IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IMPROVING HYBRID SUMMARIZATION BY USING ABSTRACT AND EXTRACT MODEL

[1]Author: Gunuputi Venkata Veenadhari, M. Tech Scholar, Department Of Computer Science & Engineering, Vishnu Institute Of Technology, Bhimavaram, India

[2]Author: B.R Bharathi, Assistant Professor, Department Of Computer Science &Engineering, Vishnu Institute Of Technology, Bhimavaram, India

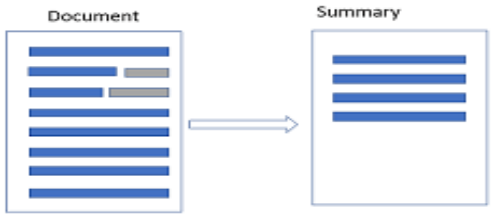
ABSTRACT: The demand for digitization has expanded in many areas, including inquiry procedures, as a result of the increasing acquisition of digitization over information storage and processing in our daily lives. To collect evidence from devices recovered from crime scenes in computer-related crimes, the best procedures must be adopted. Summarization has grown in popularity as a study subject in recent years. Researchers can produce effective findings for a variety of texts using different Natural Language Processing (NLP) methodologies. The Seq2Seq Architecture with RNN is used in the proposed work to carry out document summarizing tasks. Because of the abstract nature of the summary, This methodology can be improved upon and used continually to create strong summaries of longer materials, including legal papers.The outcomes show effective summary generation and ROUGE scores between 0.6 and 0.7.

Keywords—Natural Language Processing, Text summarization, Machine Learning, Tensor Flow, Seq2seq

I. INTRODUCTION

Natural Language is the prevalent form of human connection and communication. Such text is all around us in a variety of formats, including email messages, site content, SMS services, etc. Although the human language is constantly evolving, ambiguity is a part of it. Despite being skilled at explaining the specifics of speaking or understanding languages, performance was subpar when describing the formal rules of the same. NLP entails effectively "understanding" spoken language to provide understandable answers to it. Therefore, it is a two-way branch where naturally occurring language-based text will be part of the input and related text results will be displayed as output. Our world is currently inundated with enormous amounts of data. The demand for machine learning algorithms to automatically condense lengthy texts and produce precise summaries that can fluently pass the intended messages is driven by the vast volume of data that is being exchanged in the digital world. The manual and time-consuming summation of each case is a crucial component of the judiciary system. It takes many hours to create each of these reports, and it takes a lot of time and effort to analyse them so that they may be used as benchmarks. This task can be completed automatically thanks to machine learning algorithms. DL neural network models can be effectively taught and applied to various tasks with the correct set of data.

Develop coherent summaries of previously unseen case studies With the availability of such a module time, case study consumption for fieldwork and educational reasons will significantly decrease and
can be used to extract other hidden meaningful interpretations.



`

Fig.1: Example figure

1. LITERATURE REVIEW

2.1 Text summarization for massive data: A thorough investigation

The availability of a sizable amount of online data has made automatic text summarization necessary. Various texts summarising successes that have recently been put into practice are surveyed and discussed in this study[1]. A summary as a tool has a lot of potential in the legal realm, and the topic of summarising legal texts is given purposeful weight. The study begins with an overview of automatic text summarization, continues with a brief discussion of recent developments in extractive and abstractive text summary

approaches, conducts a literature review, and concludes with some suggestions for future research.

2.2 Text summarization from legal documents: a survey

Legal text processing is a crucial topic of study due to the vast amount of online legal material available. In this study, we make an effort to survey several text summarising methods that have been used recently. Given that it is one of the most crucial topics in the legal field, we have focused especially on the problem of summarising legal texts. We begin with an overview of the text summary before delving into extraction-based legal text summarization [2]. We also briefly discuss current developments in single- and multi-document summarization. We go over several datasets and metrics for summarising and contrasting the results of various techniques, first generally and then with an emphasis on legal material. We also list the key points of several summarising strategies and quickly go over a few computer programmes that are used to summarise legal texts. Finally, we offer some potential future study directions.

2.3 Extractive text summarization using word vector embedding

Text summarising is a current study area that aims to extract pertinent information from lengthy texts published in a variety of fields, including finance, news media, academia, politics, etc. Text summarization is a method of reducing the length of papers while keeping the key information. Through extractive and abstractive summarization, this is possible. In this article, we outline a strategy for extracting a strong set of features, which is then followed by a neural network for supervised extractive summarization[3]. The efficiency of the suggested method in comparison to other online extractive text summarizers is demonstrated by our experimental findings on the Document Understanding Conferences 2002 dataset.

2.4 Text summarization using unsupervised deep learning

To create a feature space from the term-frequency (tf) input, we offer techniques for extractive query-oriented single-document summarization. Local and international vocabularies are investigated in our experiments [4]. We examine the impact of introducing small random noise to local tf as the input representation of AE, and we suggest an ensemble of such noisy AEs that we refer to as the Ensemble Noisy Auto-Encoder (ENAE). The top phrases are chosen from an ensemble of noisy runs using an ENAE, a stochastic variation of an AE that adds noise to the input text. Every experiment in the ensemble adds a unique piece of randomly produced noise to the input representation.With this architecture, the deterministic feed-forward network used in the AE application is replaced with a stochastic runtime model. Using local vocabularies in the AE results in a more discriminative feature space and an average recall improvement of 11.2 percent, according to experiments. The ENAE still has room to grow, especially in terms of sentence construction. We do experiments on two distinct, text summarization-focused email corpora that are made available to the public to cover a wide range of topics and structures. The average ROUGE-2 recall for all tests was calculated using ROUGE, a fully automatic metric for text summarization.

2.5 Abstractive text summarization using LSTM-CNN-based deep learning

Abstractive Text Summarization (ATS) is the process of creating summary sentences by combining information from several source sentences and condensing it into a shorter representation while maintaining information content and overall meaning. For humans, manually summarising lengthy texts takes a lot of time and effort. In this paper, we offer an ATS framework (ATSDL) based on LSTM-CNN that can build new sentences by investigating more precise pieces than sentences, specifically semantic phrases [5]. ATSDL is made up of two basic steps, the first of which extracts phrases from source sentences and the second of which uses deep learning to produce text summaries. This distinguishes it from existing abstraction-based techniques. Our ATSDL framework surpasses the cutting-edge models in terms of both semantics and accuracy, according to experimental results on the CNN and DailyMail datasets.

2.6 Diversity-driven attention model for query-based abstractive summarization

The objective of an abstract summary is to provide a condensed and coherent shorter version of the material that covers all the important information. Contrarily, a query-based summary emphasises the items that are pertinent to the context of the given query. In machine translation, extractive summarization, dialogue systems, etc., the encode-attend-decode paradigm has had significant success. Its flaw, though, is that it produces a lot of phrases that are repeated. In this paper, we propose a model for the encode-attend-decode task-based query-based summarization task with two important additions: a query attention model (in addition to the document attention model) that learns to focus on different portions of the query at different time steps (instead of using a static representation for the query); and (ii) a query attention model that learns to focus on different portions of the query at different time steps.The issue of phrases recurring in the summary is addressed by (ii) a new diversity-based attention model [6]. We publish a brand-new query-based summarising dataset based on discussion media to make it possible to evaluate this technique. Our tests demonstrate that the suggested model outperforms traditional encode-attend-decode models with a gain of 28 percent (absolute) in ROUGE-L scores after these two modifications.

3.IMPLEMENTATION

The Gigaword dataset and CNN/DM dataset, which are both parts of the Harvard NLP project, are where the data is gathered. The datasets from CNN/DM were used; they included news pieces and the handwritten summaries that went with them. The news stories cover a wide range of subjects, including business, politics, sports, and finance[7]. Figure 1's flowchart outlines the steps that DL neural networks like the seq2seq architecture take to import, clean, preprocess, train, and test the data.

Importing the necessary libraries for the Python programme is the first step in streamlining some of the efforts. The dataset import comes next. The dataset needs to be preprocessed to remove extraneous input and only keep

recognised characters and meaningful words. Because the resulting text data is so extensively unstructured, the dataset needs to be cleaned. The terms could be truncated, misspelt, not standard, contain dates and other numerical information, and use legal jargon[8]. At this stage, lemmatizing or stemming words can also be useful. This is the sole focus of the preprocessing stage. Record the metrics that are used to assess the success of our model after the training and testing phases, such as the ROUGEScore.

MODULES:

1. Data Cleaning:

This module deals with cleaning the dataset, which eliminates noise from unstructured text[11]. This module will be called when loading training data, and test data, and to display a sample of the cleaning output.

2. Building Dictionary and Data Preprocessing:

The following steps are used to create the two dictionaries stated above[10]. The seq2seq method requires these stages in terms of abstraction. However, the creation of a lexicon is not necessary for ETS.

3. Building the seq2seq Abstractive Text Summarizer (ATS) model.

A recurring neural network (RNN), configured as an encoder/decoder architecture known as the seq2seq model, will be constructed in the part that follows. The seq2seq will be designed as a bidirectional structure, where the RNN cell will turn into an LSTM cell, and will also contain a beam search idea and an attention mechanism for improved encoder/decoder interaction[15].

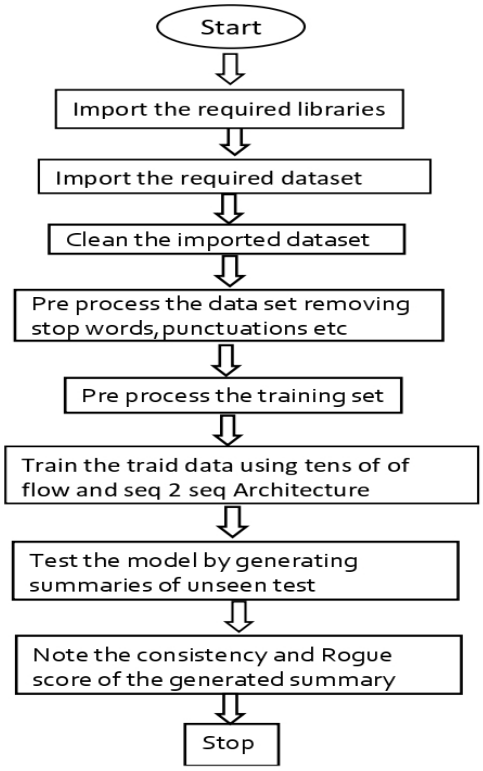


Fig.2: System architecture

4. METHODOLOGY

NLP:

Natural language processing helps computers communicate with humans in their language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important[9].Natural Language Processing (NLP) Natural language processing strives to build machines that understand and respond to text or voice data—and respond with text or speech of their own—in much the same way humans do.

When it comes to understanding language, natural language processing (NLP) combines the fields of data science, computer science, and linguistics[13]. It is advantageous to businesses because it simplifies human language for autonomous machine analysis. NLP's five phases include:lexical (structure) analysis

1. parsing
2. semantic analysis
3. discourse integration
4. pragmatic analysis.

Human language is broken down into pieces in natural language processing so that sentence structure and word meaning can be examined and understood by one another[14]. This enables computers to read and comprehend spoken or written text like that of people.

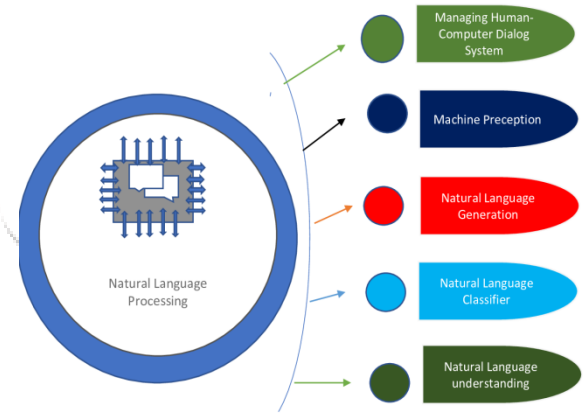


Fig.3: NLP processing

Advantages of NLP:

- NLP enables users to ask inquiries about any topic and receive a quick answer in a matter of seconds.
- NLP provides precise responses to questions, which implies it withholds unneeded and undesirable information.
- Computers can now speak in human languages thanks to NLP.
- It saves a lot of time.
- The majority of businesses employ NLP to increase the effectiveness of documentation procedures, the accuracy of documentation, and the ability to extract information from huge databases.

5. EXPERIMENTAL RESULTS

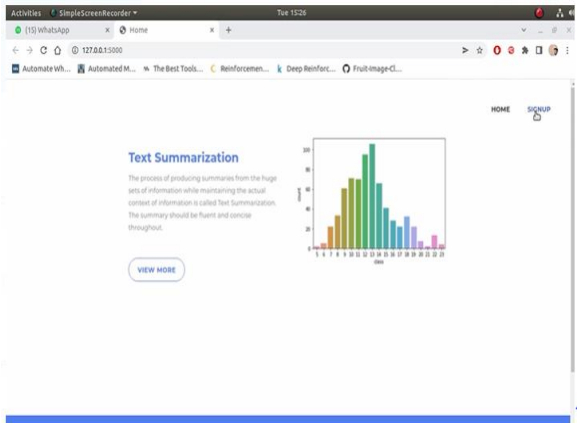


Fig.4: Home screen

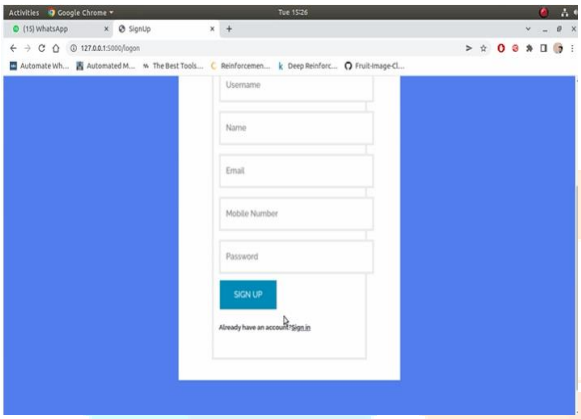


Fig.5: Signup

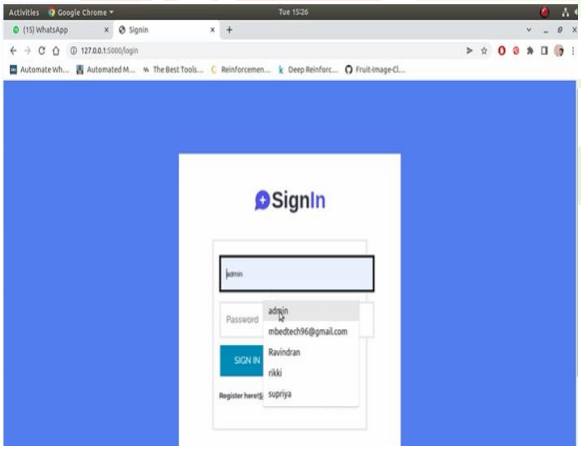


Fig.6: Sign-in

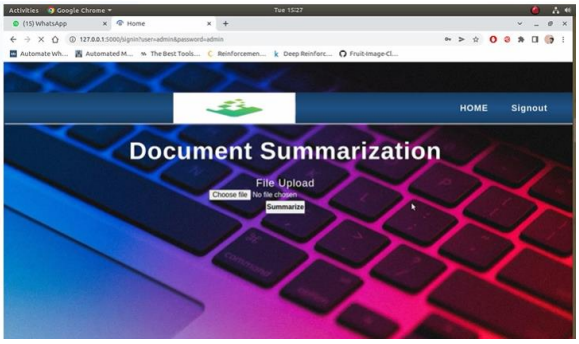


Fig.7: main screen

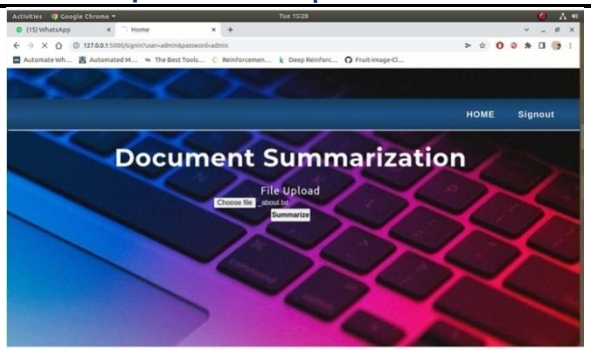


Fig.8: File uploading

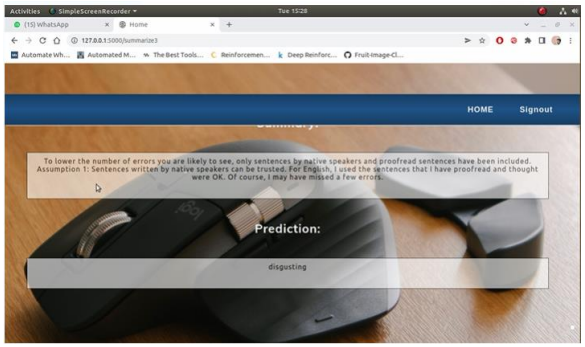


Fig.9: Prediction

6. CONCLUSION

In this study, an abstractive summarization platform is created to incorporate the potential for using legal or judicial data. Due to the lack of publicly available information regarding such sensitive matters, it was decided to use data of a similar sort, i.e. information gathered from the news, for such processes. An extractive module is also supplied in addition to the abstraction-forming module, allowing us to compare the two approaches. The seq2seq architecture is used, and additional attempts may be made using various models and architectures. The identical process of word embeddings followed by an encoder-decoder with an attention mechanism is used to incorporate the summarization findings into Hindi language and news articles.

7. FUTURE WORK

Despite only having experience with the seq2seq architecture, pointer generator networks are another potential design that can be used for summarization. The network train can be streamlined on various data and produce domain-specific summaries when new data becomes available.

REFERENCES

[1] Gupta, Vanyaa, Neha Bansal, and Arun Sharma. "Text summarization for big data: A comprehensive survey." In International Conference on Innovative Computing and Communications, pp. 503-516. Springer, Singapore, 2019.

[2] Kanapala, Ambedkar, Sukomal Pal, and Rajendra Pamula. "Text summarization from legal documents: a survey." Artificial Intelligence Review 51, no. 3 (2019): 371-402.

[3] Jain, Aditya, Divij Bhatia, and Manish K. Thakur. "Extractive text summarization using word vector embedding." In 2017 International Conference on Machine Learning and Data Science (MLDS), pp. 51-55. IEEE, 2017.

[4] Yousefi-Azar, Mahmood, and Len Hamey. "Text summarization using unsupervised deep learning." Expert Systems with Applications 68 (2017): 93-105.

[5] Song, Shengli, Haitao Huang, and Tongxiao Ruan. "Abstractive text summarization using LSTM-CNN based deep learning." Multimedia Tools and Applications 78, no. 1 (2019): 857-875.

[6] Nema, Preksha, Mitesh Khapra, Anirban Laha, and Balaraman Ravindran. "Diversity driven attention model for query-based abstractive summarization." arXiv preprint arXiv:1704.08300 (2017).

[7] Paulus, Romain, Caiming Xiong, and Richard Socher. "A deep reinforced model for abstractive summarization." arXiv preprint arXiv:1705.04304 (2017).

[8] Liao, Pengcheng, Chuang Zhang, Xiaojun Chen, and Xiaofei Zhou. "Improving Abstractive Text Summarization with History Aggregation." arXiv preprint arXiv:1912.11046 (2019).

[9] Chopra, Sumit, Michael Auli, and Alexander M. Rush. "Abstractive sentence summarization with attentive recurrent neural networks." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93-98. 2016..

[10] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).

[11] Zhang, Yizhe, Dinghan Shen, Guoyin Wang, Zhe Gan, Ricardo Henao, and Lawrence Carin. "Deconvolutional paragraph representation learning." In Advances in Neural Information Processing Systems, pp. 4169-4179. 2017.

[12] Cohan, Arman, and Nazli Goharian. "Revisiting summarization evaluation for scientific articles." arXiv preprint arXiv:1604.00400 (2016).

[13Gupta, Vanyaa, Neha Bansal, and Arun Sharma. "Text summarization for big data: A comprehensive survey." In International Conference on Innovative Computing and Communications, pp. 503-516. Springer, Singapore, 2019.

[14] Ferilli, Stefano, and Andrea Pazienza. "An abstract argumentation-based approach to automatic extractive text summarization." In Italian Research Conference on Digital Libraries, pp. 57-68. Springer, Cham, 2018.

[15] Sehgal, Sunchit, Badal Kumar, Lakshay Rampal, and Ankit Chaliya. "A modification to graph-based approach for extraction based automatic text summarization." In Progress in advanced computing and intelligent engineering, pp. 373-378. Springer, Singapore, 2018.