MSc Project Proposal

Enrolment number	30083146
MSc award title	Computer Science and Engineering
Project title	Comparative Study of Text Summarization.

Main aim of the project

Reading news or any important article is quite a hazard. As we often need to access huge online data for various purposes and time is the most crucial aspect, text summarization is proven to be very helpful to present the main essence of a huge corpus. One can save time by knowing important news or information from the article by reading the summary rather than going through it fully. Today, many text summarization models are available, but few of them can give a better summary. Our main aim will be to improve a text summarization model for generating a better summary. This project will present three machine-learning models that can generate summaries. It will compare these models to determine which one provides the best summary. Also, it will modify one of the models that can generate better summaries and implement a new method to generate even better summaries.

Project objectives

The main objective of this project is to analyse different types of text generated using different algorithms to determine which performs better for generating summarized text.

To achieve this goal, the project will specify three distinct models that follow three different techniques to generate summaries.

These models generated summaries will be evaluated by determining their F1 score and comparing it with the F1 score of human-generated summaries.

The project will then modify the model that produces the best summary and assess whether it performs better than before. By doing so, we aim to develop a text summarization model that can generate high-quality summaries efficiently and accurately.

Project description

Literature review

Ferreira (2013) conducted a thorough evaluation of 15 sentence scoring algorithms that have been published in the literature (Rafael Ferreira, 2013). The evaluation was both qualitative and quantitative and was carried out on three separate datasets, namely news, blogs, and article settings. The author also provided recommendations on how to enhance the results of sentence extraction. To evaluate the effectiveness of the algorithms, the widely used evaluation method ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was employed. ROUGE measures the degree of similarity in information between system-developed summaries and their equivalent gold summaries.

Sarker (2012) employed a simple Bengali stemmer for pre-processing and stemming, which removes suffixes based on the "longest match" using a predetermined suffix list (Sarkar, 2012). The sentences were then ranked based on features such as thematic terms, positional

value, and length. Thematic terms were selected based on a predefined threshold value of TFIDF. The top-ranked K sentences were then considered for the desired summary.

Nallapati (2017) proposed a sequence classification-based neural network model that treated the sentences of a document as having a binary form depending upon their existence in the summary (Ramesh Nallapati, 2017). The model was trained using an unsupervised approach to convert abstractive summaries to extractive labels. The approach involved maximizing the Rouge score by incrementally adding sentences to the summary set until the score was no longer improved. A GRU-RNN-based neural network model was used for training the SummaRunner model.

Isonuma (2017) introduced both sentence extraction and document classification tasks for single document summaries (Masaru Isonuma, 2017). They assumed that documents could be classified into specific subjects, and the sentences selected for the summary are extracted in relation to those subjects. Their neural network-based model was evaluated on the documents of two financial-based news publishers. A convolutional neural network (CNN) was used for sentence embedding from word embedding, while an LSTM-RNN was used for extracting summaries from the document.

Gambhir and Gupta (2016) presented a comprehensive survey of recent text summarization and extractive approaches (Gupta, 2016). The study examined the requirements, benefits, and drawbacks of various summarizing strategies, including both abstractive and multilingual techniques. The authors discussed both intrinsic and extrinsic

techniques for evaluating summaries and presented findings of extractive summarization techniques on shared DUC datasets.

Uddin and Khan (2007) surveyed different summarization techniques, mainly focusing on Bangla, as no similar work had been done on this language before (Uddin & Khan, 2007). They presented the location method, cue method, title, term frequency, numerical data, etc. methods for ranking sentences. The first 40% of higher-ranked sentences from the input text were considered as the summarized text. The authors also provided some examples of summarized text from news content and evaluated the Bangla summarizer using summary information and summary size as the evaluation parameters.

Research methodology

The purpose and objectives of this study will be clearly defined. To determine various methods for text summarizing, scholarly papers, conference proceedings, and other pertinent publications will be thoroughly reviewed. The dataset for this study will be obtained from Kaggle, containing over 300,000 news articles from CNN and the Daily Mail. The data will be pre-processed using methods similar to those used before analysis to ensure the accuracy and efficacy of the analysis.

The sentence segmentation, tokenization, stop word removal, punctuation removal, stemming, or lemmatization word count approach, similarity matrix, and TF-IDF will be used to give each sentence a weight based on how important it is to understand the text as a whole. The TF-IDF method will be used to weight sentences, which gives words a higher weight when they occur more frequently in a specific phrase than when they do in the entire corpus of

documents. To enhance the effectiveness of the summarization system, TF-IDF can be used in conjunction with other sentence weighting techniques, including sentence similarity and sentence position.

A similarity matrix will be used to visualize the text as a matrix of sentence similarities, with each cell denoting how similarly two sentences are written. Using this matrix, significant sentences that include distinctive or crucial information can be found. The word count method, where each phrase is given a weight based on the number of words it contains, will also be used. Although this method can result in less useful and less coherent summaries, it does not consider the significance of specific words or the connections between sentences.

The word2Vec model will replace the better summary provider model to evaluate if it can increase accuracy. Word2Vec is a machine learning technique that generates embedding's of words in a document, representing each word as a vector in a high-dimensional space, such that words with similar meanings are located closer to each other in this space. It uses pre-processed data to train a Word2Vec model to generate word embedding's. For each sentence in the document, a sentence embedding is generated by taking the average of the embedding's of its constituent words. The score for each sentence is calculated based on the cosine similarity between its sentence embedding and the document embedding, which is the average of all sentence embedding's. Word2Vec can capture the semantic relationships between words, resulting in more informative and coherent summaries.

Each technique will be run on the same set of documents, and the resulting summaries will be compared based on the F1 score to evaluate the performance of different text summarization techniques on the selected dataset. The summarized text generated by the three methods will be analyzed and compared to identify which generation technique provided better summaries. The sentence scoring method of the better summary generator model will be changed to the word2vec model, and the word2vec model will be evaluated based on the F1 score and compared to previously generated summaries.

Proposed methods of data collection

The project will use secondary data collection method. The CNN/Daily Mail dataset is a popular dataset for text summarization tasks. It comprises of human-written summaries of news articles from the CNN and Daily Mail websites. Over 300,000 news articles and their corresponding summaries are included in the dataset, which is divided into training, validation, and testing sets. Politics, entertainment, sports, and technology are just a few of the themes that are covered in the articles. Multiple summary sentences, each with a sentence-level summary of the article, are present with every article in the dataset.

Ethical considerations

This project will be worked on individually. The dataset that will be utilized has been authorized and obtained from the trustworthy Kaggle website. To ensure privacy and confidentiality, the data will be kept anonymous, and all sources acquired or used online will be cited appropriately. Sentence ranking methods, widely recognized for their effectiveness in weighing sentences, have been utilized, which were obtained from reputable and trustworthy

sources. Additionally, an evaluation method that is believed to produce an unbiased and precise summary has been employed.

References

1. Gupta, M. G. ·. V., 2016. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review,* 47 (1), pp. 1-66.
2. Masaru Isonuma, T. F. J. M. Y. M. a. I. S., 2017. Extractive Summarization Using Multi-Task Learning with Document Classification. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.*
3. Rafael Ferreira, L. d. S. C. ,. R. D. L. ,. G. P. e. S. ,. F. F. a. G. D. C. a. R. L. a. S. J. S. b. L. F. c., 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications,* pp. 5755-5764.
4. Ramesh Nallapati, F. Z. B. Z., 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17),* pp. 3075-3081.
5. Sarkar, K., 2012. BENGALI TEXT SUMMARIZATION BY SENTENCE. *Proceedings of International Conference on Business and Information Management(ICBIM-2012).*
6. Uddin, M. N. & Khan, S. A., 2007. A study on text summarization techniques and implement few of them for Bangla language. *2007 10th international conference on computer and information technology.*

Project plan

The project will begin with proper planning and research, which will require a total of 105 hours. This phase will involve defining the project scope and goals, conducting a literature review of existing research work, and identifying and selecting appropriate methodologies and tools. The dataset will be collected next and use a news data set for generating the summary. Data collection will require 20 hours. For data pre-processing, several techniques will be applied, including sentence segmentation, tokenization, stop word removal, punctuation removal, and stemming. This process will require 80 hours. Model development will take up the majority of the project timeline, approximately 260 hours. Three models will be used initially, and after evaluating the best summary among them, another model will be implemented to increase the accuracy of the summary. The generated summaries will be evaluated using the F1 score, a process that will require 60 hours. Ethical considerations will be incorporated into the project, which will require 15 hours. Finally, the entire project will be documented.

	Task	Start time	Time required
1	1. Project Planning and Research	0	105
2	Define project scope and goals	0	5
3	Literature review	5	60
4	Identify and select appropriate methodologies and tools	65	40
5	2. Data collection	105	20
6	3. Data preprocessing	125	80
7	Sentence segmentation	125	10
8	Tokenization	135	10
9	Stop word removal	145	20
10	Punctuation removal	165	10
11	Stemming	175	30
12	4. Model Development	205	260
13	Tf-Idf	205	60
14	Similarity matrix	265	70
15	Word count	335	50
16	Word2Vec	385	80
17	5. Evaluation	465	60
18	6. Ethical consideration	525	15