A Question-Answering Based Framework/Metric for Evaluation of Newspaper Article Summarization

Vasanth Seemakurthy and **Shashank Sundar** and **Siddharth Arvind**
and **Siddhant Jagdish** and **Ashwini M Joshi**
Department of Computer Science and Engineering
PES University
Bengaluru, India
{vasanthseemakurthy,shashanksundar16,sidarvind11,
siddhantjagdish30}@gmail.com, ashwinimjoshi@pes.edu

Abstract

Condensed summaries of newspaper articles cater to the modern need for easily digestible content amid shrinking attention spans. However, current summarization systems often produce extracts failing to capture the essence of original articles. Traditional evaluation metrics like ROUGE also provide limited insights into whether key information is preserved in the summaries.

To address this, we propose a pipeline to generate high-quality summaries tailored for newspaper articles and evaluate them using a question-answering based metric. Our system segments input newspaper images, extracts text, and generates summaries. We also generate relevant questions from the original articles and use a question-answering model to assess how well the summaries can answer these queries to evaluate summary quality beyond just lexical overlap. Experiments on real-world data show the potential effectiveness of our approach in contrast to conventional metrics. Our framework holds promise for enabling reliable news summary generation and evaluation systems.

1 Introduction and Related Work

Today's news consumption is restricted by short attention spans and limited time. Traditional evaluation metrics like ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) depend on shallow lexical analysis that poorly captures semantic understanding. They primarily reference human-written summaries, focusing on matching words between source documents and summaries through n-grams and phrases rather than by extracting content from relevant newspaper sections. Despite advances in individual components, limitations persist: integration gaps between components, evaluation shortcomings that assess lexical rather than semantic content, domain specificity issues with diverse sources, and limited customization for varying information densities. We propose an integrated pipeline with a question answering (QA)-based evaluation framework for more meaningful summary quality assessment.

Our pipeline combines newspaper page segmentation, article text extraction, summarization using current NLP models, and encoder-decoder models to derive QA pairs from original texts as well as summaries. Current systems using question-answering evaluation metrics like QuestEval (Scialom et al., 2021) have not been fully implemented on newspaper summaries. To address the limitations of traditional metrics, we present a question answering-based metric that provides a more interpretable summary quality measure by extracting context-relevant QA pairs from source articles and evaluating the summaries' ability to answer these questions. We use an array of datasets: newspaper scans from The Times of India, news stories from the CNN/Daily Mail dataset (Hermann et al., 2015; Nallapati et al., 2016), and QA pairs from the SQuAD v.20 dataset (Rajpurkar et al., 2018).

A key reference was Tomar (2022), who developed an end-to-end framework combining Mask R-CNN (He et al., 2017) for article segmentation, Tesseract OCR (Smith, 2007) for text extraction, and BERT-based models (Devlin et al., 2019) for summarization. This approach addressed complex newspaper layouts with non-rectangular articles and embedded images, achieving validation mask loss of 0.189 and bounding box loss of 0.187.

2 Proposed Methodology

The overall methodology is outlined as shown in Figure 1:

1. **Segmentation and Extraction:** We utilize

Proceedings of Recent Advances in Natural Language Processing,pages 1085–1089
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-125

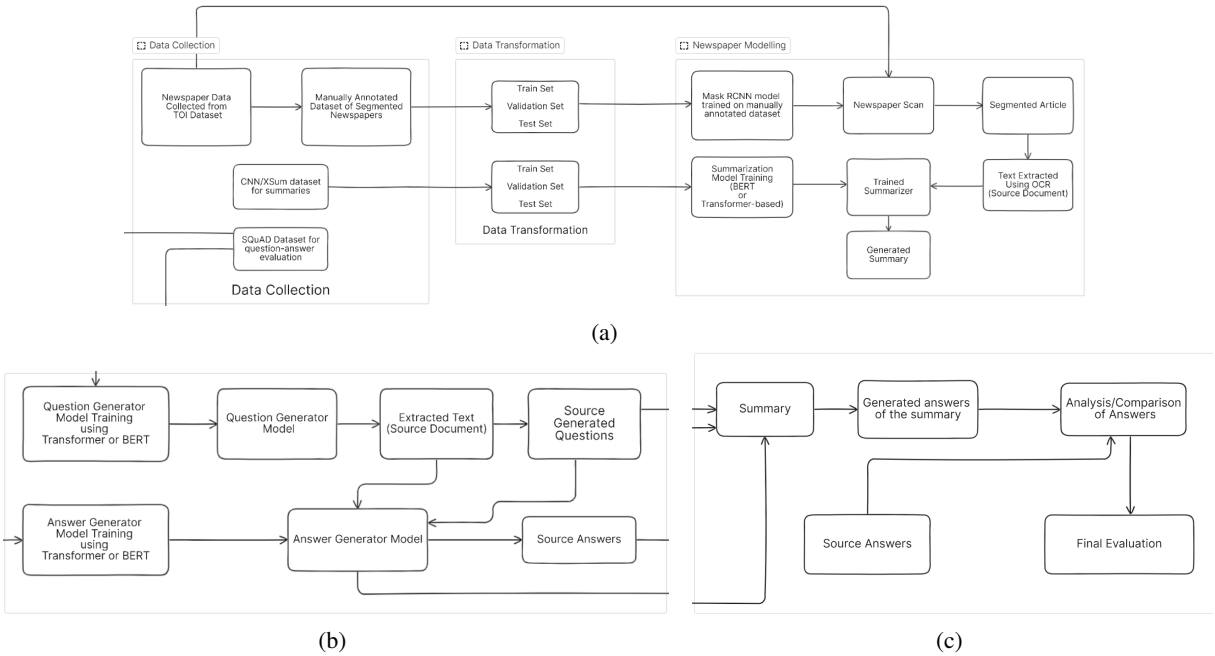

Figure 1: Proposed methodology of the system. (a) denotes the initial component of the system responsible for image procession, article segmentation, and subsequently text extraction and summarization. (b) denotes the question-answer generation component. (c) denotes the final evaluation of the summary based on sets of QA-pairs.

the Mask R-CNN model for segmenting newspapers into articles and articles into columns, and Tesseract OCR for extracting source text from segmented articles.

2. **Summarization:** Articles are summarized using Facebook's BART-Large-CNN model.

3. **Question Generation:** Questions are generated from source articles using the T5 model from the Hugging Face pipeline, fine-tuned for contextually meaningful questions.

4. **Question Answering:** Generated questions are tested on both source articles and their summaries using Google's FLAN-T5-BASE-SQuAD model, specifically designed for question-answering tasks.

5. **Evaluation:** We combine multiple metrics with different weights and penalties for comprehensive summary evaluation, using embedding similarity, BERTScore (Zhang et al., 2020), and cosine similarity. Source text is passed to a QA generator model to generate QA pairs, while summaries are passed to produce answers to source-generated questions; these pairs are compared to output the final summary evaluation result.

2.1 Evaluation Metrics

Existing metrics typically measure lexical overlap and fail to capture semantic accuracy and content completeness relevant to newspaper summarization, are more computationally expensive, or require manual tuning. Our approach aims to overcome these limitations by considering multiple metrics together with penalty functions that enforce higher priorities to contextual meanings rather than surface-text matches.

We bring in dynamic weight adjustment for embedding similarity and a scaled penalty factor. The weights change based on how strong the context of the summary is, which lets us evaluate things more flexibly, as shown below in Equation (1).

λ dynamically shifts weight based on embedding similarity: when embedding similarity is high, λ tends to favor Adjusted BERTScore; when embedding similarity is low, λ prefers weights to be biased toward semantic similarity. The penalty factor penalizes summaries that are semantically weak and lacking in critical information.

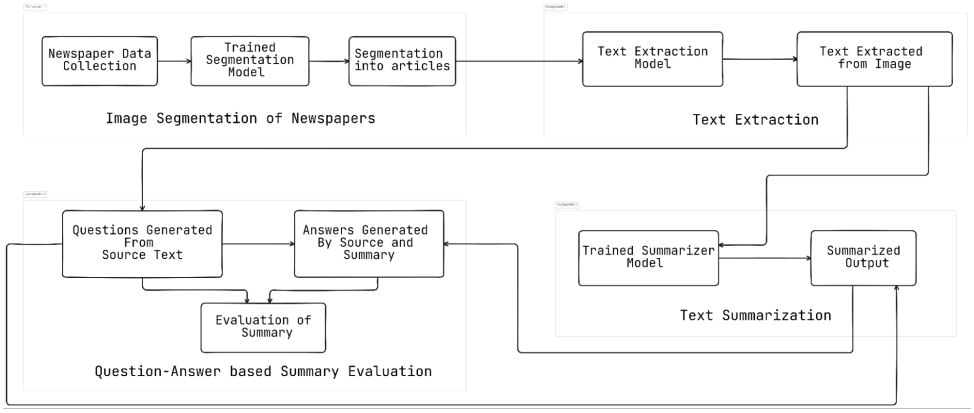

Figure 2: Architecture Diagram of the proposed system, consisting of 4 main components – (a) Image Segmentation (b) Text Extraction (c) Summarization (d) QA-based Evaluation.

$$\text{Final Combined Score} = \lambda \times \text{Adjusted BERTScore} + (1 - \lambda) \times \text{Embedding Similarity} \quad (1)$$

where,

$$\lambda = 0.5 + 0.5 \times \text{Embedding Similarity}$$
$$\text{Adjusted BERTScore} = \text{BERTScore} \times \text{Penalty Factor}$$
$$\text{Penalty Factor} = 1 - \max(0,\ 0.5 - \text{Embedding Similarity})$$

3 Implementation

3.1 Data Collection and Preparation

The newspaper scans dataset contains 9 GB of high-quality scans of the Times of India for January [1] and February 2018 [2]. Each day consists of 30-40 pages of different types; we used scans with both text and images to better model real-world usage. Our summary dataset is the CNN/Daily Mail dataset, containing original articles and their summaries. We used the SQuAD v2.0 dataset comprising question-answer pairs from Wikipedia articles for QA evaluation.

For segmentation model training, we manually parsed the dataset images to remove page scans with more than 60% image or ad content for proper segmentation model training. We developed a custom dataset by annotating scans into Rectangular and Non-Rectangular Articles. We used the open-source VGG Image Annotator (VIA) tool (Dutta and Zisserman, 2019) manually annotate 281 full-page newspaper scans to train the segmentation model, and 126 individual annotated articles to train the column segmentation model. We partitioned our custom datasets into training (75%), validation (15%), and testing (10%) sets.

Tesseract OCR is used to extract text from segmented articles and columns. Text cleaning addresses potential OCR errors, word-splitting, and misspellings.

3.2 Model Training and Transfer Learning

We use Mask R-CNN with pretrained COCO weights to speed up training and enhance model performance, thereby fine-tuning it on our custom datasets. Training involved loss correction by comparing model segmentation vs. manual segmentation, using T4 GPU resources through Google Cloud, to produce training weight files. New images could then be tested by loading these weights onto our custom Mask R-CNN model. Another Mask R-CNN model was trained for column segmentation in the same way.

3.3 Summarization

We used the pretrained Facebook–BART model for text summarization. BART produces coherent and context-relevant summaries while maintaining output quality. The BART-Large-CNN model, trained on CNN/Daily Mail, was integrated into our pipeline for high-quality summary generation.

[1] https://archive.org/details/TOIDELJAN18

[2] https://archive.org/details/TOIDELFEB18

3.4 Question Generation and Answering

Questions are generated from source text using the Hugging Face question-answering pipeline with a T5 model. The pipeline generates a variable number of questions based on source article size, with larger articles producing more questions generated for comprehensive evaluation. Answers are generated from both source and summary on identical question sets, with varying answers indicating potential summary misinterpretation to be penalized during evaluation.

3.5 Evaluation

We assess summary quality by comparing them to the original text for accuracy, relevance, and retention of essential source information. Our evaluation integrates traditional linguistic metrics (ROUGE and BLEU) to measure token- and phrase-level overlap, semantic similarity measures by word-level similarity using BERTScore and embedding-based sentence-level similarity using Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), and synonym expansion using WordNet.

4 Results and Discussion

The Mask R-CNN model produced outputs of different qualities based on scan types. Scans with fewer images achieved high segmentation quality and confidence scores, while those with images or ads gave less accurate results. We obtained similar results for article column segmentation. We implemented the BART-Large-CNN model to generate concise summaries from the extracted newspaper article text. The QA generation component successfully created relevant questions from source articles and obtained answers from both source and summary texts.

We evaluated our approach using various examples, an excerpt of which is tabulated in Table 1. The respective metrics' values (min. 0.0 & max. 1.0) in Table 2 and Table 3 are calculated between the Article Answer and Summary Answer in Table 1.

Table 2: Traditional Evaluation Metrics

S.No.	Precision	Recall	F1-Score	BLEU
Q1	0.538	0.438	0.482	0.077
Q2	0.625	0.500	0.556	0.112
Q3	0.154	0.111	0.129	0.012

Table 1: Questions and Respective Answers

S.No.	Question	Article Answer	Summary Answer
Q1	What is the Eiffel Tower known for?	The Eiffel Tower is a landmark in Paris, France, visited by millions of people every year.	The Eiffel Tower in Paris is a famous tourist spot attracting millions annually.
Q2	What factors are driving electric vehicle adoption?	Better batteries and subsidies boost EV use. Costs and charging remain challenges.	Improved batteries drive EVs. Charging is still an issue.
Q3	What causes climate change?	Climate change is caused by greenhouse gas emissions, deforestation, and industrial activities that trap heat in the atmosphere.	The stock market fluctuates due to changes in investor sentiment and economic indicators.

Table 3: Proposed Metric

S.No.	Embedding Similarity	Our Score
Q1	0.866	**0.94378**
Q2	0.874	**0.94732**
Q3	0.261	**0.509**

5 Conclusions, Limitations, and Future Work

We acknowledge several areas for future improvement. While our QA-based metric effectively identifies semantic gaps, a larger correlation study is needed to compare its performance against traditional metrics like ROUGE and modern LLM-based evaluators. Our evaluation quality depends on the relevance of T5-generated questions; future work could incorporate a filtering mechanism to keep only high-quality questions. Finally, we could explore using generated QA pairs in a feedback loop to iteratively refine summaries. Other areas of improvement include expanding our segmentation model to other newspapers, and improving the question-answering phase by testing more capable models.

We developed a metric for evaluating newspaper summaries based a pipeline from image segmentation to question-answer evaluation. Our framework and approach holds promise for enabling reliable news summary generation and evaluation systems that better match human judgments of summary quality.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Abhishek Dutta and Andrew Zisserman. 2019. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA. Association for Computing Machinery.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Venice, Italy.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, volume 1 of *NIPS'15*, page 1693–1701, Cambridge, MA, USA. MIT Press.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Shashank Sanjay Tomar. 2022. Summarizing newspaper articles using optical character recognition and natural language processing. Master's thesis, National College of Ireland, Dublin.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.