Question and answer reasoning on legal text, why and how it can be enhanced

Dissertation proposal

Contents

Introduction: 2

Objective: 2

Academic benefits: 2

Technical Benefits: 2

Business Benefits: 2

Social Benefit: 3

Research Question: 3

Research methodology: 4

Requirements and Feasibility: 5

Project plan: 5

Ethical implications: 6

Conclusion: 6

Reference: 7

Introduction:

Legal texts are complicated and heavy on legal literature. It is difficult to decipher it without any expertise in the domain. That's where natural language processing (NLP) comes in. It helps to convert unstructured data into a structured form that can be deciphered by the computational system. Nowadays the usage of NLP is heavily relied on for legal text interpretation. The reasons mainly being:

- The legal documents that are getting digitized are a plethora
- The standard of NLP programming is advancing at a higher pace
- Thereby improve the services that can be provided for the legal domain (John J. Nay, 2021)

The dissertation on which this paper emphasizes on depends on various NLP techniques and neural network construction that can aid in answering queries posted by a user in the legal domain.

Objective:

Question and answer graphical user interface (GUI), wherein the user inputs a single query regarding their concern on legal context. The model that is built internally identifies the answer to the query. The model built is mostly based on convolutional or recurrent neural network.

Academic benefits:

There seems to have been research already conducted on this topic and an article in the name of "answering legal questions by learning neural attentive text representation" by Phi Manh Kien et al. They have benchmarked their model against specific models present in the domain. What the academic aim would be, is to understand how close to it can we achieve the model and is there any possibility of improving way beyond the performance showcased by them. Once a proper framework is in place, an evaluation will be conducted by performing an ablation study. That is remove specific stages in the framework to understand the significance of each stage on the result.

Technical Benefits:

The project will introduce a method for users to punch in their query to a GUI wherein it will internally run the model based on the dataset collected to provide feedback containing the answer to the query. Since it is a GUI, it will be easier for the user to understand how to operate it. The project will also involve features that can help identify the jurisdiction of law.

Business Benefits:

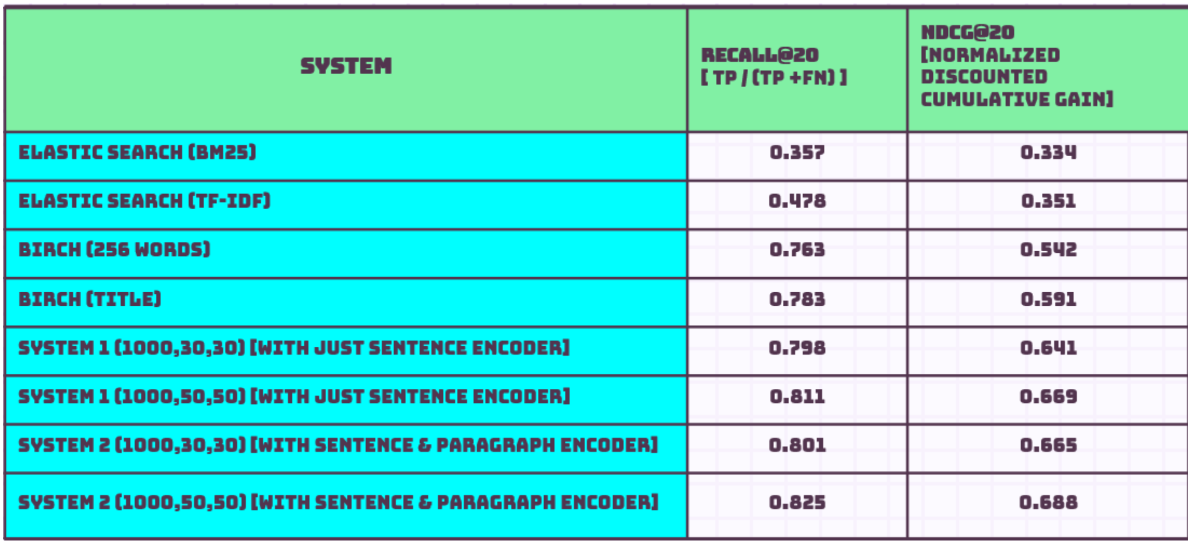
The project can be converted into a product that can be marketed in general as an educational tool for the illiterate to use. It can also be introduced as an application to big corporations that are battling legal issues as an initial go to tool to handle standard queries instead of approaching law firms which would be a waste of time and money.

Social Benefit:

Legal text interpretation using this model can help individuals battling legal wars with lawyers who are well versed in the domain. Hence providing a means to fight equally without any disparity due to lack of knowledge. This can also aid lawyers to work efficiently and reduce the time taken for them to reach a conclusion for a case. The model that is proposed can also be used as a steppingstone for future applications or models that will be researched on further in this domain.

Research Question:

Based on the recent models in this domain the model needs to be evaluated with respect to its performance. The article mentioned before used recall and normalized discounted cumulative gain (NDCG) for measuring their performance (Phi Manh Kien et al, 2020). This project can also use the same evaluation criteria to measure its performance in order to obtain a comparative analysis between the models in the article. The below table showcases the performance rating measured for different models in the mentioned article:



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

Table 1 (Phi Manh Kien et al, 2020)

In the case the model built in this project does not perform well as expected. There will be a need to deploy an ablation study wherein the model needs to be investigated in each stage to identify the core features that adds to its performance and its decline.

Another problem that needs to be addressed is the evaluation method that is deployed for measuring the legal context. It’s a fine section for research study as to what makes a model built for the legal domain determine its performance. Currently the methods employed for measurement of performance of a model is diverse, but none are standard. In the legal context the evaluation criteria vary. A standard way would be to check with individual experts in the domain against the result’s obtained. Whether they are acceptable and in context to the legal query put forward.

To give a better view of this scenario let’s take up an example. Take the case of a query that has close relation to two articles. One article contains similar wordings as that of the query whereas the latter contains legal context that has more similarity. In the case the model chooses the latter it means the model can differentiate in context to legal terms but in the case the model dose not choose the latter. It means the model is not able to interpret in terms of legal context. This can only be best identified in the presence of experts in the legal domain.

Research methodology:

The article which is in close relation to this topic, involved building of sentence and paragraph encoder followed by negative sampling paradigm to obtain the required result (Phi Manh Kien et al, 2020). In this project the plan is to create a named entity recognizer (NER) using convolutional neural network. The NER needs to identify the legal features in the statements provided. It will be deployed for both the query and the articles. For the NER the plan is to look at research studies like the one conducted by ICLR&D (the research division within the Incorporated Council of Law Reporting for England and Wales). They have implemented a module called blackstone. The modules primary framework is based on word2vec algorithm. For the NER model the researchers made it memorize the necessities by providing lots of examples. The researchers provided statements containing the entities as input. They also provided the format of the output that was expected (ICLR&D, 2022). The plan here is to use this background study when setting up the NER model.

NER will be the initial setup followed by a model that uses the output generated from the NER to identify the article which matches closely to the query. For this the foresight is to utilise concepts on which techniques like Google’s page rank, Topic model and retrieval techniques are based on.

Google’s page rank:

The basic concept is a sentence recommending another sentence. There is a similarity check between the sentence that was recommended and the initial sentence to identify in the case the sentences are a copy or not. So, the rank of the sentence is based on how many recommendations are owned by that sentence.

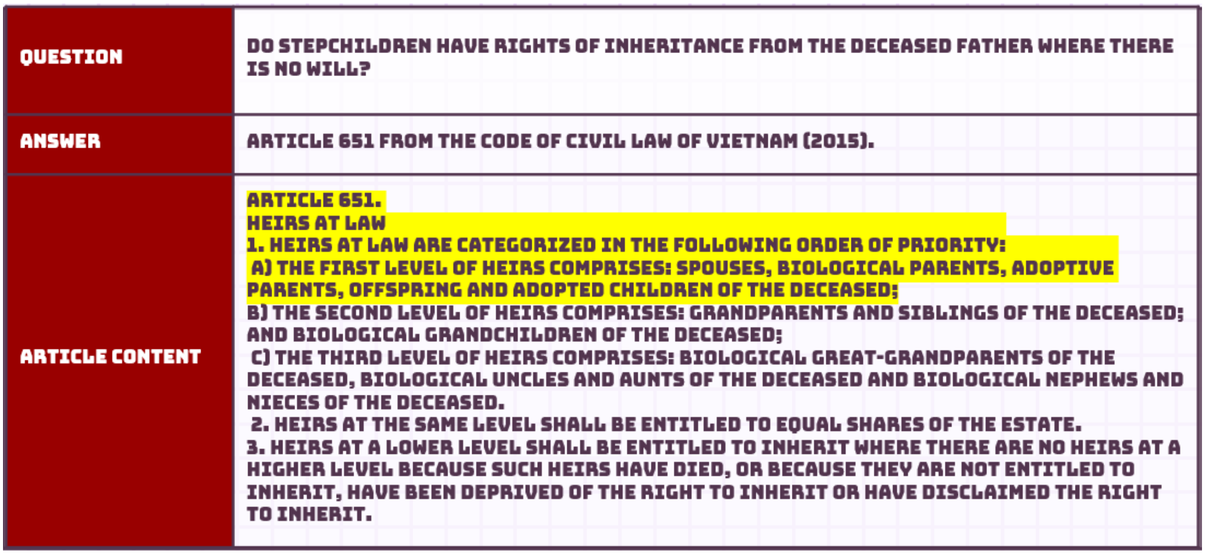
Topic model:

A flow that can produce high level abstract of complicated and heavy legal literature texts. It divided into three aspects, the subject, content meta data and the topic prevalence.

Retrieval technique:

The methodology used to retrieve collective details related to a query provided by the user (John J. Nay, 2021).

The below figure will provide a glimpse of what the representation of the model will be like:



QUESTION	DO STEPCHILDREN HAVE RIGHTS OF INHERITANCE FROM THE DECEASED FATHER WHERE THERE IS NO WILL?
ANSWER	ARTICLE 651 FROM THE CODE OF CIVIL LAW OF VIETNAM (2015).
ARTICLE CONTENT	ARTICLE 651. HEIRS AT LAW 1. HEIRS AT LAW ARE CATEGORIZED IN THE FOLLOWING ORDER OF PRIORITY: A) THE FIRST LEVEL OF HEIRS COMPRISES: SPOUSES, BIOLOGICAL PARENTS, ADOPTIVE PARENTS, OFFSPRING AND ADOPTED CHILDREN OF THE DECEASED; B) THE SECOND LEVEL OF HEIRS COMPRISES: GRANDPARENTS AND SIBLINGS OF THE DECEASED; AND BIOLOGICAL GRANDCHILDREN OF THE DECEASED; C) THE THIRD LEVEL OF HEIRS COMPRISES: BIOLOGICAL GREAT-GRANDPARENTS OF THE DECEASED, BIOLOGICAL UNCLES AND AUNTS OF THE DECEASED AND BIOLOGICAL NEPHEWS AND NIECES OF THE DECEASED. 2. HEIRS AT THE SAME LEVEL SHALL BE ENTITLED TO EQUAL SHARES OF THE ESTATE. 3. HEIRS AT A LOWER LEVEL SHALL BE ENTITLED TO INHERIT WHERE THERE ARE NO HEIRS AT A HIGHER LEVEL BECAUSE SUCH HEIRS HAVE DIED, OR BECAUSE THEY ARE NOT ENTITLED TO INHERIT, HAVE BEEN DEPRIVED OF THE RIGHT TO INHERIT OR HAVE DISCLAIMED THE RIGHT TO INHERIT.

Figure 1 (Phi Manh Kien et al, 2020)

There is also a plan to provide visual representation for the result obtained from the NER model. So that the user can understand the justification as to why a particular article was selected. In this project the insight is to build a GUI as well for the user. So that the user can input the query of interest and expect an answer. There is also a plan to add additional features that will involve further expansion of the model. For instance, as mentioned before identifying the jurisdiction of law or identifying lawyers nearby for further consultation.

The nit and grit of the code will be in the process of matching the context in the query against the articles that were previously provided during the training stage. This model either will be build using convolutional or recurrent neural network.

Requirements and Feasibility:

The language of the input information is still under consideration. The data collection for educating the model will be based on the availability of legal corpus in that specific language. Will need to access legal documents that are digitized as input for training the model. These need to be vetted for confidentiality before utilizing them.

One of the research studies discussed implemented a similar build. Which gives confidence on the feasibility of this project. If a strict timeline on building the different stages of the model is followed. The proposed project can be accomplished within the deadline provided.

Project plan:

Timeline for each task is as follows:

May: Background research on different NLP techniques and neural networks deployed recently for the legal domain. Start putting an initial framework for the models explained in the research methodology by identifying key features that adds to its foundation and documenting these details for the succeeding tasks.

June: Need to decide on the language on which the project will be built on. Collection of data that is publicly accessible for the language of interest. Run exploratory data analysis to identify the sections of the data that require pre-processing. Implementing the framework documented in the preceding task mentioned for May using train data. Then measuring the accuracy of the model by simulating it with the train data provided.

July: Identify the corner cases and bugs that were observed during the simulation with the train data. Debug the corner cases and the bugs to obtain the solutions to them. Once the solutions are obtained. Use it to modify the framework to better educate the model that is being build.

August: Run the test data on the recently build model. Measure the performance scored by the model. Try to improve the performance by using various techniques present in the field. Compare the score observed against the research model that were released recently and estimate how well the new model performs. Document the entire process and provide a report for the project.

Tasks	May	June	July	August
Background Research	🟩			
Framework initial structure	🟩			
Data Collection		🟩		
Data Analysis		🟩		
Implementing the framework		🟩		
Simulation of framework		🟩		
Identify the issues			🟩	
Debug the issues			🟩	
Modify the framework			🟩	
Run test data				🟩
Evaluate the performance				🟩
Correlation analysis against recent models				🟩
Report				🟩

Table 2

Ethical implications:

The data that is collected for training the model needs to go through the right channels before it is accessed for educating the model. Meaning if the data is not accessible publicly it is under confidentiality. So, in such cases the proper procedure would be to identify the parties in charge of the data and request access before utilising it. Cases where we can neglect requesting for access is when the benefit of the cause outweighs the penalty induced. That is also a big “if” based on the scope of the research if it is global then it might have a chance of skipping the requirement for access. That is not the case for this project so strict checking need to be adhered before picking the data for building the model.

Conclusion:

A user interface is provided for the user to interact with. The project tries to build a model which answers queries posted by the user. The answers will be in the format of legal articles that contain the answers to the query. The key academic accomplishment will be in building a robust model that doesn’t break often. Research on numerous NLP techniques will be key in identifying the efficient method for retrieval of article of interest for the query posted. The process of evaluation of the model will also be an interest for research study as it’s a field that has no standard set for measurement yet.

Reference:

[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] 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.