Legal Document Summarization using Neural Networks

Shrinivas P. Patil

California State University, Fullerton

CPSC 589 – Seminar

Prof. Kenneth Kung

8 October 2023

Table of contents

Abstract

In the realm of legal activities, the challenge of deciphering extensive legal documents has led to frustration for legal professionals, researchers, and even the general public. Legal document summarization, a facet of natural language processing, offers a solution by condensing these documents into concise summaries, saving time and resources. This survey paper explores the evolving landscape of text summarization, delving into NLP techniques, methodologies, and challenges. Legal text summarization, driven by article numbers and regulations, demands specialized attention due to variations in document interpretation based on origin and the pivotal role of citations. With unique patterns in criminal and civil judgments, the summarization process can adapt through both single and multiple-document approaches, with extractive and abstractive methods leading the way.

Keywords: Natural Language Processing, Neural Networks, Transformers

Introduction

Legal professionals, researchers, law students and normal people go through an intensive research work in legal activities. Most of the times, they feel helpless when they try to study legal documents and end up with frustration. Some of them hire other people to understand the documents and spend time and money both. Legal domain requires accuracy and clarity to understand the case and evident. Legal document summarization is a part of natural language processing which overcomes this problem and provides the summary of large documents in short with less time. Legal text summarization is technique of collecting extensive legal texts such as case studies, contracts, legal opinions into a concise summary. This summary consists of legal

facts, arguments, judgements help lawyers to understand the case clauses and assist them to decide the next steps in particular direction.

This survey paper explores the text summarization techniques which is rapidly evolving field of natural language processing. It delves in the NLP techniques, methodologies, metrics that shape the scope of text summarization. This paper also seeks the challenges and key issues that should be in the consideration to succeed these technologies for future use from technical, business, and ethical perspective.

Legal text summarization generates summaries that mainly contains the article no., regulations. The importance of document is determined by their origin because the same document interpreted differently if it occurs in higher court that the opinion of lower court. Citations in a document decided the next direction of research (4). These are some reasons; legal text summarization needs special attention to make its application successful. For legal text summarizer, every case is different. The criminal judgements differ in their patterns from civil ones. Most of the times, civil cases are different from each other, and criminal cases tends to be similar in nature. Thus, single document and multiple document approach can be applied (2). Summarization can be done in multiple ways but extractive and abstractive are well known.

Background

Most of the papers focused on legal text summarization using traditional approaches like statistics method, graph-based method, and classical machine learning model. The deep learning era and availability of huge amount of data with computational facility provides an edge in NLP field. In current time, there have been few efforts are taken by researchers using deep learning architectures. Though there is a big challenge to generate coherent and concise summary. The architecture of neural network changes as per the different types of summarization methods.

There are basically four categories of text summarizer a) based on the output form b) based on the input form c) based on the context d) based on the purpose.

Extractive and abstractive are the output-based text summarizers. Extractive summarization extracts important sentences and phrases from the document and obtains summary and the words used in summary is subset of original text. Extractive summaries use term frequency and inverse document frequency but there are chances that it will drop critical information (6). On the other hand, abstractive summarization creates a new summary by itself, and it is similar to human generated one. Also, it is complicated as it gets large amount of data and complex algorithms to compress the sentences. Abstractive summaries are new trending topic among all researchers. These techniques implement rule-based methods by identifying important segments. Tree based methods and ontological methods are used in abstractive summaries (1). Considering the number of documents is given to summarizer, a summarizer can be single document or multiple documents. Since, summarizers available in market used multiple resources to generate a text, so researchers have started working multiple legal document summarization. This multi-document summarizer can cause coherence and redundancy issue (5). Context-based summarization is divided into two parts generic and user-oriented. The generic summary represents the general view of writers towards topic. In contrast, user-oriented considers information relevant to user query and user search. To execute this summarizer, NLU techniques are required.

The purpose-based summarizer generates summary based on entities and important objects in document. An indicative summary represents only topics mentioned in the topic. An informative summary makes shorthand version of document's content (6). Critical summary is used for critical evaluation of the document, and it compares your work with other literatures on

the same topic. The graph-based approach is used for legal text summarization where sentences are represented as nodes of an asymmetric weighted graph. The sentences are represented in node and the node which has higher value will be selected for summary. This approach helps to achieve diversity and ensures cohesive flow in the document (1). Some researchers have adopted a different approach in which authors divided segmentation task into two steps – segmentation of document using rhetorical role identification using conditional random field and generate a summary from identifies segments. The method is based on identifying semantically similar text segments and exploits the structure of whole legal document. The one drawback about this approach is it fully rely on labelled data for segmentation and annotation (3).

Challenges and issues

The first thing is legal texts are different from newspaper articles or any scientific texts. There are some key factors like size of document, structure, vocabulary, ambiguity, and citations that plays important role in implementation decision. The summaries can be divided into generic, or query based (4). The structure of legal documents depends on the country of origin of a case and the heavy use of citations make text summarization more challenging. Here are the two structures of legal documents. Figure (1) shows the legal documents from United states and from India. The legal document from Unites States is hierarchical in structure where the document from India is sequential. The two documents are very different in structure that creates a serious difficulty in developing generalized legal text summarizer (5).

The vocabulary of legal documents is different than the vocabulary used in regular life. Legal documents follow their own domain specific legal-based terminology besides standard language. So, it is very difficult to understand the meaning of keywords and their semantics

before giving it to model. Another issue is ambiguity issue. The legal documents contain phrases and clauses that may have different meaning as per the different legal organizations.

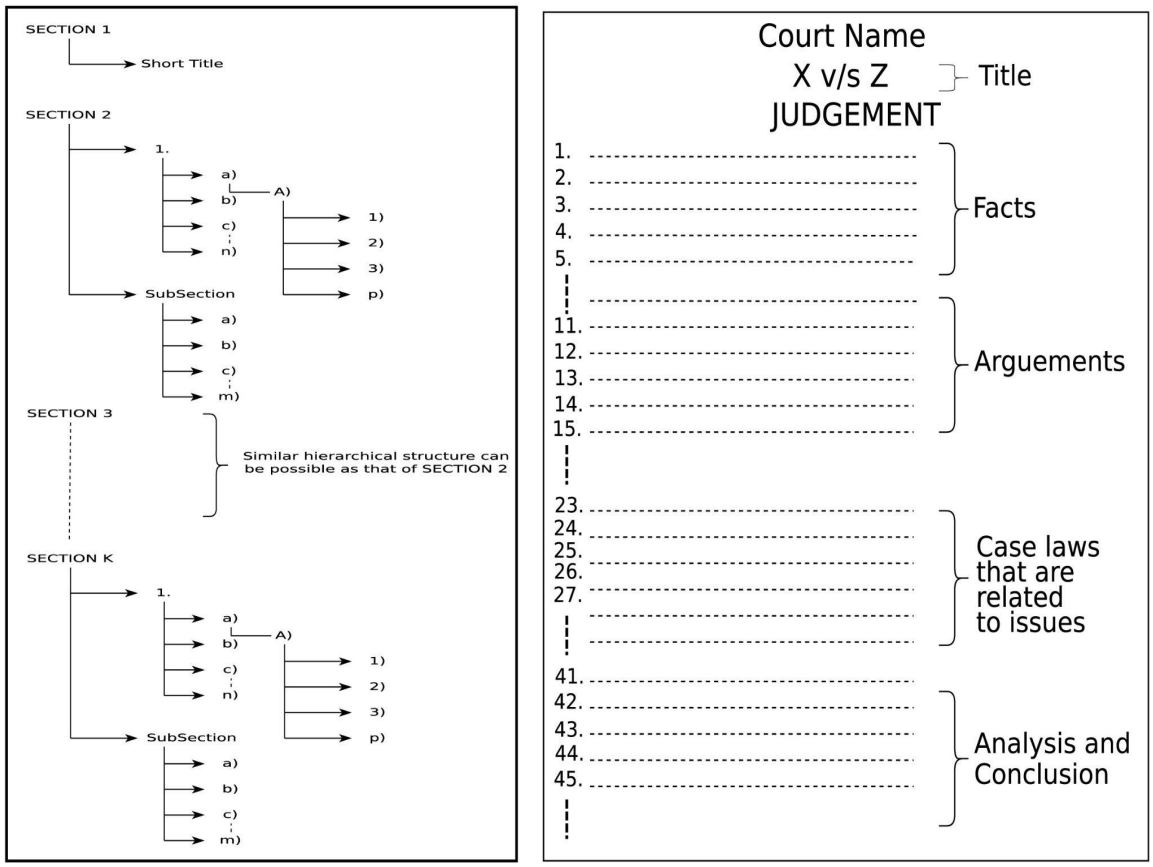

Figure 1. Legal structures of two countries.

Secondly, the main challenge is to find the proper datasets which contains the legal documents and citations. There are several open sources that provides dataset without any permission. But most of them are not providing the datasets of legal documents. The resources which have datasets containing legal documents are not allowed to use them. So, it is very crucial to find the proper dataset from legit sources. Additionally, the datasets should be verified from lawyers and professional writers to maintain quality content. The datasets can be obtained

by web-scrapping. To retrieve data, a web crawler was created in python using selenium and beautiful soup libraries (2). The main issue in web scrapping is data volume. When large amount of data is scrapped from the website, it may cause storage and processing challenges. Storing and managing such datasets is extensive work. While web scrapping it will raise ethical and legal concerns. It is important to respect websites terms and policies to not accessing sensitive information without consent. Many websites have anti-bots such as CAPTCHAs to block automated access. Dealing with CAPTCHAs programmatically can be challenging.

The one of the challenging parts to execute NLP problem is choosing right and precise approach. Neural network process each sentence in the form of vectors and there are several methods to create vectors from chunk of sentences. One of the famous methods are one-hot encoding, glove and word2vec. However, sometimes neural network is unable to identify those vectors and end with wrong prediction. So, creating vectors in data preprocessing stage is more questionable. The next concerning point is architecture of neural network and hyperparameters of each layer. Different architecture is suitable for different dataset. As dataset changes the same architecture may not be suitable for a new dataset. In that case, a new neural network can be built from the scratch and researchers needs to test each combination on the dataset. It needs constant monitoring and respective hardware support in terms of GPU and RAM. The length of neural network should be determined in a way such that it should not be that short that it will miss important features. Also, it should not be that lengthy so that it will take more time to process and extract data. If it has more layers, then there may arise issue of memory size which will turn into big problem at the time of deployment.

The evaluation metrics is very important to ensure the quality of summaries generated by system. The performance of automatic summarization is often measured by

ROUGE score. ROUGE consider the number of overlapping units such as word pairs, n-grams and word sequence between system generated summary and professional human generated summary (5). However, this evaluation compares system generated summary and human generated summary, if there are more than one references, and average is considered in that case. The score takes specific words in account and the human generated summaries are generally written by professionals and experts. So, there is a chance that both do not match. Also, there is length difference between because one is working as extractive while other is working as abstractive (2). Apart from that, the score does not evaluate linguistic qualities, it relies on lexical overlaps between terms and phrases (4). The model should be tested to get minimum ROUGE score which means the system generated text match with professional writers. Human evaluation is needed to check quality of summaries. If there are multiple.

Comparisons and Observations

In feature-based approach term frequency and inverse document frequency catches the catchphrases and presents the summary depending on the context. The absence of freely available legal dictionary can create hindrances semantic analysis of legal documents. Linguistic feature-based represents two concepts- sentence description and term description. It also considers the concept of neighbour weight and utilize reinforcement between neighbour sentence to create term sentence matrix. Language independent approach use three extractive features word frequency, sentence frequency, and sentence position. Each feature computes values for the sentences of the text. Later these values aggregated, and top scoring selected for the summary according to threshold provided by user (4). Statistical and sematic feature approach deals with both statistical and semantic feature like term frequency, world frequency, inverse sentence, stop words filtering, resolved anaphora, word senses and textual entailment to generate extractive

summary. Probabilistic approach finds salient sentences, key concepts, and relationship among the words. The models such as Bayesian model and hidden Markov model is used to find such main features from the text. Rhetorical role-based approach uses rhetorical role associated with each sentence. It aligns different sentences that are associated with the rhetorical role in the final summary generation (5).

Approach	ROUGE-1		ROUGE-2		ROUGE-L	
	F1	R	F1	R	F1	R
CaseSummarizer [85]	0.3632	0.3338	0.1551	0.1372	0.2947	0.2586
RBM [65]	0.3166	0.4049	0.1007	0.1350	0.2469	0.2755
LSTM with w2v [84]	0.4073	0.4638	0.1883	0.2093	0.3312	0.3588
LSTM with glove [84]	0.4071	0.4596	0.1863	0.2056	0.3322	0.3576
Lexrank[34]	**0.4144**	0.4529	0.1936	0.2083	0.3406	0.3531
Textrank [33]	0.4069	0.5055	**0.2015**	**0.2461**	**0.3457**	0.3848
LSA [60]	0.3363	0.3145	0.1313	0.1203	0.2970	0.2840
Reduction [62]	0.3996	0.4870	0.1843	0.2214	0.3255	0.3632
Luhn [17]	0.4112	**0.5112**	0.1981	0.2423	0.3447	**0.3871**

Figure 2: Test results of existing software tools and respective ROUGE score.

Latent Semantic Analysis (LSA) is automated unsupervised statistical-algebraic summarization technique that follows extractive way to analyse documents. LSA is widely used in the field of NLP for text mining, information retrieval, and document comparison. LSA assumes semantically similar terms at the same piece of text. LSA uses singular value decomposition to reduce the number of rows in the matrix to perform cosine similarity (2). In extractive summarization, neural network follows two stages – 1) generating labelled datasets and 2) extracting the essential components from training data. The author has used the

combination of LSTM and CNN to capture long dependencies and generate pattern on multiple input. In addition, a pretrained word2vec and glove embeddings used to feed in as input to LSTM. In abstractive summary generation, the author has used pointer generator approach which is trained on OpenNMT. The standard sequence to sequence model is implemented which utilizes a pointer generator network that can copy the words from the input text. Transformers models are used with embedders size of 512. NMTSmall uses 2 layers, unidirectional LSTM with 512 units and it has converged in 26,000 steps. Transformer uses AAN (Average Attention Network) uses cumulative average attention network in the decoder side (6).

There are several software available in the market and respective ROUGE score is shown in above table. From the observation, it is claimed that Lexrank has good ROUGH-1 score that is 0.4144 and 0.5112. ROUGE-1 represents unigram overlap between generated summary and referencing summary. Textrank has goof ROUGE-2 score among all of them that is 0.2015 and 0.2461. ROUGE-2 measures the word pair among that matches with referencing document. For ROUGE-L Textrank and LUHN gives best result among all of them. ROUGE-L emphasizes the longest common subsequence, which considers the order of words in the summary.

Conclusion

The difficulty of navigating through lengthy legal papers has long been a problem for legal experts, scholars, and law students. They frequently feel overwhelmed and use a lot of time and energy trying to understand difficult legal documents. A branch of natural language processing called legal document summarizing comes to the rescue by providing succinct summaries of long papers, saving time and effort. These summaries encompass legal arguments,

facts, and rulings, enabling attorneys to understand the particulars of each case and reach well-informed conclusions.

With a focus on legal text summarization, this survey article examined the quickly developing subject of text summarization. It examined the methods, approaches, and metrics influencing the text summarization market in the field of natural language processing. It also emphasized the difficulties and crucial problems that must be addressed from technical, commercial, and ethical viewpoints to guarantee the success of these technologies. Legal text summary has some challenges of its own. Case studies and contracts included, legal papers display distinctive features such as variable volumes, structures, specialist vocabulary, ambiguities, and the significant use of citations. To provide accurate and comprehensible summary, these complexities necessitate specialized methodologies.

The survey divided text summarizers into categories based on their output, input, context, and goal, illuminating the range of methods applied in the industry. Different methods for extracting and abstracting information from legal documents, context- and purpose-based summarizers, and graph-based approaches are available.

References

Anand, D., & Wagh, R. (2022). Effective deep learning approaches for summarization of legal texts. *Journal of King Saud University-Computer and Information Sciences, 34*(5), 2141-2150.

Merchant, K., & Pande, Y. (2018, September). Nlp based latent semantic analysis for legal text summarization. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1803-1807). IEEE.

Saravanan, M., Ravindran, B., & Raman, S. (2008). Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*

Kanapala, A., Pal, S., & Pamula, R. (2019). Text summarization from legal documents: a survey. *Artificial Intelligence Review, 51*, 371-402.

. Jain, D., Borah, M. D., & Biswas, A. (2021). Summarization of legal documents: Where are we now and the way forward. *Computer Science Review, 40*, 100388.

Sheik, R., & Nirmala, S. J. (2021, November). Deep learning techniques for legal text summarization. In *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (pp. 1-5). IEEE.