Extractive Summarization of Law Domain Documents

Saksham Balyan[1], Shrawan Kumar[2], Saurav Dubey[3], Praneetha[4]
[1, 2, 3] Dept of Information Science and Engineering
[4]Assistant professor, Dept of Information Science and Engineering
[1, 2, 3, 4] Visvesvaraya Technological University Sapthagiri College of Engineering, Bangalore,India

Abstract- *The objective of Extractive Summarization of Law Domain Documents project is to produce a summary of the whole document which can present the information portrayed in the document in less number of words. This work aims at producing a high-quality article summary by taking into account the generic components of an article within a specific domain, which is law domain in our project. The amount of information available on the Web in the form of pages is increasing exponentially. Users use this information to be updated with day to day activity, to gain knowledge, and for learning purpose. Since so much information is available about a topic, user finds it difficult to select the best option among the available options. This, in turn, is a time-consuming process. Only some reports, articles, etc. come with author written summaries which help in deciding whether the article is suitable for gaining in-depth knowledge about the topic or not. Since, the remaining articles, reports, etc. do not come with their summaries; it increases the difficulty for the user to check their relevancy. This approach uses sentence feature extraction to assign scores to each and every sentence of the document. A final summary is prepared by normalizing the scores generated and taking into consideration the entire summary worthy sentences.*

Keywords- Law Domain, Extractive Summarization, Normalizing.

I. INTRODUCTION

The amount of information available on the Web in the form of pages is increasing exponentially day by day. Users use this information to be updated with day to day activity and also the articles or paragraphs published in the law domain in the past. This information is also used to gain knowledge and for learning purposes. Since so much information is available about a topic, users find it difficult to select the best option among the available options. This, in turn, is time consuming. Even after selecting the best option, the user might face the difficulty in understanding the novelty and validity of the information. The above problem is also valid in law domain for people associated with law and order disciplines. With the latest amendments or modifications in the law domain field, the information about the same topic is available from a variety of sources. The law domain information available in the articles, reports, magazines, old books, paragraphs, judicial documents and news forums is one of the important source of information for lawyers, judges, advocates, practitioners and law makers of the country. Only a few articles or reports come with author written summaries which help in deciding whether the article is suitable for gaining in-depth knowledge about the topic or not. Since, the remaining articles, reports, etc. do not come with their summaries, it increases the difficulty for the user to check their relevancy. The main aim of the project is to provide a best suitable extractive summary of a given article which can accurately present the overall meaning or gist of the article. This approach can also be extended to bills, amendments, old articles converted to digital formats, or anything related to law domain.

II. LITERATURE SURVEY

The previous proposed models for document summarization were based on few or all of the features related to any sentence like sentence position, relevancy of sentences with the title of the document, sentence position in the document, term frequencies, standard keywords or cue phrases related to the topic discussed in the document and acronyms. The final scores computed using these above mentioned features are sometime needed to generate a short and efficient summary consisting of only summary worthy sentences using the extractive summary model. The very first breakthrough in the field of legal/law domain text summarization was the work by Farzindar & Lapalme (2004) [1]. They proposed a system called LetSum, which was a combination of using thematic structures and the document's architecture to produce table style summaries. It uses features like inverse document frequency. The major drawback of this approach was that it uses hand-written dictionaries. Several other strategies were proposed in later years without the use of hand-crafted dictionaries. In 2006, Saravanan et.al. [2] presented an approach for applying probabilistic graphical models based on conditional random fields for text summarization. In later year 2010, the approach was applied to different sub-domains of court documents, for example, sales tax, income tax and rent control. It was observed that the performance of the presented approach did not change across different domains. Yousfi-Monod et. al. (2010) [3] used a Naïve Bayes algorithm with

some heuristic features to identify sections such as introduction, context, reasoning, conclusion, etc. and finally create a summary. It was found that the quality of summary for different sections was different from each other.Most of the previous work only focused on text extraction in the legal area. Recent research by Merchant & Pande (2018) [4] proposed an automated text summarization system that makes use of latent semantic analysis (LSA) to capture concepts in a legal document. LSA is an unsupervised learning technique that is similar to sentence embedding. Elnagger et. al. (2018) [5] used the Multi Model algorithm for translation, summarization and classification through transfer learning. For summarization, the authors made use of a dataset containing around 20K legislative documents of the European Parliament since 1958. Each document is labelled with a short description, 1 to 3 sentences highlighting the core. Results showed that when first trained on another task, translation or classification, the model also performed better on summarization, as opposed to starting from scratch. Table 1 gives an overview of different studies done in the field of legal text summarization which includes the technique proposed by the researchers name, scores and the corresponding dataset used.

Table-1 An overview of studies on legal text summarization

Technique	**Accuracy (ROUGE Scores)**	**Dataset**
Farzinder& Lapalme (2004)	58.00%	3.5K Canadian Law Cases
Saravanan et al. (2006)	80.00%	200 Indian Law documents
Yousfi-Monod et al. (2010)	64.70%	4K Canadian Law Cases
Galgani et al. (2012) [6]	29.10%	3K Australian Federal Court cases
Merchant & Pande (2018)	58.00%	50 Indian Court documents
Elnagger et al. (2018)	82.00%	20K European Parliament legislative documents

The summarization techniques that are used in other domains are used as it is for law domain articles to summarize the articles. For example, in the medical domain, MiTAP (MITRE Text and Audio Processing) [7] is one of the software that is used to summarize the legal domain articles by extractive approach. Cluster signature of the document is used to rank the extracted sentences. The Journal of the American Medical Association's articles and abstracts as well as the full texts were used to carry out this project. Another system called TRESTLE (Text Retrieval Extraction and Summarization Technologies for Large Enterprises) [8] is also used to generate single sentence summaries of the newsletters related to pharmaceutical field. It produces summaries with the help of Information extraction process to fill out the templates. Some other approaches used for extractive text summarization along with advantages and limitations are listed in table 2.

Table-2. Advantages and limitations of various summarization approaches

S No.	Methodology	Advantages	Limitations
1	Machine Learning Approach Bayes Rule	Large set of training data improves the sentence selection for summary.	Human interruption required for generating manual summaries.
2	Artificial Neural Network	The network can be trained according the styles of human reader	Neural network is slow in training phase and also in application phase
3	Extractive Text Summarization using Sentence Ranking [9]	Automatic text summarization using frequency of words.	It does not uses LSA model.
4	Information Extraction Framework for Legal Documents [10]	Evaluation through gist information extraction with 80-90% accuracy	Applied on major laws which are frequently used in court.
5	Graph Based Approach	Capture redundant information and improves coherency.	Doesn't focus on issues such as dangling anaphora problem.

III. PROPOSED METHODOLOGY

In this section, the approach of performing document summarization is presented. We have used extractive document summarization approach in which the important step is to identify summary worthy sentences from the source document and at the same time reducing the redundancy from the original text so that the final summary generated is relevant to the users. In text summarization there are three parts: 1) Pre-Processing 2) Sentence Feature Extraction 3) Normalization of Scores. The system architectural design is presented in figure 1.

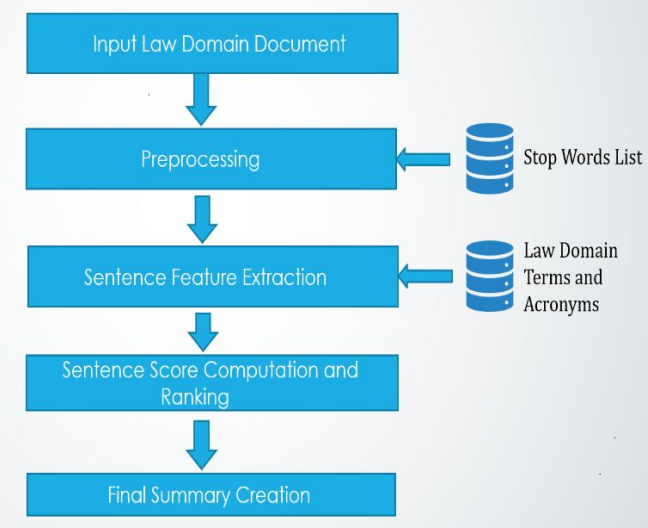

Figure 1. Architectural design of extractive document summarization system.

The summary can be classified on the basis of information content as indicative and informative summary. An indicative summary provides an insight about the document like the purpose and the approach of the document for easy selection by the user in order to gain in-depth knowledge about the topic. An informative summary is a summary that covers and provides all the important features in the document with some level of detailing

1. Pre-Processing

Data pre-processing is one of the most important step in any natural language processing system and it should be performed before carrying out any other tasks on the dataset. It is important to have a consistent and a proper dataset to have quality results. In our system of extractive document summarization, there are three tasks in this step.

1.1 Redundancy Removal

The task of figuring out the similarity between sentences is considered to be one of the most important tasks which have a wide range in many text based applications. The main objective of this system is to generate a summary out of n sentences in a document related to law domain by finding a subset of sentences that are summary worthy and contain the most important information of the document. Thus, removing redundancy by identifying the similar sentences is important.

The similarity between two sentences has been calculated using cosine similarity formula which uses the angle of vectors of the two sentences [12]. Let Sa and Sb be the two sentences, then the calculations are performed using the below equation

$$sim_{\cos}(S_a, S_b) = \frac{\sum_{c=1}^{m} w_{ac} w_{bc}}{\sqrt{\sum_{k=1}^{m} w_{ac}^2 \cdot \sum_{k=1}^{m} w_{bc}^2}}$$

where m is the total number of terms in the document, wac refers to the weight of the term c in the sentence Sa and wbc is the weight of the term c in the sentence Sb.

1.2 Sentence Segmentation

Sentence segmentation is the process of dividing thedocument into parts based on the delimiter. The most commonly used delimiter in articles are full stop (.) and question marks (?).

1.3 Removal of Stop Words

The most frequently used words that provide very less meaning to the content of the document are called ‘Stop Words’. They can be removed in the initial stages. For example, ‘a’, ‘an’, ‘the’, ‘above’, etc.

2. Sentence Feature Extraction

In this step, all the segmented sentences are made to go through test that checks the features related to it. We have laid emphasis on four important features because according to the law domain context, these features are enough to categorize a sentence into “summary worthy” or “not summary worthy” sentence

2.1 Sentence Position

Each segmented sentence is given a rank according to the following equation [13], where ‘a’ indicates the position of the sentence in the document:

$$P_a = \frac{1}{\sqrt{a}}$$

2.2 Sentence Length

The length of the sentence is calculated by counting the number of words in a sentence before stop word removal. This is done because short sentences that contains proper nouns only are beneficial for the summary.

2.3 Number of Law Domain Related Terms

The number of law domain related terms or words in a sentence are counted. This will make sure that the sentences that contains more such words are summary worthy [14]. Online law dictionaries are used for this purpose.

2.4 Number of Law Domain Related Acronyms

The number of law domain related acronyms in a sentence are counted. This will make sure that the sentences that contains more such acronyms are summary worthy.

3. *Sentence Score Computation and Normalization*

The summation of all feature values of a sentences gives its sentence score. Since, the range of scores can be varying, data normalization is done using min-max normalization to bring the scores in a range of 0 to 1. Following formula is used:

$$y' = \frac{y - \min_B}{\max_B - \min_B}(new\_\max_B - new\_\min_B) + new\_\min_B$$

where, minB and maxB are minimum and maximum sentences score of document B and new_maxB = 1 and new_minB = 0.

4. Final Summary Creation

The original document can have N number of sentences. We would like to create a summary with a lesser number of sentences, say K number of sentences, where K << N (K is much lesser than N).

The system will prompt the user to enter the number of words, he/she would like to have in the final summary. Normally, the summary is made up of 80 to 150 words such that it can be read in less than 30-40 seconds.

IV. CONCLUSION

The text summarization in law domain on the basis of sentence feature extraction and score normalization has been discussed in this report. There are so many summarizers proposed in this domain that uses a few or more sentence features for summary generation. In development of this summarizer, we have opted for few best sentence features such as sentence length, sentence position in the document, number of law domain related terms and number of law domain related acronyms, and finally by normalizing the score to provide high acceptance summary to user. This is tested on two documents (Titled “Human Rights” and “RTI”) and two summaries for each of the document was obtained. A summary having around 100 words has proven to be more efficient overall than the one having around 150 words. Also, it is observed that the idea of the document can be well understood with around 100 words summary with a much less reading time. Hence, it can be concluded that an ideal summary should be around 80-120 words with a reading time of around 25-30 seconds at most.

REFERENCES

[1] Farzindar, A. & Lapalme, G. (2004). Legal Text Summarization by Exploration of the Thematic Structure and Argumentative Roles. In Marie-Francine Moens, S. S. (Ed.), Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, (pp. 27–34)., Barcelona, Spain. Association for Computational Linguistics.

[2] Saravanan, M., Ravindran, B., & Raman, S. (2006). Improving Legal Document Summarization Using Graphical Models. In Proceedings of the 2006 Conference on Legal Knowledge and Information Systems: JURIX 2006: The Nineteenth Annual Conference, (pp. 51–60)., Amsterdam, The Netherlands, The Netherlands. IOS Press.

[3] Yousfi-Monod, M., Farzindar, A., & Lapalme, G. (2010). Supervised Machine Learning for Summarizing Legal Documents. In Farzindar, A. & Keˇselj, V. (Eds.), Advances in Artificial Intelligence, Lecture Notes in Computer Science, (pp. 51–62). Springer Berlin Heidelberg.

[4] Merchant, K. & Pande, Y. (2018). NLP Based Latent Semantic Analysis for Legal Text Summarization. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), (pp. 1803–1807).

[5] Elnaggar, A., Gebendorfer, C., Glaser, I., & Matthes, F. (2018). Multi-Task Deep Learning for Legal Document Translation, Summarization and Multi-Label Classification. arXiv:1810.07513 [cs, stat]. arXiv: 1810.07513.

[6] Galgani, F., Compton, P., & Hoffmann, A. (2012). Combining Different Summarization Techniques for Legal Text. In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, HYBRID ’12, (pp. 115– 123)., Stroudsburg, PA, USA. Association for Computational Linguistics.

[7] L. D. Day, L. Hirschman, R. Kozierok, S. Mardis, T. McEntee, et al., “Real users, real data, real problems: the MiTAP system for monitoring bio Events”, In the Proceedings of the Conference on Unified Science &

Technology for Reducing Biological Threats & Countering Terrorism (BTR 2002), 2002, PP 167—77.

[8] R. Gaizauskas, P. Herring, M. Oakes, M. Beaulieu, P. Willett, H. Fowkes, et al., "Intelligent access to text: integrating information extraction technology into text browsers", In Proceedings of the Human Language Technology Conference (HLT 2001), 2001, PP. 189—93. Extractive Summarization of Law Domain Documents Dept. of I.S.E., S.C.E. 2019-20 39

[9] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

[10] Kowsrihawat, Kankawin and Peerapon Vateekul. "An information extraction framework for legal documents: A case study of Thai Supreme Court verdicts." 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE) (2015): 275-280.

[11] J.-M. Torres-Moreno: Automatic Text Summarization, John Wiley & Sons, Inc. , Hoboken, NJ, Sep. 2014, 348 pages.

[12] Achananuparp, P., Hu, X., & Shen, X. (2008). The evaluation of sentence similarity measures. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5182 LNCS, 305–316. doi:10.1007/978-3-540-85836-2_29 [13] M.A. Fattah and Fuji Ren, "Automatic Text Summarization" In proceedings of World Academy of Science, Engineering and Technology Volume 27. pp 192-195. February 2008.

[13] M.A. Fattah and Fuji Ren, "Automatic Text Summarization" In proceedings of World Academy of Science, Engineering and Technology Volume 27. pp 192-195. February 2008.

[14] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer" Addison-Wesley Publishing Company, 1989.