ISSN 2047-3338

Legal Document Summarization using Latent Dirichlet Allocation

Ravi Kumar V.[1] and K. Raghuveer[2]

[1]Research Scholar, Department of Information Science and Engineering, The National Institute of Engineering, Mysore, India
[2]Faculty, Department of Information Science and Engineering, The National Institute of Engineering, Mysore, India
[1]ravikumarv@nie.ac.in, [2]raghunie1967@gmail.com

***Abstract*—Online access to the legal judgments of cases for both past and present, from most of the courts around the world creates an opportunities and challenges for both the legal community and for information technology researchers. Due to this increased availability of legal judgments, it has become very much essential to provide a mechanism to extract information quickly and to present it in the form of a short summary from the given legal judgment. In this paper we propose an approach to generate a short summary from the given legal judgment using the topics obtained from Latent Dirichlet Allocation (LDA). The developed topic based document summarization model is capable of generating short summary in effective manner. As per our experience this is the first approach for Indian legal judgment summarization using LDA topic model.**

***Index Terms*– Latent Dirichlet Allocation (LDA), Legal Judgments and Legal Document Summarization**

I. INTRODUCTION

IN the information era, the text summarization is one of the most interesting and challenging task. The goal of a text summarization is to provide the reader an accurate and complete idea of the content of the source [1]. The phenomenal growth of legal documents always creates an undefeatable situation to read and digest all the information, this leads to the requirements of text summarization. One of the form of information selection can achieved by finding the summarization using unconstrained vocabulary manually with no artificial linguistic limitations[2].Text summarization is in many ways an encompassing subfield of NLP. Researchers in the area often make use of part-of-speech (POS) tagging, named entity recognition (NER), language modeling, and many other techniques in NLP and machine learning. Despite our plentiful access to these state-of-the-art tools and research, however, most complex Automatic Text Summarization (ATS) approaches rarely surpass the results achieved with simple statistics-based methods grown principally out of 60-years-old ideas of term frequency analysis [3], [4]. Nevertheless, more structured statistical approaches, based on Blei, et al.'s latent Dirichlet allocation (LDA) [5], have recently been showing promising results through the use of topic- or content-modeling [6], [7].

The task of document summarization can be achieved by breaking the documents in to seven topics and use these topics effectively in summarization. Thus document can be viewed as mixture of topics, which we have to infer, and these topics are mixture of words as data that are visible variables from documents. The score for each sentence in a document is obtained by combining score for all the words in a sentence based upon their relevance to a topic. The summary is top two sentences for each topic from the sorted sentences.

Here we have proposed an approach to perform legal judgment summarization using the latent Dirichlet allocation (LDA) topic model. Here we are interested in extraction based summarization i.e. it extracts the entire sentence without any modification to the original sentence.

II. RELATED WORK

Atefeh Farzindar et al. describes method for summarization of legal documents (proceedings of the federal court of Canada) by exploring the document's architecture and its thematic structure to provide a table style summarization, that help a legal expert in determining the key ideas of the judgment and to improve coherency and readability of the system [8].

Marie-Francine Moens et al. automatically summarizes Belgian criminal cases in order to improve access to the large number of existing and future court decisions. In this project a double methodology was used. First, the case category, the case structure and irrelevant text units are identified based on a knowledge base represented as a text grammar. Consequently, general data and legal foundations concerning the essence of the case are extracted. Secondly, the system extracts informative text units of the alleged offences and of the opinion of the court based on the selection of representative objects [9].

Claire Grover et al. examined the use of rhetorical and discourse structure by finding the main verbs in the sentence of legal cases. The methodology is based on [14], where argumentative roles are considered to classify the sentences [10].

Ben Hachey et al. conducted a set of experiments to classify sentences for rhetorical status using a wide range of machine learning techniques. The task of classifying sentences forms

part of a sentence extraction-based automatic summarization system in the legal domain. With sentences classified in this manner, different kinds of summaries can be generated with prominence given to particular kinds of sentence [11].

Rachit Arora et al. used Latent Dirichlet Allocation to find summary by extracting weighted sentences using some weighting mechanism from the given text documents based on the different events being covered by them [12], [13].

III. OUR APPROACH

Extraction based Legal judgment Summarization is a process of extracting sentences from the given document based on the sentence weight and grouping them to form a summary. In this work we are planning to generate summary for the given Legal judgment based on Latent Dirichlet Allocation (LDA) topic model that helps the lawyers and judges (legal domain) to find the different points about the case in a short duration of time instead of reading the entire judgment.

The architecture of our approach to generate summary for the given Indian legal judgment document based on LDA topic module is shown in Fig. 1. The steps involved in this process are explained below:

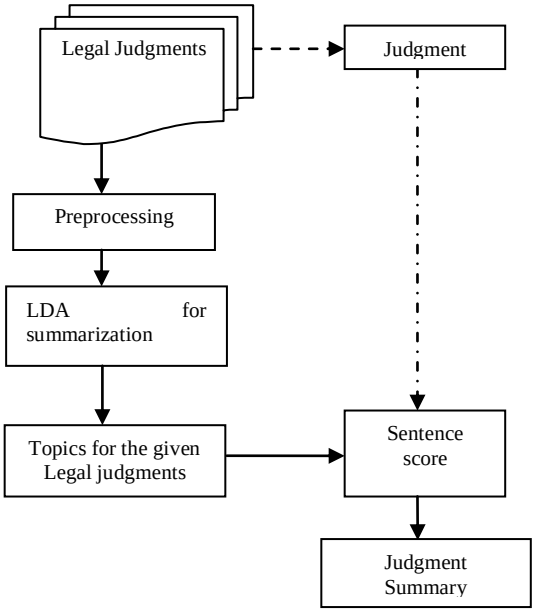

Fig. 1: Architecture to generate Legal Judgment summary

A. *Legal judgments*

The characteristics of Legal judgments are different when compared with scientific articles reporting various research papers and other simple domains related to the identification of basic structure of a document. Based on Teufel & Moen's [14] and Farzindar [15] a legal document consists of different basic rhetorical roles and are shown in Table 1. According to [17] these general classifications have been further classified into seven different labeled elements for a more structured presentation and are shown in Table 2. With this we make an assumption that each judgment can be break down in to seven topics.

Table 1: Description of the basic rhetorical scheme for a legal domain

Labels	Description
Facts of the case	The sentence that gives detail descriptions about the case.
Background	The sentence contains the generally accepted background knowledge (i.e., legal details, summary of law, history of a case)
Own	The sentence contains statements that can be attributed to the way judges conduct the case.
Case relatedness	The sentences contain the details of other cases coded in this case.

Table 2: The rhetorical annotation scheme for legal judgments [17]

Rhetorical Status	Description
Identifying the case	The sentences those are present in a judgment to identify the issues to be decided for a case. Courts call them as "Framing the issues".
Establishing facts of the case	The facts that are relevant to the present proceedings/litigations that stand proved, disproved or unproved for proper applications of correct legal principle/law.
Arguing the case	Application of legal principle/law advocated by contending parties to a given set of proved facts.
History of the case	Chronology of events with factual details that led to the present case between parties named therein before the court on which the judgment is delivered.
Arguments (Analysis)	The court discussion on the law that is applicable to the set of proved facts by weighing the arguments of contending parties with reference to the statute and precedents that are available.
Ratio decidendi (Ratio of the decision)	Applying the correct law to a set of facts is the duty of any court. The reason given for application of any legal principle/law to decide a case is called Ratio decidendi in legal parlance. It can also be described as the central generic reference of text.
Final decision (Disposal)	It is an ultimate decision or conclusion of the court following as a natural or logical outcome of ratio of the decision.

B. *Preprocessing*

The judgment part of the legal judgment document is similar to other documents consists of stop words like is, of, an, etc. We remove these stop words to avoid in getting these stop words as topic terms.

C. *LDA for summarizatio*

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data such as text documents [4]. The documents are represented as a finite mixture over an underlying set of topics which, in turn, are representation of an infinite mixture over a fundamental set of word probabilities. Thus the topics probabilities provide an explicit symbol of the documents. Topics for the given documents and corpora can be obtained using this simple algorithm for Natural Language Processing. These topics are used as basic elements for summarization by extracting topic related sentences from the given document. LDA effectively attempts to get better data modeling over other techniques by allowing documents within corpora to be represented as collections of topics. The idea behind this unique and

revolutionary model is that the topic variable in the model is selected repeatedly within each document to be consisting of multiple topics.

D. Legal Judgments into topics

Given a vocabulary of W distinct words, a number of topics K, two smoothing parameters α and β, and a prior distribution (typically Poisson) over document lengths, this generative model creates random documents whose contents are a mixture of topics. With the use of LDA we break down the set of documents into topics. LDA uses a Dirichlet distribution to represent these topics and under the Dirichlet distribution these variables are independent of each other. After the preprocessing we give all the documents to LDA as bag of words and we get different topics based on this probabilistic model. Here we made an assumption that, the number of topics we get from LDA is equal seven that represents the different rhetorical role of each judgment [17] in the corpus used for summarization.

E. Sentence Score

Now we have topics in hand for the given corpus using LDA topic model. Consider each judgment from the corpus and find the sentences present in the document using [16] Sentence Boundary method. The algorithm to find the sentence score for each sentence from the given judgment is shown in Fig. 2.

Consider all the sentences S_r, r Є {1,...,R} in the documents and all the Topics T_j, j Є {1,...,K} and we calculate the probability of the Sentence S_r given the Topic T_j i.e. $P(S_r|T_j)$.Thus we are calculating the probability that the sentence S_r belongs or represents the topic T_j. Let the words of the sentence S_r be {W1,W2, ... Wq}.

1. Algorithm Sentence_Score()
//Input: D= {d_1, d2,…,dm} //Documents in the corpus for summarization
T_j ={ T_1, T_2,…,T_j } //Topics from LDA
//Output: S= {s_1, s_2,…,s_m} // Sentence score for each sentence for each document
2. for each document d_i ∈ D doT_j
3. for each Topic T_j do
4. for each sentence s_r ∈ d_i do
5. $P(S_r|T_j) = \prod_{Wq \in Sr} P(Wq|Tj)\, P(Tj|di) \rightarrow Si$
6. endfor
7. endfor
8. endfor

Fig. 2: Algorithm to find sentence score each sentence based on each topic for the entire corpus

The algorithm takes each document from the given corpus as input and for each sentence in the document the score based on the probability of occurrence of each word with respect to each topic is found and are sorted in the ascending order.

F. Judgment summary

Now we have score for each sentence based on each topic, the next step is to find the summary, consisting of maximum of two sentences from each topic. The algorithm to find the summary is shown in Fig. 3.

1. Algorithm Judgment_Summary()
//Input: D= {d_1, d2,…,dm} //Documents in the corpus for summarization
T_j ={ T_1, T_2,…,T_j } //Topics from LDA
//Output: Summary= {sm_1, sm_2,…,sm_m} // Summary of the each document in the set
2. for each document d_i ∈ D do
3. for each Topic T_j do
4. S_i=Sentence_Score(d_i, T_j)
5. Arrange the sentences in the descending order based on the sentence score
6. endfor
7. For each topic T_j do
8. Select top 2 sentences from S_{mij} whose score is greater than or equal to average score considering all the sentence in that document
9. if (any of the sentence appears already with respect to previous topic)then
10. Select the next sentence.
11. endfor
12. Arrange the sentences according to the sentence number of d_i to sm_i
13.endfor

Fig. 3: Algorithm to find legal judgment summary using LDA topics

The algorithm takes each document from the given corpus as input and sentence score for each sentence is calculated using Sentence_Score with respect each topic. The top 2 sentences with respect to each topic are selected for final summary by eliminating redundancy. The final summary is obtained for each legal judgment by arranging the extracted sentences according to the sentence number of the original document.

IV. EXPERIMENTAL SETUP

A. Data set

The data set consists of 116 documents from 5 different sub domains pertaining to civil case in India collected from [18]. It has taken from a larger corpus of 250 documents of different sub-domains related to civil court judgments. The documents in the dataset consisting of judgments are dated up to the year 2012. The judgments belongs to different sections like Sales Tax, Rent Control, Motor Vehicle, Family Law, Patent, Trademark and Company law, Taxation, Property and Cyber Law, etc.

B. Parameter for Gibbs sampling

According to [17], a judgment can be divided into seven segments, each represent one rhetorical role, hence we set the K = 7 to match the number of anticipated topics in the corpus.

Following Blei et al. [15], we use $\alpha = 50/K$ and $\beta = 0.1$. Two additional parameters for the Gibbs sampling are the number of sampling and burn-in iterations, which we set to 30 and 200, respectively.

V. EXPERIMENT AND RESULT

Experiment was conducted on the above mentioned dataset to generate summary for the given legal judgment. We have evaluated the performance of our system by comparing the summary generated by our system with the summary generated by legal experts. Sample of judgments from each domain were given to 3 legal domain experts without giving any information about the system generated summary.

To evaluate the results we have used precision, recall and F-measure that are commonly used in information retrieval tasks. The precision, recall and F-measure are calculated using the equation 1.0, for each document using manually extracted summary denoted as S_{ref} and system generated summary denoted as $S_{sys.}$

$$P=\frac{|Sref \cap Ssys|}{Ssys} \qquad R=\frac{|Sref \cap Ssys|}{Sref} \quad F1=\frac{2*P*R}{P+R} \qquad \text{----} \quad 1.0$$

Table 3 shows the mean scores of recall, precision and F-measure of the summary generated for the above said dataset for different domain of civil case.

Table 3: The mean scores of recall, precision and F-measure of the summary generated

Domain	Precision	Recall	F-Measure
Income Tax	0.604	0.587	0.595
Rent control Act	0.589	0.568	0.578
Motor Act	0.551	0.542	0.546
Negotiable Instrument Act	0.526	0.513	0.519
Sales Tax	0.532	0.553	0.542

VI. CONCLUSION AND SCOPE FOR FUTURE

An attempt has been made to generate summary for the given Indian Legal judgment, which very much essential for the legal domain to know the gist of the case with in a short duration of time. The experimental results are very encouraging and this can be further improved by using hierarchical approach to bring the summary in to form that matches the rhetorical structure of the legal document.

VII. ACKNOWLEDGMENT

We would like to thank Mr. Shivakant Aradya, advocate, for the assistance and guidance given to us in preparing legal summary and legal fraternities Miss. Depu and her colleague for their help in preparing the reference summary.

REFERENCES

[1] Inderjeet Mani. "Automatic Text Summarization". John Benjamins Publishing Company.2001

[2] Claire Grover, Ben Hachey, lan Hughson and Chris Korycinski. "Automatic summarization of legal documents", In. Proc. of International Conference on Artificial Intelligence and Law, Edinburgh, UK, 2003.

[3] William M. Darling. "Multi-document summarization from first principles". In Text Analysis Conference, 2010.

[4] Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. "A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization". In SIGIR 2006, New York, NY, USA, 2006. ACM, pages 573-580.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation". J. Mach. Learn. Res., 3:993-1022, 2003.

[6] Hal Daume, III and Daniel Marcu. "Bayesian query-focused summarization". In ACL 2006, Morristown, NJ, USA, 2006. Association for Computational Linguistics, pages 305-312.

[7] Aria Haghighi and Lucy Vanderwende. "Exploring content models for multi document summarization". In NAACL 2009, Morristown, NJ, USA, 2009. Association for Computational Linguistics, pages 362-370.

[8] Atefeh Farzindar and Guy Lapalme. "Legal Texts Summarization by Exploration of the Thematic Structures and Argumentative Roles" .In Text Summarization Branches Out Conference held in conjunction with ACL 2004, pages 27-34

[9] Marie-Francine Moens, C. Uyttendaele, and J. Dumortier. "Abstracting of legal cases: the potential of clustering based on the selection of representative objects". Journal of the American Society for Information Science, 1999, 50(2):pages151–161.

[10] Claire Grover, Ben Hachey, and Chris Korycinski. "Summarising legal texts: Sentential tense and argumentative roles". In HLT-NAACL 2003 Workshop: Text Summarization (DUC03), Edmonton, Alberta, Canada, May 31 – June1, pages 33–40.

[11] Ben Hachey and Claire Grover. "Sequence Modelling for Sentence Classification in a Legal Summarization System". In Proceedings of the 2005 ACM Symposium on Applied Computing.

[12] Rachit Arora and Balaraman Ravindran. "Latent Dirichlet Allocation Based Multi-Document Summarization", In. Proc. of the Second Workshop on Analytics for Noisy Unstructured Text Data, AND 2008, Singapore, July 24, 2008, pages 91-97.

[13] Rachit Arora and Balaraman Ravindran. "Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization". In. Proc. of Eighth IEEE International Conference on Data Mining (ICDM '08) 2008, pages 713-718.

[14] Simone Teufel and Marc Moens. "Summarising scientific articles - experiments with relevance and rhetorical status". Computational Linguistics, 2002, 28(4): pages 409–445.

[15] Atefeh Farzindar and Guy Lapalme, "Letsum, an automatic legal text summarizing system". Legal Knowledge and Information System, Jurix 2004: The seventeenth Annual Conference, Amsterdam: IOS Press, pages. 11-18, 2004.

[16] http://alias-i.com/lingpipe/demos/tutorial/sentences/

[17] M. Saravanan, B. Ravindran, and S. Raman. "Improving Legal Document Summarization using Graphical Models". In Proc. of 19th International Annual Conference on Legal Knowledge and Information Systems, JURIX 2006, pages. 51-60, Paris, France, December, 2006.

[18] http://www.keralawyer.com/asp/sub.asp?pageVal=judgements