International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228361

Text Summarization using Extractive and Abstractive Techniques

Chintan A. Shah, Prof. Neelam Phadnis
Department of Computer Engineering, SLRCTE, University of Mumbai, Maharashtra, India

Article Info
Volume 8, Issue 3
Page Number : 236-241

Publication Issue :
May-June-2022

Article History
Accepted: 01 June 2022
Published: 05 June 2022

ABSTRACT

There have been considerable advancements in Text Summarization over the last few years. There are two ways to text summarization: one is based on natural language processing (NLP), and the other is based on deep learning. In the realm of NLP, text summarization is the most intriguing and challenging task. NLP stands for Natural Language Processing, which studies and manipulates human language by computers. Because of the massive increase in information and data, it has become critical. Text summarization is creating a thorough and meaningful summary of text from various sources such as books, news stories, research papers, and tweets, among others. Large text documents that are difficult to summarize manually are the research subject. The job of condensing a piece of text into a shorter version, minimizing the size of the original text while maintaining key informative components and content meaning, is known as summarising. Because human text summarising is a time-consuming and intrinsically tiresome process, automating it is gaining popularity and hence acts as a potent stimulus for academic study.

Keywords: Abstractive Summarization, BART, BERT, Extractive Summarization, NLP, RoBERTa, T5, Pegasus

I. INTRODUCTION

Summarization is compressing a piece of text into a shorter version, lowering the size of the original text while keeping vital informative aspects and content significance. Because human text summarising is a time-consuming and typically arduous activity, automating the work is gaining popularity and consequently acts as a significant impetus for academic study. Text summarising has vital applications in various NLP tasks such as text categorization, question answering, legal text summarization, news summary, and headline creation. Furthermore, the development of summaries may be implemented into these systems as an intermediary step, which helps to decrease the document's length. The volume of text data from diverse sources has expanded considerably in the significant data age. This volume of material provides a lot of information and expertise that must be adequately summarised to be helpful. The increased availability of documents has demanded substantial study in the NLP discipline for automated text summarization. The job of creating a concise and fluent summary without the intervention of a person while keeping the sense of the original text material is known as automated text summarization. It is exceedingly tough since, when we humans summarise a piece of material, we usually read it thoroughly to expand our knowledge before creating a summary emphasizing its essential elements. Because computers lack human understanding and linguistic competence, automated text summarization is a complex and time-consuming operation. For this goal, numerous machine learning methods have been suggested. The bulk of these techniques represents this challenge as a classification problem that outputs whether or not a sentence should be included in the summary. Topic information, Latent Semantic Analysis (LSA), Sequence to Sequence models, Reinforcement Learning, and Adversarial processes have all been applied in different techniques.

There are two approaches to summarization in general: Extraction and Abstraction summarization.

II. LITERATURE REVIEW

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right contexts in all layers. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5%[3].

Language model pretraining has led to significant performance gains but the careful comparison between different approaches is challenging. Training is computationally expensive, often done on private datasets of different sizes and hyperparameter choices have a significant impact on the final results. We find that BERT was are significantly undertrained, and can match or exceed the performance of every model published after it. Our best model achieves state-of-the-art results on GLUE, RACE andSQuAD[4].

BART is trained by corrupting text with an arbitrary noising function and learning model to reconstruct the original text. It matches the performance of RoBERTa with comparable training resources on a range of dialogue, question answering, and summarization tasks, with gains of up to 6 ROUGE. BART also provides a 1.1 BLEU increase over the back-translation system for machine translation, with only target language pretraining[2].

Transfer learning, where a model is first pre-trained on a data-rich task before being finetuned on downstream tasks, has emerged as a powerful technique in natural language processing. In this paper, we explore the landscape of transfer learning techniques for NLP by introducing a unified framework that converts all text-based language problems into a text-to-text format. We achieve state-of-the-art results on many benchmarks covering summarization, question answering, text classification and more[1].

In PEGASUS, important sentences are removed/masked from an input document and are generated together as one output sequence from the remaining sentences, similar to an extractive summary. In this work, we propose pre-training large Transformer-based encoder-decoder models on massive text corpora with a new self-supervised objective. Our model shows surprising performance on low-resource summarization, surpassing previous state-of-the-art results on 6datasets with only 1000 examples. We evaluated 12 downstream summarization tasks spanning news, science, stories, instructions, emails, patents, and legislative bills[5].

III. EXTRACTIVE SUMMARIZATION

Using a scoring system, extractive summarization extracts relevant phrases from a text and stitches them together to generate a single, cohesive summary. Condensing the material is done by recognizing significant chunks of the text and then trimming and sewing them together. Because of this, they depend only on extracting sentences from the source text. There has been a significant increase in extractive summarising research in recent years, mainly because it is simpler and provides more natural grammatical summaries. As a result, extractive summaries comprise the most relevant phrases from the input, which might be a single document or a collection of papers.

A transitional function is included in the input text to help discover relevant stuff. In most cases, the TF performance measurements for each phrase in the matrix are computed using the matrix itself. Based on the interpretation, each phrase is assigned a relevance level that indicates how likely it is to be included in the summary. An executive summary is generated by using the top k most significant sentences. Some research has employed latent semantic analysis (LSA) to find semantically relevant phrases. The old-fashioned way, in which key passages from the book are scoured for significance and then tacked on to the summary. In addition, it is essential to point out that the summary is based on the original text data.

A. BERT

In BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova proposed the BERT concept. The Toronto Book Corpus and Wikipedia were used to train a bidirectional transformer using a mix of masked language modelling and next sentence prediction. The right side of the inputs should be padded rather than the left since BERT embeds values in absolute positions. BERT was trained using the MLM and NSP goals (masked language modelling and next sentence prediction, respectively). However, it is excellent for NLU in general and anticipating masked tokens.

B. RoBERTa

Although RoBERTa's architecture and training method are essentially similar to BERT's, the authors made a few simple design tweaks to improve the results of BERT architecture. The following changes are in effect: It is our goal to get rid of the Next Sentence Prediction (NSP). Predicting whether the observed document segments are from the same or distinct documents using an auxiliary Next Sentence Prediction (NSP) loss is one of the primary goals of next sentence prediction. According to the researchers, erasing the NSP loss results in downstream job performance that is on par with or slightly better than before. Larger groups and lengthier training sessions: A batch of 256 sequences was used to train BERT at a rate of 1 million steps per batch. In this study, the researchers used 125 steps of 2K sequences and 31K steps of batch size 8k sequences to train the model. This has two advantages: it makes the masked language modelling aim easier to understand, and it improves the accuracy of the final product. Using distributed parallel training, large batches may be parallelized more simply.

Dynamic masking: Masking is done once during data preparation, resulting in a single static mask in the BERT architecture. Rather than using a single static mask, ten copies of the training data are made, and 40 epochs of masking are performed, each time using a different method to produce the same mask on four occasions. Dynamic masking, in contrast, creates a fresh mask each time data is fed into the model.

IV. ABSTRACTIVE SUMMARIZATION

Abstractive summarising approaches try to construct summaries by understanding the text using sophisticated natural language techniques in order to develop a new shorter text that may not be present in the original document but communicate the most important information from the original text, necessitating rephrasing phrases and combining information from the complete text to generate summaries, as is normally the case with a human-written abstract to be generated. To be considered an adequate abstractive summary, it must encompass all of the relevant information in the input and be grammatically sound in its presentation. As a consequence, people aren't confined to just picking and selecting from the original text.

Recent breakthroughs in deep learning are used in abstractive approaches. Abstractive approaches make use of the recent success of sequence-to-sequence models since the job of mapping the source text to the target summary may be seen as a sequence mapping one. For these models, there is a neural network that reads the text and encodes it, and then outputs the target text.

Making abstract summaries is, in general, a tough process that requires more advanced language modelling than data-driven alternatives like paragraph extraction. There have been substantial advancements made by applying neural networks inspired by neural machine translation and sequence models, although they are still far from human-level summaries.

With a revolutionary approach to the identification of significant portions, context interpretation and the reproduction of the text in a novel fashion, the advanced technology guarantees that the essential information is presented in the minimum feasible quantity of text. However, it's important to note that the sentences, in this case, were created by the model rather than merely taken from the original text input.

C. BART

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension was published on October 29, 2019, by Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer.

BERT-like bidirectional encoder and left-to-right decoder are both parts of Bart's standard seq2seq/machine translation architecture (like GPT).

During pre-training, the source words are randomly rearranged, and text spans are replaced with a single mask token.

In terms of text generation, BART is more effective when fine-tuned, but it also does well in terms of comprehension challenges. New state-of-the-art results on abstractive discourse, question answering, and summarization tasks with gains of up to six ROUGE were achieved with similar training resources to RoBERTa's performance on GLUE and SQuAD.

D. T5

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu introduced the T5 model in their work Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In contrast to BERT models, T5 is an end-to-end transformer model that is trained utilizing text input and transformed text output, rather than only a class label or a span of the input.

Natural language processing (NLP) issues are encoded and decoded using T5, an encoder-decoder paradigm. Instructor compulsion is used to train it. As a result, in order to do training, we always need a set of inputs and a set of outputs. Input ids are used to provide the model with the input sequence. Using decoder input ids, the target sequence is prepended with a start-sequence token before being sent to the decoder. Later, the EOS token adds a goal sequence that, in a teacher-forcing way, correlates to the labels. In this situation, the PAD token serves as the start sequence token. Training and fine-tuning T5 may be done in both supervised and unsupervised environments.

E. PEGASUS

PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization was published on December 18, 2019, by Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu.

As a pre-training exercise for Pegasus, a job akin to an extractive summary is created by removing or masking major sentences from an input text and then creating a single output sequence from the remaining phrases.

Pegasus achieves SOTA summary performance on all 12 downstream tasks, as determined by ROUGE and human eval.

V. EVALUATION AND RESULTS

We've evaluated the results on the ROUGE, which is a set of metrics. There are various sets of ROUGE metrics, but we've ROUGE-1, ROUGE-2 and ROUGE-L.

TABLE 1. RESULTS FOR EXTRACTIVE TECHNIQUES

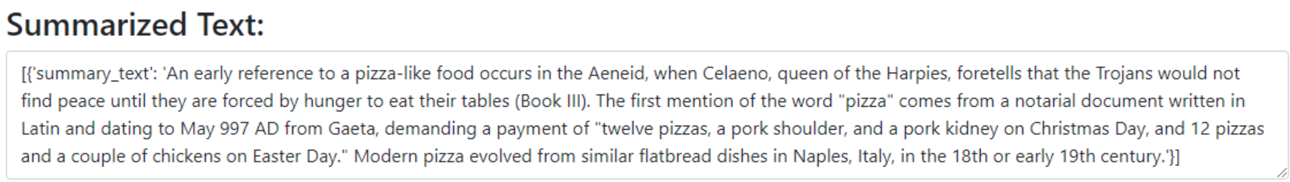
MODEL	ROUGE-1	ROUGE-2	ROUGE-L
BERT	41.04	18.69	33.58
RoBERTa	43.32	20.47	35.58

TABLE 2. RESULT FOR ABSTRACTIVE TECHNIQUES

MODEL	ROUGE-1	ROUGE-2	ROUGE-L
BART	44.16	21.28	40.9
T5	43.52	21.55	40.69
PEGASUS	43.9	21.2	40.76

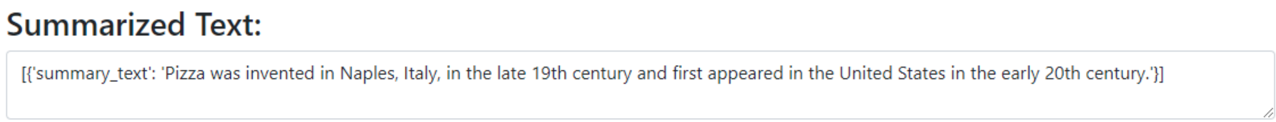
II. CONCLUSION

In this paper, extractive and abstractive techniques were implemented using the state of the models to perform very well after the fine-tuning. The extractive techniques are fast as compared to abstractive techniques, but for better accuracy in the NLP which has been evaluated with ROUGE. The RoBERTa model performs better than BERT for the extractive summarization technique. And the Pegasus model performs the best ROUGE score for the summarization task in abstractive. The T5 models have some advancements in the field of NLP with more tasks like question answering, Text classification, text generation, translation, and sentence similarity.


Summarized Text:

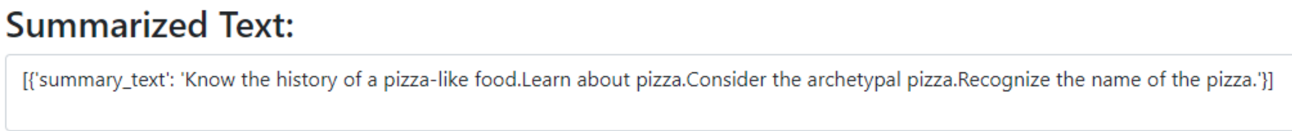
[{'summary_text': 'An early reference to a pizza-like food occurs in the Aeneid, when Celaeno, queen of the Harpies, foretells that the Trojans would not find peace until they are forced by hunger to eat their tables (Book III). The first mention of the word "pizza" comes from a notarial document written in Latin and dating to May 997 AD from Gaeta, demanding a payment of "twelve pizzas, a pork shoulder, and a pork kidney on Christmas Day, and 12 pizzas and a couple of chickens on Easter Day." Modern pizza evolved from similar flatbread dishes in Naples, Italy, in the 18th or early 19th century.'}]

Fig 1: Summary output for Pegasus Model


Summarized Text:

[{'summary_text': 'Pizza was invented in Naples, Italy, in the late 19th century and first appeared in the United States in the early 20th century.'}]

Fig 2: Summary output for BART Model


Summarized Text:

[{'summary_text': 'Know the history of a pizza-like food.Learn about pizza.Consider the archetypal pizza.Recognize the name of the pizza.'}]

FIg 3: Summary output for T5 Model

Fig 4: Summary output for BERT Model

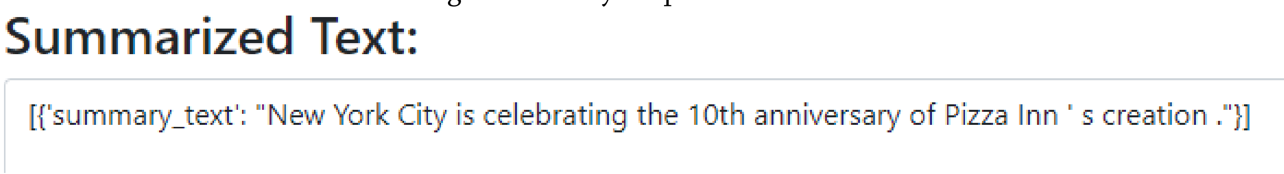


Fig 5: Summary output for RoBERTa Model

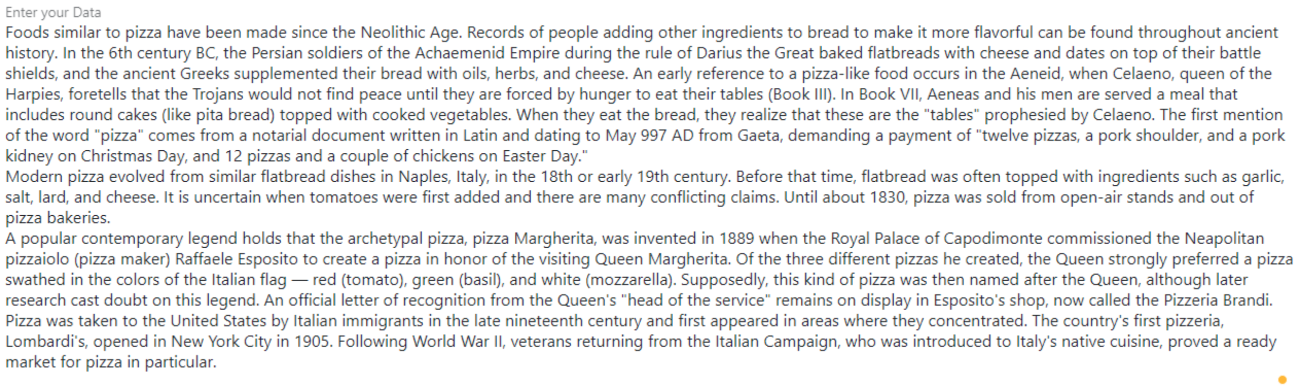


Fig 6: Input data for Summarization on both techniques

VI. REFERENCES

[1]. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"

[2]. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer on 29 Oct 2019, " BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension "

[3]. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova , " BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

[4]. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. It is based on Google's BERT model released in 2018,"RoBERTa: A Robustly Optimized BERT Pretraining Approach"

[5]. Jingqing Zhang, Yao Zhao, Mohammad Saleh and Peter J. Liu on Dec 18, 2019, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

Cite this article as :

Chintan A. Shah, Prof. Neelam Phadnis, "Text Summarization using Extractive and Abstractive Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 3, pp. 236-241, May-June 2022. Available at doi : https://doi.org/10.32628/CSEIT228361
Journal URL : https://ijsrcseit.com/CSEIT228361