Automatic Extraction of Legal Citations using Natural Language Processing

Akshita Gheewala[1], Chris Turner[2] and Jean-Rémi de Maistre[1]
[1]*Jus Mundi, Paris, France*
[2]*Surrey Business School, University of Surrey, Guildford, U.K.*

Keywords: Natural Language Processing (NLP), International Law, Legal Citations, Java Annotation Patterns Engine (JAPE).

Abstract: The accessibility of legal documents to the different actors of the judicial system needs to be ensured for the implementation of a strong international rule of law. The gap of such accessibility is being addressed by the Jus Mundi multilingual search-engine for International Law. The data updated on this platform is qualified by skilled lawyers. However, the interconnection of references within such documents, is a key feature for lawyers since, a major part of the legal research is analysing such citations to support their arguments. The process of interconnecting such references can prove to be expensive as well as time-consuming, if completed manually. Hence, the purpose of this research is to automatically extract such legal citations within international law, using Natural Language Processing (NLP), enabling the interconnectivity of documents on Jus Mundi. This study also discusses and addresses research gaps within this subject, especially in the domain specific to International Law. The method followed to achieve the automation is building an adaptable model through Regular-Expression based annotation language named JAPE (Java Annotation Patterns Engine). This set of automatically extracted links are then to be integrated with the search engine, having direct implication in the enablement of smoother navigation, making the law more accessible. This research also contributes to the state of the art bringing closer the eventual use of NLP in applications used to interact with International Law documents.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Computer and cognitive technologies make it possible today to understand the quantity and diversity of legal data constituting International law. Therefore, Jus Mundi, a search engine developed for International Law, relies on this technological progress to make legal data more accessible worldwide in order to enhance the application of international law and thus, strengthen the international rule of law. The purpose of this research is to automate the identification of references within International Law through Natural Language Processing. This research is conducted as a supporting activity for Jus Mundi.

These documents entail International Treaties, Cases, Awards, Decisions, Opinions and various other documents. Case-law plays a critical role in decision-making and legal reasoning pertaining to International Law. Furthermore, when lawyers identify certain precedents that support their arguments in a legal dispute, they make a citation towards that particular precedent. Various other citations towards the law are made for such purposes. Case law citations are referred to as a part of legal analysis as they are used to support legal argumentation. Currently, such links to the citations are being added to the search engine manually through a highly systematic state-of-art methodology. It can be interesting and value additive to support Jus Mundi with contributing to building a solution to this problem through development of tools which support legal professionals in their activities. The approach towards resolving this problem is to build a model for automatic identification of links within legal texts through regular-expression based pattern recognition.

It is absolutely necessary to develop tools for automatic identification of Legal texts, if the Law has to remain available, manageable and realistic (Moens et al., 2000). There is an existing need to make the information in legal texts easily and automatically accessible (Nguyen et al., 2018). We are presently confronted with a large number of legal texts, being constantly recorded in an electronic format.

Gheewala, A., Turner, C. and de Maistre, J.
Automatic Extraction of Legal Citations using Natural Language Processing.
DOI: 10.5220/0008052702020209
In *Proceedings of the 15th International Conference on Web Information Systems and Technologies (WEBIST 2019)*, pages 202-209
ISBN: 978-989-758-386-5
Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

Moreover, there has been a substantial development specifically of international law over the last seventy years. New treaties are signed and ratified every day. The number of international courts and tribunals has also seen a remarkable increase over a similar time period. This evolution is not only quantitative. The scope of international law is also widening. It no longer concerns only nation states but also individuals, companies, associations and administrations. Thus, the number of documents among which judges, lawyers, professors, etc. are led to conduct their research has increased considerably. Legal professionals usually rely on a combination of information from primary, secondary or tertiary sources. It is important to note that under pressing conditions of information overload, the quality of analysis by humans that read, collate and draft such legal documents may have a detrimental effect.

When adhering to such legal texts, analysts need to follow the citations within the texts. Enabling easier navigation requires automated support for the identification of the natural language expressions used in citations. The approach for this study is to build a Regular-expression based model for automatic extraction of links from legal texts. This set of links will be extracted in an appropriate format to enable later integration with a search engine. This task can be obtained through GATE, an open-source, Natural Language Processing based platform (Cunningham, 2017). The overall approach for achieving this task is intended to be simple, concise and at the same time, adequate and efficient.

2 RELEVANT RESEARCH

In industries such as Banking, Insurance, Law, Retail, Wholesale, Transportation and Manufacturing, a large proportion of industry related data exists in textual form (Inzalkar et al., 2015). Hence, text mining forms one of the major fields of data mining for information extraction for both structured and unstructured databases.

Text mining in the legal industry has been explored by many practitioners right from the very first attempts by Frank Shephard in 1873, to more recent entities such as LexisNexis (2019) and Westlaw (2019). Studying analysis specific to citations, one of the key contributors (Garfield, 1955) is often cited towards such analysis in science, who argued that Shepard's methodologies can also be applied to scientific citations. Data from citation indexes created by Garfield were also used for extracting, aggregating and analysing quantitative aspects of bibliographic information (Moed, 2005). It can be said that the lawyers have placed the foundations for citation analysis (Ogden, 1993), however such research in the domain of scientific reports has been developing much faster than in the field of common law, despite the increasing size of the legal database. A continuously increasing amount of information available, along with widely improved means of storing such data, may have led to information overload. The constantly evolving legal knowledge base including the growing amount of available citations is no exception to this (Elliott et al., 2013). Organisations such LexisNexis and Westlaw have developed numerous resources that have organized such information in a structured format providing a streamlined research experience. This includes their electronic citators like Shepard's Citations or KeyCite respectively. However, an interesting study reveals that despite providing a similar service, there can be found notable and unexplained inconsistencies between the results produced by both these applications (Mart, 2013). Moreover, it is pointed out that it would be a challenge for a researcher to gain full benefit of such a huge amount of unfiltered data. Publishers are still relying upon specifically trained humans to provide case analysis (Geist, 2009). It can be concluded that compilation of such citators would yet rely upon human research of the case law, irrespective of the progress made in automated data management and retrieval. A manual annotation approach using Regular Expressions, supervised by qualified lawyers is hence apt to attain a balanced solution addressing such gaps between technology and the Law.

There also exist studies which take up approaches not related to regular expression for automated reference resolution in Legal texts. Tran et al. (2014) demonstrate a Four-step framework to deal with the task: mention detection, contextual information extraction, antecedent candidate extraction, and antecedent determination. A significant component of Legal references is the cross-reference type of citation (reference to a paragraph or a section within the same document). There are many different approaches towards such link resolution. However, it is argued that the majority of the aforementioned works fail to sufficiently address the subtleties that arise during the resolution task (Sannier et al., 2017). Existing research does not clearly distinguish between simple and complex cross referencing. A more flexible framework is developed by Sannier et al. (2017) for automated detection as well as a resolution of cross references as an attempt to address the aforementioned research gap. The framework is

based on structuring the database into text schemas followed by the resolution task executed through GATE. Building on GATE workbench Sannier et al. (2017) implement their approach into a tool named LeCA (Legal Cross Reference Analyzer), primarily built in JAPE. It consists of 13k lines of code within 114 JAPE scripts, with additionally supporting 5k lines of JAVA code. This tool supports transforming text into mark-up format, detection and resolution of cross references as well as visualization and analysis. Such an intricately built tool can however only be applied towards cross-reference type of citations.

Deep learning approaches for text mining of legal documents have gained in interest over recent years. Moreno and Redondo (2016) highlight this interest in deep learning and point to its potential for use with legal documents. In applied work Chalkidis and Androutsopoulos (2017) compare deep learning with sliding windows, for the purpose of contract element extraction, and conclude that in relation to token classification such a method provides an improvement in performance with fewer errors over the latter approach. Ashley and Walker (2013) have also investigated the use of deep learning in the mining of legal decisions relating to a defined set of medical cases, utilising the DeepQA architecture from IBM Watson.

Recent developments in legal data analytics have resulted in the creation of unprecedented tools for identifying statistical, semantic as well as citation-based patterns in large corpuses of legal repositories. Such dramatic growth in data-centric approaches have paved the way for new algorithmic approaches to legal analysis and problem solving (Conrad et al., 2018). With the increasing interest in recent times in the adoption of machine learning, an approach utilising such technology to resolve the automation task is a valid target for further research.

3 METHODOLOGY

The focus of this research is to achieve automatic identification of links within legal texts, using Natural Language Processing. For such a purpose, it is necessary to attain a precise knowledge of the structure of International Law as well as the type of references within each document. International Law practitioners not only refer to Treaties and Conventions between countries, but also Cases, Decisions and Awards related to the Treaties and Conventions. Consequently, the links within each of the above documents are mainly external links with legal instruments as their target. International Law mainly consists of Investment Arbitration Law, comprising of Treaties, Conventions, Cases, Decisions, Awards and Opinions.

These documents are structured with a systematic state-of-the art methodology within a UML class text schema. Each document is hierarchically structured into HTML mark-up defining Levels (Titles and metadata), Articles (Number or Title of the Article) and Elements (Paragraphs and Sub-paragraphs). Decisions, Awards and Opinions are documents released by the tribunals as an outcome for investment arbitration cases. If “MINE v. Guinea” is the short title for a case name, and “Decision of 22 December 2016” is the Decision by the tribunal, then “ICSID Case of Maritime International Nominees Establishment v. Republic of Guinea, Decision of 22 December 2016” is the complete text that will be used to make a reference to this case in any other Awards, Opinions or Decisions. The above can also be mentioned along with its case number such as, “ICSID Case No. ARB/ 87 /3”. At many points within these legal texts, references are made to Treaties or other related Cases to support legal arguments. These links can be found to be present at any part of the document including footnotes. For the current task, links from each part of the document are to be automatically extracted. Consequently, the whole of the document including the footnotes are to be processed through the model for complete extraction.

The nature of links, unlike cross-references, is said to be external, in view of references made to Articles not within the same document but a different Award or Decision of a Case or another Treaty altogether. As a result, the nature of a simple link is, for example, “Article 24(a) of the BIT”. Here, ‘the BIT’ is the name of the Treaty (Level) and ‘24(a)’ is the destination of the Article -> Element. Links to Cases increase the complexity for identification due to the unique names of the parties involved. A combination of Regular Expression rules will be written to correctly identify such links. Taking into consideration the magnitude of the database, it is practical to conduct a pilot study of the end-to-end process to test the feasibility. It will additionally provide an estimate of the timeline required to build a model that should be applicable to a variety of structures of documents and links. The sample model, built to extract a basic link to an Article of a Treaty will be tested on two documents. One of these is a Decision wherein the links are already manually identified by a qualified Lawyer, in order to compare the output results with correct links. The second document is a Multilateral Treaty, which is selected for the pilot study due to the large number of articles,

hence the length of a document, providing a diverse set of links for tests. The approach for automatic identification of links within legal texts is demonstrated in (Fig 1).

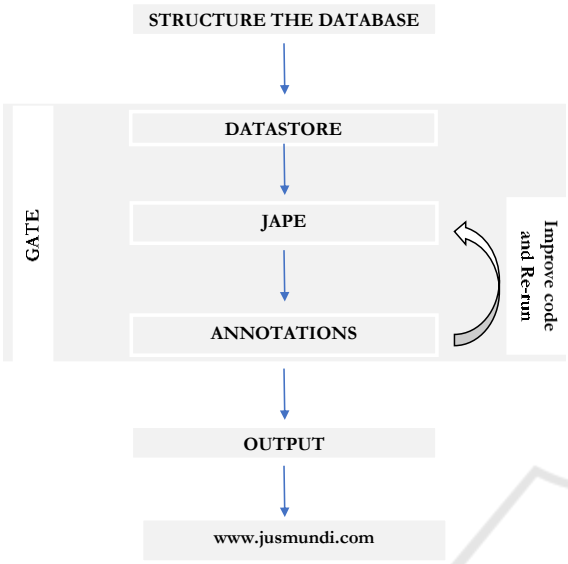

Figure1: Overall Methodology.

The approach consists of three major steps. Beginning with structuring the database for the datastore to be loaded into GATE, followed by NLP tasks performed within GATE ending with integration of the output with the search engine. The second step can be further divided into pre-processing, running the model on the datastore and improving and re-running the model until accuracy is achieved. Accuracy in such a task can be in accordance with the maximum number of links that the model is able to correctly identify, as well as the ability of the model to identify an appropriate start position and the end position of the link. This can be attained through JAPE rules, which will be illustrated in Section 4 of this paper. Before JAPE can be applied to the data, pre-processing tasks will be carried out on structured data using English Tokeniser, Gazetteer, Sentence Splitter, POS Tagger and Ortho Matcher. Each step is executed in the respective order as above. The tokenizer splits every single text into tokens with an individual token kind suchlike, word, number or punctuation. Each token kind is further allotted a token-orthography (Allcaps, Uppercase or Lowercase) in case of words, or allotted a punctuation-kind in case of punctuation, respective to the characteristics of the particular text. The Gazetteer enables users to lookup predefined categories of text including geographical data such as name of countries, or a date including months and years. Case names, Awards and Decisions within International Law, consist of the name of at least one country and the date on which it was implemented. The Gazetteer is hence particularly applicable to identify Case names, Awards and Decision types of links. The execution of Sentence Splitter will result in identifying parts of the data as sentences within the text. These tasks are a pre-requisite to write Regular Expression grammar utilizing the unique annotation types for customizing the identification of more complex legal links.

4 ANALYSIS OF THE CITATION EXTRACTION APPROACH

It is important to begin with an understanding of the structure of Legal texts pertaining to International Law, as well as how they may be tailored into an appropriate format to achieve the extraction of links. The analysis begins with structuring the data using Postgres SQL, followed by integrating the schema with GATE platform, through the Datastore. Understanding each type of reference made between Legal documents is a definite pre-requisite to enable following tasks. Furthermore, the analysis in GATE consists of pre-processing the datastore by running numerous NLP algorithms on the same. This step is performed as a pre-requisite for the following main step of the analysis, i.e. formulating JAPE rules that are able to identify all the links mentioned in each document in the datastore. In this vein, the construction of codes building these rules is explained in detail. Various challenges faced while performing the analysis are also brought forward. Considering the nature of the analysis on an unsupervised database, the procedure of measuring the accuracy requires manual supervision by skilled lawyers. However, the manual tasks are only required for a one-time examination of the output. This task is followed by obtaining the output as a set of annotations in the required format for further integration with the search engine. The overall analysis mainly focuses on the use of JAPE on the Legal texts for automatic extraction of each of the mentioned references within the data as an annotation that can be exported to the search engine; this will then be treated as a link to the document that the reference is made to. The legal data is collected from public institutions, international organizations and ministries of foreign affairs. The following institutions have already agreed to share their data with Jus Mundi: The International Court of Justice, the International Centre for Settlement of

Investment Disputes, the Permanent Court of Arbitration and the Energy Charter Secretariat. Secondly, this data is structured. Mainly, it consists of extracting the text of the document from its original format (often a PDF), checking it, and entrusting it to several algorithms which will recognize its structure (articles, paragraphs, table of contents, footnotes and more). This also enables the next steps of qualifying the data. Qualifying the data means identifying legal concepts held within such a given set. Such steps result from artificial intelligence but also from the participation of lawyers. This is followed by enabling interconnection, which leads to this research study. The interconnection of references is a key-feature for lawyers: in one click, they should be able to access all documents relating to an article of a treaty or a paragraph of a judgment/award.

Such qualified data is obtained from the schema of the search engine through a Postgres SQL query. As a result of this query execution, the data can be exported as one file for one document. Hence, each Treaty, Decision, Award or Opinion will be described by an individual file. The data is originally obtained in an HTML format. For the purpose of this analysis, the data is loaded as HTML text in a ".txt" format. Due to this, the text also includes HTML tags, which will need to be included in the analysis and JAPE rules. There exists an alternative of dealing with such data in ".html" format, which is also a format supported by GATE. However, in such a case the data would appear in a text format without any HTML tags. The start position and end position numbers of an identified link is counted as per the number of letters in the document. This position will differ in both formats, due to the HTML tags present in the text format which are also included in the count of position number. It is vital to include the tags in the count for detecting the position to support the later integration of the output with the back-office of the search engine, wherein the data is stored in the original HTML text format. It can hence be justified to operate with the HTML data in a text format for further analysis.

The database consists of 3778 Treaties, 41 Acts, 1277 Cases, 1570 Opinions, 1937 Decisions as well as 6091 other documents with a total of 14,694 documents. Moreover, each of the above documents exist in more than just English, with a separate version for each of the other languages. As an example, a Treaty between the United Kingdom and France will exist in English as well as French. In order to build a model to be applied to such a magnitude of data, it is practical to conduct a pilot study, as stated in the methodology section of this paper. The results of the short-analysis within a stipulated timeline assured the feasibility of conducting further analysis. The documents selected for the pilot study along with a part of the datastore with maximum diversity, are treated as a Gold Standard documents, in order to train the model. The training dataset is curated under the guidance of qualified lawyers, to obtain maximum applicability of the JAPE rules on the entire dataset.

The basic pre-processing tasks for information extraction are performed as follows. Running the Tokenizer on Legal texts will help us differentiate between Names beginning with an upper Initial or abbreviations with 'AllCaps' types of word token. Similarly, the Gazetteer will help us identify dates of Awards and Decisions as well as the name of countries as mentioned in the Case names. This is followed by a sentence splitter which will complete the pre-processing of the database.

The above steps allot labels to each text unit in the data. Hence, the output after the above modules as executed, is an annotated document. In order to extract a combination of such annotations, we instruct the software to match a certain pattern of particular annotations. This in GATE, is achieved through JAPE Transducer. JAPE Transducer runs a Java Annotation Patterns Engine (JAPE) File, which is a regular expression rule-based language. As a result of executing this file on the data, GATE will create a set of annotations named "Sample", which will consist of every text region starting with Article, followed by a number. A JAPE rule is a set of rules which are executed sequentially, matching regular expression patterns over already created annotations by pre-processing as well as strings. In the above example Article is matched via a string which will only identify the word "Article". At the same time, the following number is matched through an annotation of 'kind = number'. This will identify any number followed by the string Article and not just a particular number. A JAPE rule consists of an LHS (Left Hand Side) which is all the rules applied for identification, mentioned before the "-->". Everything after the "-->" is the RHS (Right Hand Side) part of the JAPE rule, which is typically related to the output of the matched text. The name for the set of annotations, as well as the pattern in the output annotation can be customized on the RHS part of the rule. A set of rules on the LHS is referred to on the RHS by an allotted label on both sides. A combination of such rules for matching annotation and strings are carefully written in a particular sequence, to identify a complete text of a legal reference. Additionally, JAPE also provides the classic regular expression operators like Repetitions: (+,?,*), Alternative: "|", Negation: "!" or

a range "[]" and more. Moreover, JAPE also provides functions to deal with overlapping annotations in case an identified text contains a complete additional set of annotations within itself. This will be explored later on in this chapter. For more advanced or specific operations, the RHS of the JAPE rule can be written in Java, since the entire rule is ultimately translated into Java. An additional feature that can be applied to matching patterns are the control options such as brill or appelt style. The appelt control option is used for this analysis due to the nature of the output required, with the complete set of identified text as one annotation. For example, given a reference such as "Article 42(b) of the BIT", various matches would be identified such as "Article 42", "Article 42(b), "Article 42(b) of the BIT". However, the reference is identified correctly only if there is one complete identified output that "Article 42(b) of the BIT". A JAPE file is loaded on the JAPE Transducer to run on the datastore. GATE provides a JAPE Transducer by default. However, considering the nature of this analysis affecting the length of the code, JAPE Plus Transducer is certainly a preferred executor. JAPE Plus Transducer is an updated version of JAPE Transducer, available as a plug-in. This not only saves upon the execution time but also enables the user to run the code on a large corpus at once. The length of the code increases proportionately with the number of words to be identified as a set. This can be explained through the following details. The methodology followed here describes a method intended to achieve simple, and most common matches first, followed by building upon the existing code to match with more complex and unique types of links. The code in the above example would identify any text beginning with the word "Article" which is followed by a number. However, the same reference is also discovered in different forms like "Art." Or "Articles". Such instances can be deal with using the Alternative "|" function of Regular Expressions. There also exists the possibility of "Article" being followed by more than just a simple number. For drafting rules to match all such patterns, a study was conducted to understand each possible pattern of different type of citations. A brief summary of variety of such patterns are observed as follows:

"Article III(b) of the Treaty and Article 4.a of the HCL"
"paragraph 5, sub-para 3 of the Hydrocarbon's Law"
"Tecmed S.A. v. The United Mexican States, (ICSID Case No. ARB (AF)/00/2), Partial Award of 13 November 2000"

Each of the above can be mentioned in more than one form. For example, dates, case numbers, paragraph numbers, may or may not be mentioned in brackets. Also, while building a rule for complex cases, it is important to keep the original identifications in place, as an example, if the rule identifies the whole of the above-mentioned link towards the treaty, it should also still be able to identify the instance with just "Article III(b)". In such a case, for everything following the number is treated as "?", which matches the patterns in case of zero or one occurrence. Similar approach is built for all the optional occurrences such as "of the", "and" or brackets "(","(". Different types of references will have a different pattern. In total, 10 JAPE files are created for 10 different types of references. Rules for each individual type of reference are based on different foundations from each other. For example, the Case-citations are identified around the occurrence of "v" followed by a ".". In order to make the execution of such complex codes more organized, macros are created within JAPE which are then simply referred to within the LHS of the code. The RHS deals with the output of the identified text. The default output provides the start and end position of the words. However, the RHS is further improved to also include the text and the length of the identified citations. This additionally simplifies the review process for the Lawyers, which will be discussed further in the following section of the paper.

5 VALIDATION OF THE APPROACH

It is established that due to the nature of the task on an unsupervised Database, the method of evaluation of the model involves manual validation. It is important to remember that although such a review is a manual activity, it is a one-time investment of resources to evaluate a small set of data, to achieve the automation for all links on a significantly larger dataset. Once the procedure is completed, the model can be applied even to all the additional data that would be added in the future. A question to ask at this point: is the model effective in its ability to automatically extract citations? For the purpose of evaluating the model following the training, 15 diverse set of International Treaties, Cases, Acts and other documents with a total of 3206 pages were selected as test data. Noting the scale-free nature of citations, 25% of these pages were selected at random. As a result, the model was run on a total of

800 pages. The total number of links within the test data as validated by supervisors are 1317. Table 1 shows the summary of the results after evaluation. The model fails to identify 7 links, and incorrectly identifies 2 links with partial identification possible for 10 links. As a total, the number of links correctly identified are 1298, giving 98.5% of correct identifications. Construct validity is measured in terms of precision and recall which for the following results is 99.23% and 98.5% respectively. Construct validity is measured in terms of precision and recall which for the following results is 99.23% and 98.5% respectively. It can be noticed that the maximum errors have occurred with Act and Case type links. As discussed earlier, the above types of links are based on names of Rules or Parties respectively, hence they may not follow a specific pattern. However, it can be said that the written rules cover all the commonly occurring patterns as well as most of the complex and unique patterns. A case may be cited in multiple ways, which is ultimately drafted under human supervision. Hence, irrespective of the defined framework for citations, slight discrepancies may always be expected. It can be concluded that resolving such a task with a rule-based approach, the completeness of the rules will always be undefined. As an alternate approach to overcome such issue, rules can be based on strings containing the list of names of all documents. However, such a rule will neither identify anything beyond the list nor the alias of any names existing in the list. Hence, current analysis is preferred in order to widen the application of the model. Another challenge faced in such analysis is the incorrect identification of links due to the issue of Break-line wherein the identified link is extended even to the following sentence in cases where the input patterns are matched. This problem would occur if the data is directly loaded as a PDF. However, this problem can be avoided by loading the data in an HTML format, where the split is created as a break-line after each sentence. This can also point out the importance of structuring the data in a mark-up format before processing the same. This also addresses a challenge faced during similar research of text deviating from its originally intended schema (Sannier et al., 2017).

6 CONCLUSIONS

Making International Law more accessible would not only enhance its utilization but also strengthen its effectiveness. The underlying purpose of this research is to create an automatic process for extracting links within Legal Texts, further enabling such documents to be interconnected within an International Law Search Engine, Jus Mundi. Building a search engine involves a vast range of NLP tasks including Text Mining. With a rapid advancement in technology enabling increased accessibility and navigation, it can be interesting and value adding to pursue such research. It is established that although research is present within this domain, the task of developing such a model specifically towards all of International Law is explored for the first time in this research.

The outcome of this study is a model that automatically identifies various types of citations between Legal documents. This enables easier navigation, supporting the activities of lawyers to improve the quality of legal analysis on the relation between different laws. This task is achieved using Natural Language Processing techniques within

Table 1: Results of Citation Extraction Approach.

SR.NO	Type of Citation	Correctly Identified (True Positives)	Partially Identified (False Positive)	Missed (True Negative)	Incorrectly Identified (False Negative)
1	Article	208	0	0	0
2	Treaty	219	1	0	0
3	Convention	129	0	0	0
4	Act	194	3	5	1
5	Case	389	4	7	0
6	Clause	91	1	0	0
7	Paragraph	68	1	0	1
TOTAL		**1298**	**10**	**7**	**2**

GATE workbench. This involves loading the data extracted using Postgres SQL database, loaded as a Datastore within GATE, followed by some pre-processing steps. As a result of running the model on test data, 98.5% of the citations are correctly identified with a precision of 99.23%. The evaluation suggests that this can certainly be a very concise and less time-consuming approach to resolve such a challenging problem. The results contribute to similar existing studies by successfully resolving more than just a set of cross references. However, there exists a possibility of improving the output of such a model through using JAVA on the RHS of the JAPE code. This can provide further customizations also with regard to detection of context within citations with unknown contexts. This step will be implemented in the future work in order to identify the targets for such citations. This study will be extended to numerous other language sets and in addition forthcoming research with evaluate deep learning approaches in comparison with the NLP application presented in this paper. It is also the case that domains such as social media feeds or scientific research paper databases present further opportunities for legal domain specific analysis with the NLP approach.

REFERENCES

Ashley, K.D. and Walker, V.R., 2013, June. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law (pp. 176-180). ACM.

Bhardwaj, R.K. and Madhusudan, M., 2016. Online Legal Information System (OLIS) Leveraging Access to Legal Information Resources in Indian Environment. *DESIDOC Journal of Library & Information Technology*, *36*(1).

Chalkidis, I. and Androutsopoulos, I., 2017, December. A Deep Learning Approach to Contract Element Extraction. In JURIX (pp. 155-164).

Conrad, J.G. and Branting, L.K., 2018. Introduction to the special issue on legal text analytics. *Artificial Intelligence and Law*, *26*(2), pp.99-102.

Cunningham, H., Maynard, D., Bontcheva, et al. 2017, Developing language processing components with GATE version 8. *University of Sheffield, Dept. of Computer Science*.

de Maat, E., Winkels, R. and van Engers, T., 2009. Making sense of legal texts. Formal Linguistics and Law, 212, p.225.

Elliott, C. and Quinn, F., 2013. *English legal system*. Pearson.

Garfield, E., 1955. Citation indexes for science. *Science*, *122*, pp.108-111.

Geist, A., 2009. Using citation analysis techniques for computer-assisted legal research in continental jurisdictions.

Inzalkar, S. and Sharma, J., 2015. A survey on text mining-techniques and application. *International Journal of Research In Science & Engineering*, *24*, pp.1-14.

Katayama, T., 2007. Legal engineering-an engineering approach to laws in e-society age. In *Proc. of the 1st Intl. Workshop on JURISIN*, *2007*.

Katayama, T., Shimazu, A., Tojo, S., Futatsugi, K. and Ochimizu, K., 2008. e-Society and legal engineering. *Journal of the Japanese Society for Artificial Intelligence*, *23*(4), pp.529-536.

LexisNexis (2019) Lexis ® Library, [online] https://www.lexisnexis.com/uk/legal/ accessed 08/04/19.

Mart, S.N., 2013. The case for curation: the relevance of digest and citator results in westlaw and lexis. *Legal Reference Services Quarterly*, *32*(1-2), pp.13-53.

Moed, H.F., 2005. Citation analysis of scientific journals and journal impact measures. *Current Science*, pp.1990-1996.

Moens, M.F., Uyttendaele, C. and Dumortier, J., 2000. Intelligent information extraction from legal texts. *Information & Communications Technology Law*, *9*(1), pp.17-26.

Moreno, A. and Redondo, T., 2016. Text analytics: the convergence of big data and artificial intelligence. IJIMAI, 3(6), pp.57-64.

Nguyen, T.S., Nguyen, L.M., Tojo, S., Satoh, K. and Shimazu, A., 2018. Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. *Artificial Intelligence and Law*, pp.1-31.

Ogden, P., 1993. Mastering the lawless science of our law: a story of legal citation indexes. *Law Libr. J.*, *85*, p.1.

Sannier, N., Adedjouma, M., Sabetzadeh, M. and Briand, L., 2017. An automated framework for detection and resolution of cross references in legal texts. *Requirements Engineering*, *22*(2), pp.215-237.

Tran, O.T., Ngo, B.X., Le Nguyen, M. and Shimazu, A., 2014. Automated reference resolution in legal texts. *Artificial intelligence and law*, *22*(1), pp.29-60.

Westlaw (2019) Westlaw UK: The Leading Legal Research Service [online] https://legalsolutions.thomsonreuters.co.uk/en/products-services/westlaw-uk.html accessed 08/04/19

Wyner, A., Mochales-Palau, R., Moens, M.F. and Milward, D., 2010. Approaches to text mining arguments from legal cases. In *Semantic processing of legal texts* (pp. 60-79). Springer, Berlin, Heidelberg.