International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Text Summarization Using Deep Learning

R Vanisha[1],Shashani Akanksha[2], Vishnumolakala Venkata Mahalakshmi[3] ,R Sirisha[4]

Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

Abstract. In the big data era, there has been an explosion in the amount of text data from a variety of sources. This volume of text is an inestimable source of information and knowledge which needs to be effectively summarized to be useful. There are plenty of text material available on the internet. Earlier traditional approaches for extractive text summarization have been heavily dependent on human engineered features. However, it is a laborious and tedious task. In this paper, a data-driven approach has been used to generate extractive summaries using deep learning. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it. Thus, the need for a solution emerges, that transforms this vast raw information into useful data which a human brain can understand. One such common technique in research that helps in dealing with enormous data is text summarization. We must first comprehend what a summary is before moving on to text summarization. A summary is a text created from one or more texts that delivers key information from the source material in a condensed style. The most significant benefit of adopting a summary is that it cuts down on reading time. Text summarization is the process of extracting the most important meaningful information from a document or set of related documents and compressing it into a shorter version while retaining its overall meanings. Automatic summary is a well-known method for refining a document's important points. It works by providing a reduced version of the text that preserves significant information. There are two types of text summarization methods: Extractive and Abstractive.

Keywords: Automatic Text Summarization, Extractive, Abstractive, Summary, Comprehend.

1. Introduction

1.1 About Project

Post the advent of the World Wide Web the amount of data and information accessible has increased tremendously. The extent of information is such that it has now become practically impossible for any single entity to process all the data and more so summarize it. Consumers are not interested in reading a long piece of text and hence, tend to skip important portions of the text frequently. This has increased the demand of automation of text summarization. The technique of text summarization can be classified into two major categories, abstractive and extractive. However, in past other classification categories have also been defined on various other parameters such as single vs. multi document classification and mono-lingual vs multi-lingual summarization. Abstractive text summarization approaches aim to

achieve the task of generating summaries, which present the gist of the text, generally the way humans do post reading any text. It uses generative approaches which can generate meaningful sentences and at the same time preserve the semantics of the original text fed to them. This is viewed as a difficult problem to solve and many new approaches are being proposed for it post the boost gained by Deep Learning.

Extractive text summarization is a relatively simpler and robust way of generating summaries by selecting salient sentences from the given text and presenting it to the user. Each sentence is attached with score and highest scored sentences are chosen to be the part of the extract. This is relatively simpler in contrast to abstractive summaries which involve generating phrases and words and organizing them to form meaningful sentences and at the same time presenting an interpreted gist of the text. It would involve high degree of natural language processing and hence, is a much more difficult task. This work aims to achieve the goal of text summarization by generating extractive summaries using data driven approach, through deep learning techniques. This includes processing the bunch of text and generating list of sentences which might be the most useful and contain the major gist of the text. Although humans generally do not perceive summaries as sentences extracted verbatim from text but rather try to summarize it in a format that conveys the same meaning as the given text. Though extractive summaries are not very intuitive, they serve the purpose of giving the most important piece of text, which to a great extent can provide an idea of what the text is about and at the same time, certain sentences which can be used to quote or refer to for some other purpose. Hence, this paper tries to capture the essence of extractive summaries using paraphrase detection. Another motivation that drives this work is to try out this approach of extract generation using data driven approaches for Indian languages.

1.2 Objectives of the Project

1. The main objective of a text summarization system is to identify the most important information from the given text and present it to the end users.

2. Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning.

3. Automatic text summarization aims to transform lengthy documents into shortened versions, something which could be difficult and costly to undertake if done manually.

4. Implementing summarization can enhance the readability of documents, reduce the time spent in researching for information, and allow for more information to be fitted in a particular area.

5. The approach is to identify the important sections, interpret the context and reproduce in a new way. This ensures that the core information is conveyed through shortest text possible. Note that here,

the sentences in summary are generated, not just extracted from original text.

6. A summary can be defined as a text that is produced from one or more texts, that contains a significant portion of the information in the original text and that is no longer than half of the original text .

7. Text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user.

1.3 Scope of the Project

Text summarization is defined as the process of refining the most useful information from the source document to provide an abridged version for the specific task. This paper focuses on developing a multi-document text summarization approach.

Text summarization has become a significant part of the life of researchers, students, and people who go through huge textual documents every day. With the development of technology and AIs, it is possible that one day, automatic text summarization will be just as good and clear as manual text summarization. Automatic text summarization is a well known solution for the problem of information overload. Text summaries are essential guide to the user to form an opinion to the relevance of the document. In other words Summaries save time for Internet users in their daily work and produce reader-friendly summary. The application of text summarization reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area**.**

Traditional print publications are also available online. It's impossible for anyone to keep track of recent publications even if limited to one domain. This is where text summarization can help.

2. Literature Survey

2.1 Existing System

Approaches based on machine learning also began with training simple classifiers like naive-bayes classifiers, decision-trees, clustering and hidden markov models with the feature vector hand-engineered based on some of the parameters mentioned above. Certain work also explored the use of

Genetic algorithms, these are algorithms which model optimization problems and solve them by using techniques often observed in nature in natural selection procedures like mutation, cloning, cross-overs.

Text Rank Algorithm :

Is a graph based ranking model for text processing which can be used in order to find the most relevant sentences in text and also to find keywords.

Latent Semantic Analysis :

LSA is for computer modelling and simulation of the meaning of words and passages by analysis of representative corpora of natural text.

Existing systems include text summarizations. Text summarization has traditionally been focused on text input. Various methods for summarization were proposed which include extraction-based, abstraction-based, maximum entropy-based and aided summarization. These methods use linguistic and natural language processing techniques. The steps followed are interpretation of the source text to obtain a text representation, transformation of the text representation into a summary representation, and finally, generation of the summary text from the summary representation.

Most of the existing methods for document summarization use the information present in the given document. These methods have explored various techniques for summarization process based on the assumption that the specified document is independent of any other documents. But the few topic related documents can be helpful for producing summary. The alternate approach for document summarization will make use of neighbor documents which provide the neighborhood knowledge that makes the summarization process efficient. Comments left by readers on Web documents contain valuable information that can be utilized in various information retrieval tasks including document search, visualization and summarization. In this project, we study the problem of comments-oriented blog summarization and aim to summarize a blog post by considering not only its content, but also the comments left by its readers. Much existing research on blog summarization focused on posts only, ignoring their comments. Reading comments does change one's understanding about blog posts. In this project, we aim to extract representative sentences from a blog post that best represent

2.2 Proposed System

A data-driven approach has been used to generate extractive summaries using deep learning. Approach proposed uses paraphrasing techniques to classify sentences as a candidate sentence for inclusion in summary or not. The proposed solution first derives representative words from comments and then selects sentences containing representative words. Significant differences between the sentences labelled before and after reading comments can be observed. By considering these

comments, the generated summary can better capture the input from the readers, as opposed to the author of the blog only. That is, a comments-oriented summary provides balanced views from both author and readers. Second, most websites present a blog post together with its comments. Also readers treat comments associated with a post as an inherent part of the post. A comment -oriented summary hence better matches one's understanding of the blog post as readers often read the post together with its comments.

Long Short – Term Memory (LSTM) is an artificial Recurrent Neural Network (RNN) architecture used in the field of deep learning**.**

LSTM is a novel recurrent network architecture training with an appropriate gradient-based learning algorithm.The attention mechanisms is a part of neural architecture that dynamically highlight relevant features of the input data which in the Natural Language processing (NLP) is typically a sequence of textual elements.LSTM is designed to overcome error back-flow.

3. Proposed Architecture

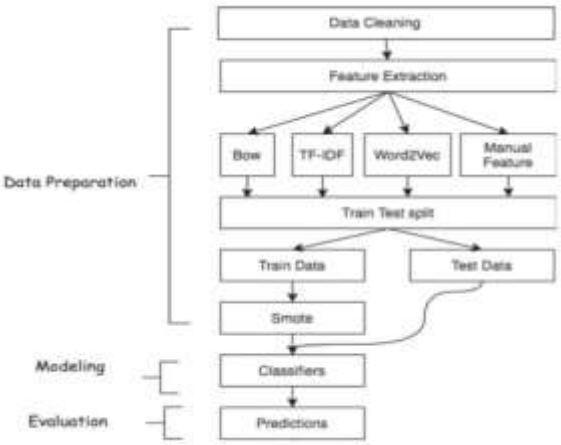


Fig.1. System Architecture

DATA FLOW DIAGRAM

- The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

- DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
- DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

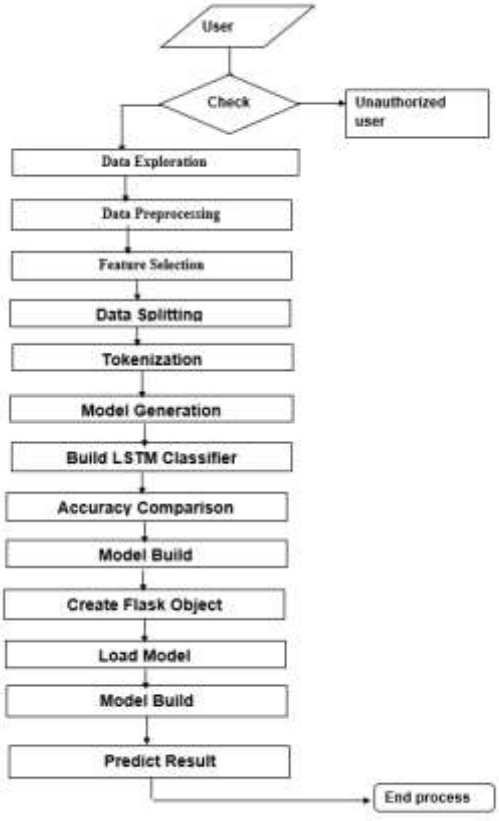


Fig.2.Data Flow Diagram

4. Implementation

4.1 Algorithm

Long short-term memory (**LSTM**) is an artificial recurrent neural network (RNN) architecture[1] used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition,[2] speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).

A common LSTM unit is composed of a **cell**, an **input gate**, an **output gate** and a **forget gate**. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications

Training:

An RNN using LSTM units can be trained in a supervised fashion, on a set of training sequences, using an optimization algorithm, like gradient descent, combined with backpropagation through time to compute the gradients needed during the optimization process, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network) with respect to corresponding weight.

A problem with using gradient descent for standard RNNs is that error gradients vanish exponentially quickly with the size of the time lag between important events. However, with LSTM units, when error values are back-propagated from the output layer, the error remains in the LSTM unit's cell. This "error carousel" continuously feeds error back to each of the LSTM unit's gates, until they learn to cut off the value.

4.2 Code Implementation

Tensorflow. TensorFlow is an amazing information stream in machine learning library made by the Brain Team of Google and made open source in 2015.It is intended to ease the use and broadly relevant to both numeric and neural system issues just as different spaces. Fundamentally, TensorFlow is a low level tool for doing entangled math and it targets specialists who recognize what they're doing to construct exploratory learning structures, to play around with them and to transform them into running programs.

Python 3.7. Python is broadly utilized universally and is a high-level programming language. It was primarily introduced for prominence on code, and its language structure enables software engineers to express ideas in fewer lines of code. Python is a programming language that gives you a chance to work rapidly and coordinate frameworks more effectively.

Anaconda3 5.3.1. Anaconda is a free and open-source appropriation of the Python and R programming for logical figuring like information science, AI applications, large-scale information preparing, prescient investigation, and so forth. Anaconda accompanies in excess of 1,400 packages just as the Conda package and virtual environment director, called Anaconda Navigator, so it takes out the need to figure out how to introduce every library freely. to Anaconda appropriation that enables clients to dispatch applications and oversee conda packages, conditions and channels without utilizing command line directions.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

5. Result

After implementing the algorithm we have summarized the long summary into

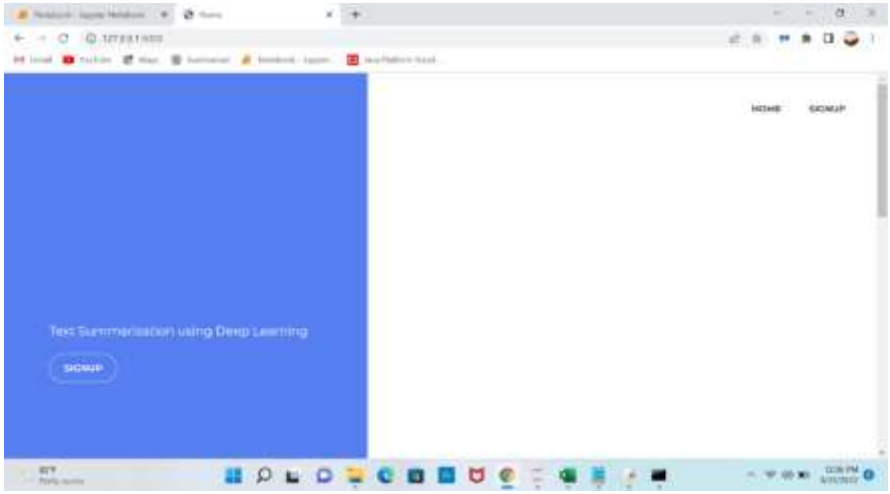


Fig:3 Home Page

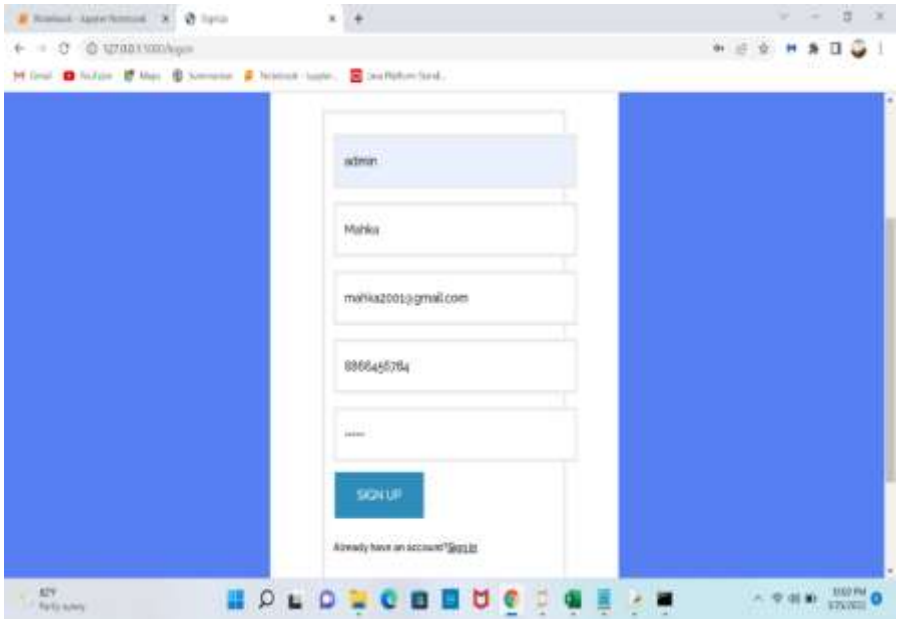


Fig:4 Sign U

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Fig:5 Sign in Page

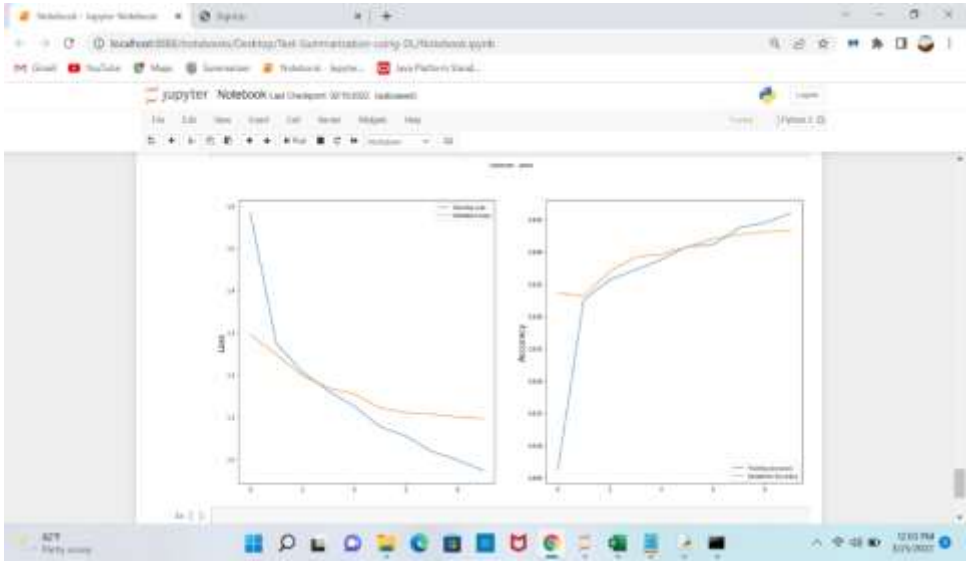
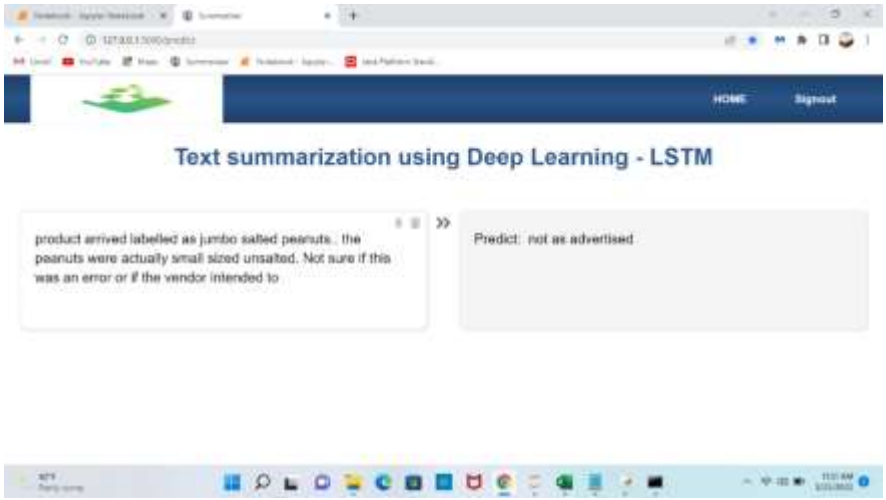


Fig: 6 Accuracy Graphs



International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Fig:7 Review 1

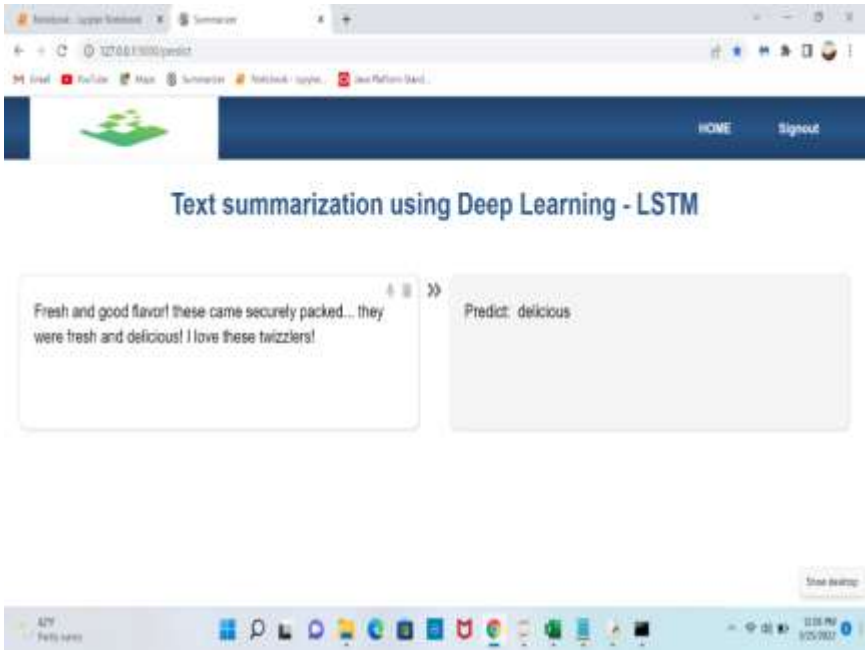


Fig:8 Review 2

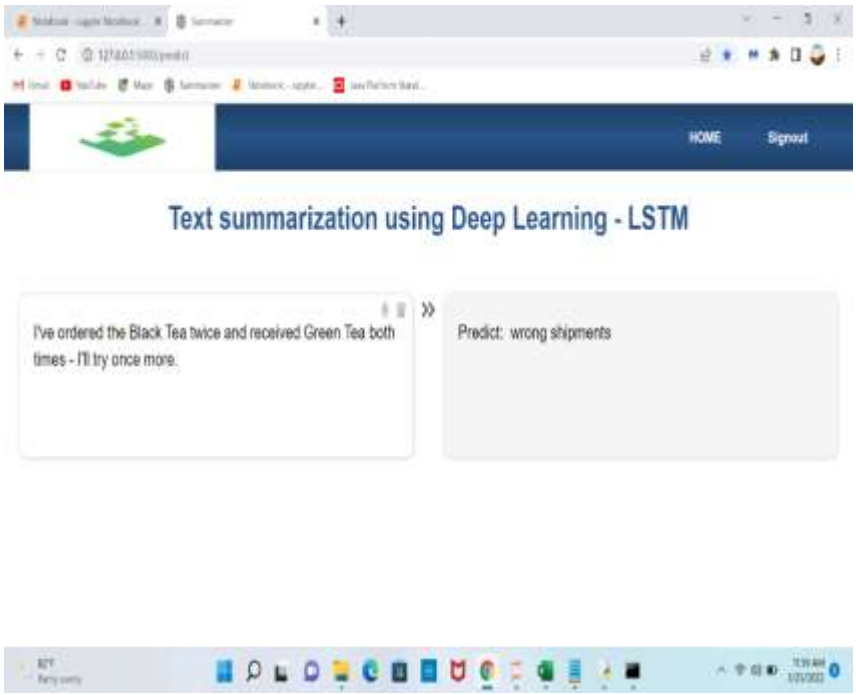


Fig:9 Review 3

6. Conclusion

One facet in which the approach could be taken forward is to consider about increasing the scope of feature extraction from sentence-level to higher abstraction like paragraphs and in case of multi-document summarization to documents. Approach based on Recurrent Neural Networks and their attention mechanisms could be focused upon. Since vanilla RNNs and their modifications like LSTMs have been seen to be good models in cases where we require some sort of memory their use in this kind of approach could prove out to be really useful as it would help to classify a sentence to belong to a summary by considering the contributions from a few previous sentences as well. Using attention mechanisms along with gated memory cells help them to select which inputs to focus more upon. A good measure for evaluation of the generated summaries possibly human scoring is a part on which the work may be extended and, the most important part where contributions to this work are welcome is in the field of data set collection/generation. A large enough to carry out the data driven approach for learning is a necessary requirement.

7. Future Scope

Automatic text summarization, the computer-based production of condensed versions of documents, is an important technology for the information society. Without summaries it would be practically impossible for human beings to get access to the ever growing mass of information available online. Although research in text summarization is over 50 years old, some efforts are still needed given the insufficient quality of automatic summaries and the number of interesting summarization topics being proposed in different contexts by end users (“domain-specific summaries”, “opinion-oriented summaries”, “update summaries”, etc.)

1. 8. References

2. [1] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K., 2017. Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268 .
3. [2] Bhargava, R., Sharma, G., Sharma, Y., 2017. Deep paraphrase detection in indian languages, in: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, ACM. pp. 1152–1159.
4. [3] Bhargava, R., Sharma, Y., Sharma, G., 2016. Atssi: Abstractive text summarization using sentiment infusion. Procedia Computer Science 89, 404–411.
5. [4] Carenini, G., Cheung, J.C.K., Pauls, A., 2013. Multi-document summarization of evaluative text. Computational Intelligence 29, 545–576.
6. [5] Chopra, S., Auli, M., Rush, A.M., 2016. 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.
7. [6] Cohan, A., Goharian, N., 2017. Scientific article summarization using citation-context and article's discourse structure. arXiv preprint arXiv:1704.06619 .
8. [7] Dalal, V., Malik, L., 2013. A survey of extractive and abstractive text summarization techniques, in: 2013 6th International Conference on Emerging Trends in Engineering and Technology, IEEE. pp. 109–110.
9. [8] Fattah, M.A., 2014. A hybrid machine learning model for multi-document summarization. Applied intelligence 40, 592–600.
10. [9] Fornito, A., 2016. Graph theoretic analysis of human brain networks, in: fMRI Techniques and Protocols. Springer, pp. 283–314.
11. [10] Gambhir, M., Gupta, V., 2017. Recent automatic text summarization techniques: a survey. Artificial Intelligence Review 47, 1–66.
12. [11] Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 .
13. [12] LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.
14. [13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

Accuracy Graphs