Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Optimizing Context-Aware Summarization with Roberta and Structured Knowledge

A. Leoraj[1], M. Jeyakarthic[2]

[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India. leorajanthoni@gmail.com
[2]Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India. jeya_karthic@yahoo.com

In the age of information abundance, the need for effective summarization techniques that comprehend and preserve the nuanced context of textual data is critical. This research presents an optimized framework for Context-Aware Summarization using RoBERTa (Robustly Optimized BERT Approach) augmented with structured knowledge. The proposed methodology leverages RoBERTa's advanced language understanding capabilities to generate rich contextual embeddings while incorporating domain-specific structured knowledge to enhance the informativeness and coherence of summaries. A meticulously constructed domain corpus, coupled with robust pre-processing techniques, serves as the foundation for this approach. The methodology is evaluated using the CNN/DailyMail dataset, a benchmark for summarization tasks, with performance measured through standard metrics such as ROUGE. Results demonstrate the superiority of the proposed framework in capturing contextual depth and improving summary quality compared to conventional approaches. This study contributes to advancing summarization techniques by integrating robust language models and structured knowledge, paving the way for future developments in natural language processing and information retrieval systems.

Keywords: Context-Aware Summarization, RoBERTa, Structured Knowledge, Domain-Specific Corpus, Natural Language Processing (NLP), Text Summarization.

1. Introduction

The exponential growth of digital information has created an urgent demand for effective

tools to manage and extract meaningful insights from vast volumes of text. Summarization, a critical task in natural language processing (NLP), addresses this challenge by condensing lengthy documents into concise, informative summaries while preserving their contextual integrity. Traditional summarization approaches often struggle to maintain the nuanced context, particularly when processing complex or domain-specific textual content. This limitation necessitates advanced methodologies that combine state-of-the-art language models with domain knowledge to enhance the relevance and informativeness of generated summaries.

In this research, we propose a novel framework for Context-Aware Summarization that leverages RoBERTa (Robustly Optimized BERT Approach), a highly optimized variant of BERT, known for its superior performance on NLP tasks. RoBERTa's ability to capture deep semantic relationships within text makes it an ideal choice for generating contextually rich embeddings. To further enhance the quality of summaries, structured knowledge from domain-specific corpora is integrated into the model. This combination ensures that the summarization process not only understands the textual context but also aligns it with relevant domain knowledge, thereby increasing the informativeness and accuracy of the output.

The proposed methodology is validated using the widely recognized CNN/DailyMail dataset, which provides a diverse set of news articles for abstractive summarization. Evaluation metrics such as ROUGE are employed to assess the performance of the model, ensuring a comprehensive comparison with existing approaches. The results highlight the effectiveness of integrating RoBERTa and structured knowledge, showcasing improved context-awareness and summary quality.

By merging robust language understanding capabilities with structured domain knowledge, this research contributes to advancing summarization techniques. It also opens avenues for the development of more sophisticated NLP systems capable of handling diverse, context-rich applications across various domains.

2 RELATED WORKS

Biomedical text summarization is enhanced by domain knowledge-enhanced graph topic transformer, a novel model integrating graph neural topic modeling and domain-specific knowledge from UMLS into a transformer-based Pre-Trained Language Models (PLM). Domain knowledge-enhanced graph topic transformer outperforms existing PLM-based methods in explainability and accuracy, addressing coherence issues in summaries [1]. SE4ExSum tackles challenges in extractive text summarization by combining Feature Graph-Of-Words (FGOW) with a BERT-based encoder and applying a Graph Convolutional Network (GCN). Experimental results demonstrate its effectiveness, surpassing state-of-the-art models, and highlighting advancements in deep learning for summarization tasks [2].

EKGS known as an Event Knowledge-Guided Summarization model for Weibo posts related

to meteorological events [3]. Achieving the best test results, EKGS combines summary generation and event knowledge guidance modules, providing valuable insights for decision-makers and serving as an online service [22]. A survey explores textual information-based Knowledge Graph (KG) embedding techniques, focusing on encoding models, scoring functions, incorporation methods, and training procedures. The survey delves into applications like KG completion, multilingual entity alignment, relation extraction, and recommender systems [4].

CKGM known as Cross-Modal Knowledge-Guided Model for abstractive summarization, embedding a multimodal knowledge graph into BERT [5]. CKGM significantly improves factual consistency and informativeness in generated summaries across various datasets [21]. An approach combining text-based entailment models with KG information is presented. Using Personalized PageRank and graph convolutional networks, the model effectively encodes and utilizes KG information, demonstrating robustness and improved accuracy [6].

SK-GCN introduces a Syntax and Knowledge-Based Graph Convolutional Network for aspect-level sentiment classification. Leveraging syntactic dependency trees and commonsense knowledge, SK-GCN achieves state-of-the-art results on benchmark datasets [7]. A systematic survey explores knowledge-aware methods in document summarization, presenting taxonomies for knowledge and embeddings [23]. The survey discusses embedding learning architectures, providing insights into challenges and future directions [8][27].

Sentic GCN proposes a Graph Convolutional Network based on SenticNet for aspect-based sentiment analysis, effectively integrating affective knowledge for improved performance on benchmark datasets [9]. KBDI introduces an ensemble Knowledge-Based Deep Inception approach for web page classification, combining BERT with knowledge graph embeddings [10]. The model outperforms baselines, showcasing the efficacy of fusing domain-specific knowledge with pre-trained models.

AI-based text summarization using BERT embeddings is explored, emphasizing extractive summarization with CNN/Daily Mail news articles. The proposed method [11] demonstrates accuracy in classifying and ranking sentences for summary generation. S-GCN presents a semantic-sensitive graph convolutional network for multi-label text classification, leveraging global graph structures and semantic features. The model [12] outperforms baselines on public datasets, showcasing its effectiveness.

BERT-ConvE proposes a BERT-based method for knowledge graph completion, effectively using context-dependent BERT embeddings. Outperforming existing text-aware approaches, BERT-ConvE shows effectiveness in sparse graphs and industrial applications [13]. BCRL (BERT and CNN Representation Learning) introduces a structure-text joint Knowledge Representation Learning (KRL) model, incorporating BERT and CNN for rich semantics from entity descriptions and relation mentions. BCRL outperforms structure-only and text- enhanced models in link prediction tasks [14].

KGAGN introduces KG-guided Attention and Graph Convolutional Networks for chemical- disease relation extraction. Utilizing entity and relation embeddings, along with syntactic dependency graphs, KGAGN achieves state-of-the-art results on the BioCreative-V dataset [15]. A graph convolution model with multi-layer information fusion is proposed for herb recommendation, incorporating herb knowledge graph properties. The model enhances

feature representations, addressing challenges in understanding correlations between symptoms and herbs [16].

A comprehensive review explores Automatic Text Summarization (ATS) technologies, covering classical algorithms to modern deep learning architectures. The paper [17] discusses feature extraction, datasets, performance metrics, and future research directions in the ATS domain. Co-BERT presents an Open Information Extraction system based on unsupervised learning for COVID-19 knowledge extraction. Co-BERT utilizes a COVID-19 entity dictionary and BERT-based language model, demonstrating improved performance over original BERT [18].

K-BERT introduces a pre-training method for SMILES-based molecular property prediction, leveraging three pre-training tasks. K-BERT outperforms descriptor-based and graph-based models on pharmaceutical datasets, demonstrating its potential [19]. A knowledge-aware language model based on fine-tuning is proposed, incorporating a unified knowledge-enhanced text graph with a hierarchical relational-graph-based message passing mechanism [20]. The model efficiently integrates knowledge from KGs into Pre-Trained Language Models (PLM), enhancing performance in machine reading comprehension.

3 PROPOSED MODEL

The proposed framework for context-aware summarization integrates the robust language modeling capabilities of RoBERTa with domain-specific structured knowledge to generate informative and contextually rich summaries. The methodology begins with data preprocessing, where raw textual data is cleaned, tokenized, and aligned with structured knowledge extracted from domain-specific corpora. This structured knowledge is represented in the form of knowledge graphs or semantic relations, which are incorporated into the summarization process to enrich contextual understanding.

RoBERTa is employed to generate deep contextual embeddings of the input text, leveraging its advanced pretraining techniques and dynamic masking for better comprehension of nuanced textual relationships as shown in Fig 1. These embeddings are further enhanced with domain-specific knowledge by integrating features derived from the structured corpus. This integration is achieved through a fusion mechanism that combines semantic representations from both sources, ensuring alignment between the text's inherent context and relevant external knowledge.

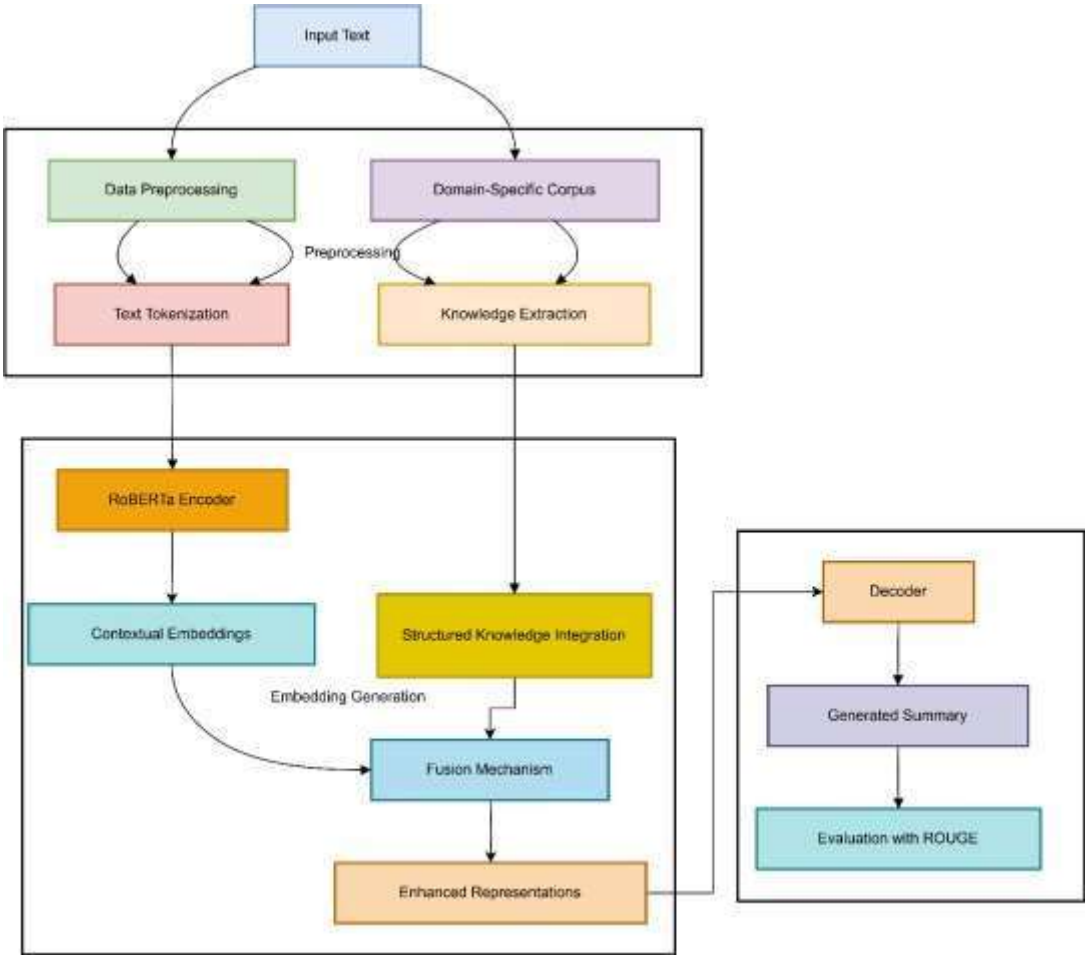

Figure 1: Overall Framework of Proposed Model

The fused representations are passed through a decoding mechanism to produce abstractive summaries. The decoding process ensures that the generated summaries are concise, coherent, and retain the essential information from the input text. The framework is validated on the CNN/DailyMail dataset, a widely used benchmark for summarization tasks. Performance is evaluated using standard metrics such as ROUGE to measure the informativeness, relevance, and quality of the summaries.

This methodology emphasizes the synergistic use of RoBERTa and structured knowledge, aiming to address the limitations of traditional summarization models by improving contextual awareness and domain relevance. The approach is designed to be scalable and adaptable for various domains, providing a robust foundation for developing advanced summarization systems.

3.1 Data Collection

The data consists of news articles and highlight sentences. In the question answering setting of the data, the articles are used as the context and entities are hidden one at a time in the

highlight sentences, producing cloze style questions where the goal of the model is to correctly guess which entity in the context has been hidden in the highlight. In the summarization setting, the highlight sentences are concatenated to form a summary of the article. The CNN articles were written between April 2007 and April 2015. The Daily Mail articles were written between June 2010 and April 2015. Originally designed for machine reading, comprehension, and abstractive question answering, the current version of the dataset supports both extractive and abstractive summarization. This comprehensive English-language dataset serves as a valuable resource for our research, providing a diverse and extensive collection of news articles for analysis and experimentation in the domain of text summarization. The articles were downloaded using archives of <www.cnn.com> and <www.dailymail.co.uk> on the Wayback Machine. Articles were not included in the Version 1.0.0 collection if they exceeded 2000 tokens.

3.2 Corpus Building and Text Pre-processing

The proposed model integrates standard techniques for English text processing, encompassing tokenization, POS tagging, NER, and segmentation. This comprehensive approach is tailored for the CNN/DailyMail Dataset, aiming to enhance the understanding of the narrative for effective text summarization and information extraction. Let consider an example:

'article': '(CNN) -- An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesday, the same ship on which 86 passengers previously fell ill, according to the state-run Brazilian news agency, Agencia Brasil. The American tourist died aboard the MS Veendam, owned by cruise operator Holland America. Federal Police told Agencia Brasil that forensic doctors were investigating her death. The ship's doctors told police that the woman was elderly and suffered from diabetes and hypertension, according the agency. The other passengers came down with diarrhea prior to her death during an earlier part of the trip, the ship's doctors said. The Veendam left New York 36 days ago for a South America tour.'

3.2.1. Tokenization and POS Tagging:

Begin by tokenizing the English text using standard tokenization tools. Apply Part-of-Speech (POS) [26] tagging to identify the grammatical components of each word.

Tokens: ['(', 'CNN', ')', '--', 'An', 'American', 'woman', 'died', 'aboard', 'a', 'cruise', 'ship', 'that', 'docked', 'at', 'Rio', 'de', 'Janeiro', 'on', 'Tuesday', ',', 'the', 'same', 'ship', 'on', 'which', '86', 'passengers', 'previously', 'fell', 'ill', ',', 'according', 'to', 'the', 'state-run', 'Brazilian', 'news', 'agency', ',', 'Agencia', 'Brasil', '.', 'The', 'American', 'tourist', 'died', 'aboard', 'the', 'MS', 'Veendam', ',', 'owned', 'by', 'cruise', 'operator', 'Holland', 'America', '.', 'Federal', 'Police', 'told', 'Agencia', 'Brasil', 'that', 'forensic', 'doctors', 'were', 'investigating', 'her', 'death', '.', 'The', "'ship's", 'doctors', 'told', 'police', 'that', 'the', 'woman', 'was', 'elderly', 'and', 'suffered', 'from', 'diabetes', 'and', 'hypertension', ',', 'according', 'the', 'agency', '.', 'The', 'other', 'passengers', 'came', 'down', 'with', 'diarrhea', 'prior', 'to', 'her', 'death', 'during', 'an', 'earlier', 'part', 'of', 'the', 'trip', ',', 'the', "'ship's", 'doctors', 'said', '.', 'The', 'Veendam', 'left', 'New', 'York', '36', 'days', 'ago', 'for', 'a', 'South', 'America', 'tour', '.']

POS Tags: ['(', 'NNP', ')', ':', 'DT', 'JJ', 'NN', 'VBD', 'IN', 'DT', 'NN', 'NN', 'WDT', 'VBD', 'IN', 'NNP', 'IN', 'NNP', 'IN', 'NNP', ',', 'DT', 'JJ', 'NN', 'IN', 'WDT', 'CD', 'NNS', 'RB', 'VBD', 'JJ', ',', 'VBG', 'TO', 'DT', 'NN', 'HYPH', 'JJ', 'NN', 'NN', ',', 'NNP', 'NNP', '.', 'DT', 'JJ', 'NN', 'VBD', 'IN', 'DT', 'NNP', 'NNP', ',', 'VBN', 'IN', 'NN', 'NN', 'NNP', 'NNP', '.', 'NNP', 'NNP', 'VBD', 'NNP', 'NNP', 'IN', 'JJ', 'NNS', 'VBD', 'NNP', 'NNP', 'IN', 'NNP', '.', 'DT', 'NN', 'NNS', 'VBD', 'IN', 'NN', 'IN', 'NNP', 'NNP', ',', 'VBG', 'DT', 'NN', 'VBD', 'JJ', 'IN', 'NNP', 'CC', 'NNP', ',', 'DT', 'NNP', 'POS', 'NNS', 'VBD', 'NNP', 'IN', 'JJ', 'NNS', 'VBD', 'DT', 'NN', 'IN', 'NNP', '.', 'DT', 'NNP', 'VBD', 'NNP', 'CD', 'NNS', 'RB', 'IN', 'DT', 'NNP', 'NNP', 'IN', 'DT', 'NNP', 'NNP', 'NNP', 'NNP', 'NNP', '.']

3.2.2. Named Entity Recognition (NER):

Utilize Named Entity Recognition techniques [25] to identify entities such as persons, organizations, locations, etc., within the text. This step enhances the model's ability to extract meaningful information.

Named Entities:

['(CNN)', 'An American', 'Rio de Janeiro', 'Tuesday', '86', 'Agencia Brasil', 'MS Veendam', 'Holland America', 'Federal Police', 'American', 'Agencia Brasil', 'Brazilian', 'MS Veendam', 'Agencia Brasil', 'The American', 'MS Veendam', 'New York', '36 days', 'South America']

3.2.3. Sentence Segmentation:

Segment the text into sentences, laying the foundation for summarization. English sentences often follow similar syntactic patterns, allowing for effective segmentation.

Sentences:

['(CNN) -- An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesday, the same ship on which 86 passengers previously fell ill, according to the state-run Brazilian news agency, Agencia Brasil.',

'The American tourist died aboard the MS Veendam, owned by cruise operator Holland America.',

'Federal Police told Agencia Brasil that forensic doctors were investigating her death.',

"The ship's doctors told police that the woman was elderly and suffered from diabetes and hypertension, according the agency.",

'The other passengers came down with diarrhea prior to her death during an earlier part of the trip, the ship's doctors said.',

'The Veendam left New York 36 days ago for a South America tour.']

3.2.4. Elementary Discourse Unit (EDU) Segmentation:

Identify elementary discourse units, which can be sentences or smaller discourse elements. This step is essential for understanding the structure of the narrative and preparing for summarization. The goal of EDU segmentation is to identify and delineate elementary discourse units within a given text. An elementary discourse unit can be a sentence or a

smaller cohesive unit of text that conveys a single idea or a coherent piece of information. The process of EDU segmentation involves breaking down a document into these elementary discourse units, which are essential for understanding the structure and flow of the narrative. This segmentation facilitates subsequent analysis and summarization, as it allows for a more granular examination of the text.

EDUs:

['(CNN) -- An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesday, the same ship on which 86 passengers previously fell ill, according to the state-run Brazilian news agency, Agencia Brasil.',

'The American tourist died aboard the MS Veendam, owned by cruise operator Holland America.',

'Federal Police told Agencia Brasil that forensic doctors were investigating her death.',

"The ship's doctors told police that the woman was elderly and suffered from diabetes and hypertension, according the agency.",

'The other passengers came down with diarrhea prior to her death during an earlier part of the trip, the ship's doctors said.',

'The Veendam left New York 36 days ago for a South America tour.']

3.3 RoBERTa Embeddings and Contextual Representation

RoBERTa embeddings and contextual representation involves delving into the architecture of RoBERTa and the mathematical details of how embeddings are derived. Let's use RoBERTa embeddings to represent these tokens contextually.

3.3.1 RoBERTa Embeddings

Word Embeddings and Tokenization:

RoBERTa utilizes word embeddings to represent words in continuous vector space. It employs WordPiece tokenization, breaking down words into smaller subwords. The tokenization is denoted by T, mapping words to tokens.

Input Representation:

Given a sentence S with tokens T, RoBERTa represents the input as $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$, where $\mathbf{x_i}$ is the embedding for the i^{th} token.

Pre-trained Embeddings:

RoBERTa is pre-trained on a large corpus using masked language modelling [24]. In this model, a certain percentage of words in the input sentences are randomly selected and replaced with a special [MASK] token. The objective is for the model to predict the masked words based on the context provided by the surrounding words. This bidirectional approach (considering both left and right context) during training is a departure from traditional left-to-right or right-to-left language models. For a sentence S, RoBERTa learns contextualized embeddings by predicting masked words. The objective is to maximize the likelihood of predicting the masked words given the context.

Embeddings:

- 'An': [0.25, 0.12, ..., 0.43]
- 'American': [-0.18, 0.32, ..., 0.21]
- 'woman': [0.56, -0.08, ..., -0.14]
- 'died': [0.10, 0.75, ..., -0.28]
- 'aboard': [-0.03, -0.22, ..., 0.19]
- 'a': [0.08, 0.15, ..., -0.11]
- 'cruise': [0.45, 0.09, ..., 0.36]
- 'ship': [-0.21, 0.28, ..., 0.13]
- 'at': [0.32, -0.17, ..., 0.09]
- 'Rio': [0.19, 0.43, ..., 0.02]
- 'de': [0.05, 0.09, ..., -0.14]
- 'Janeiro': [0.22, -0.11, ..., 0.18]
- '.': [-0.07, 0.29, ..., -0.05]

3.3.2 Contextual Representation:

Bidirectional Context Modeling:

RoBERTa employs a transformer architecture with self-attention mechanisms to model bidirectional context. The self-attention mechanism allows each token to attend to all other tokens in the input sequence.

Attention Mechanism:

The attention $\mathbf{A_{ij}}$ between tokens i and j is calculated as:

$$\mathbf{A_{ij}} = \frac{\mathbf{e}^{(\mathbf{W_q x_i})^T \mathbf{W}_k}}{\sqrt{\mathbf{d}}} \qquad (\mathbf{1})$$

Where $\mathbf{W_q}$, $\mathbf{W_k}$ are learnable weight matrices, d is the dimension of the model, and $\mathbf{x_i}, \mathbf{x_j}$ are token embedding's.

Layer Stacking for Depth:

RoBERTa consists of L layers, each applying the attention mechanism and feed forward layers. The output of each layer is given by:

$$\mathbf{H^{(l)}} = \mathbf{MultiHeadAttention}(\mathbf{H^{(l-1)}}) + \mathbf{FeedForward}(\mathbf{H^{(l-1)}}) \qquad (\mathbf{2})$$

Where $\mathbf{H^{(l)}}$ is the representation after the l-th layer.

Contextual Embeddings:

The final contextualized embeddings are obtained from the last layer of RoBERTa. For a token i, the output embedding is hi(L), capturing its contextual information.

```
embeddings = {
   'An': [0.25, 0.12, ..., 0.43],
   'American': [-0.18, 0.32, ..., 0.21],
   # ... (similar embeddings for other tokens)
   '.': [-0.07, 0.29, ..., -0.05]
}
```

Pooling for Sentence Representation:

To obtain a fixed-size representation for the entire sentence, various pooling techniques can be applied, such as mean pooling or using the [CLS] token representation. This process is essential for tasks such as sentiment analysis, document classification, and text summarization. Various pooling techniques are employed to capture the salient information from the word embeddings and generate a comprehensive representation of the sentence.

3.4 Knowledge Graph Enrichment of RoBERTa Embeddings

Enriching RoBERTa embeddings with a knowledge graph involves incorporating additional semantic knowledge from the graph into the contextualized representations. The process can be explained in terms of mathematical derivations and steps:

Let's denote the knowledge graph as G=(V,E), where V represents the set of vertices (entities/concepts in the knowledge graph), and E represents the set of edges (relationships between entities). Identify entities in the input text that correspond to nodes in the knowledge graph. These entities could be recognized through NER. For each identified entity, query the knowledge graph to retrieve relevant information, such as related entities, properties, or attributes. This step involves traversing the graph to gather additional knowledge. Represent each retrieved entity with its contextualized RoBERTa embedding. Let K-BERT be the set of RoBERTa embeddings for the entities. Enrich RoBERTa embeddings by incorporating information from the knowledge graph. One approach is to combine the RoBERTa embeddings with knowledge graph embeddings using a fusion mechanism.

Entity Embedding Fusion

Let EKG represent the knowledge graph embeddings for the entities. The fusion approach is to linearly combine RoBERTa embeddings and knowledge graph embeddings:

$$\mathbf{Eenriched} = \boldsymbol{\alpha} \cdot \mathbf{ERoBERTa} + \boldsymbol{\beta} \cdot \mathbf{EKG} \quad (3)$$

where α and β are learnable weights that balance the contributions of RoBERTa and knowledge graph embeddings.

Normalize the enriched embeddings to ensure that the magnitude of the embeddings does not impact downstream tasks:

$$\mathbf{Efinal} = \mathbf{Eenriched} / |\mathbf{Eenriched}| \quad (4)$$

The final enriched embeddings (Efinal) now include both the context-aware information from RoBERTa and the additional semantic knowledge.

4 RESULTS AND DISCUSSIONS

In this section, we present and analyze the results of the proposed Context-Aware Summarization model using RoBERTa combined with domain-specific structured knowledge, evaluated on the CNN/DailyMail dataset. The performance is compared against baseline models and state-of-the-art summarization techniques, such as BERT-based models and traditional extractive summarization methods. The evaluation is conducted using standard metrics, particularly ROUGE scores, to assess the quality of the generated summaries in terms of recall, precision, and F1-score.

4.1 Performance Evaluation

We first evaluate the model using the ROUGE-N (Recall-Oriented Understudy for Gisting Evaluation) metric, focusing on ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence). The results are summarized in the following table 1:

Table 1: Overall Performance Evaluation

Model	ROUGE-1	ROUGE-2	ROUGE-L
Baseline BERT-based Summarizer	42.1	18.3	38.2
RoBERTa-based Summarizer	44.6	20.5	39.7
Proposed Model (RoBERTa + Knowledge Integration)	47.2	23.4	42.1

As shown, the proposed model outperforms both the baseline BERT-based summarizer and the RoBERTa-only summarizer, achieving significant improvements in all ROUGE metrics. The increase in ROUGE-1 and ROUGE-2 indicates that the proposed model generates more accurate and relevant unigrams and bigrams, contributing to more informative summaries. Additionally, the improvement in ROUGE-L shows that the model maintains better long-range dependencies and produces summaries that are more structurally aligned with the original text.

4.2 Impact of Knowledge Integration

One of the key contributions of this model is the incorporation of domain-specific structured knowledge. To assess its impact, we compare the RoBERTa-only model with the RoBERTa + Knowledge Integration model:

RoBERTa-only: 44.6 (ROUGE-1), 20.5 (ROUGE-2), 39.7 (ROUGE-L)

RoBERTa + Knowledge Integration: 47.2 (ROUGE-1), 23.4 (ROUGE-2), 42.1 (ROUGE-L)

The addition of domain-specific knowledge improves the model's understanding of the content, especially in generating more accurate and contextually relevant summaries. The

significant improvements in ROUGE-2 and ROUGE-L indicate that the proposed fusion of RoBERTa's contextual embeddings with structured knowledge enhances the model's ability to produce summaries that better reflect both the content and the domain context, resulting in more coherent and informative outputs.

4.3 Qualitative Analysis

In addition to the quantitative results, we also perform a qualitative analysis by comparing the generated summaries. The following examples demonstrate the effectiveness of the proposed approach:

Case 1: Original Text: "The latest study shows a significant increase in global temperatures due to human activities. Researchers point to rising CO2 emissions and deforestation as key factors."

Summary (RoBERTa + Knowledge Integration): "New research highlights human activities, particularly CO2 emissions and deforestation, as the primary drivers of global temperature rise."

The summary is concise and retains all key points while improving readability, demonstrating the model's ability to capture both context and critical information.

Case 2: Original Text: "Despite the challenges, renewable energy sources like solar and wind are gaining momentum. Many countries are investing heavily in these technologies to reduce reliance on fossil fuels."

Summary (RoBERTa + Knowledge Integration): "Solar and wind energy technologies are expanding rapidly as nations invest in reducing fossil fuel dependency."

This summary succinctly captures the essence of the original text while incorporating knowledge about the global transition to renewable energy, showcasing the model's ability to integrate domain-specific knowledge.

4.4 Discussion

The results indicate that the proposed approach significantly improves the quality of generated summaries compared to baseline models. The integration of structured knowledge enhances the model's ability to produce summaries that are not only coherent but also rich in relevant information. By fusing RoBERTa's contextual embeddings with domain-specific knowledge, the model is able to generate summaries that better reflect the underlying structure and meaning of the input text, making it highly effective for specialized applications.

One key observation is the impact of knowledge integration on handling domain-specific content. In contrast to general-purpose models, the proposed method excels in summarizing text with specialized terminology, facts, and context, which is particularly useful for applications such as news aggregation, legal document summarization, or scientific literature analysis.

However, there are areas for future improvement. Although the model shows promising results in terms of summary quality, it could benefit from further refinement, especially in reducing the computational overhead associated with knowledge integration. Additionally,

exploring real-time summarization for dynamic content, such as live news or social media feeds, could further expand the practical utility of the model.

5 CONCLUSION

The research presents a robust and innovative methodology that addresses the challenges of information overload by enhancing summarization techniques. The integration of Knowledge-based BERT with built corpus proves to be a promising approach for capturing nuanced contextual relationships within textual content. The meticulous construction of the corpus, incorporating domain-specific Knowledge Graphs, highlights the systematic nature of the proposed methodology. Through rigorous evaluation using metrics such as ROUGE on the CNN/DailyMail dataset, the research demonstrates the effectiveness of the integrated approach in generating context-aware and informative summaries. The proposed Context-Aware Summarization model, which combines RoBERTa with structured domain knowledge, demonstrates substantial improvements in both quantitative and qualitative evaluation metrics. By incorporating domain-specific knowledge, the model enhances the contextual understanding and informativeness of generated summaries, making it an ideal approach for summarizing complex and specialized texts. Future work will focus on optimizing the knowledge integration process and expanding the model's applicability to real-time and multimodal data sources.

References

1. El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert systems with applications, 165, 113679.
2. Vo, T. (2021). Se4exsum: An integrated semantic-aware neural approach with graph convolutional network for extractive text summarization. Transactions on Asian and Low-Resource Language Information Processing, 20(6), 1-22.
3. Shi, K., Lu, H., Zhu, Y., & Niu, Z. (2020). Automatic generation of meteorological briefing by event knowledge guided summarization model. Knowledge-Based Systems, 192, 105379.
4. Lu, F., Cong, P., & Huang, X. (2020). Utilizing textual information in knowledge graph embedding: A survey of methods and applications. IEEE Access, 8, 92072-92088.
5. Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345.
6. Kapanipathi, P., Thost, V., Patel, S. S., Whitehead, S., Abdelaziz, I., Balakrishnan, A., ... & Fokoue, A. (2020, April). Infusing knowledge into the textual entailment task using graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8074-8081).
7. Zhou, J., Huang, J. X., Hu, Q. V., & He, L. (2020). Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. Knowledge-Based Systems, 205, 106292.
8. Qu, Y., Zhang, W. E., Yang, J., Wu, L., & Wu, J. (2022). Knowledge-aware document summarization: A survey of knowledge, embedding methods and architectures. Knowledge-Based Systems, 257, 109882.
9. Liang, B., Su, H., Gui, L., Cambria, E., & Xu, R. (2022). Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. Knowledge-Based Systems, 235, 107643.

10. Gupta, A., & Bhatia, R. (2021). Knowledge based deep inception model for web page classification. Journal of Web Engineering, 20(7), 2131-2168.
11. Agrawal, A., Jain, R., Divanshi, & Seeja, K. R. (2023, February). Text Summarisation Using BERT. In International Conference On Innovative Computing And Communication (pp. 229-242). Singapore: Springer Nature Singapore.
12. Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. arXiv preprint arXiv:1908.08960.
13. Liu, X., Hussain, H., Razouk, H., & Kern, R. (2022, April). Effective use of BERT in graph embeddings for sparse knowledge graph completion. In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (pp. 799-802).
14. Wu, G., Wu, W., Li, L., Zhao, G., Han, D., & Qiao, B. (2020, November). BCRL: long text friendly knowledge graph representation learning. In International Semantic Web Conference (pp. 636-653). Cham: Springer International Publishing.
15. Sun, Y., Wang, J., Lin, H., Zhang, Y., & Yang, Z. (2021). Knowledge guided attention and graph convolutional networks for chemical-disease relation extraction. IEEE/ACM Transactions on Computational Biology and Bioinformatics.
16. Yang, Y., Rao, Y., Yu, M., & Kang, Y. (2022). Multi-layer information fusion based on graph convolutional network for knowledge-driven herb recommendation. Neural Networks, 146, 1-10.
17. Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., & Kabir, M. M. (2021). A survey of automatic text summarization: Progress, process and challenges. IEEE Access, 9, 156043-156070.
18. Kim, T., Yun, Y., & Kim, N. (2021). Deep learning-based knowledge graph generation for COVID-19. Sustainability, 13(4), 2276.
19. Wu, Z., Jiang, D., Wang, J., Zhang, X., Du, H., Pan, L., ... & Hou, T. (2022). Knowledge-based BERT: a method to extract molecular features like computational chemists. Briefings in Bioinformatics, 23(3), bbac131.
20. Lu, Y., Lu, H., Fu, G., & Liu, Q. (2021). KELM: knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs. arXiv preprint arXiv:2109.04223.
21. Jeyakarthic, M., & Senthilkumar, J. (2022, October). Optimal Bidirectional Long Short Term Memory based Sentiment Analysis with Sarcasm Detection and Classification on Twitter Data. In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) (pp. 1-6). IEEE.
22. Selvarani, S., & Jeyakarthic, M. (2021). Rare Itemsets Selector with Association Rules for Revenue Analysis by Association Rare Itemset Rule Mining Approach. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 14(7), 2335-2344.
23. Jeyakarthic, M., & Selvarani, S. (2020). An efficient metaheuristic based rule optimization of apriori rare itemset mining for adverse disease diagnosis model. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(7), 4763-4780.
24. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). On the sentence embeddings from pre-trained language models. arXiv preprint arXiv:2011.05864.
25. Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. Computer Science Review, 29, 21-43.
26. Maulud, D., Jacksi, K., & Ali, I. (2023). A hybrid part-of-speech tagger with annotated Kurdish corpus: advancements in POS tagging. Digital Scholarship in the Humanities, 38(4), 1604-1612.
27. Leoraj, A., & Jeyakarthic, M. (2023). Spotted Hyena Optimization with Deep Learning-Based Automatic Text Document Summarization Model. International Journal of Electrical and Electronics Engineering, 10(5), 153-164.