JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text Summarization in Natural Language processing(NLP)

1)Omkar Garud 2) Vishnu Shendge 3)Vishal fulsundar 4) Prof.Poonam Bhawake Department Of Technology

Savitribai Phule Pune University,Pune,India

Abstract - The text data online is increasing massively; hence, producing a summarized text document is essential. We can create the summarization of multiple text documents either manually or automatically. A manual approach may be tedious and a time-consuming process. The resulting composition may not be accurate when processing lengthy articles; hence the second approach, i.e., the automated summary generation process, is essential. Training machine learning models using these processes makes space and time-efficient summary generation possible. There are two widely used methods to generate summaries, namely, Extractive summarization and abstractive summarization. The extractive technique scans the original document to find the relevant sentences and extracts only that information from it. The abstractive summarization technique interprets the original text before generating the summary. This process is more complicated, and transformer architecture-based pretrained models are used for comparing the text & developing the outline. This research analysis uses the BBC news dataset to evaluate and compare the results obtained from the machine learning models. Index Terms—Summarization, Natural Language Processing, Transformers, Deep Learning.

Introduction

Every day, we overwhelm with the maximum amount of information. Most of the information available is in the form of text. We may read multiple articles daily, and to understand the article's general information, we must read it thoroughly. If the article is excessively long, then time consumption in reading and understanding the document is very high. When we compare the other articles on a similar subject, most information will be quite the same. Hence, reading all the related articles may need extended time. We may need to spend a lot of time extracting only the required information from the individual report. A good text processing tool is essential to read the entire document and get a summary of only the required information. The primary goal of text summarizing is to develop a method that uses natural language processing to process the article's data. This method helps reduce time consumption in reading lengthy articles and helps the reader read multiple reports in a short period, saving time in the long term.

DATA INTRODUCTION –

A) DATA CONTAIN –

- **Data contains Three Columns id , dialogue, summary**
- **Train Dataset contain 14732 number of rows**
- **Test Dataset contain 819 number of rows**
- **Validation Dataset contain 818 number of rows**

id	dialogue	summary
13818513	Amanda: I baked cookies. Do you	Amanda baked cookies and will bring Jerry some tomorrow.
13728867	Olivia: Who are you voting for in	Olivia and Olivier are voting for liberals in this election.
13681000	Tim: Hi, what's up?	Kim may try the pomodoro technique recommended by Tim to get more stuff done.
13730747	Edward: Rachel, I think I'm in ove	Edward thinks he is in love with Bella. Rachel wants Edward to open his door. Rachel is outside.
13728094	Sam: hey overheard rick say	Sam is confused, because he overheard Rick complaining about him as a roommate. Naomi thinks Sam should talk to Rick. Sam is not sure what to do.
13716343	Neville: Hi there, does anyone	Wyatt reminds Neville his wedding anniversary is on the 17th of September. Neville's wife is upset and it might be because Neville forgot about their anniversary.
13611672	John: Ave. Was there any	John didn't show up for class due to some work issues with his boss. Cassandra, his teacher told him which exercises to do, and which chapter to study. They are going t
13730463	Sarah: I found a song on youtube	Sarah sends James an instrumental song he might like. James knows the song. The brain connects the songs to the context they were played in and brings to mind the a
13809976	Noah: When and where are we	Noah wants to meet, he quit his job, because his boss was a dick.
13809912	Matt: Do you want to go for date?	Matt invites Agnes for a date to get to know each other better. They'll go to the Georgian restaurant in Kazimierz on Saturday at 6 pm, and he'll pick her up on the way to
13727633	Lucas: Hey! How was your day?	Demi got promoted. She will celebrate that with Lucas at Death & Co at 10 pm.
13729168	Mark: I just shipped the goods	Mark just shipped the goods and he will send George the tracking number tomorrow.
13864825	Anita: I'm at the station in Bologna	Anita is at Bologna station.
13729567	Leon: did you find the job yet?	Arthur is still unemployed. Leon sends him a job offer for junior project manager position. Arthur is interested.
13864634	Macca: i'm so exited today	Macca has done ice climbing for the first time today, close to Reykjavik. He enjoyed it very much.
13815560	Isabella: fuck my life, I'm so not	Isabella feels bad after the Christmas party. She got drunk. She is ashamed to go back to work.
13731403	Tina: I'd only like to remind you	Lucy owes Tina 50 dollars. She made a transfer but it is Sunday so the payment will be on Tina's account on Monday. Tina needs the money because she has been having
13729191	Betty: Please remind me next time	Betty feels remorse she got drunk last night and went out of control.
13827937	Mary: Hi Mike!	Mike and Mary are going to visit Mike's grandma tonight. Mary will buy her some chocolate.
13828064	Laura: ok , I'm done for today-)	Laura will pick up Kim from work around 7, and they will come back home together.
13716048	Ashley: Guys, you have to read this	Erin is convinced by Ashley's book recommendations, while Seamus and Marcus aren't.

samsum-train

Exploratory data analysis(EDA)

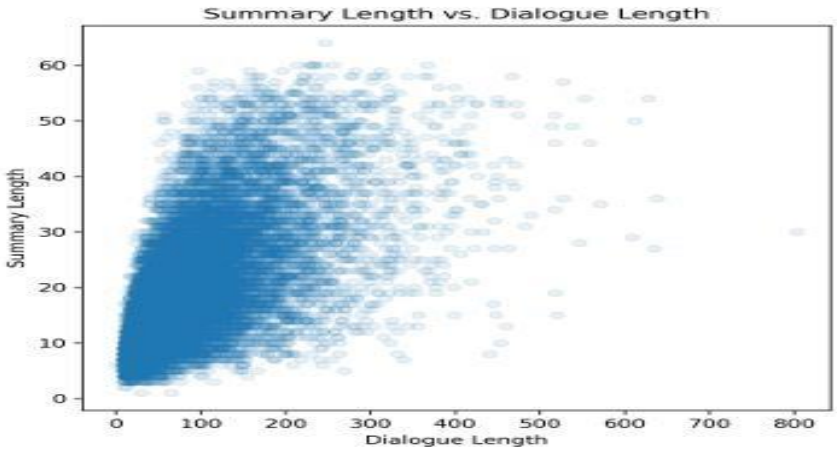

Figure 5.3: Summary Length vs. Dialogue Length

Text summarization is a natural language processing (NLP) technique used to generate concise summaries of longer texts.

Text summarization is a technique in NLP used to create short summaries of longer texts.

It involves processing and understanding the content of a text to extract the most important information.

NLP algorithms are employed to identify key sentences or phrases that capture the essence of the text.

There are two main approaches to text summarization: extractive and abstractive.

Extractive summarization involves selecting and combining important sentences from the original text.

Abstractive summarization goes beyond extraction by generating new sentences that capture the meaning of the original text.

Text summarization can be applied to various types of texts, including articles, documents, news stories, and academic papers.

It has numerous applications such as creating automatic summaries for news articles, generating abstracts for research papers, or providing condensed versions of lengthy documents for easier reading.

Text summarization using NLP has gained popularity due to its ability to save time and effort in information retrieval and comprehension tasks.

However, it still poses challenges such as maintaining coherence and preserving important details while condensing large amounts of information into concise summaries.

NLP algorithms are trained to understand and extract the most important information from a given text, allowing them to create summaries that capture the main points.

NLP algorithms are useful for text summarization.

These algorithms are trained to understand and extract important information from a text.

By using NLP, summaries can be created that capture the main points of the original text.

Text summarization using NLP can save time and effort by providing concise summaries of lengthy texts.

NLP algorithms analyze the structure and context of a text to generate accurate summaries.

Summaries generated through NLP can be customized to meet specific requirements or criteria

Text summarization using NLP is widely used in various applications such as news aggregation, document management systems, and content curation platforms.

NLP-based summarization techniques have evolved over time, utilizing advanced machine learning models like neural networks and transformer architectures for better results.

The accuracy and effectiveness of NLP-based text summarization depend on factors such as the quality of training data, algorithm selection, and fine-tuning parameters

Overall, text summarization using NLP has significant potential in automating information extraction tasks and enhancing productivity in various fields.

There are two main approaches to text summarization: extractive and abstractive.

Extractive summarization: This approach involves selecting and extracting the most important sentences or phrases from the original text to create a summary. It relies on identifying key information and preserving the original wording of the text.

Abstractive summarization: Unlike extractive summarization, abstractive summarization involves generating new sentences or phrases that capture the meaning of the original text, without necessarily using identical words or phrases. This approach relies on natural language processing techniques to understand and generate human-like summaries.

Extractive vs abstractive: Extractive summarization is generally considered more straightforward as it directly pulls information from the source text, but it may suffer from redundancy and lack of coherence in some cases. Abstractive summarization, on the other hand, has more flexibility but can be challenging due to its requirement for understanding context and generating coherent summaries.

Challenges in text summarization: Some common challenges in text summarization include handling ambiguous language, maintaining coherence and readability in generated summaries, dealing with long documents or multiple document sources, and ensuring accurate representation of key information without bias.

NLP techniques for text summarization: Natural Language Processing (NLP) plays a crucial role in both extractive and abstractive methods of text summarization. Techniques such as sentence scoring based on importance metrics (eg. TF-IDF). clustering similar sentences together, applying deep learning models leg. recurrent neural networks) for sequence generation are commanly used in NLP-based approaches.

Evaluation metrics for text summarizers: To measure the performance of different approaches to text summarization, various evaluation metrics are used including ROUGE (RecallOriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering), etc.

Future directions: Ongoing research in text summarization is focused on improving the quality of abstractive summaries by incorporating more context understanding and generating more coherent and human- like summaries. Additionally, domain-specific or task- specific text summarizers are being developed to cater to specific needs in industries like finance, healthcare, legal documents, etc.

Extractive summarization involves selecting and combining sentences or phrases from the original text, while abstractive summarization involves generating new sentences that convey the same meaning as the original text.

Extractive summarization involves selecting and combining sentences or phrases from the original text.Abstractive summarization involves generating new sentences that convey the same meaning as the original text.

NLP models for text summarization often use techniques such as word frequency analysis , sentence clustering, and deep learning algorithms like recurrent neural networks (RNNs) or transformers.

Word frequency analysis: NLP models for text summarization often start by analyzing the frequency of words in the text. This helps identify important keywords and phrases that can be used to summarize the content effectively.

Sentence clustering: Another technique used in NLP- based text summarization is sentence clustering. This involves grouping similar sentences together based on their semantic similarity or topic relevance. By clustering sentences, redundant information can be eliminated, and only representative sentences are retained for summary generation.

Deep learning algorithms: NLP models for text summarization also leverage deep learning algorithms like recurrent neural networks (RNNs) or transformers. RNNs are particularly useful for sequential data processing, making them suitable for generating summaries by considering the context of previous words and sentences. Transformers, on the other hand, excel at capturing long-range dependencies and have been widely adopted in state-of-the-art summarization models like BERT or GPT-2.

Extractive vs abstractive summarization: Text summarization models can be categorized into extractive or abstractive methods. Extractive methods involve selecting important sentences directly from the original text

to form a summary, while abstractive methods generate new phrases or rewrite existing ones to create a concise summary that may not exist verbatim in the source material.

Evaluation metrics: Evaluating the quality of generated summaries is crucial in NLP-based text summarization research and applications. Common evaluation metrics include ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which measures overlap between model- generated summaries and human-written references based on n-gram matching.

Challenges in text summarization: Despite advancements in NLP techniques, there are still challenges in achieving high-quality automatic text summarization due to issues such as understanding nuanced language nuances, handling diverse document types (e.g., news articles versus scientific papers), generating coherent summaries with proper grammar and structure, as well as addressing biases and ensuring the preservation of key information.

Applications of text summarization: Text summarization has a wide range of applications across industries, including news aggregation, document summarization for research or legal purposes, social media analysis. chatbot responses, and more. By condensing large volumes
of text into concise summaries, NLP-based text summarization can save time and improve efficiency in various information processing tasks.

Text summarization can be applied in various domains, including news articles, research papers, legal documents, and social media posts.

News articles: Text summarization can help extract key information from news articles, providing a concise summary for readers who may not have time to read the full article.

Research papers: NLP techniques can be used to automatically generate summaries of complex research papers, making it easier for researchers to quickly grasp the main findings and conclusions.

Legal documents: Summarizing legal documents such as contracts, court cases, and legal briefs can save time for lawyers and legal professionals who need to review large volumes of text.

Social media posts: With the abundance of social media content, text summarization can help users quickly understand the main points or sentiments expressed in lengthy posts or comments.

Email communications: Summarizing emails can be useful for busy professionals who receive a large number of messages daily, allowing them to prioritize important information without reading every email in detail.

Audio and video transcripts: NLP-based summarization techniques can also be applied to audio and video transcripts, enabling users to get a condensed version of spoken content without listening or watching the entire recording

Book summaries: Generating brief summaries of books or lengthy texts can provide readers with an overview before deciding whether they want to invest time in reading the full work.

The benefits of text summarization include time savings for readers who need to quickly understand the main points of a document or article and improved accessibility for people with limited reading abilities

Time savings: Text summarization using NLP allows readers to quickly grasp the main points of a document or article without having to read through the entire text. This is especially useful for busy individuals who need to extract information efficiently.

Improved accessibility: Summarized texts are easier to understand and comprehend, making them more accessible for people with limited reading abilities or those who struggle with lengthy texts. It enables a wider audience to access and benefit from the information contained in a document.

Efficient information retrieval: When dealing with large volumes of text, such as research papers or news articles, text summarization helps in retrieving relevant Information quickly and effectively. By providing concise summaries, it allows users to identify relevant content without having to go through every detail.

Effective decision-making: Text summarization aids in decision-making processes by presenting key insights and important details in a concise form. Decision-makers can save time by focusing on summarized information that captures the essence of a document or article.

Language translation assistance: Summarizing texts using NLP can be particularly useful when working with translated documents or articles from different languages. It provides an overview of the content, helping readers understand key points even if they are not fluent in the original language.

Content organization: Summaries serve as an effective way to organize and categorize large amounts of textual data into digestible chunks, making it easier for users to navigate through complex documents or datasets.

Automatic summarization tools: With advancements in NLP technology, automatic summarization tools have become increasingly accurate and efficient at generating summaries that capture essential information accurately while maintaining coherence.

Learning aid: For educational purposes, text summarization can assist students by condensing lengthy academic papers or textbooks into shorter summaries that highlight important concepts and key takeaways.

Scalability: Text summarization using NLP can be applied to various domains and industries, including journalism. research, legal documents, customer feedback analysis, and more. Its scalability makes it a valuable tool for processing large volumes of text-based information efficiently.

However, challenges in text summarization include handling ambiguity in language understanding, capturing context-specific information accurately, and maintaining coherence in generated summaries.

Handling ambiguity in language understanding: Text summarization involves understanding the meaning of a given text and condensing it into a shorter form. However, natural language is often ambiguous, with words or phrases having multiple interpretations. NLP techniques need to address this challenge by accurately disambiguating the intended meanings.

Capturing context-specific information accurately: Effective text summarization requires capturing important information while filtering out irrelevant details. This becomes particularly challenging when dealing with context-specific information that may be crucial for understanding the text's meaning. NLP models need to be able to recognize and extract such information accurately.

Maintaining coherence in generated summaries: A good summary should maintain coherence and readability. ensuring that the condensed version conveys the main points of the original text coherently Achieving this is difficult because summarization models need to select relevant sentences or phrases from a longer document while preserving logical flow and coherence.

Dealing with domain-specific knowledge: Texts can cover various domains, each having its own specific vocabulary and knowledge base To generate accurate summaries, NLP models must be trained on diverse datasets. encompassing different domains and utilize domain- specific knowledge effectively

Addressing data scarcity issues: Training robust summarization models often requires large amounts of annotated data from which they can learn patterns and extract meaningful features for generating summaries effectively.

A) Limited availability of high-quality labeled datasets poses a challenge in training accurate NLP models for text summarization.
B) Researchers are exploring techniques like transfer learning, weakly supervised learning, or leveraging pre-trained language models to overcome data scarcity issues

Evaluating summary quality objectively: Assessing the quality of generated summaries objectively is an ongoing challenge in NLP research.

A) Automatic evaluation metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) are commonly used but have limitations in capturing semantic similarity or overall coherence accurately.
B) Developing better evaluation metrics that align more closely with human judgments of summary quality is an active area of research.

Addressing ethical considerations: Text summarization can raise ethical concerns, such as potential bias in generated summaries or the risk of spreading misinformation if summaries are not factually accurate.

A) Researchers and developers need to ensure fairness.transparency, and accountability in text summarization systems, taking steps to minimize biases and improve the reliability of generated summaries.

Real-time summarization: Generating summaries in real- time poses additional challenges due to time constraints.

A) NLP models need to be efficient enough to process large volumes of text quickly and produce concise summaries within seconds or milliseconds.
B) Balancing speed with accuracy becomes crucial for real- time applications like news or social media feeds where timely updates are essential.

8.Evaluating the quality of generated summaries is also an ongoing challenge in NLP research

Automatic evaluation metrics: Various metrics such as ROUGE, BLEU, and METEOR are commonly used to evaluate the quality of generated summaries. These metrics compare the generated summary with reference summaries, measuring similarity in terms of n-grams or semantic alignment.

Human evaluation: In order to measure the true quality of a summary, human evaluators are often involved. They assess factors like coherence, informativeness, and overall readability. This approach provides more nuanced insights into the strengths and weaknesses of a summarization system.

Corpus-based evaluation: Another method is to compare the generated summaries against gold standard summaries within a corpus. This allows researchers to assess how well their system performs relative to existing state-of-the-art approaches.

User feedback: Collecting feedback from end users can provide valuable insights into whether the generated summaries meet their expectations and needs in real- world scenarios.

Domain-specific evaluation: Given that different domains may have specific requirements for summarization tasks (e.g.. news articles vs scientific papers). It is important to evaluate systems within those specific domains using relevant criteria.

Comparisons with baselines: Comparing the performance. of a new summarization model against existing baseline models can help determine if any improvements have been made in terms of summary quality.

Inter-annotator agreement: When multiple human evaluators are involved in assessing summaries, calculating inter-annotator agreement helps determine the reliability and consistency of evaluations.

Task-specific evaluation benchmarks: Creating standardized datasets with annotated summaries can facilitate fair comparison between different summarization systems by providing a common ground for benchmarking performance across various tasks (e.g.. singledocument vs multi-document summarization).

Adaptation to user preferences: Evaluating how well a summarization system can adapt its output based on user preferences or personalized requirements is crucial for measuring its practical utility and effectiveness in real- world applications.

CONCLUSION –

```
Dialogue:
Hannah: Hey, do you have Betty's number?
Amanda: Lemme check
Hannah: <file_gif>
Amanda: Sorry, can't find it.
Amanda: Ask Larry
Amanda: He called her last time we were at the park together
Hannah: I don't know him well
Hannah: <file_gif>
Amanda: Don't be shy, he's very nice
Hannah: If you say so..
Hannah: I'd rather you texted him
Amanda: Just text him 🙂
Hannah: Urgh.. Alright
Hannah: Bye
Amanda: Bye bye

Reference Summary:
Hannah needs Betty's number but Amanda doesn't have it. She needs to contact Larry.

Model Summary:
Hannah is looking for Betty's number. Amanda can't find it. Larry called Betty last time they were at the park together. Hannah
wants Amanda to text Larry.
```

References

Here are some references for text summarization using NLP:

- Text Summarization Using Natural Language Processing[1]
- A Comprehensive Guide to Natural Language Processing Text Summarization[2]
- Text Summarization for NLP: 5 Best APIs, AI Models, and AI[3]
- Text Summarization with Natural Language Processing (NLP)[4]
- An abstractive text summarization technique using[5]
- Text Summarization Approaches for NLP - Machine Learning Plus
- Shi T, Keneshloo Y, Ramakrishnan N, Reddy CK (2018), ”Neural Abstractive Text Summarization with Sequence-to-Sequence Models”, CoRR abs/1812.02303:
- Zhang J, Zhao Y, Saleh M, Liu PJ (2019), ”PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”, CoRR abs/1912.08777:
- Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019), ”HuggingFace’s Transformers: State-of-the-art Natural Language Processing”, CoRR abs/1910.03771.
- Plug and Play Machine Learning APIs, https://huggingface.co/inference-api