CAWESumm: A Contextual and Anonymous Walk Embedding Based Extractive Summarization of Legal Bills

Deepali Jain, Malaya Dutta Borah and **Anupam Biswas**
Department of Computer Science and Engineering
National Institute of Technology Silchar, India
jaindeepali010@gmail.com,{malayaduttaborah,anupam}@cse.nits.ac.in

Abstract

Extractive summarization of lengthy legal documents requires an appropriate sentence scoring mechanism. This mechanism should capture both the local semantics of a sentence as well as the global document-level context of a sentence. The search for an appropriate sentence embedding that can enable an effective scoring mechanism has been the focus of several research works in this domain. In this work, we propose an improved sentence embedding approach that combines a Legal Bert-based local embedding of the sentence with an anonymous random walk-based entire document embedding. Such combined features help effectively capture the local and global information present in a sentence. The experimental results suggest that the proposed sentence embedding approach can be very beneficial for the appropriate representation of sentences in legal documents, improving the sentence scoring mechanism required for extractive summarization of these documents.

1 Introduction

Automatic summarization of lengthy legal documents has several benefits regarding the quick understanding of these documents for various types of users like lawyers, judges, lawmakers, and the general public (Jain et al., 2021). One of the most popular ways of performing such automatic summarization is via extractive summarization approaches, where the main idea is to extract summary-worthy sentences directly from the original documents. This process involves the appropriate representation of the individual sentences of the document, followed by the summary worthiness scoring of these sentences. The quality of sentence representation and the subsequent scoring mechanism greatly impact the overall extractive summarization performance. This motivates the need for effective sentence embedding approaches that can capture both the meaning of the individual sentences and their global context with respect to the entire document. This work proposes an improved sentence embedding approach that combines domain-specific sentence embedding with feature-based anonymous walk embeddings (AWE) of a document (Ivanov and Burnaev, 2018), which can help represent a sentence more effectively for extractive summarization.

Several research works have explored the problem of extractive summarization in the legal domain (Jain et al., 2021). CaseSummarizer (Polsley et al., 2016) is a tool which is specifically developed for summarizing legal judgment documents. In this approach, extractive summary is produced based on the frequency of words. In addition to the frequency, it also uses domain specific knowledge.

In the recent years, neural networks based approaches have shown to be very effective for extractive summarization. Most of these approaches have formulated the summarization task either as binary classification problem (Eidelman, 2019; Nallapati et al., 2017) or classification followed by ranking (Zhou et al., 2018; Narayan et al., 2018) of the sentences inside the documents. In order to perform such classification based summarization, researchers have extensively explored the problem of finding appropriate embeddings or representations of the individual sentences (Diao et al., 2020; Liu and Lapata, 2019). In this work also, our main focus is on finding the appropriate sentence representations for the summarization of legal bills.

Representing a document in terms of a sentence connectivity graph is an idea which has been very popularly applied in the general text summarization domain (Mihalcea and Tarau, 2004; Baralis et al., 2013; Li et al., 2020), because it captures the document-level global context of each sentence effectively. This suggests that if local semantics of a sentence can be combined with a graph-based

Proceedings of the 18th International Conference on Natural Language Processing, pages 414–422
Silchar, India. December 16 - 19, 2021.

global context, appropriate summarization-centric sentence embeddings can be obtained.

The key contributions of this work are given below:-

- For better sentence representation, a combination of domain-specific local embedding and graph-based anonymous random walk embedding approach is proposed.
- A detailed empirical analysis of different anonymous random walk settings are explored for finding appropriate sentence embeddings.
- A Multilayer Perceptron (MLP) based sentence summary worthiness prediction approach is presented which can make use of the improved sentence embeddings in the extractive summarization process.

Following the introduction, the organization of the rest of the paper is done as follows: A brief description of the related work is given in Section 2. A detailed description of our proposed method is given in Section 3. The evaluation strategies are presented in Section 4. The experimental results are given in Section 5 along with a detailed discussion. Finally Section 6 concludes our work, by summarizing the key findings and the potential future research directions.

2 Related Work

There are popular classical unsupervised extractive approaches in general text summarization which either utilizes the frequency-based methods (Nenkova and Vanderwende, 2005) or graph-based methods (Mihalcea and Tarau, 2004; Jing, 2000) for scoring sentences. Finally, the top scoring sentences are picked up to form an extractive summary. Several research works also find important sentences in a document, based upon the idea of Singular Value Decomposition (SVD), such as LSA (Steinberger et al., 2004). There is yet another popular classical approach in which sentences are added in a greedy manner into the summary as long as the Kullback-Lieber (KL) divergence keeps on decreasing between the document set and the summary set (Haghighi and Vanderwende, 2009). Recently a Bayesian Optimization (BO) based approach BO-Textrank has been proposed by Jain et al. (2020), in which the authors improve the Textrank algorithm for extractive summarization.

A neural network based supervised approach is proposed by Eidelman (2019), in which scores are assigned to each of the sentences of the document and the best among them are selected. The authors have formulated the sentence scoring task as a sentence classification problem for which the random ensemble and Bert models are used as classifiers to predict the important sentences for summary formation. In (Nallapati et al., 2017), authors have proposed a novel approach called SummaRuNNer, which is a Recurrent Neural Network (RNN) based approach in which the summarization task is formulated as the sequence classification task for extractive summarization of documents. Another unsupervised neural network approach is proposed by Verma and Nidhi (2017) where summary creation is done by firstly extracting the Restricted Boltzmann Machine (RBM) based features followed by a feature enhancement step.

Several random walk based approaches have been proposed in the literature for the summarization of documents represented in terms of a graph. Wang et al. (2017) have proposed an affinity-preserving random walk for the multi-document summarization problem. The summary sentences are extracted once the random walk reaches a stationary state for the purpose of summary generation. In another work, Wang et al. (2014) have proposed a random walk model in which utterances are the nodes and the relationship between two utterances is determined with the help of topic relevance, opinion relevance and structure relevance features. Finally, PageRank algorithm-based global ranking is done to select the relevant utterances to form an opinion summary. Otterbacher et al. (2005) have proposed a topic-sensitive version of Lexrank method (Erkan and Radev, 2004) where the sentence score is calculated based on the concept of random walks. The sentence score is determined by considering sentence's relevance to the query as well as it's similarity to other high scoring sentences. Apart from these works, several efficient attention mechanisms have also been proposed in the literature for handling long documents and thus achieving better performance in downstream tasks such as summarization (Zaheer et al., 2020; Beltagy et al., 2020).

From the literature review, it has been observed that most of the works ignore either the importance of domain specific knowledge or the capability to handle long documents. To deal with

these shortcomings, a novel sentence representation approach is proposed in this work, which utilizes domain specific Legal-Bert embeddings of sentences along with AWE based document graph embeddings. Such combined sentence representation scheme can capture both the local sentence level information as well as the global document level information, thereby achieving better summarization of lengthy legal documents. The reason for utilizing AWE is to find the accurate vectorized representation of the entire document graph. This vectorized representation can be efficiently found using anonymous walk distribution, as proven by Micali and Zhu (2016).

3 Proposed approach

The basic steps of our proposed approach CAWE-Summ (Contextual Anonymous Walk Embedding Summarizer) to automatically generate the extractive summary of legal documents is presented in this section. Our training dataset (D^{Tr}) consists of $\{(d_1, s_1), (d_2, s_2), ..., (d_m, s_m)\}$, where (d_i, s_i) corresponds to the i^{th} document-summary pair in the dataset. The overall methodology has been depicted in Fig. 1.

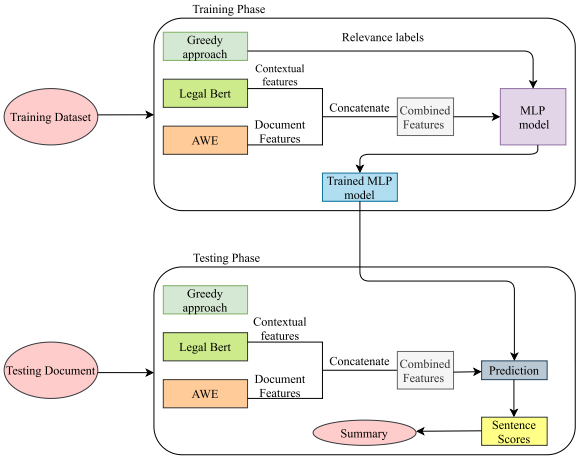

Figure 1: Overall Proposed Methodology

In the training phase, firstly, the Legal-Bert (Chalkidis et al., 2020) based sentence embeddings and Anonymous Walk Embeddings of the entire document graph are combined via concatenation operation. After this, the combined representation is used in an MLP model for learning the binary classification of summary worthiness for each of the sentences. Once the training phase is complete, we obtain a trained MLP model, which can be used at test time for predicting whether a sentence is summary-relevant or not. Each of the individual steps in Fig. 1, which depicts the proposed summarization approach, is discussed in a detailed manner in the following subsections.

3.1 Contextual Representation

The very first step is to create sentence embeddings for each of the sentences present in each document. This sentence representation is achieved through a pre-trained model. Since Bert has achieved state-of-the-art performances on several tasks (Devlin et al., 2018), researchers have started exploring the application of Bert to domain specific legal tasks as well. But the adaptation of general Bert could not perform well in legal specific tasks. Hence Chalkidis et al. (2020) has developed a legal specific Bert known as Legal-Bert. In this work, we use Legal-Bert for sentence representation which consists of 12-hidden layers, where each layer consists of 768 units. To get the contextual representation/features of sentences, the average of all the tokens in a sentences is taken. In this way, we get a vector of 768 features representing the input sentences. If the i^{th} document consists of k sentences, then it is represented as shown below:

$$\begin{aligned} d_i &= \{LB(s_1), LB(s_1), ...LB(s_k)\} \\ &= \{[s_1^1, s_1^2, ..., s_1^{768}], [s_2^1, s_2^2, ..., s_2^{768}], \\ &\quad [s_k^1, s_k^2, ..., s_k^{768}]\} \end{aligned}$$

where $LB(s_k)$ is a pretrained Legal-Bert model, applied on each sentence of a document to obtain the corresponding sentence representation.

Thus, we get the contextual representation of each sentence present in each of the documents.

3.2 Feature-based Anonymous Walk Embeddings

After obtaining the contextual representation for the input sentences, we try to enhance the representation with the help of graph representation. For this, we convert every document into Graph $G_i = (V_i, E_i)$, where V_i consists of sentences from d_i and E_i consists of direct edges between all sentences or nodes in V_i with edge weights as the similarity values between each pair of vertices. We consider cosine similarity metric for finding the similarity between each pair of vertices. The adjacency matrix representation of G_i is shown in Fig. 2.

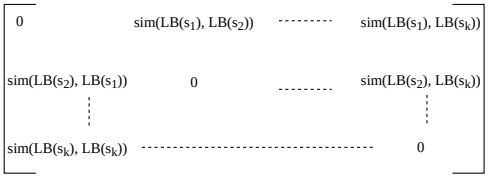

Figure 2: Adjacency matrix representation of G_i

Once the graph representation G_i is available, a p- dimensional embedding for the graph G_i can be obtained using feature-based Anonymous Walk Embeddings approach as shown below:

$AWE(G_i) = [a_1, a_2, ..., a_p]$

The main idea of AWE is to represent the random walks as a sequence of times when node in a graph was visited first, and not as a sequence of nodes (Ivanov and Burnaev, 2018). In order to understand feature-based AWE, let's first try to understand AWE. Consider the random walks as shown in Fig. 3:

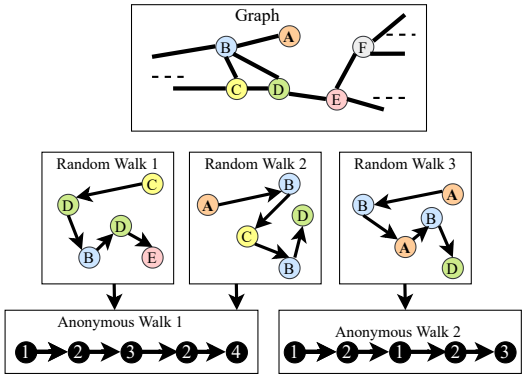

Figure 3: Illustration of anonymous walk in a unweighted directed graph

From Fig. 3, it can be observed that random walks are not represented as a sequence of nodes but as the index of a node when it appears first. These walks are known as anonymous walks because they are agnostic to the identity of nodes visited. It means that, random walks that have visited different nodes but in the same order, get the same anonymous walk representation (for example, look at random walk 1 and 2). Micali and Zhu (2016) theoretically justified that AWE, allows to encapsulate and reconstruct the structure of the entire graph irrespective of global information and therefore can be used to represent feature based embeddings for the entire network. Based on this, (Ivanov and Burnaev, 2018) came up with the feature representation of the entire network. There is a exponential growth in the number of anonymous walk with length l. For example, there are five anonymous walks w_j of length 3: $w_1 = 111$, $w_2 = 122$, $w_3 = 121$, $w_4 = 112$, $w_5 = 123$. The j^th coordinate of $AWE(G)[j]$ is the probability of anonymous walk w_j in Graph G, i.e., the probability that the anonymous walk of type j occurs in graph G. Since it is infeasible to count all the anonymous walks in a large graph, Ivanov and Burnaev (2018) have proposed an efficient sampling approach to approximate the true distribution. In this work also, we have considered the same sampling approach for finding the AWE of length 3, 4, 5 and 6, on the document graph G_i mentioned above.

3.3 Combined features

To enhance the representation of each input sentence, we propose to concatenate AWE (G_i) to each sentence embedding for the i^{th} document to obtain final d_i, thereby capturing the entire graph information as well as the local contextual information. The concatenation is done as shown below:

$$d_i = \{[LB(s_1; AWE(G_i))], \\ [LB(s_2; AWE(G_i))], \\ [LB(s_k; AWE(G_i))]\}$$

$$d_i = \{[s_1^1, s_1^2, ..., s_1^{768}, a_1, a_2, ..., a_p], \\ [s_2^1, s_2^2, ..., s_2^{768}, a_1, a_2, ..., a_p], \\ [s_k^1, s_k^2, ..., s_k^{768}, a_1, a_2, ..., a_p]\}$$

where, p is the all possible anonymous walk of length 3, 4, 5, and 6. More specifically, the possible values of p is 5, 15, 52 and 203 for lengths 3, 4, 5, and 6 respectively.

3.4 Extractive dataset building

After having all the d_i's $(i = 1, 2...., m)$ for D^{Tr}, we then build an extractive training dataset for summarization using the greedy approach as proposed in (Nallapati et al., 2017). In this way, we get the dataset in the form:

$D^{TrExt} = \{[d_1, y_1], [d_2, y_2], ..., [d_m, y_m]\}$ where, each d_i is a collection of $(768 + p)$-dimensional vectors, representing each sentence of a document and y_i is a binary vector representing the relevance of each sentence in d_i for summary formation.

3.5 Summary worthiness classification task

For the purpose of training, we use an MLP (details shown in Table 1) that takes sentence embedding

as an input and outputs its summary relevancy or worthiness. In order to train an MLP model, we flatten D^{TrExt} by one level to obtain $D^{TrExtMLP}$ as:

$$D^{TrExtMLP} = \{([x_1, x_2, ..., x_{768+p}]^{(1)}, y^{(1)}), \\ ([x_1, x_2, ..., x_{768+p}]_{(2)}, y^{(2)}), \\ \vdots \\ \vdots \\ ([x_1, x_2, ..., x_{768+p}]^{(q)}, y^{(q)})\}$$

where q is the total number of sentences across all the documents ($\approx$ 10M for BillSum dataset).

We then train the model on $D^{TrExtMLP}$, which takes $(768 + p)$-dimensional sentence embedding as input and outputs its summary worthiness. For training purpose, we consider four dense layers of nodes 768, 128, 64, 32 followed by one output layer with a sigmoid activation function. The batch size is chosen as 32, adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001.

MLP Layers	# nodes/ Pr
FC Layer 1	768
Dropout	Pr=0.4
FC Layer 2	128
Dropout	Pr= 0.4
FC Layer 3	64
Dropout	0.4
FC Layer 4	32
Dropout	Pr=0.4
Prediction Layer(sigmoid)	1

Table 1: Number of nodes/Dropout Probabilities per layer in the MLP classification model.

3.6 Summary generation

At the time of inference, for every test document, firstly $(768 + p)$ dimensional embedding of each sentence is found, followed by the MLP based prediction of summary worthiness. The final summary formation is done by taking the top 15% sentences which is the ratio of number of words in the training documents to the number of words in the training summaries based on their summary worthiness in the order they appear in the original document.

4 Evaluation strategy

4.1 Dataset

Our proposed approach is evaluated on the BillSum dataset which is a legal specific benchmark dataset introduced by Eidelman (2019). It consists of United States (US) Congressional bills which has been divided into 18,949 training documents and 3,269 testing documents. Along with the US Congressional bills, the BillSum dataset also contains 1,237 California (CA) state bills so that the models build upon US legislatures can be tested upon new legislature as well. The training documents contain 150 sentences on an average while the training summaries contain 20 sentences on an average. The US testing dataset contains an average of 100 and 12.5 sentences in the documents and summaries respectively. The CA test dataset contains an average of 75 and 20 sentences for the CA test documents and summaries respectively.

4.2 Evaluation metric

For the purpose of evaluating the automatically generated summaries, a very popular metric known as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is considered (Lin, 2004). It counts the overlapping of n-grams between reference summaries and system generated summaries. In this work, three variants of ROUGE are considered-ROUGE-1, ROUGE-2 and ROUGE-L.

4.3 Baselines and state-of-the-art methods

We consider 8 baseline methods: Textrank (Mihalcea and Tarau, 2004), Sumbasic (Nenkova and Vanderwende, 2005), Latent Semantic Analysis (LSA) (Steinberger et al., 2004), KLSum (Haghighi and Vanderwende, 2009), Reduction (Jing, 2000), Restricted Boltzman Machines (RBM) (Verma and Nidhi, 2017), CaseSummarizer (Polsley et al., 2016) and 2 state-of-the-art methods: DOC+Sum (Eidelman, 2019) and BO-Textrank (Jain et al., 2020) (see Section 2 for brief descriptions of these methods).

Apart from these methods, we also compare our proposed approach with a Legal-Bert based summarization approach which we refer to as Contextual, in Table 2. In this approach, the Legal-Bert based local sentence embeddings are considered, in which sentence importance is predicted using the MLP model described in Section 3.5.

4.4 Experimental setup

The pretrained Legal-Bert implementation has been acquired from the Hugging-Face package [1]. Whereas, the implementation for finding AWE-based representations is publicly available from Ivanov and Burnaev (2018) and in this work the default parameter settings are utilized for the experimentation purposes. The MLP model has been implemented using the Tensorflow package (Abadi et al., 2016). Finally, the experimental results are reported in terms of the F1-Scores of the ROUGE-1, ROUGE-2 and ROUGE-L metrics, using the rouge 1.0.1 package [2].

We have run all the experiments on a Linux machine with i7 processor and RTX 2070 GPU (8GB RAM).

5 Summarization results and discussion

The ROUGE metric based summarization results for the proposed as well as the baseline and state-of-the-art approaches are depicted in Table 2. The results shows that the proposed CAWESumm approach outperforms all the baselines and state-of-the-art approaches for the extractive summarization on the BillSum test datasets. Importantly, even with the smallest length of anonymous walk embeddings, results have been improved significantly comparing to even with the legal-specific summarization baseline (CaseSummarizer) as well as with the state-of-the-art approaches. More specifically, the CAWESumm approach can obtain the best ROUGE-1, ROUGE-2 and ROUGE-L scores of 0.42827, 0.25288 and 0.41319 on the US test set respectively and 0.43120, 0.21762 and 0.36445 on the CA test set respectively.

It is important to note here that, in case of the US testing data, even though the best performances have been observed when we consider anonymous walk of length 6, still the difference between the other walk lengths are not that significant. In case of the CA testing dataset, the proposed approach has been able to outperform all the baselines and state-of-the-art methods; however here also we can see that the specific walk lengths do not have very significant impact on improving the summarization performance.

Based on the number of sentences in a document, the summarization performance of the document changes. This change in performance is depicted with the help of line charts in Fig. 4. From the line charts it can be clearly observed that when we consider small-sized documents(# of sentences ≤ 50), we can achieve much better summarization performances, across different anonymous walk lengths. On the other hand, as we get medium-sized($51 \leq$ # of sentences ≤ 100) and large-sized($100 <$ # of sentences) documents, we see that the summarization performance decrease significantly, for both the US test and CA test datasets. This is due to the fact that, always the top 15% of the high scoring sentences are picked for summary formation, and when a larger document comes as an input, the top 15% will include some low confidence predictions as well. Moreover, this causes large-sized summary formation which might be detrimental by itself.

Studying the inference time of the proposed approach can give appropriate insights into its real-time applicability. This analysis is presented in Fig. 5, with the help of a line chart diagram. From Fig. 5, it can be observed that for both the test sets, the average inference time for generating summaries is in the range of $(1 - 7)$ seconds. Another important observation is that, as the sentence embedding dimension increases, the inference time also increases sharply. This is to be expected, since larger embedding size are due to the presence of longer anonymous random walks with more number of samplings. Repeated simulation of such long walks is bound to increase the total inference time, as during inference also the AWE vectors for each sentence is needed to be calculated.

One of the key findings of this work is that the inclusion of anonymous random walk based document graph embeddings as part of the sentence embedding itself can significantly improve the overall quality of sentence representation. Such improved representation of sentences can help in the subsequent summary worthiness prediction process, as these sentences are aware of their global context in the document. The intuition behind such performance improvement is that through the learning of different anonymous walks, the embeddings are much more informative than only contextual embeddings since the anonymous walk embeddings can efficiently model the entire document. This effectiveness of AWE based graph embeddings is supported by (Micali and Zhu, 2016), where the authors prove that under sufficient samplings of anonymous random walks, entire subgraphs of the

[1] https://huggingface.co/nlpaueb/legal-bert-base-uncased

[2] https://pypi.org/project/rouge/

	Method type	Approach	US test data			CA test data		
			ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	Unsupervised	Textrank (Mihalcea and Tarau, 2004)	0.32698	0.17939	0.33835	0.40693	0.20159	0.34574
		Sumbasic(Nenkova and Vanderwende, 2005)	0.23979	0.08106	0.22749	0.32799	0.12773	0.29769
		LSA (Steinberger et al., 2004)	0.32771	0.12888	0.28909	0.33635	0.13136	0.2971
		KLSum (Haghighi and Vanderwende, 2009)	0.26383	0.0927	0.21385	0.28002	0.10348	0.22647
		Reduction (Jing, 2000)	0.34728	0.17574	0.33046	0.39962	0.18439	0.3256
		CaseSummarizer (Polsley et al., 2016)	0.34019	0.14488	0.28507	0.36321	0.15515	0.29476
		RBM (Verma and Nidhi, 2017)	0.29710	0.10796	0.23970	0.31660	0.10074	0.24697
	Supervised	Contextual	0.38043	0.22336	0.3886	0.42100	0.20827	0.34982
		SummaRunner (Nallapati et al., 2017)	0.41604	0.22454	0.39148	0.38616	0.17467	0.32814
Proposed	Supervised	CAWESumm ($l = 3$)	0.42247	0.25002	0.41068	0.43104	**0.21762**	0.36325
		CAWESumm ($l = 4$)	0.42465	0.25058	0.41151	**0.43120**	0.21653	**0.36445**
		CAWESumm ($l = 5$)	0.42739	0.25246	0.41309	0.42730	0.21420	0.36067
		CAWESumm ($l = 6$)	**0.42827**	**0.25288**	**0.41319**	0.42998	0.21671	0.36186
State-of-the-art	Supervised	DOC + SUM (Eidelman, 2019)	0.4080	0.2383	0.3373	0.3965	0.2114	0.3405
	Unsupervised	BO-Textrank (Jain et al., 2020)	0.356	0.172	0.312	0.404	0.194	0.327

Table 2: Comparison of proposed CAWESumm approach for different AWE lengths (l), with baseline and state-of-the-art approaches.

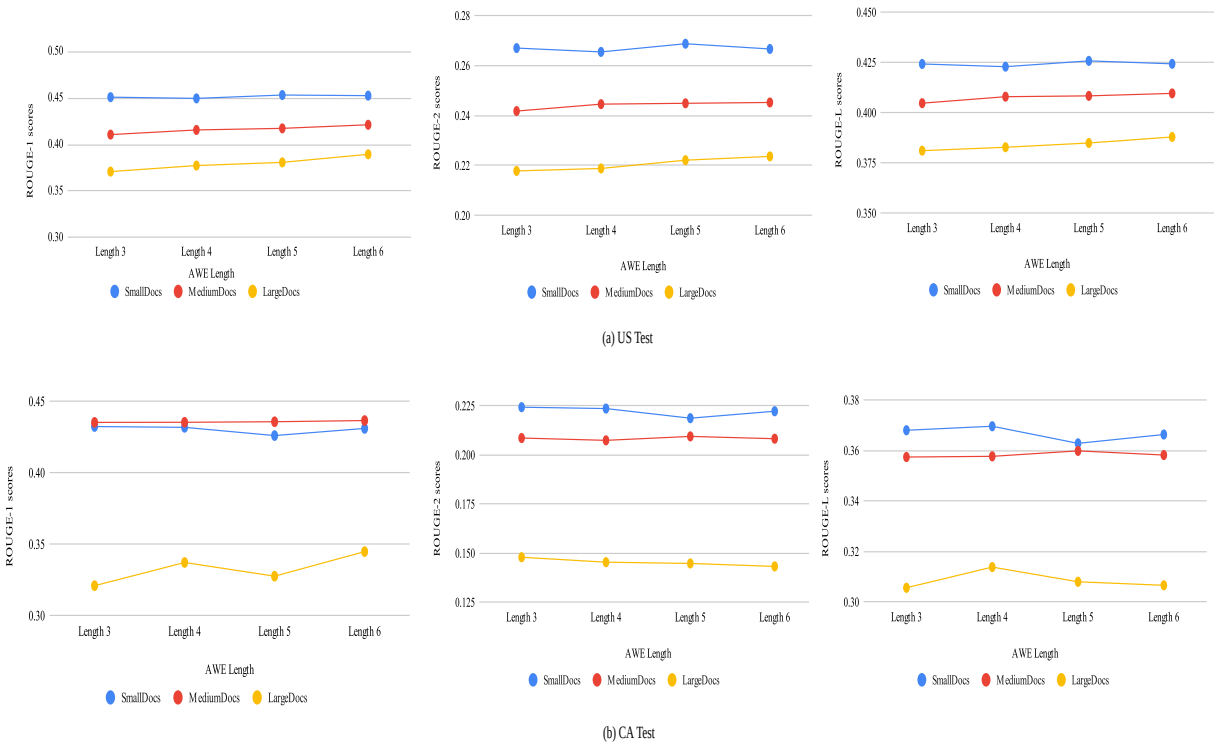

Figure 4: Document length wise ROUGE scores

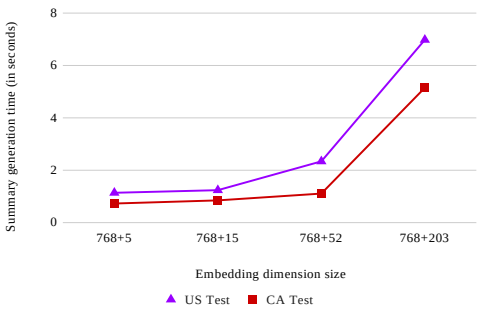

Figure 5: Average running time to generate summaries for BillSum testing dataset where 5,15,52 & 203 are all possible anonymous walks for l=3, 4, 5 & 6 respectively.

underlying graph can be reconstructed with the help of such walks in that region. In our case, this result ensures that the AWE obtained on the document graph can effectively represent the original input document. Moreover, since the MLP model has been trained to recognize summary worthy embeddings during the training process, the interaction between the contextual and document graph embedding features can be effectively modeled for summarization.

6 Conclusion

Due to the lengthy nature of legal bills, it becomes very difficult to capture the important information of the documents. To overcome this difficulty, extraction of summary worthy sentences for automatic summarization has been explored in the literature. However, in order to efficiently identify summary worthy sentences, they need to be appropriately represented in the form of numerical vectors. In this work, we propose to capture a sentence's local as well as global context information in the form of embeddings via contextual and anonymous walk embeddings. From the experimental results, it has been found that, when we incorporate the global document level information with the sentence's local information, a significant improvement can be obtained in terms of ROUGE scores. The experimental results suggests that the anonymous walk embeddings are very effective in capturing the entire document graph information, and can enhance the representation of a sentence for its summary worthiness prediction. Such improved sentence representation are able to significantly improve the extractive summarization of the document.

To further improve the representation learning, leveraging embeddings in the form of hierarchical structure of the entire legal documents will be part of our future work. Moreover, a more end-to-end graph based approach can also be studied, by considering the emerging area of graph neural networks.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Elena Baralis, Luca Cagliero, Naeem Mahoto, and Alessandro Fiori. 2013. Graphsum: Discovering correlations among multiple terms for graph-based summarization. *Information Sciences*, 249:96–109.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yufeng Diao, Hongfei Lin, Liang Yang, Xiaochao Fan, Yonghe Chu, Di Wu, Dongyu Zhang, and Kan Xu. 2020. Crhasum: extractive text summarization with contextualized-representation hierarchical-attention summarization network. *Neural Computing and Applications*, 32(15):11491–11503.

Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Sergey Ivanov and Evgeny Burnaev. 2018. Anonymous walk embeddings. In *International conference on machine learning*, pages 2186–2195. PMLR.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2020. Fine-tuning textrank for legal document summarization: A bayesian optimization based approach. In *Forum for Information Retrieval Evaluation*, pages 41–48.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Sixth Applied Natural Language Processing Conference*, pages 310–315.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. *arXiv preprint arXiv:2005.10043*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries acl. In *Proceedings of Workshop on Text Summarization Branches Out Post Conference Workshop of ACL*, pages 2017–05.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Silvio Micali and Zeyuan Allen Zhu. 2016. Reconstructing markov processes from independent and anonymous experiments. *Discrete Applied Mathematics*, 200:108–122.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Jahna Otterbacher, Gunes Erkan, and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 915–922.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.

Josef Steinberger, Karel Jezek, et al. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100.

Sukriti Verma and Vagisha Nidhi. 2017. Extractive summarization using deep learning. *arXiv preprint arXiv:1708.04439*.

Kexiang Wang, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2017. Affinity-preserving random walk for multi-document summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 210–220.

Zhongqing Wang, Liyuan Lin, Shoushan Li, and Guodong Zhou. 2014. Random walks for opinion summarization on conversations. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 430–437. Springer.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. *arXiv preprint arXiv:1807.02305*.