TEXT SUMMERIZATION USING DEEP LEARNING

Mr.P.V. HariPrasad[1], Annam Komali Radha Sri Mounika[2], Ravuri Vasanthi[3], Suri Sailaja[4], Shaik Areef[5],[1]Associate Professor,[2,3,4,5]Students

Dept. Of Computer Science & Engineering, Dhanekula Institute of Engineering and Technology, A.P., India.

ABSTRACT: Many applications now have the ability to intelligently classify text thanks to the emergence of deep learning techniques. Today, research is focused on developing automated text summarization utilizing deep learning techniques. Traditional approaches to extractive text summarization were primarily reliant on human-engineered features in the early days. It is, however, a time-consuming chore. A data-driven technique was utilized to construct summaries using deep learning in this article. Paraphrasing techniques are used in the suggested approach to determine whether or not a sentence should be included in the summary.

Keywords:
deep learning, text summarization, data-driven approach, support vector machines (LSTM).

1. INTRODUCTION

Summarization has been done manually by humans for a long time. The amount of information available on the internet and from other sources is steadily expanding. Text summary is crucial to combat the information overload in order to solve this problem. As a matter of thumb, text summarization helps to keep the text data organized by adhering to certain norms and laws. As an example, one can extract a summary from a single document or multiple papers in order to extract a specific piece of content. According to Wikipedia, "text summarizing" is the process of extracting content from a textual document and delivering it in an easily digestible form to the user or application. Text interpretation, which is a prerequisite for automatic summarization, comes with its own set of complexities, including a wide range of text formats, phrases, and editions. From natural language processing, statistical and machine learning approaches, text summarization researchers have taken on this topic. Text analysis is the essential difficulty in determining which parts of the text are most important.

Abstractive summarizing and extractive summarization are the two types of text summarization. The technology of Natural Language Processing (NLP) is utilized in abstractive summarizing for parsing, word reduction, and text summarization. Low-cost NLP is currently lacking in precision. As compared to abstractive summarization, extractive summarization is more versatile and takes less time. If you regard all the sentences as a matrix of some kind, extractive summarization considers just those that are necessary or important. It is possible to represent an item as an n-dimensional feature vector. Text summarization based on the extraction approach aims to select the most relevant sentences for a user depending on their specific needs.

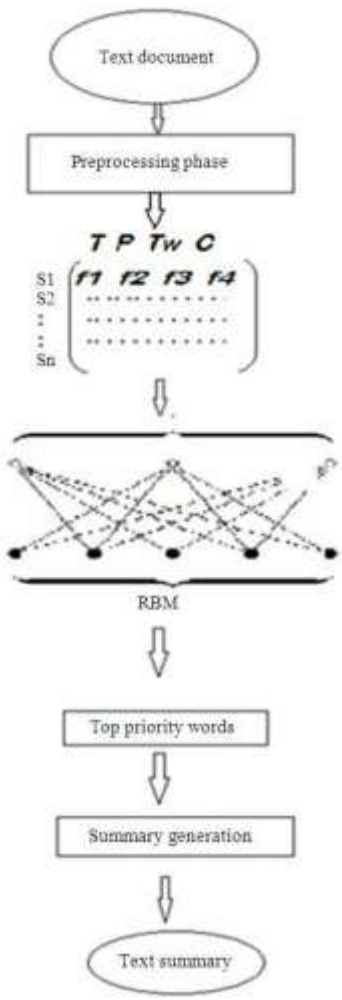

Fig.1: Example figure

Abstractive and extractive summarization methods are the most frequent. Various other characteristics have been used in the past to create classification groups such as single vs. multi-document classification and mono-lingual vs. multi-lingual summarization. Summary sentences can be derived from text using extractive text summarizing, which identifies and delivers to the user the most significant passages from a document. In each paragraph, the best-scoring sentences are taken and used as the basis for the extracts. While abstractive summaries need the creation of phrases and words and the subsequent organization into intelligible sentences in order to convey the gist of the text, this method is much simpler. This is a considerably more complex task because it requires a high level of natural language processing. The goal of this work is to generate extractive summaries utilizing a data-driven method, using deep learning techniques, to achieve text summarization. One way to do this is to sort through the content and come up with the most valuable phrases and ideas from it. There are several techniques to achieving the objective of summarizing a text in the same way as humans often do after reading a piece of writing. It uses generative methodologies that can generate meaningful sentences while preserving the original text's semantics. As a result of the success of Deep Learning in solving this challenge, numerous novel ways have been offered.

2. LITERATURE REVIEW

2.1 Text summarization techniques: a brief survey:

In recent years, the volume of text data from various sources has increased dramatically. This volume of literature contains significant information and knowledge that must be appropriately summarized in order to be of any use. The most common methods for summarizing large amounts of text are outlined in this article. When summarizing data, we look at how well the various strategies work as well as how they might be improved upon.

2.2 Deep Indian language paraphrase detection:

To handle the challenge of paraphrase identification in English and Indian languages, we used Convolutional Neural Networks (CNNs) and Recurrent Neural Network (RNNs) (RNN). With POS taggers, dependency parsers, and more, machine learning approaches used English language resources in the past. Lack of resources for paraphrase detection in Indian languages has been a deterrent to the use of this technology. Deep learning can overcome the disadvantages of standard machine learning techniques. In this study, a basic CNN that uses word embeddings as input, a simple CNN that uses WordNet scores as input, and an RNN-based approach using both LSTM and bi-directional LSTM were all presented.

Incorporating sentiment infusion into an abstract text summarizing method, Atssi

It is feasible to condense a large amount of text into a smaller amount of information by using text summarization. It has become more vital to analyze social media data in order to find information and use it for the benefit of many apps and people... Natural Language Processing and Text Mining communities have been interested in automatic summary for a few years now, especially when it comes to summarizing opinions. Decisions are made based on people's opinions in society. When it comes to making decisions, an individual or business relies heavily on the input and suggestions of others. Summaries of duplicated opinions are generated using a graph-based technique that uses sentiment analysis to integrate the assertions in this study. An abstraction-based summation of the text is used to construct these summaries.

Evaluative text summary using multiple documents:

This paper compares and contrasts two methods for distilling evaluative arguments. A sentence extraction strategy and a language generation approach are both possible approaches. User research found that both approaches performed equally well in terms of quantitative performance. In terms of their performance, however, we feel that they complement each other effectively. To summarize evaluative arguments effectively, we believe a method must successfully combine the two approaches.

2.5 A survey of text summarizing methods that include extraction and abstraction:

Using text summarization, the original text can be reduced in length while still retaining its original meaning and content. Manually summarizing lengthy texts is nearly impossible for us mere mortals. Extractive and abstractive summarization are two types of summarizing approaches. Using extractive summarizing, the most important sections of a long document can be condensed into a single sentence, paragraph, or section. According to the statistical and linguistic qualities of sentences, they are ranked in importance. Abstractive summaries aim to convey the essence of a piece of writing in a concise manner. In order to better understand the text, it employs linguistic methods of analysis and interpretation before creating a new, condensed version that summarizes the key points of the original. The process of condensing a large amount of text into This research looked on extraction methods.

Fig.2: Extractive technique.

2.6 Probabilistic document modeling for syntax removal in text summarization:

Statistics-based automated text summarizing continues to beat more complex summarizing methods. Before usable frequency statistics can be computed, semantically meaningful terms must be separated from low-content function words. When syntactical phrases are regarded semantically important, but content-related words are removed, stop word lists can result in both under coverage and over coverage. We suggest employing a generative probabilistic modeling technique where syntax phrases are learned directly from the data using a Hidden Markov Model and deemphasized when computing term frequency statistics for constructing content distributions for statistical multi-document summarization. Using the ROUGE metric, we compare our approach to the stop word-list strategy as well as to the POS-tagging approach. Our method exhibits better coverage on both datasets.

Extracting sentences from a variety of documents in order to provide a concise summary

In order to improve multi-document summarizing, text extraction approaches that expand on single-document summarizing techniques can be used to gather information about the complete document collection and its connections. When summarizing multiple documents rather than just one, the issues of compression, speed, redundancy, and passage selection are critical to the generation of effective summaries. It is our goal to solve these problems by employing domain-independent techniques based on fast and statistical processing, a measure to reduce redundancy and maximize selection diversity, and a modular framework that allows easy parameterization for different genres, corpora characteristics, and user requirements.

Sentence annotation can be used to improve the development of semantic summaries from a variety of publications.

Using document summarization, you may speed up the process of going through a lot of material. A technique called as sentence extraction can be used to extract the most relevant and crucial sentences from a manuscript. Thus, a more effective method for picking the most important sentences is required. This study employs semantic analysis at the phrase level to produce both an initial and an updated summary of the data. In order to select the most important information from a large amount of text, each sentence is annotated with aspects, prepositions, and named entities. Each word in a document is weighted according to the term synonym concept frequency-inverse sentence frequency metric (TSCF-ISF). For each sentence, its score is calculated and the most highly ranked sentences are used to build up the summary. An intrinsic measurement used to evaluate the quality of a summary based on the coverage between a machine summary and a human summation is Precision and Recall. A summary's precision is measured by its accuracy, whereas recall is measured by its completeness. To compare our findings, we employed the Semantic Summary Generation method and the Lex Rank Update summarizing job. The ROUGE-1 metric is used to evaluate summaries created by computers.

3. IMPLEMENTATION

An effective text summarization technique is one that captures the most important points and conveys them in a succinct and accurate manner. Techniques like page rank algorithms and other natural language processing algorithms are used to summarize text. While these algorithms accomplish the goal of text summarization, they are unable to generate new phrases that are not already in the document. Also, they may have grammatical mistakes. With the use of deep learning, we can overcome this challenge. Using deep learning, text summarization can be done quickly and efficiently. Summaries created using deep learning methods can include new terms and sentences, as well as grammatical errors.

THE SYSTEM IN USE:

For naive bayes classifiers, decision trees, clustering, and hidden Markov models, the feature vector was handcrafted based on some of the properties listed above. A number of research have also studied genetic algorithms, which are algorithms that express optimization problems and solve them using strategies observed in nature in natural selection procedures such as mutations, cloning, and cross-overs, among others

DISADVANTAGES:

Optimization problems can be modelled and solved utilizing strategies that have been seen in nature by machine learning algorithms.

The following is the proposed system:

A data-driven technique was utilized to construct summaries using deep learning in this article. Paraphrasing techniques are used in the suggested approach to determine whether or not a sentence should be included in the summary.

ADVANTAGES:

As a result of the success of Deep Learning in solving this challenge, numerous novel ways have been offered.

Preparation of data:

Before we start developing the model, it's critical that we complete some basic preprocessing tasks. It's a bad idea to work with filthy and untidy text data. In other words, we'll get rid of anything in the text that isn't directly related to solving our problem. Our data will undergo the following preprocessing tasks:

• Lowercase all letters and numbers.

"s" should be removed.

Text inside the parenthesis should be deleted ( )

Empty out punctuation and special characters in your writing.

Do away with filler words

• Cut down on the number of words.

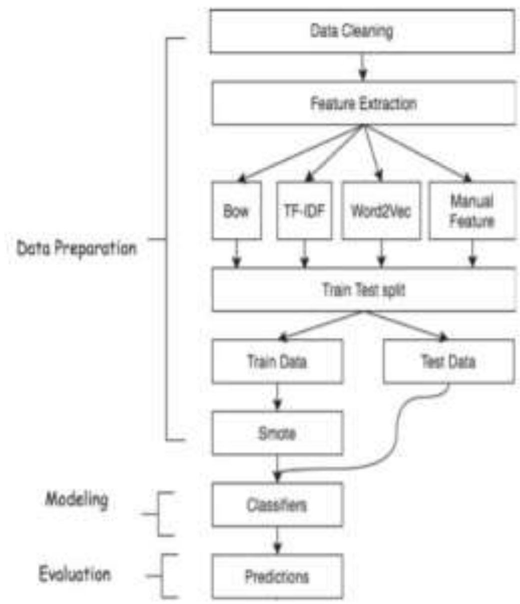

Fig.3: System architecture

MODULES:

1. Data Collection: Emotional speech dataset for humans. And the dataset's structure. how to connect the dots between various aspects. Core features, as well as the dataset as a whole, are shown on a graph. A third of the dataset is used to train the algorithms, while the remaining third is used to test them. A representative sample must be obtained by ensuring that each class in the whole dataset is represented in about the same proportion in both the training and testing datasets. The different ratios of the datasets used for training and testing in the paper.

There may be inconsistencies in the data that was collected because of missing values. Preprocessing data is necessary to improve the algorithm's performance. Also, variable conversion is required to remove outliers from data. We utilize the map function to get around these problems.

Third, model selection: Machine learning is all about finding the best way to anticipate or recognize patterns in order to provide the best results. Algorithms for machine learning (ML) look for and learn from patterns in data. Each time an ML model is used, it will get better and better. To test a model's performance, you must first divide the data into training and test sets. Prior to training our models, the data was split into a training set of 70% and a testing set of 30%. After that, it was critical to tie our model's predictions to a variety of performance criteria.

When the planned system is put through its paces on the test bench, the results are confirmed. Evolutionary analysis relates to the description and modeling of regularities or trends for objects that change their behavior through time. Confidence matrix metrics include Precision and Accuracy. For the purpose of creating a predictive model using an ordinary logistic regression neural network, these are the most critical features to consider.

4. ALGORITHMS

Long-term memory RNN architecture is used for deep learning (LSTM). LSTM neural networks, unlike feedforward networks, have feedback connections. A single data point (such as an image) isn't the only thing that can be processed when dealing with data (such as speech or video). [2] Handwriting recognition,[3] speech recognition,[4] and anomaly detection in network traffic or IDSs can all benefit from the LSTM (intrusion detection systems). Cell, input gate, output gate, and forgetting gate make up a typical LSTM. The cell stores values over time, and the three gates govern the flow of information in and out of the cell.

LSTM networks are useful for classification, processing, and prediction because to the fact that significant occurrences in a time series can have lags of undetermined duration. The vanishing gradient problem may arise when training typical RNNs. This is why LSTMs were created. Gap-insensitive sequence learning methods, such as LSTMs, have an advantage over RNNs, hidden Markov models, and other sequence learning algorithms.

Training: This method uses gradient descent and backpropagation to calculate the gradients needed to adjust each LSTM network weight in proportion to the derivative of error (at the LSTM network output layer) with respect to the respective weight during the optimization process, which can then be used for unsupervised training of an RNN. Error gradients quickly disappear as the time latency between key events rises in gradient descent for ordinary RNNs. It's a concern of mine. LSTM units convey error values backwards from the output layer, while the error itself remains in the LSTM unit's cell. Errors in the "error carousel" gradually teach the LSTM units to cut off the value they're getting.

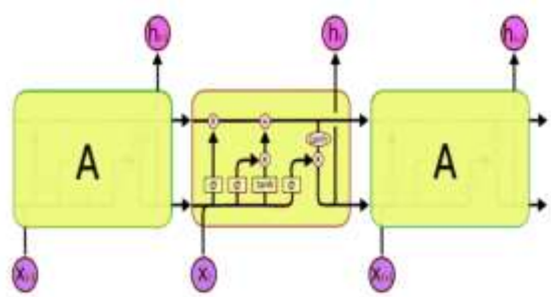

Fig.4: LSTM model

5. EXPERIMENTAL RESULTS

Fig.5: Signing screen

Fig.6: Admin login screen

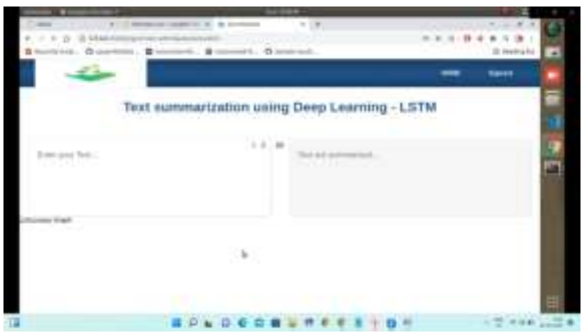

Fig.7: Home screen

Fig.8: Text data screen

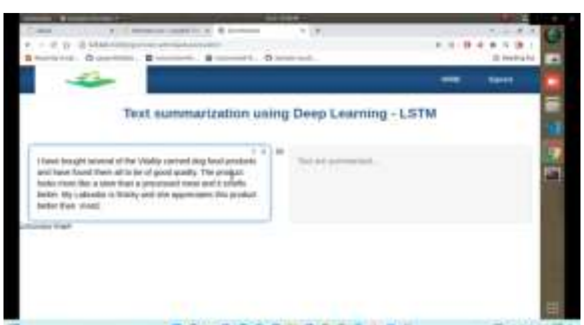

Fig.9: Input text

Fig.10: Prediction result

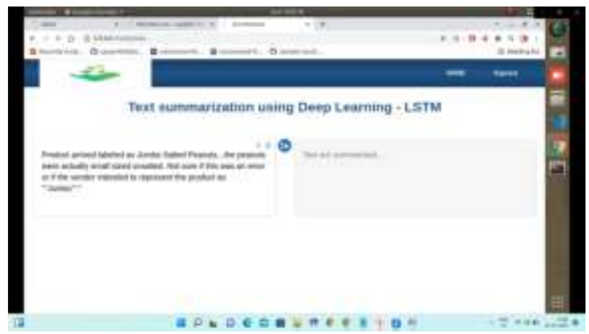

Fig.11: Prediction result

6. CONCLUSION

When summarizing multiple texts, feature extraction can be increased from sentences to paragraphs and even entire pages, according to this approach. Recurrent Neural Networks and its attention mechanisms could potentially draw attention. By taking into account the contributions of several previous sentences, LSTMs and other RNN modifications have been shown to be good models in situations where we need some sort of memory. As such, their use in this approach could be extremely useful in classifying a sentence as belonging to a summary." text-to-speech (speech bubble) Using attention mechanisms and gated memory cells, individuals may select which inputs to pay more attention to. Human scoring may be a useful metric for evaluating the summaries that have been prepared, as the collection and development of data sets is a critical component of this effort. There must be a sufficient amount of data to enable data-driven learning to be effective.

This is the seventh thing to consider when looking into the future:

The ability of a human to understand and extract significant characteristics from a written text in order to describe the document in our own words is typically quite good. Automatic ways to text summarizing are crucial in today's world, where there is an abundance of data but a lack of labor and time to examine it.

REFERENCES

[1] U. Hahn, I. Mani, "of Automatic Researchers are investigating summarization tools and methods that", IEEE Computer33. 11, pp. 29-36, November 2000.

[2] E. Lloret, M. Palomar, "Text summarization in progress: a literature review" in Springer, Springer, pp. 1-41, 2012

[3] K. Spärck Jones, "Automatic summarizing: The state of the art", Information Processing & Management, vol. 43, pp. 1449-1481, nov 2007

[4] A. Khan, N. Salim, "A review on abstractive summarization methods", Journal of Theoretical and Applied Information Technology, vol. 59, no. 1, pp. 64- 72, 2014

[5] E. Baralis, L. Cagliero, A. Fiori, P. Garza, "MWI-Sum: A Multilingual Summarizer Based on Frequent Weighted Itemsets", ACM Transactions on Information Systems, vol. 34, no. 1, pp. 1-35, 2015

[6] M. Kamal, "Text Summarization in Bioinformatics", 15th International Conference on Computer and Information Technology, pp. 592-597, 2012

[7] Darling, W.M. and F. Song, 2011. Probabilistic document modeling for syntax removal in text summarization. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, (CL' 11), ACM Press, Stroudsburg, PA., pp: 642-647.

[8] Goldstein, J., V. Mittal, J. Carbonell and M. Kantrowitzt, 2000. Multi-document summarization by sentence extraction. Proceedings of the NAACL-ANLP Workshop on Automatic Summarization, (WAS' 00), ACM Pres, Stroudsburg, PA, USA., pp: 40-48. DOI: 10.3115/1117575.1117580

[9] Kogilavani, A. and P. Balasubramanie, 2012. Sentence annotation based enhanced semantic summary generation from multiple documents. Am. J. Applied Sci., 9: 1063-1070. DOI: 10.3844/ajassp.2012.1063.1070

[10] Mani, I., 2001a. Automatic Summarization. 1st Edn., John Benjamins Publishing, Amsterdam, ISBN-10: 9027249865, pp: 285.

[11] Mani, I., 2001b. Recent developments in text summarization. Proceedings of the 10th International Conference on Information and Knowledge Management, Nov. 06-11, ACM Press, McLean, VA, USA., pp: 529-531. DOI: 10.1145/502585.502677

[12] Ou, S., C.S.G. Khoo and D.H. Goh, 2008. Design and development of a concept-based multi-document summarization system for research abstracts. J. Inform. Sci., 34: 308-326.

[13] Patil, K. and P. Brazdil, 2007. Text summarization: Using centrality in the pathfinder network. Int. J. Comput. Sci. Inform. Syst., 2: 18-32

[14] Sharef, N.M., A.A. Halin and N. Mustapha, 2013. Modelling knowledge summarization by evolving fuzzy grammar. Am. J. Applied Sci., 10: 606-614. DOI: 10.3844/ajassp.2013.606.614

[15] Zhang, Y., D. Wang and T. Li, 2011. iDVS: An interactive multi-document visual summarization system. Mach. Learn. Know. Disco. Databases, 6913: 569-584. DOI: 10.1007/978-3-642-23808- 6_37