Review of Topic Modeling and Summarization

Chinmay Patil[1], Parag Wayangankar[2], Pranay Yadav[3], Shweta Sharma[4]

[1] ,[2], [3]Student, [4] Professor, Department of Computer Engineering, Atharva College of Engineering, Mumbai

---***---

Abstract - *Topic Modeling is a technique of unsupervised machine learning which is used in discovering topics that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is one of the most used algorithm for topic modeling. It considers that documents are mixture of topics and each topic is a mixture of different tokens or words. While considering many documents, one can think that the topics extracted by the LDA algorithm relate to all of the documents together. But if we consider only one text document, if we try to extract topics from it using LDA algorithm, we can say that these are the keywords of the text document as it summarizes the entire idea of the document in a concise form. This can be useful in summarization of the document. Summarization with respect to text is shortening of the text document such that it highlights all the important points of the text document. In this paper, we represent a LDA model which helps to identify the dominant topics in the text document, then identifies sentences that reflect these dominant topics and stiches them together to formulate a human readable summary.*

Key Words: Natural Language Processing, Text Summarization, Latent Dirichlet Allocation, Topic Modeling

1.INTRODUCTION

As there is an ever-increasing amount of data available, it has become important for extracting only important or only meaningful information from this data since every bit of data is not useful. This is where topic modeling and summarization can be of use. Due to the fact that the algorithm we used here is unsupervised, it eliminates the need for structured data to be provided to the model for it to work Motivation for developing this is to reduce the time required for reading or analyzing a text document. Text documents come in a variety of form including news reports, Research papers, legal documents and many more, the task can become tedious and some important information might slip out if not done carefully. The advantage with such a model doing the task is that one can decide the number of topics or points one wants to discover in the text. Based on that, the extraction would be done automatically, thus reducing the time required for the same task is done manually. Text summarization has two approaches namely Abstractive and Extractive. We have chosen the extractive summarization approach.

2. Literature Survey

Barde et al. [1] discusses various methods and tools used for topic modeling with their features and limitation. Some of the methods discussed are Vector Space Model (VSM), Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). Some tools discussed are Gensim, Standford topic modeling toolbox, MALLET, BigARTM.

Surabhi Adhikari et al. [2] discusses different methods that have been used for text summarization. Mainly, the paper discusses two methods- Abstractive (ABS) and Extractive (EXT) summarization. Also query based summarization is discussed. The paper mostly discusses about the structured based and semantic based approaches for summarization of the text documents. Various datasets were used to test the summaries produced by these models, such as the CNN corpus, DUC2000, single and multiple text documents etc.

Kenli Li et al. [3] use Latent Dirichlet Allocation (LDA) algorithm which is used to automatically generate text corpora topics, and applied to sentences extraction based multi-document summarization algorithms. The approach is to combine the traditional summary generation algorithm and the abstract generation algorithm based on deep learning.

David Alfred Ostrowski [4] uses Latent Dirichlet Allocation algorithm is used which is a generative probabilistic model for a collection of discrete data. Evaluating this technique from the perspective of classification as well as identification of noteworthy topics as it is applied to a filtered collection of Twitter messages. Experiments show that these methods are effective for the identification of sub-topics as well as to support classification within large-scale corpora.

Jinqiang Bian et al. [5] In their paper based on LDA Model, a new method of sentence-ranking is proposed. The method combines topic-distribution of each sentence with topic-importance of the corpus together to calculate the posterior probability of the sentence, and then, based on the posterior probability, it selects sentences to form a summary. Topic-distribution of each sentence represents the likelihood of sentence belonging to each topic and topic-importance represents the degree that the topics cover the significant portion of the corpus. The method highlights the latent topics and optimizes the summarization. Experiment results on the dataset DUC2006 show the advantage of the multi�document summarization algorithm proposed in the paper document

J. N. Madhuri et al. [6] proposes a system for summarizing documents using sentence ranking algorithms. Sentence are given weights and then ranked based on these weights. The sentences with the highest rank are selected in the summary. The sentences are ranked on the basis of the preprocessed text and the weights are given by frequency of terms divided by the total number of terms in the document.

Shohreh Rad Rahimi et al. [7] explores many methods of text mining and text summarization. Text summarization can be performed on the basis of various criteria. Some discussed criteria are based on output summary, based on details, based on contents, based on limitation, based on number of input texts and based on language acceptance. It also discusses various similarity measures which are used in text summarization

Hirohata et al. [8] presents automatic speech summarization techniques and its evaluation metrics. It mainly focuses on sentence extraction based summarization methods for making abstracts from some spontaneous presentations. Some metrics that have been discussed are summarization accuracy, sentence F-measure, ROUGE-n and some more.

Aditya Jain et al. [9] proposes a Neural Network based approach for text summarization. The paper proposes an approach to extract a good set of features followed by neural network for supervised extractive summarization. It assigns a predictive score to each sentence and the sentences with the highest predictive scores are added to the summary.

Liu Na et al. [10] present a system that use Latent Dirichlet Allocation topic model for multi�summarization. It extracts title and content for each document provided and creates a topic model for title and content. In the end it calculates sentence weights according to the topic model and forms a summary based on these sentence weights.

Mahsa Afsharizadeh [11] propose a technique of summarization which is query oriented. Most important sentences are extracted from the document based on a feature extraction process where some features like sentence length, normalized sentence length, sentence position in the document, topic frequency etc. are used. 11 unique features are extracted. Based on these 11 every sentence is scored and top ranked sentences are selected for creating the summary.

Shweta Ganiger and K.M.M Rajashekhariah [12] discuss implementation of some keyword extraction algorithms. These algorithms were used to find how effective they are when it comes to extracting important keywords from a document. The 3 algorithms discussed here are TF-IDF (Term frequency - Inverse Document Frequency), TextRank and RAKE (Rapid Automatic Keyword Extraction).

3. CONCLUSIONS

Topic modeling and topic summarization are two important tasks in natural language processing. With the help of LDA algorithm for extracting keywords, the need for structured data was eliminated which helped in reducing the time required for creating the summary. Also, the extraction of keywords or dominant topics can help in categorization purpose which can increase the scope of the project where suggestions can be made based on the similarity of different topics with the given document.

4. ACKNOWLEDGEMENT

We owe a sincere thanks to our college Atharva College of Engineering, especially our Head of Department, Dr. Suvarna Pansambal, our guide, Prof. Shweta Sharma for their kind cooperation and guidance which helped us in the completion of this project which would have seemed difficult without their motivation, constant support and valuable suggestions. Moreover, the completion of this research would have been impossible without the cooperation, suggestions and help of our family and friends.

5. REFERENCES

[1] Bhagyashree Vyankatrao Barde and Anant Madhavrao Bainwad, “An Overview of Topic Modeling Methods and Tools”, International Conference on Intelligent Computing and Control Systems, ICICCS 2017.

[2] Rahul, Surabhi Adhikari, Monika, “NLP based Machine Learning Approaches for Text Summarization”, Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020).

[3] Ying Zhong, Zhuo Tang, Xiaofei Ding, Li Zhu, Yuquan Le, Kenli Li, Keqin Li, “An Improved LDA Multi-Document Summarization Model Based on TensorFlow”, 2017 International Conference on Tools with Artificial Intelligence.

[4] David Alfred Ostrowski, “Using Latent Dirichlet Allocation for Topic Modelling in Twitter”, Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)

[5] Jinqiang Bian, Zengru Jiang, Qian Chen, “Research On Multi-document Summarization Based On LDA Topic Model”, 2014 Sixth International Conference on Intelligent Human�Machine Systems and Cybernetics

[6] J. N. Madhuri and R. Ganesh Kumar, “Extractive Text Summarization Using Sentence Ranking,” 2019 Int. Conf. Data Sci. Commun.

[7] Shohreh Rad Rahimi, Ali Toofanzadeh Mozhdehi, Mohamad Abdolahi, “An Overview on Extractive Text Summarizzation”, 2071 IEEE 4th International Conference on knowledge-Based Engineering and Innovation (KBEI)

[8] Hirohata, M., Shinnaka, Y., Iwano, K., & Furui, S. (n.d.). “Sentence extraction-based presentation

summarization techniques and evaluation metrics", Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.

[9] Jain, A., Bhatia, D., & Thakur, M. K. (2017), "Extractive Text Summarization Using Word Vector Embedding", 2017 International Conference on Machine Learning and Data Science (MLDS).

[10] Na, L., Ming-xia, L., Ying, L., Xiao-jun, T., Hai�wen, W., & Peng, X. (2014), "Mixture of topic model for multi-document summarization", The 26th Chinese Control and Decision Conference (2014 CCDC).

[11] Ebrahimpour-Komleh, H., Afsharizadeh, M., & Bagheri, A. (2018), "Query-oriented text summarization using sentence extraction technique.", 2018 4th International Conference on Web Research (ICWR).

[12] K. M. M. Ganiger, S., and Rajashekharaiah, (2018), "Comparative Study on Keyword Extraction Algorithms for Single Extractive Document", 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS