Extracting New Zealand's Case Law Metadata to Link Related Cases

Jelena Matković*, Marko Marković*, Stevan Gostojić*
* University of Novi Sad, Faculty of Technical Sciences, Serbia
matkovic.jelena@uns.ac.rs, markic@uns.ac.rs, gostojic@uns.ac.rs

***Abstract*—Court decisions represent outcomes of court cases and their number constantly increases. These decisions are usually published online allowing lawyers and other interested citizens to be informed on the work of courts. The size of published datasets can bring some difficulties to the retrieval and analysis of court decisions on a particular legal issue. Extracting data from court decisions and identification of cited legislation, regulation, and case law would allow better data organization to more efficiently retrieve, browse and further process the data for statistical analysis. These data could also help in detecting potential inconsistencies in decisions delivered by courts. This paper proposes an approach consisting of various NLP techniques for extracting relevant data from New Zealand's court decisions and using them to construct a system of linked legal documents for better navigation through legal cases.**

I. INTRODUCTION

Publishing court decisions online on a regular basis improves the availability and diversity of case law. Sharing these documents is important for the transparency of the judiciary, especially with the emergence of the open data concept. Reusing the published data can create new value for the judiciary and the public, but it also brings some technical issues that should be resolved.

Machine readability of published documents is a major obstacle for processing judicial data. To improve case law management, data extraction from legal texts should be performed. Sometimes, the structure and layout of these documents differ with some inconsistencies in document formats making data extraction a challenging task.

Retrieval and browsing of case law on the web is usually provided by a search engine or by using a classification scheme. The complexity of supported search criteria is limited by these web portals. Using legal citation data for linking related documents could improve the browsing of case law databases. Furthermore, the extracted data could be used for advanced search of court decisions and could also provide useful data for statistical analysis.

The main goal of this research is to provide a system for improving search, navigation and exploring of court decisions and the related legislation and regulation using citations to determine which legal documents are relevant to particular court decisions.

Court decisions are usually citing case law precedents or legislation and regulation by referencing a particular legal document relevant to the current case. Since appealed decisions of first instance courts are cited in decisions of appellate courts, these references are also important for determining the final outcome of a case.

This paper proposes linking between court decisions on the basis of citations to enable navigation through case law within a particular legal issue. This approach also provides a chronological overview of the outcomes of related court cases.

Our plan is to identify legislative acts relevant to the delivery of a court decision. This information can be useful for the retrieval of case law related to the particular legislation.

There is numerous research proposing various methods for data extraction from legal texts.

In [1] the authors use rules based on logical and positional relations between parts of the text for extracting relevant information. They propose a method for processing inconsistently structured legal opinions in PDF format to extract metadata and provide searching and filtering of case law.

In [2] the authors propose the software library LexNLP for processing natural language in legal documents. The library provides document segmentation and tokenization, recognition of key parts, extraction of structural information and named entities. Furthermore, the library provides text classification for legal and non-legal documents.

In [3] the authors explore how usage of word embeddings and word order can improve text classification. Authors noted that this vector representation of words instead of bag-of-words and TF-IDF approach gives promising results with recurrent and convolutional neural networks. Similar to this approach, our method combines both, pre-trained word vectors and TF-IDF, with neural networks.

The rest of the paper is organized as follows. Section 2 explains the methodology utilized in the paper. Section 3 presents the results. Section 4 gives the conclusion and future work.

II. METHODOLOGY

Our method consists of several phases briefly explained in the following subsections. First, we introduce how the dataset is obtained and how the data is preprocessed. Next, we describe the method for metadata extraction. Then, we explain how related cases are linked together. After that, we describe how legislation relevant for deciding the case can be identified using text classification.

A. Data collection

Court decisions are collected from New Zealand's legal web portal [4] which publishes case law, legislation, legal journals and other legal acts. All documents are available in PDF format, with a length from a few to tens of pages.

Decisions are organized by court hierarchy and year of publishing. Overall, roughly 39000 decisions are collected from three levels of court instances of general jurisdiction i.e. High Court of New Zealand, District Court of New Zealand and Court of Appeal of New Zealand. The decisions are delivered between 2010 and 2021.

The content of these documents is obtained in two ways. The first is the simple extraction of complete text regardless of how it is displayed in the document, and the second method preserves the layout of the text.

The software tools PyPDF2 [5] and PDFMiner [6] are used for extracting the raw text of decisions from PDF documents. It is suitable for text summarization because it does not depend on structural information. The text obtained this way can be used for the extraction of citations explicitly stated in the documents.

However, for some documents the text obtained by these tools is retrieved in an inconsistent order. Therefore, this approach is not reliable for metadata retrieval because the extraction of metadata is based on the position of data in the document layout.

Analyzing court decisions from the dataset we noticed that the first two pages in these documents contain most of the metadata relevant to this research. The textual content of these pages is first converted into high-resolution images and then the text is obtained using optical character recognition (OCR). This way the data extraction is reliable but significantly slower than using the PDF parsers.

A list of legislation titles is obtained from New Zealand's official legislation website [7] and consists of 1808 items. The legislation titles are provided in two ways, in alphabetical order and by the year of publishing. Processing of the HTML documents containing these titles is performed on the basis of the HTML markup elements in these documents.

B. Metadata Extraction

To develop the search engine of court decisions by multiple attributes, the first task is finding and obtaining relevant metadata from legal documents.

Metadata is extracted either by its position in the document layout or by using regular expressions. Because the structure of documents differs for each level of the court hierarchy, three separate sets of patterns are formed for extracting data i.e., one set of patterns for each court type.

While extracting metadata from text obtained by OCR, few errors are detected in character recognition and resolved by extending the regular expressions to include those misreadings. Some of the frequent errors are confusion between the letters 'l', 'I' and number '1' and also between the number zero and the letter 'O'.

Besides legal citations and legislation titles, other information is extracted from the first two pages of documents. For all three levels of court instances, the set of metadata includes case title, file number, date of hearing, date of the judgement delivery, names of judges and names of parties (e.g., prosecutors, plaintiffs, defendants and solicitors). Additionally, the metadata of the District Court decisions contains the court seat, and the metadata of decisions delivered by the Court of Appeal contains a briefly explained outcome of the appeal. Figure 1 shows an example of extracted data from decisions of the District Court.

New Zealand courts have their unique identification marks e.g.: NZDC for the District Court of New Zealand, NZCA for the Court of Appeal of New Zealand and NZHC for the High Court of New Zealand. A single document is identified by this mark and the given case number also called the neutral citation identifier and by the file number. Legal citations can be found in the documents either as the file number or as the neutral citation. These identifiers are extracted using regular expressions.

To extract all cited legislation, every sentence containing the word *act* is retrieved. Then we determine if some of these sentences contain any of the 1808 legislation titles.

Every paragraph in the documents starts with its ordinal number enclosed in square brackets. After a detailed analysis of the documents, we concluded that information relevant for text classification can be found in the first two paragraphs and includes a brief explanation of cases, names of participants and a summary of charges.

Figure 2 shows a sample of extracted data needed for text classification. The column *Acts* contains whole sentences that include the word *act.* The column *Cleaned_acts* shows extracted legislation titles and the column *Summary* represents the summarization of text using tools from the NLTK library [8]. The column *Introduction* represents the first two paragraphs of the court decisions.

	index	Case_name	Case_ID	Case_link	Hearing	Judgment	Plaintiff	Defendant	Judge	Court
0	1	Police v Jackson [2017] NZDC 11268 (26 May 2017)	[2017] NZDC 11268	CRI-2016-070-001435	26 May 2017	26 May 2017	NEW ZEALAND POLICE	ANTHONY KARAURIA JACKSON	T R INGRAM	TAURANGA
1	4	Police v Crown [2017] NZDC 1628 (27 January 2017)	[2017] NZDC 1628	CRI-2017-096-000163	27 January 2017	27 January 2017	NEW ZEALAND POLICE	MELVIN J IMOLE RANGI CROWN	P J BUTLER	HUTT VALLEY
2	8	Southland Regional Council v Woutersen [2015] ...	[2015] NZDC 18079	, CRI-2015-025-000913	8 September 2015	8 September 2015	SOUTHLAND REGIONAL COUNCIL	CORNELUS RUDOLFES MARIA WOUTERSEN|MURRAY WILLI...	B P DWYER	INVERCARGILL
3	16	R v Hohepa [2017] NZDC 12980 (16 June 2017)	[2017] NZDC 12980	CRI-2016-092-014730	16 June 2017	16 June 2017	THE QUEEN|NEW ZEALAND POLICE	TEHURA DAMON HOHEPA	R L B SPEAR	HAMILTON
4	17	R v Carter [2017] NZDC 27112 (29 November 2017)	[2017] NZDC 27112	CRI-2016-019-007714	29 November 2017	29 November 2017	THE QUEEN	PAMELA JOY CARTER	R L B SPEAR	HAMILTON

Figure 1. Metadata extracted from the District Court decisions

	Case_name	Case_ID	Acts	Cleaned_acts	Summary	Introduction
1	Taranaki Regional Council v Fonterra Limited [...	[2015]NZDC14962	Fonterra Limited appears for sentence on one c...	Resource Management Act 1991	The intermittent discharges between March and ...	Fonterra Limited appears for sentence on one ...
2	Otago Regional Council v Cockroft [2015] NZDC ...	[2015]NZDC20608	Mrs Cockroft, you appear for sentence on two c...	Resource Management Act 1991--Sentencing Act 2002	Mrs Cockroft, you appear for sentence on two c...	Mrs Cockroft, you appear for sentence on two ...
3	Invercargill City Council v Perkins [2015] NZD...	[2015]NZDC20845	Mr Perkins, you appear for sentence on one cha...	Resource Management Act 1991--Sentencing Act 2002	Mr Perkins, you appear for sentence on one cha...	Mr Perkins, you appear for sentence on one ch...
4	Thames-Coromandel District Council v Kiwi Bobc...	[2015]NZDC19450	Zane Beckett Construction Ltd and Kiwi Bobcats...	Local Government Act 1974--Resource Management...	Introduction [1] Zane Beckett Construction Ltd...	Zane Beckett Construction Ltd and Kiwi Bobcat...
5	Auckland Council v Motukaha Investments Limite...	[2015]NZDC23105	In addition, the purposes of the Sentencing Ac...	Sentencing Act 2002	Mr Petrou, you appear today as the managing di...	Mr Petrou, you appear today as the managing d...

Figure 2. Extracted data from the District Court decisions for text classification

C. Linking related cases

Court decisions may contain references to other legal documents. Referenced documents can be cited as a source of law (i.e., legislation or a legal precedent) or as a decision under appeal. There is no distinction between the structure of these two types of citations in the documents.

To distinguish between those references, first we retrieve the cited decisions and then compare the names of participants in the citing documents with the names of participants in the cited documents. If matches are found, those documents are linked together as decisions in the subsequent court proceedings. If no matches are found, referenced document is considered to be a legal precedent. Documents found to be part of the subsequent proceedings are organized by the date of hearing indicated in the text and chronologically aligned.

D. Text classification by regulations

Some documents do not contain titles of legislation. Support vector machines (SVM) and neural networks (NN) are used in those cases to determine which legal acts are most relevant for those cases. As already observed, we found that the first two paragraphs in court decisions contain a brief explanation of the case indicating the legal area of the court case. We use paragraphs from documents containing legislation titles as training data for text classification, whereas legislation titles are used as labels.

Initially, the idea of using summarization of document text as training data gave unsatisfactory results. In most cases, summarized text completely disregards parts of the text with an explanation of the factual background and focuses on participants and the outcome of the case.

Text is preprocessed using named entity recognition (NER), part-of-speech (PoS) and other techniques such as stemming and lemmatization. Personal names, dates and numbers are excluded from the text. GloVe dictionary [9] is used for representing words as vectors. The dictionary of relevant words for text classification is constructed using TF-IDF to omit words irrelevant for the classification and to reduce the volume of the text. Words that have a final score less than the empirically established threshold are discarded. All words from the input text that do not exist in the dictionary are ignored. This significantly reduces the size of the input data.

As each document could be related to multiple legislation, we use a sequential neural network and SVM for binary classification with one classification model for each legal act.

III. RESULTS

The method for data extraction, allows us to successfully extract over 87% of metadata from the documents. Obtaining this data enables document search by any of 14 predefined attributes.

Linking the documents on the basis of citations produced roughly 27 thousand groups of interrelated documents, whereas the largest group has eight documents. Extraction of cited case titles and case identifiers in some cases points to the documents that are not available for download at the web portal. As a consequence, some court decisions cannot be linked to the cited documents because these documents are not available in the dataset. Thus, every case identifier found in the documents is verified if the document with that case number is available in the dataset.

More than 400 different legislation titles are extracted from court decisions. Because some of these titles are named only once in the dataset, we focused on 16 legislations, each named in more than 50 documents, and used their titles for training. In the dataset 79% of documents originally contain titles of legislation indicated directly in the text. After document linking, around 86% of documents became part of groups with at least one explicitly named legislation.

Both models, SVMs and NNs achieve an F1-score between 0.91 and 0.97, whereas NNs give better results for the documents with greater volume. To obtain this score, the crucial step is the removal of irrelevant words from the text using TF-IDF while the removal of named entities, dates, and numbers is performed using NER. Without removal of these data, the F1-score drops below 0.6. Once titles of laws cited in a court decision are found, they can be used for retrieval of case law from the same legal area.

IV. CONCLUSION

In this paper, we present the method for extracting data from decisions delivered by the New Zealand courts to enable the search of the case law and to create links between related decisions.

Regular expressions proved to be a useful asset in extracting data from documents with different structures. Consistency in writing style and the use of machine-readable formats for semantic annotations of data in legal documents could improve data extraction.

Although most decisions delivered by New Zealand courts are available online, there are still documents which are not published. This makes a chronological line for some subsequent court proceedings incomplete.

Furthermore, the unavailability of legal documents impairs the potential for statistical analysis.

Usage of pre-trained word embeddings with neural networks gives promising results. However, text classification on the basis of relevant legislation when the legislation title is not clearly indicated in the sufficient number of documents for creating training models remains a challenging task.

Further research could focus on using the extracted metadata and the links between related court decisions to facilitate the establishment of an integrated legal web portal combining case law and legislation.

The proposed method for metadata extraction could be applied to other types of legal documents e.g., complaints, indictments and appeals. This could extend the network of legal acts and improve navigation through judicial documents.

REFERENCES

[1] B. M. Oliviera, R. V. Guimarães, L. Antunes, &. P. P. Rodrigues, "Sifting Through Chaos: Extracting Information from Unstructured Legal Opinions", In *MIE* (pp. 441-445), 2018, January.

[2] M. J. Bommarito II, D. M. Katz & E. M. Detterman, "LexNLP: Natural language processing and information extraction for legal and regulatory texts", In *Research Handbook on Big Data Law*, Edward Elgar Publishing, 2021.

[3] M. J. Berger, Large Scale Multi-label Text Classification with Semantic Word Vectors, *Department of Computer Science*, 2017.

[4] "New Zealand Legal Information Institute", [Online]. Available: http://www.nzlii.org/ [Accessed 10 January 2022].

[5] "PyPDF2", [Online]. Available: https://pypdf2.readthedocs.io/en/latest/ [Accessed 28 May 2022].

[6] "PDFMiner", [Online]. Available: https://github.com/euske/pdfminer [Accessed 28 May 2022].

[7] "New Zealand Legislation", [Online]. Available: https://www.legislation.govt.nz/ [Accessed 22 May 2022].

[8] "NLTK library", [Online]. Available: https://www.nltk.org/ [Accessed 28 May 2022].

[9] R. Socher, C. D. Manning, J. Pennington, "GloVe: Global Vectors for Word Representation", *I Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

.