performance. Overall, adding context greatly improves the performance in detecting FACT and CONCLUSION sentences, reaching an overall Macro F1 of 0.76 and Micro F1 of 0.79, a 0.10 and 0.07 improvement over the single sentence model, respectively. Interestingly, adding context does not seem to affect REASONING sentences much, with a small decrease in Recall. This could be evidence that REASONING sentences can be detected only by local content within the sentence, without necessarily requiring extra-sentential context.

Effect of Undersampling We also investigated the impact of undersampling the training data. Our motivation for undersampling is the unbalanced nature of the dataset, where around half of the sentences are labelled as REASONING. This is an issue since, as explained in Section 3, legal experts are mostly interested in FACT and CONCLUSION sentences.

In Table 3 we present the results for $\text{XLNet}_{\text{context}}$ without undersampling, to compare against the

[5] We refer the reader to the original papers from each model for details of the architecture and model pre-training.

original results in (the bottom half of) Table 2 with undersampling. The results show a drop in recall for CONCLUSION, which was expected, while improving the recall for REASONING. FACT, however, was largely unaffected. Note that recall is particularly critical in our use case, in highlighting potential FACT and CONCLUSION sentences to our legal expert.

5 Discussion and Future Work

In this paper, we have presented the preliminary investigations of our interdisciplinary collaboration. The main focus was to scope out the areas in which NLP can assist in the task of interpreting legal judgments — a task that every lawyer must do in researching a case. The main contribution of this paper is developing and testing the annotation schema. In future work, we aim to extract trends over time for a given aspect of the annotation, e.g. how the presentation REASONING changes over time as new cases are judged with each new CONCLUSION. Given that Australia has a common law system, these judgements in effect shape the interpretation and understanding of the law and set a precedence for subsequent cases.

The results of the sentence-level text classification are promising despite the inherent confusability within the REASONING class: even professional lawyers with years of training can disagree in ascertaining whether a sentence is indeed a REASONING rather than a CONCLUSION or in some cases a REASONING or a FACT sentence, as there can be elements of either within a REASONING sentence. Although the results do show promise, in future work, we intend to experiment with the annotation schema to explore more detailed sub-categories under REASONING. This will assist us in identifying more targeted zones within the judgements, which may better assist in legal information extraction tasks, and in better characterising the structure of these legal documents.

From an application perspective, we plan to test the newly released LegalBERT (Chalkidis et al., 2020) and compare this to our adaptation of a domain-specific BERT and XLNet for legal texts. We note that LegalBERT was pre-trained on a variety of legal texts that are different from the legal texts in our database, which consisted solely of Australian court judgments. The data used to pre-train LegalBERT included legislation and contracts, which are different to court judgments in terms of structure and content. Also, the data used to pre-train LegalBERT was from multiple legal jurisdictions, being the United States, United Kingdom and Europe, with each jurisdiction having unique nuances to the language used in its legal texts. Given these differences between our data and the training data of LegalBERT, it remains an open question as to whether LegalBERT would have any advantage over BERT, and whether a custom-tuned BERT for our purposes may be more advantageous.

Acknowledgements

This work was supported in part by the Melbourne Data Analytics Platform. We are grateful for the discussions and suggestions from Priyanka Pillai, Emily Fitzgerald, Daniel Russo-Batterham, Kim Doyle, and Andrew Turpin in the shaping of this project. The authors would also like to thank the reviewers for their constructive feedback.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Legal NERC with ontologies, Wikipedia and curriculum learning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 254–259, Valencia, Spain. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019b. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

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.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 99–107, Uppsala, Sweden. Association for Computational Linguistics.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Jerrold Soh, How Khang Lim, and Ian Ernst Chai. 2019. Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, Jeju Island, Korea. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.