Interpreting Legal Texts Using Natural Language Processing

Nithin Mathew Joseph
College of Engineering and Physical Sciences
Aston university
Birmingham, UK
210199723@aston.ac.uk

***Abstract*— This article first introduces the basic structure of NLP, that is the NLP pipeline. Followed by various methods currently used for data mining from unstructured texts. The paper discusses three research studies and one suggestive idea. It also deep dives into the challenges posted for NLP in general. Suggestive solutions for them as well. The ethical implications posted and finally a conclusion on how the NLP techniques can contribute to the society.**

Keywords—Expertise, legal, affairs, provisions, case, law, NLP, domain, techniques, token, baseline, training, entity, relevance

I. INTRODUCTION

Legal texts are long and complex. Without certain expertise in that domain, it is difficult to interpret or even decipher the legal lingo. Natural language processing (NLP) is the main ingredient that can be used to simplify or forecast law. NLP is the process of converting unstructured text to standard format that can be interpreted using computational systems [1]. NLP pipeline consist of various techniques to extract key information at different stages, this will be discussed in detail in this paper.

There are three main reasons for the adoption of NLP for legal texts:

1. The quantity of legal documents that get digitized and stored in repositories are rapidly growing.
2. NLP programming has advanced over the decade
3. Thereby increasing the chances of improving the legal services [1].

Legal texts are rich with latent features. This paper discusses how they can be mined, which can then be used to create beneficial applications in the legal domain:

1. Development of search engines to better educate the public to understand legal context
2. Design a system which learns wider social trends and stigmas concerning law
3. Assisting reporters to intercept law-changing cases [2].

II. NLP BASICS

To deep dive into NLP for legal text we must understand how NLP works. NLP pipeline is divided into five segments, that are as follows:

1. Language
2. Punctuation
3. Morphology
4. Syntax
5. Semantics

A. Language Identification

NLP detects the language used from the input text provided. How this is possible? it is because statistical models trained on character sequences helps recognize the language stored in it. If we are to take an example to showcase this, ‘th’ is a character sequence that has high relative frequency in English whereas ‘sch’ is usually observed in high frequency in German language. Hence these character sequences based on the frequency of occurrence helps identify the language they are originating from.

B. Puctuation

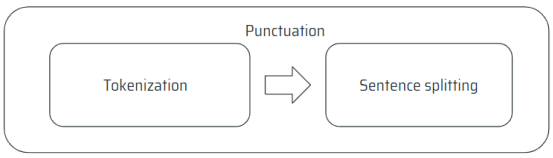

Fig. 1. Punctuation block diagram

It is divided into two stages. Tokenization and sentence splitting.

Tokenization: The text provided is split into individual words. This is obtained by splitting on the basis of white spaces.

Sentence splitting: The text that gets tokenized is split into sentences based on punctuations.

These two stages are valuable cause latter helps to evaluate the sentence construction whereas the primary helps to identify entities.

C. Morphology

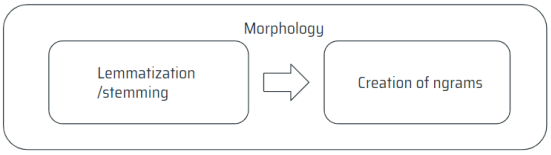

Fig. 2. Morphology block diagram

It is divided into Stemming or lemmatization and creation of ngrams.

Lemmatization or stemming: Both the process is used to obtain the normalized form of a word. For example, removing plurals or breaking the word into its simplest form like pupils to pupil or running to run. Lemmatization involves the usage of dictionary to map the words to their simpler form. The con is that its slow whereas the pro is that its 100 percent accurate. Stemming involves using rules like regular expressions used in python. Its fast but prone to inaccuracy.

Creation of ngrams:

Ngrams are sequences of tokens with length n. For example:

Unigrams (n=1): The, bunny, ate, the, carrot.

Bigrams (n=2): The bunny, bunny ate, ate the, the carrot.

Trigrams (n=3): The bunny ate, bunny ate the, ate the carrot

D. Syntax

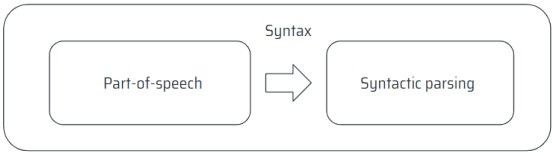

Fig. 3. Syntax block diagram

It is divided into Part-of-speech and Syntactic parsing.

Part-of-speech (POS): Characterizing each token as a part of speech specifically noun, verb, adjective, pronoun, etc. The tags that are used for these are more refined that those used in school grammar. The features surrounding the tokens and their POS tags can be extracted using a machine learning model, which will be the essence of NLP.

Syntactic parsing: Mapping each token to its syntactic role specifically to subject, verb, and object. It also establishes links dependent token. The output of this stage will be syntactic role labels and dependency links.

E. Semantics

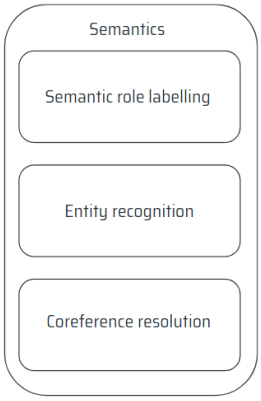

Fig. 4. Semantics block diagram

It is divided into three stages namely semantic role labelling, entity recognition and coreference resolution.

Semantic role labelling: Segmenting each token to its semantic roles namely agent, patient, instrument, location etc.

Entity recognition: Categorize each noun as person, organization, or location. The machine learning model used here is trained on a specific context that enables it to extract key aspects related to that topic.

Coreference resolution: It's the capability to recognize different text in a running text refers to the same entity. For example: Boris Johnson is the prime minister. The prime minister held an after party in London during lockdown. Here both Boris Johnson and the prime minister refers to the same entity.

III. TECCHNIQUES IN NLP

Understanding the basics won't suffice for the requirement of extracting information from legal texts. Hence introducing various techniques that provide efficient results when mining for information in any given texts.

A. Textual summarising techniques

It is divided in two types namely extractive and abstractive summarization.

Extractive summarization: The algorithm identifies important sections of the provided input text. Then it combines the relevant sections to generate a new text. This technique in general, is considered easy to compute.

Abstractive summarization: This technique is quite complex when it comes to computation. It creates entirely new sentence from the summary extracted from the input text.

When compared against one another in the case of a single paragraph or two. It is observed that abstractive summarization outclasses extractive summarization [1].

B. Text rank

This method is a graphical approach where in the significance of the sentence is plotted as a graph to determine the ranking of the sentence. The key factor which helps distinguish each sentence ranking is its similarity of each sentence against one another [3].

C. Google's PageRank

The most popular graph ranking method is Google's PageRank. The concept revolves around the fact where a sentence recommends another. The similarity of the recommended sentence is also checked against the sentence that recommends it. In a high-level perspective, the recommendation count is what determines the ranking of the sentence [1].

D. Topic model

An algorithm that can generate high-standard abstract of long and complex texts. On the contrary it is designed to summarize only one document at a time. It is a combination of three aspects, which are the topic, document meta data and the topic prevalence [4].

E. Attribute extraction

"This technique is used to obtain pre-specified attributes from a text". An example for this can be currency value specified in a judicial statement [5].

F. Relational extraction

"This technique is used to extract relationship within the attributes obtained from the text" [1].

G. Retrieval techniques

It is the algorithm used to retrieve documents or information's related to a query provided by the user. It is of two types, which is as follows:

Boolean retrieval: Unranked retrieval with no significance to the question posted by the user.

ML retrieval: Ranked retrieval with significance to the question posted by the user.

The retrieval performance is evaluated on two primary measures:

1. Precision: how important is the retrieved document to the user's need.
2. Recall: the quantity of significant documents that were retrieved [6].

H. Term frequency-inverse document frequency

It is a summation of two processes namely term frequency and inverse document frequency.

Term frequency: "How frequently the term is observed in a document".

Inverse document frequency: "Logarithm of total number of documents by the number of documents that contain the term in question" [6].

IV. WORKING AND RELATED WORKS

The methods mentioned above can be used for effective mining of texts. Combining the techniques mentioned in a structured and logical sense is one solution for extracting the context in legal texts. Research have gone into on how effectively this can be achieved. One of the studies mentions that there are three main steps in the case the model designed is like question and answer (Q&A) format. The steps are as follows query representation, document representation and similarity check to identify the significance between the query and the document.

A. CNN model for legal text

In this study the researchers utilized convolutional neural network (CNN) and attention mechanism to extract relevant information and match similarity present in the Q&A. They use the CNN to create sentence and paragraph encoders that extract key features present in legal texts. When the researchers trained the CNN model, negative sampling paradigm was used. Wherein the mechanism labels the article that is related to the query as positive whereas the article that has no significance to the query as negative.

Once this model was built, its performance was evaluated against the models present in the current decade. It performed exceptionally well against Bert's counterpart for legal domain called Birch. It still couldn't beat a version of Birch that was trained on title of articles. The researchers also did an ablation study to understand the significance of each stage they built. They were able understand that each stage of their model contributed to the overall performance [7]. This approach seems to have been effective as per the result the study has shown.

B. RNN model for legal text

It still isn't the only likely solution; we could take a recurrent neural network (RNN) as an NLP model that can be utilized to extract key information from law context. The idea is to create a neural network that can predict strings from all the previous ones. This process of prediction is repeated several times. So, if we come back to basics and remove the concept of Q&A, we can build a system that can be trained on legal context to predict information that contains legal value. This can be another way of building an NLP model powerful in the legal domain.

One way of utilizing this model is by building a named entity recognizer (NER) that can specifically decipher legal articles to their respective entities. Entity identification will help breakdown a legal text jargon into simpler format that can be interpreted by a computational system. The application for this is countless.

The model can be built into an app by further research. Where it takes in input from user as a query and provide a detailed output based on the article of interest. Here the RNN discussed is gated recurrent unit (GRU) as it's the most efficient model that provides best performance when it comes to NLP till date. Whereas Long short-term memory (LSTM) is also an area that can be investigated in RNN for legal texts, but it is not as effective as GRU as per studies.

C. Blackstone model

Another research study which investigated on NLP for legal text is "ICLR&D which is the research division within the Incorporated Council of Law Reporting for England and Wales ("ICLR")". They designed a model called blackstone that has multiple methods designed in it for extracting key aspects in the legal texts. The blackstone language model was educated using the word2vec algorithm. One of the methods is named entity recognition. It's very interesting as to how the researchers trained the blackstone model for entity recognition. They made it memorize the requirement that they were interested in by providing lots of examples of those scenarios. That is, they provided a sentence which contained the entity as an input and the format of the output that is expected out of the model. For this 70 percent of such data was taken up for training the model whereas 30 percent for validation. The present status of the NER model is not perfect, but it gets the job done as shown in table 1 [2]:

TABLE I.

Entity	Name	Examples
CASENAME	Case names	e.g. Smith v Jones, In re Jones, In Jones' case
CITATION	Citations (unique identifiers for reported and unreported cases)	e.g. (2002) 2 Cr App R 123
INSTRUMENT	Written legal instruments	e.g. Theft Act 1968, European Convention on Human Rights, CPR
PROVISION	Unit within a written legal instrument	e.g. section 1, art 2(3)
COURT	Court or tribunal	e.g. Court of Appeal, Upper Tribunal
JUDGE	References to judges	e.g. Eady J, Lord Bingham of Cornhill

[2]

The module blackstone still contains other features that can help extract key aspects from legal text. The blackstone module requires installation before it can be used and certain requirements the system needs to adhere, so that it can be installed. This module is a prototype and some features of it are still lacking. It acts as a steppingstone for those interested in NLP for the legal domain. The module blackstone code execution is simple and refined, which made it a study worth mentioning [2].

D. Graph reasoning model

This research involved predicting legal provision for legal artificial intelligence (LegalAI). LegalAI is basically artificial intelligence techniques or technologies used in the legal domain [9]. The goal of this model is to predict the nearest legal provision in a given text of interest. This model can play a significant role in the legal domain as it can help avoid redundancies and extract only the required information. In this method they regard affairs and provisions as entities. They have a well-defined schema that clearly separates affairs from their provision and provide the legal linkage between them in the form of a knowledge graph. Hence legal provision prediction (LPP) is a link prediction algorithm in the knowledge graph. For this model to work it requires text understanding and legal reasoning. In text understanding the model needs to identify the significant sentence constructions and filter out the unwanted details. When it comes to legal reasoning, legal text usually follows a set of rules. Hence the model must be smooth in traversing through these rules and identify the basis of the legal context. This helps the model spit out the legal provisions effortlessly. To satisfy the requirements of text understanding and legal reasoning the model uses BERT to represent entities in low dimension vectors followed by utilizing graph neural network to obtain inference to perform legal reasoning. Their research has shown that they were able to obtain better performance when they compared it against the baseline [10].

V. CHALLENGES

The challenge in this domain is that there is no standard method to evaluate the performance of NLP on legal text. The first discussed study CNN for legal text utilized recall and normalized discounted cumulative gain (NDCG) to measure the performance of their model against the current models. Recall is true positive divided by total of true positive with false negative. NDCG measures the relevance or gain of a document based on the position in the result list. The gain is obtained from the peak of the result list to the bottom, with the gain of each output discounted at lower ranks [8]. The figure below shows the comparison study the first research obtained against the current models in the industry:

TABLE II.

SYSTEM	RECALL@20 [TP / (TP +FN)]	NDCG@20 [NORMALIZED DISCOUNTED CUMULATIVE GAIN]
ELASTIC SEARCH (BM25)	0.357	0.334
ELASTIC SEARCH (TF-IDF)	0.478	0.351
BIRCH (256 WORDS)	0.763	0.542
BIRCH (TITLE)	0.783	0.591
SYSTEM 1 (1000,30,30) [WITH JUST SENTENCE ENCODER]	0.798	0.641
SYSTEM 1 (1000,50,50) [WITH JUST SENTENCE ENCODER]	0.811	0.669
SYSTEM 2 (1000,30,30) [WITH SENTENCE & PARAGRAPH ENCODER]	0.801	0.665
SYSTEM 2 (1000,50,50) [WITH SENTENCE & PARAGRAPH ENCODER]	0.825	0.688

[7]

The second one regarding RNN, is a model that is not built yet and hence there is no evaluation about its performance studied. On the contrary the study on legal text by ICLR&D is evaluated using precision. Precision is the fraction of predicted positives that are actual positives. The problem with these measures is that they are built for classifying problems. Since a text can have different context the ranking on which the text is evaluated on cannot be considered even optimal. It clearly cannot provide a basis of understanding on the performance the model is able to achieve.

Let's consider an example where we take the case of a query that has close relation to two articles but since one article contains similar wordings as that of the query whereas the latter contains legal context that has more similarity. In the case the latter is discarded from the solution. Then it means the evaluation of the model needs to indicate a lower performance, if it does not then it means the evaluation method is not proper. So clearly the evaluation method that needs to be implemented need to judge the model performance based on the context of the text that is in question. For the graph reasoning model, they used Mean Rank (MR) and Mean Reciprocal Rank (MRR). Where MR is the average of the ranks for all observations within each sample and MRR is the score for the reciprocal rank for the first relevant item [10].

The other challenge that is more common in NLP tools is the availability of train data. As all languages do not have datasets that are readily available. Based on the language of interest the train dataset may not be sufficient. So, training a model with insufficient dataset can result in the model misinterpreting the information and producing low performance. This will be evident when evaluating the same model for a different language that has better train dataset.

The first discussed study CNN for legal text, "the researchers utilized two built datasets. Legal document corpus, which contains Vietnamese legal documents and Q&A dataset that contains quite a bit of legal queries and a set of significant articles for each query". "The raw legal documents were extracted from official online sites and the queries were obtained from legal advice websites". "The queries contained title and content". The researcher identified that the content was long and puzzling, so they decided to keep good titles, modified uninformative titles, and removed some sections of the content. Multiple versions of the law regulations were present in the legal corpus which were removed to filter out the redundancy. This was done with the help of experts in the legal domain [7].

As mentioned, their data set was not of good quality due to which they had to modify and filter out irrelevant matters to obtain better performance. These sorts of preprocessing are always required when it comes to text mining.

The study on legal text by ICLR&D train dataset consisted of two which are namely "ICLR's archive of law reports, dating back to 1865 and ICLR's archive of unreported judgments, dating back to 2000". The primary focus was in preprocessing the train dataset. "The requirement was to obtain each sentence as a segment for processing which involved removing leading, trailing, and excess white spaces from the train dataset" [2]. This study also required a stage were in they had to modify the dataset meaning there will not be a case where we can obtain train dataset directly for use without any preprocessing involved. Extracting information is only possible if we can provide dataset in a standard format as input which the model can take up to execute the specific task that results in the output that is desired.

The study on knowledge graph collected data from Guangdong government service website. The researchers performed preprocessing and data analysis to filter out the below mentioned

issues. Non-standard texts, similar affairs, and no legal provisions. For non-standard texts the legal texts usually contain abbreviations, missing specific formats which makes it difficult to map the affairs against the provisions. So, the alternate way around that they have implemented is a dictionary mapping of provisions to standardize the non-standard texts. Similar affairs were found in the dataset which makes it difficult to map all the affairs against provisions. In the preprocessing stage they merged the similar affairs together to obtain the unique ones. No legal provisions issues are due to outdated provision. This makes affairs impossible to link towards. It mainly occurs when legal provisions change over a period meaning to say the older provision get removed off as a result. There are also cases where the legal provisions do not have the required format. In such cases also the legal provisions are not linked to the affairs. In the preprocessing stage these legal provisions are removed out [10].

VI. SOLUTIONS

A suggestive idea would be to analyze the weightage of strings in the legal text. Use a baseline as the benchmark on the type of analysis that provides best accuracy when it comes to legal domain. Then tune that measurement to precisely interpret at how much capacity the resulting NLP can generate a result in relation to the context of the legal text. Then the measurements like recall, NDCG, precision, MR and MRR can be discarded when it comes to evaluating the performance of an NLP model.

Regarding unavailability of dataset, the three-research study has shown preprocessing the obtainable data can be valuable to obtain an optimal performance on the model of interest. Another option would be to let technological advancement take over. Meaning as more and more data gets digitized the chances for building a robust and durable NLP model increases.

A simpler option would be to get hold of individuals who have expertise in the field to obtain intel on as to where the dataset for the model that is in need of implementation can get its input from. Still, it will have loophole as the legal context gets updated each year. For example, the law that is considered important right now might not have any significance in the coming years. This can damage the interpreter functionality of the model as the dataset will be corrupt and will require update with the new laws in place. Therefore, perfect solution for unavailability of dataset can never be attained if there is change in society.

Another workaround might be to train more with the dataset that is available. The con would be overfitting, so proper care needs to be taken to find the fine margin of the dataset for training and testing. To solve the issue of overfitting in the case of using more data for training would be to use a learning curve plot to identify the point at which the plot leads to overfitting. Then use it as inference to identify the optimal dataset that will provide best performance for the model of interest with the available data.

VII. ETHICAL IMPLICATIONS

Ethical issues start with the dataset that is been collected. If the data that has been retrieved for training a model is of confidential information and has not gone through the right channels. It is unethical to use such data. The researchers in charge of the models also need to be aware of what data they can consider and those they cannot. It's a very fine line and every data point in question need to be only considered after careful thought. The case where it can be ignored is when the benefit of the cause outweighs the damage that will be done. This does not mean that every chance the benefit of the cause outweighs the damage to the subject it will not be deemed accessible. It still must undergo several checks.

Let's recognize what private data protection means: "one in which the publicity of court trials as a rule of law guarantee trumps the individual's wish to hide oneself from others in public space" [11]. This law is old but still according to this concept any data which does not have the consent of the subject cannot be used for any purpose and it is considered confidential. There is a contradicting statement to that law where convicts or terrorist that undergo trial have the right to privacy [12]. So,

what's the contradictory implication of this example, well for one the public must be made aware of such criminal trials so that they can keep a lookout and secondly the consequences of the action of the convict or terrorist need to be brought to light for the public to be made aware to reduce the chances of bring up such individuals in the society. As per the statement in [11] the implications made above will not come to fruition.

So, when it comes to ethical implication in the case of legal NLP research, the research is scrutinized cause of the high standard for privacy protection [13]. This isn't to say that the data that is available in the public can't be considered accessible even if it contains some but not all confidential data. In such cases the dataset can be accessed as there is no one subject in play there. "The primary moral duty of legal NLP researchers, like all researchers, is to the disinterested pursuit of truth as they understand it, and not to substantive ends which are extrinsic to that pursuit" [13].

VIII. CONCLUSION

This paper refreshes the basics of NLP at first. Then it addresses various techniques in the industry that can be used to mine text datasets. It discusses three research study and one suggestive idea on how to implement a model with RNN for legal text interpretation. The research studies are discussed briefly to give an idea of the mechanism models utilize to interpret the legal texts.

The paper also sheds light on the challenges faced by NLP models when it comes to evaluation and input data set. There are no perfect evaluation criteria that is set in place for legal text yet, it is our duty as researchers to keep searching for measurement concept that best explains the performance of the model for legal text. For the challenge of unavailability of input dataset, the only thing that can be done is give time and preprocess the data that is currently available to meet the requirements to its best.

The ethical implication of NLP for legal text is complex and it threads on a fine line. So, it would be best to find justification for every action that is carried out for the model. Request for permissions that require access through the right channels so that any subject in jeopardy can be contacted.

Interpretation of legal texts using NLP can be utilized to build applications that can help illiterate people battle legal wars with literate adversaries. Thereby providing a means to fight tooth and nail for justice with a click of a button. This doesn't mean we try abolishing the individual's expert in the domain. It is done so that we can help support those individuals to work smart and reduce the time taken for justice to act. The research studies discussed can also be used as benchmark for the coming-of-age inventions in this domain.

IX. ACKNOWLEDGEMENT

I had like to thank my module tutor Abinaya Sowriraghavan. She has been helpful in making me come to terms with the different AI techniques that are currently present in the industry. Her classes were exciting to be in apart from some of the guest lectures. She also helped me with my dissertation topic as well as in this article as she nudged me in the right directions whenever I went off track. Some of the journals she recommended opened my eyes to the effort and dedication the researchers made to make their research a success.

X. REFERENCES

[1] Nay, J. J., 2021, "Natural Language Processing for Legal Texts," in Katz, D. M., Dolin, R., and Bommarito, M. J. (eds) Legal Informatics. Cambridge: Cambridge University Press, pp. 99–113. doi: 10.1017/9781316529683.011.

[2] ICLR&D, 2022, Open source NLP and machine learning for legal texts. What is Blackstone and how did we build it? — ICLR&D. Available at: https://research.iclr.co.uk/blog/blackstone-goes-live (Accessed 12 April 2022)

[3] Rada M. and Paul T., 2004, "TextRank: Bringing Order into Texts," in Proceedings of the Conference on

Empirical Methods in Natural Language Processing (EMNLP), pp. 404–411

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, 2003, "Latent Dirichlet Allocation," Journal of Machine Learning Research 3, pp. 993–1022.

[5] Stuart J. Russell and Peter N., 2009, Artificial Intelligence: A Modern Approach, 3rd ed

[6] Christopher D. Manning, Prabhakar R., and Hinrich S., 2008, Introduction to Information Retrieval, supra note 23

[7] Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen and Tu Minh Phuong, 2020, Answering Legal Questions Neural Attentive Text Representation, Proceedings of the 28th International Conference on Computational Linguistics, pp. 988-998.

[8] Kalervo J. and Jaana K., 2002, Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), pp. 422–446

[9] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M., 2020, How does NLP benefit
legal system: A summary of legal artificial intelligence. In: Jurafsky, D., Chai, J.,
Schluter, N., Tetreault, J.R. (eds.) Proceedings of ACL. pp. 5218–5230

[10] Li, L., Bi, Z., Ye, H., Deng, S., Chen, H. and Tou, H., 2021, November, Text-guided Legal Knowledge Graph Reasoning. In China Conference on Knowledge Graph and Semantic Computing, pp. 27-39

[11] Ian Langford, 2009, Fair trial: The history of an idea. Journal of human rights, 8(1), pp. 37–52.

[12] Judith Resnik, 2011, Bring Back Bentham: Open courts, terror trials, and public sphere(s). Law & Ethics of Human Rights, 5(1), pp. 4–69.

[13] Tsarapatsanis, D. and Aletras, N., 2021, On the ethical limits of natural language processing on legal text. arXiv preprint arXiv:2105.02751.