# From Abstract to Highlight: Automatic Research Highlight Generation

Anindita Bhattacharya[1,*], Tohida Rehman[1]

[1]Jadavpur University, Kolkata, India.

**Abstract**

Research paper *highlights* are short bullet-point summaries that help readers quickly grasp the key contributions and findings of a study. Although typically written by authors, the process can be time consuming and may vary in quality. In this work, we explore an automatic approach to generate *highlights* directly from research paper *abstracts*.

For this, we fine-tuned pre-trained `T5-base` and `BART-base` models for abstractive summarization, aiming to generate short, compact, meaningful and concise *highlights* that capture the essential ideas of each paper. The quality of the generated *highlights* is evaluated using standard metrics, including ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BERTScore, and SciBERTScore. Our fine-tuned T5-base model achieves the best performance, with the system proposed by our team *The NLP Explorers* attaining a ROUGE-L F1 score of 22.94% and securing the 5th position in the SciHigh track of FIRE 2025.

These results demonstrate that transformer-based models can serve as effective tools for enhancing scientific communication by automatically generating reliable and easy-to-read research paper *highlights*.

**Keywords**

Research Paper Highlights, Pre-trained Language Models, Fine-tuning, Natural Language Generation, Abstractive Summarization, Evaluation

## 1. Introduction

Research paper highlights are short bullet points that give readers a quick idea of the main contributions of a study. They make papers easier to understand and help readers quickly decide whether a paper is relevant to their work. However, highlights are usually written by authors, which takes time and can vary in quality.

With the rapid growth of scientific publications, there is increasing interest in automating this task. Recent progress in Natural Language Processing (NLP), especially transformer-based models, has shown strong results in text summarization. Since abstracts already summarize the paper, they provide a good starting point for automatically generating highlights.

In this work, we fine-tuned the T5-base model and the BART-base model to generate research paper highlights directly from abstracts. The models were trained to generate short, clear, and meaningful research highlights that captures the main ideas. We evaluated the models' performance using standard metrics such as ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BERTScore, and SciBERTScore which measure the similarity between generated and reference highlights.

The main contributions of this study are as follows:

1. Fine-tuning the **T5-base** model and **BART-base** model for the task of highlight generation from abstracts.
2. Evaluating the models' performance using metrics like **ROUGE**, **METEOR**, **BERTScore**, and **SciBERTScore**.
3. Showing that transformer-based summarization can support faster and more consistent highlight generation.

*Forum for Information Retrieval Evaluation, December 17-20, 2025, India*

*Corresponding author.

aninbhtry.01@gmail.com (A. Bhattacharya); tohidarehman.it@jadavpuruniversity.in (T. Rehman)

0009-0002-4623-1333 (A. Bhattacharya); 0000-0002-3578-1316 (T. Rehman)

## 2. Literature Review

Automatic text summarization has been studied for decades, beginning with some of the earliest extractive approaches. Luhn et al. [1] pioneered a method in 1958 that selected sentences based on the frequency of significant words while discarding common terms. This early work established the foundation for extractive summarization, where the task is to select existing sentences from a document to represent its content. Over time, extractive methods evolved with more sophisticated heuristics and statistical models to better identify important sentences.

The field took a major turn with the introduction of neural sequence-to-sequence models, which allowed abstractive summarization by generating new sentences rather than simply extracting them. Techniques such as attention-based encoders and pointer-generator networks further improved the quality of summaries by addressing long-range dependencies, handling out-of-vocabulary (OOV) words, and reducing the problem of repetitive phrase generation [2]. These innovations enabled models to generate summaries that were not only concise but also fluent and semantically aligned with the input.

The next breakthrough came with the transformer architecture [3], which has since dominated Natural Language Processing (NLP). Transformers made it possible to build very large pre-trained models that could be fine-tuned for specific downstream tasks. Models such as T5 [4], BART [5], and PEGASUS [2] demonstrated state-of-the-art performance across multiple summarization datasets. These pre-trained language models captured rich linguistic patterns during large-scale pre-training, which translated into strong results in domain-specific fine-tuning tasks. Rehman et al. [6] developed a GRU-based encoder-decoder with Bahdanau attention to generate concise English news summaries, achieving improved performance for headline style outputs. Rehman et al. [7] further evaluated pre-trained models such as Pegasus-CNN-DailyMail, T5-base, and BART-large-CNN across multiple datasets, including CNN-DailyMail, SAMSum, and BillSum, to benchmark summarization performance. Bhattacharya et al. [8] conducted a comparative analysis of abstractive summarization models for summarizing clinical radiology reports using the MIMIC-CXR dataset. Their study evaluated models such as T5-base, BART-base, PEGASUS-x-base, ChatGPT-4, LLaMA-3-8B, and a Pointer Generator Network with a coverage mechanism, assessing their performance using ROUGE, METEOR, and BERTScore metrics to identify the strengths and limitations of each in generating concise medical summaries.

While text summarization has been widely studied, generating research paper highlights is a more recent and specialized task. Unlike traditional summaries, highlights are typically short bullet points that emphasize the most important contributions of a paper. Early attempts to address this include Collins et al. [9], who used supervised learning with a binary classifier to identify highlight-worthy sentences, and Cagliero and Quatra [10], who employed regression methods to select the top-k most relevant sentences.

Rehman et al. made significant contributions in this area with a series of studies. They first proposed an abstractive approach using pointer-generator networks [11] to generate highlights directly from research abstracts. This method was important because it went beyond extraction and attempted to generate highlights that were concise, coherent, and aligned with the abstract. To further improve this approach, they incorporated Named Entity Recognition (NER) [12] into the highlight generation pipeline, showing that domain-specific knowledge could enhance the informativeness of generated highlights. Rehman et al. [13] also explored different types of embeddings with a pointer-generator model to evaluate highlights generation using various input combination, such as the abstract only or a combination of sections including the abstract, introduction, and conclusion, employing ELMo embeddings. Later, they carried out a comprehensive study comparing different deep learning models with SciBERT word embeddings for highlight generation [14], offering one of the most detailed benchmarks in this emerging field and contributed MixSub dataset. Together, these works form the basis for current research in automated highlight generation.

## 3. Dataset

We used the **MixSub** dataset contributed by Rehman et. al. [14] for the highlight generation task. This dataset was built by collecting research articles from ScienceDirect, comprising 19,785 articles published in 2020 across multiple domains. Each article is paired with its corresponding author-written research highlights, making it well-suited for training and evaluating highlight generation systems. Every entry in the dataset contains an abstract along with the highlights section. The dataset is divided into training, validation, and test splits using an 80:10:10 ratio. 10,000 samples from the training set, 1,985 samples from the validation set, and 1,840 samples from the test set were supplied for the SciHigh track at FIRE 2025 experiments. Figure 1 presents an illustrative example from the MixSub dataset. The author-written highlights are not direct extracts from the abstract but instead these are paraphrased and condensed content, exemplifying abstractive summarization.

**Abstract:** "The current study introduces the flexible approach of mixture components to model the spatiotemporal interaction for ranking of hazardous sites and compares the model performance with the conventional methods . In case of predictive accuracy based on in sample errors the Mixture 5 demonstrated superior performance in majority of the cases indicating the advantage of mixture approach to accurately predict crash counts. LPML was also calculated as a cross validation measure based on out of sample errors and this criterion also established the dominance of Mixture 5 further reinforcing the superiority of the mixture approach from different perspectives ."

**Author-written research highlights:**

- "A comprehensive evaluation was conducted for 9 spatiotemporal crash frequency models."
- "The model performance was evaluated based on both in sample and out of sample errors."
- "The site ranking performance of the proposed models was assessed using three criteria."
- "A flexible approach was proposed which accommodates the variations of time trend across space."
- "The research findings indicated the advantage of the proposed mixture approach to accurately predict crash counts."

**Figure 1:** An example of an abstract and its corresponding author-written research highlights from the MixSub dataset. Colored boxes indicate the semantic correspondence between phrases in the abstract and author written highlights. Input and author-written research highlights taken from https://www.sciencedirect.com/science/article/abs/pii/S0001457518312326.

## 4. Model Used

### 4.1. T5-Base

The Text-to-Text Transfer Transformer (T5) is built on the encoder–decoder framework and represents a refined adaptation of the transformer architecture introduced by Vaswani et al. [3]. One of the key contributions of T5 is its unified text-to-text paradigm, where a wide range of NLP tasks including machine translation, question answering, text classification, and summarization are cast into a single format: transforming an input text into an output text. This flexibility allows T5 to be applied across diverse tasks without requiring significant changes in the model architecture.

During pre-training, T5 is trained on a span-corruption objective, where random spans of text are masked and the model is required to reconstruct the missing parts. This enables the model to learn both syntactic and semantic dependencies across different contexts. The T5-base variant used in this study contains 220 million parameters, making it computationally efficient compared to larger variants while still offering strong performance in summarization tasks.

The model's encoder is responsible for creating contextual embeddings of the input sequence, while the decoder generates the corresponding output sequence in an autoregressive manner. This encoder–decoder synergy makes T5 particularly effective for abstractive summarization tasks, such as research highlight generation, where the goal is not just to extract key sentences but to generate coherent, concise, and human-like summaries.

### 4.2. BART-Base

Bidirectional and Auto-Regressive Transformers (BART) is a sequence-to-sequence model introduced by Lewis et al. [15], designed to combine the strengths of bidirectional encoder models like BERT and autoregressive decoder models like GPT. Built upon the standard transformer encoder–decoder architecture, BART is particularly well-suited for text generation tasks such as abstractive summarization, paraphrasing, and dialogue modeling.

BART is pre-trained using a denoising autoencoder objective, where the original text is corrupted using various noise functions—including token masking, token deletion, sentence shuffling, and document rotation—and the model is trained to reconstruct the clean text from the noisy input. This flexible corruption strategy allows BART to learn robust representations of linguistic structure and semantics, enabling strong downstream performance across diverse natural language processing tasks.

The BART-base variant used in this study contains 139 million parameters, making it lighter than its larger counterparts while still retaining competitive generative capabilities. Its encoder captures bidirectional contextual information from the input text, and its decoder generates output sequences autoregressively, predicting one token at a time based on previously generated tokens. This architecture makes BART highly effective for abstractive summarization tasks, where the model must integrate information across sentences and generate fluent, coherent, and human-like summaries rather than simply extracting content from the source.

In applications such as research highlight generation, BART-base offers an optimal balance between computational efficiency and summarization quality, allowing it to generate concise, contextually rich summaries that maintain the essential meaning of the input document.

## 5. Performance Evaluation Metrics

To assess the quality of the generated research highlights, we employed widely used automatic evaluation metrics from the field of text summarization. These metrics are designed to compare the model-generated summaries with human-written reference highlights, thereby providing an objective measure of accuracy and fluency.

We primarily focused on ROUGE, METEOR, BERTScore, and SciBERTScore, these standard benchmarks in summarization research.

1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [16]:
   ROUGE is one of the most commonly used evaluation measures for summarization tasks. It calculates the overlap between the generated summary and the reference text based on n-grams, word sequences, and sentence-level structures. In our experiments, we used three key variants:
   a) ROUGE-1: Measures unigram (single word) overlap, which reflects the model's ability to capture the core words from the reference highlights.
   b) ROUGE-2: Considers bigram (two consecutive words) overlap, offering insight into the fluency and coherence of the generated highlights.
   c) ROUGE-L: Based on the longest common subsequence, this metric evaluates the similarity of sentence structures and captures how well the generated highlights preserve the ordering and organization of information.
2. METEOR (Metric for Evaluation of Translation with Explicit ORdering) [17]:

METEOR complements ROUGE by focusing on semantic similarity and sentence-level alignment between the generated and reference highlights. Unlike ROUGE, which relies heavily on surface-level n-gram matches, METEOR accounts for synonyms, stemming, and word order variations. This makes it more sensitive to the actual meaning conveyed in the generated output, ensuring that highlights are not only word-accurate but also semantically faithful to the reference.

3. BERTScore (Bidirectional Encoder Representations from Transformers Score) [18]:
   BERTScore evaluates the semantic similarity between the generated and reference highlights using contextual embeddings from the BERT model. Instead of relying on exact word matches, it measures token-level cosine similarity in the embedding space, allowing it to capture deeper semantic relationships and paraphrased expressions. This makes BERTScore particularly effective for abstractive summarization tasks, where meaning preservation is more important than surface-level overlap.
4. SciBERTScore (Scientific BERTScore):
   SciBERTScore is an adaptation of BERTScore that uses embeddings from the SciBERT model, which is pre-trained on scientific and biomedical corpora. This domain-specific version enhances the evaluation of summaries in scientific texts—such as radiology or biomedical reports—by better understanding specialized terminology and contextual nuances. It provides a more accurate measure of semantic fidelity for summaries in technical and research-oriented datasets.

## 6. Experimental Setup

In this section, we describe the data pre-processing steps and the implementation details used in our experiments.

For pre-processing, we first removed extra spaces from the text and kept only those samples that had enough content: abstracts with at least 11 tokens and highlights with at least 14 tokens. To keep the data consistent and easier to train on, we set a maximum length of 512 tokens for abstracts and 100 tokens for generated highlights.

For fine-tuning, we used the T5-base[1] and BART-base[2] pre-trained language models from Hugging Face. We fine-tuned them on the given Scihigh-MiXSub dataset for 5 epochs, with a batch size of 8 and a learning rate of 2e-5. These settings were chosen to achieve good performance while maintaining stability during training.

## 7. Results

In this section, we present the performance of the fine-tuned T5-base and BART-base models on the highlight generation task. Table 1 shows the F1-scores (%) for ROUGE, METEOR, BERTScore, and SciBERTScore obtained by both fine-tuned models on the test set. ROUGE and METEOR focus on n-gram overlap and sentence-level alignment between the generated and reference highlights, while BERTScore and SciBERTScore capture semantic similarity using contextual embeddings. The fine-tuned T5-base model achieved the best results on all metrics except BERTScore and SciBERTScore, where BART-base performed slightly better.

Table 2 presents the ROUGE-L F1 score of all participating teams in the SciHigh track at FIRE 2025, where our team, *The NLP Explorers*, achieved a ROUGE-L F1 score of 22.94% and secured the 5th position.

### 7.1. Case Study

To better understand the quality of the generated highlights, we present a case study in Figure 2. This shows an example in which the author-written highlight is compared with the highlight generated by

[1] https://huggingface.co/google/t5-base
[2] https://huggingface.co/facebook/bart-base

**Table 1**
Performance of fine-tuned T5-base and BART-base for highlights generation on the MixSub test set across multiple evaluation metrics. All F1-scores (ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BERTScore, and SciBERTScore) are reported as percentages (%).

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | METEOR | BERTScore | SciBERTScore |
|---|---|---|---|---|---|---|
| T5-base | **34.25** | **12.09** | **22.94** | **29.57** | 85.95 | 62.76 |
| BART-base | 31.60 | 11.56 | 22.52 | 28.81 | **86.84** | **65.05** |

**Table 2**
Best ROUGE-L F1 scores achieved by participating teams in the SciHigh track of FIRE 2025. Only the highest-scoring run for each team is shown. All scores are presented as percentages.

| Group Name | Run Submission (Best Run) | ROUGE-L F1 | Rank |
|---|---|---|---|
| Text_highlights_gen | run1 | 23.45 | 1 |
| AiNauts | run1 | 23.24 | 2 |
| SVNIT_CSE | run1 | 23.02 | 3 |
| NLPFusion | run2 | 22.96 | 4 |
| **The NLP Explorers** | **run2** | **22.94** | **5** |
| NIT_PATNA_2025 | run1 | 22.42 | 6 |
| MUCS | run1 | 22.08 | 7 |
| JU_CSE_PR_KS | run1 | 22.06 | 8 |
| SCaLAR | run1 | 20.33 | 9 |
| Ayanika | run1 | 17.91 | 10 |
| Shilpo | run1 | 16.75 | 11 |
| TJP | run1 | 06.73 | 12 |

the fine-tuned T5-base model and BART-base model. This comparison helps illustrate how closely the generated text matches the style and content of the reference highlight.

From the case study, it can be seen that the fine-tuned T5-base model captures the main idea of the author-written highlight while using slightly different wording. The generated highlights correctly identify the role of csGRP78 as a molecular chaperone, its surface expression on cancer and angiogenic endothelial cells, and its function as a promising biomarker and therapeutic target. These align closely with the central ideas presented in the author-written highlights, showing that the T5-base model is capable of extracting and paraphrasing domain-specific information with notable accuracy. However, while the generated highlights successfully convey factual precision and topical relevance, they exhibit a narrower scope by omitting certain advanced details such as the integration of targeting moieties into nanoparticles, differential impacts of antibody epitopes, and recommendations for future clinical translation. This suggests that the model tends to prioritize high-level biomedical concepts over nuanced mechanistic or experimental insights.

In comparison, the fine-tuned BART-base model identifies relevant biomedical concepts, but its first two generated points do not directly align with the author-written highlights. Instead, they draw on broader and more descriptive ideas from the abstract, such as the general challenges of targeting heterogeneous cancer cells and the detailed molecular role of csGRP78 within the heat shock protein family. Although these points are factually accurate, they represent background information rather than the focused, actionable insights emphasized by the authors. The third generated point, however, directly aligns with the author-written highlights by correctly identifying the overexpression of GRP78 on the surface of cancer and angiogenic endothelial cells. Moreover, while BART-base partially captures the significance of csGRP78 surface expression, it does not fully articulate the biomarker or therapeutic implications with the same level of specificity demonstrated by the T5-base model.

Overall, the comparative analysis suggests that T5-base produces highlights that more closely align with the author's intended summary structure and biomedical emphasis, whereas BART-base tends to generate broader, more explanatory content. Both models exhibit strong semantic understanding, but T5-base demonstrates greater fidelity to the concise, insight-focused nature of scientific highlights,

**Abstract:**"As one of the deadliest diseases cancer frequently resists existing therapeutics because they do not target all cells within a progressing tumor for example both tumor stem and proliferating cells. This frequently results in enrichment of invasive and metastatic drug resistant tumor cells subpopulations cancer recurrence and eventually patient mortality. Thus there is an urgent need to identify specific markers by which the targeted imaging and or therapeutic guided missile like agents can specifically detect and or eradicate all cancer cells within a heterogeneous tumor while leaving the normal cells intact. As a member of heat shock protein 70 superfamily glucose regulated protein 78 has been documented as a molecular chaperone in the endoplasmic reticulum which mainly responds to ER stresses in normal cells. There is over expression of GRP78 on the surface of cancer cells and angiogenic endothelial cells which makes it a promising target for different types of peptides and antibodies that can be employed for targeted cancer therapy or imaging. In this review we discuss the biological processes functional importance and translocation mechanisms of cell surface GRP78 in tumor cells. As a cancer biomarker we also review the potential applications of csGRP78 targeted therapy and imaging and finally we suggest a brief roadmap ahead of csGRP78 targeting for targeted theranostic implications."

**Author-written highlight:**

- csGRP78 is expressed principally by cancer cells and angiogenic endothelial cells.
- csGRP78 is a promising cancer biomarker and target for imaging and therapy.
- csGRP78 targeting moieties can be integrated into nanoparticles.
- Anti csGRP78 antibodies targeting different epitopes differ in therapeutic impact.
- New generation csGRP78 targeting agents should encourage clinical testing.

**Fine-tuned T5-base model generated highlight:**

- csGRP78 is a molecular chaperone in the endoplasmic reticulum.
- GRP78 is expressed on the surface of cancer cells and angiogenic endothelial cells.
- csGRP78 is a promising target for different types of peptides and antibodies.
- csGRP78 is a potential cancer biomarker for targeted therapy and imaging.

**Fine-tuned BART-base model generated highlight:**

- Cancer is a heterogeneous disease and it is difficult to target all cancer cells within a tumor while leaving the normal cells intact.
- As a member of heat shock protein 70 superfamily glucose regulated protein 78 has been documented as a molecular chaperone in the endoplasmic reticulum which mainly responds to ER stresses in normal cells.
- There is over expression of GRP78 on the surface of cancer cells and angiogenic endothelial cells

**Figure 2:** Case study comparing author-written highlights with T5-base and BART-base generated highlights. Color shading indicates conceptual correspondence between the author-written highlights and the model-predicted highlights. Input and author-written research highlights taken from https://www.sciencedirect.com/science/article/pii/S0168365920306362.

while BART-base reflects a tendency toward verbose, context-heavy summarization.

## 8. Conclusion and Future Scope

In this paper, we explored the task of generating research paper highlights directly from abstracts using the pre-trained T5-base and BART-base models. The models were fine-tuned on the SciHigh Track FIRE 2025–provided SciHigh-MixSub dataset and evaluated using standard summarization metrics such as ROUGE, METEOR, BERTScore, and SciBERTScore. The experimental results show that these models are capable of generating highlights that are semantically relevant and stylistically close to author-written highlights. Both the quantitative evaluation and the qualitative case study confirm the potential of using transformer-based models for highlight generation, thereby reducing the manual effort required

by researchers and publishers.

Although the results are promising, there are several directions for future research. First, more advanced transformer models such as PEGASUS or LLaMA could be explored and compared with T5-base and BART-base to assess improvements in highlight quality. Second, incorporating domain-specific pre-training, especially for scientific articles in fields such as medicine or computer science, may improve relevance and factual accuracy. Third, human evaluation could be included alongside automatic metrics to better assess the usefulness of generated highlights for end users. Finally, developing lightweight and energy-efficient approaches will be important to address the environmental concerns associated with training large-scale models.

## Declaration on Generative AI

Generative AI tools were employed solely to aid in language polishing and formatting for specific sections of this manuscript. All scientific content, experimental design, data collection, analysis, and interpretation were independently developed and verified by the author(s). The AI tools did not participate in experiment planning, coding, data processing, or drawing conclusions.

## References


[1] H. P. Luhn, The automatic creation of literature abstracts, IBM Journal of Research and Development 2 (1958) 159–165. doi:10.1147/rd.22.0159.

[2] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proc. 55th ACL, Vol. 1: Long Papers, 2017, pp. 1073–1083.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[6] T. Rehman, S. Das, D. K. Sanyal, S. Chattopadhyay, Abstractive text summarization using attentive gru based encoder-decoder, in: Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2021, Springer, 2022, pp. 687–695.

[7] T. Rehman, S. Das, D. K. Sanyal, S. Chattopadhyay, An analysis of abstractive text summarization using pre-trained models, in: Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2021, Springer, 2022, pp. 253–264.

[8] A. Bhattacharya, T. Rehman, D. K. Sanyal, S. Chattopadhyay, Comparative analysis of abstractive summarization models for clinical radiology reports, arXiv preprint arXiv:2506.16247 (2025).

[9] E. Collins, I. Augenstein, S. Riedel, A supervised approach to extractive summarisation of scientific papers, in: Proc. 21st Conf. on Computational Natural Language Learning (CoNLL 2017), ACL, Vancouver, Canada, 2017, pp. 195–205.

[10] L. Cagliero, M. La Quatra, Extracting highlights of scientific articles: A supervised summarization approach, Expert Systems with Applications 160 (2020) 113659.

[11] T. Rehman, D. K. Sanyal, S. Chattopadhyay, P. K. Bhowmick, P. P. Das, Automatic generation of research highlights from scientific, in: 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2021), collocated with JCDL 2021, 2021.

[12] T. Rehman, D. K. Sanyal, P. Majumder, S. Chattopadhyay, Named entity recognition based automatic generation of research highlights, in: Proceedings of the Third Workshop on Scholarly

Document Processing (SDP 2022) collocated with COLING 2022, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 163–169.

[13] T. Rehman, D. K. Sanyal, S. Chattopadhyay, Research highlight generation with elmo contextual embeddings, Scalable Computing: Practice and Experience 24 (2023) 181–190.

[14] T. Rehman, D. K. Sanyal, S. Chattopadhyay, P. K. Bhowmick, P. P. Das, Generation of highlights from research papers using pointer-generator networks and scibert embeddings, IEEE Access 11 (2023) 91358–91374. doi:10.1109/ACCESS.2023.3292300.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 7871–7880.

[16] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81.

[17] S. Banerjee, A. Lavie, 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, 2005, pp. 65–72.

[18] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, (ICLR 2020), 2020, pp. 1–43.