Technical Journal

Deep Learning Methods for Text Summarization

Rajesh Kamar [1*], Saroj Giri [2], Shiva Ram Dam [3] Suraj Basant Tulachan[4]

[1,2,3] *Department of Information Technology, Gandaki University, Nepal*

[4]*Department of Electronics and Computer Engineering, IoE, Paschimanchal Campus,Nepal*

**Corresponding Email: rajesh.kunwar@gandakiuniversity.edu.np*

Received: May 27, 2025; Revised: August 06, 2025; Accepted: September 11, 2025

doi : https://doi.org

Abstract

This review evaluates recent advancements in deep learning methods for text summarization, a key Natural Language Processing (NLP) task driven by the explosion of textual data. The goal is to generate concise summaries while preserving the core meaning of original texts. We analyze key deep learning architectures including Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Transformers using a systematic literature review (SLR) approach. Additionally, attention mechanisms, pointer-generator networks, and beam search techniques are explored, along with benchmark datasets such as CNN/Daily Mail and Gigaword. Transformer-based models consistently achieve superior performance in abstractive summarization, as evidenced by higher ROUGE-1 (43.9), ROUGE-2 (20.3), and ROUGE-L (40.1) scores on benchmark datasets like CNN/Daily Mail, compared to RNN and LSTM models. This performance gain is attributed to the self-attention mechanism and parallel sequence processing capabilities of Transformer architectures. Deep learning has transformed summarization, with Transformer models demonstrating clear advantages over earlier approaches. Future work should focus on hybrid models that integrate extractive and abstractive techniques, alongside developing more robust evaluation metrics that align closely with human judgment beyond n-gram overlap.

Keywords: *Text summarization, Deep learning, RNN, LSTM, Transformer, Attention mechanism, Evaluation metrics.*

1. Introduction

In today's digital era, the volume of textual information generated through news media, academic publications, corporate documents, and social media platforms is expanding exponentially. This deluge of data has created a pressing need for automated methods to distill essential information efficiently. Text summarization, a critical task in Natural Language Processing (NLP), aims to generate concise versions of longer texts while preserving their key ideas and meaning (Allahyari et al., 2017).

Text summarization has become increasingly important across a wide range of applications. In journalism, it enables news aggregation services such as Google News to deliver condensed updates. In academia, it supports literature review automation, helping researchers stay updated with recent developments. Social media platforms utilize summarization for real-time event monitoring and

content curation, while organizations use it to produce executive summaries of technical reports, legal documents, or customer feedback (Rush et al., 2015; See et al., 2017).

The key motivation behind developing summarization systems is to tackle information overload and facilitate quick decision-making. With the sheer volume of available text, manually reading and extracting insights becomes impractical. Automated summarization enables users to gain a high-level understanding quickly, supporting knowledge management, time-sensitive decision-making, and enhanced accessibility for diverse user groups and languages (LeCun et al., 2015). Traditional summarization techniques based on rule-based heuristics or statistical extraction—often produce disjointed or redundant outputs, especially for longer or complex documents. In contrast, deep learning models have enabled abstractive summarization, where the system generates novel sentences rather than copying from the source. This capability more closely mimics human summarization but introduces challenges such as repetition, factual inaccuracies, and handling out-of-vocabulary (OOV) words (See et al., 2017; Vaswani et al., 2017).

The advent of attention mechanisms, pointer-generator networks, and the Transformer architecture has significantly improved the quality of generated summaries. Transformers, in particular, can process entire sequences in parallel and capture long-range dependencies more effectively than sequential models like RNNs or LSTMs (Vaswani et al., 2017; Devlin et al., 2018). Given these developments, this paper provides a comprehensive review of deep learning methods for text summarization. It explores various model architectures, reviews commonly used datasets and evaluation metrics, highlights persistent challenges, and outlines potential future directions in this evolving field.

2. Materials and Methods

2.1 Deep Learning in Natural Language Processing : Deep learning has fundamentally reshaped the field of Natural Language Processing (NLP). By leveraging neural networks with multiple hidden layers, these methods automatically learn complex, hierarchical representations of language that were previously difficult to capture using traditional approaches. Early advances such as word embeddings (e.g., Word2Vec, GloVe) provided continuous vector representations that capture syntactic and semantic relationships between words. These word embedding techniques laid the foundation for subsequent deep learning architectures in NLP (Mikolov et al., 2013). The multi-layered nature of deep models allows the abstraction of high-level concepts from the raw text – lower layers capture local word dependencies, while higher layers synthesize broader contextual information. This capability is particularly beneficial for text summarization, where understanding both phrase-level meaning and document-level coherence is crucial. Notably, the general success of deep learning in NLP has been widely recognized (LeCun, Bengio, & Hinton, 2015).

2.2 Text Summarization Techniques

Text summarization techniques are commonly divided into two major approaches: extractive and abstractive summarization. In extractive summarization, the system selects key sentences or phrases directly from the source text and assembles them into a summary. This approach preserves the original phrasing and ensures grammatical correctness, but the resulting summaries may lack cohesion, especially in lengthy documents. In abstractive summarization, the system generates new sentences that capture the meaning of the source text in a condensed form. This requires a deeper understanding of the content and the ability to paraphrase. Abstractive methods, often powered by sequence-to-sequence neural networks, can produce more fluent and human-like summaries but are more complex

to implement. They also face challenges such as avoiding repetitive output, handling OOV words, and maintaining factual accuracy. Advances in encoder–decoder architectures with attention mechanisms have substantially improved abstractive summarization, helping to overcome early difficulties like limited vocabulary and context handling.

2.3 Evolution of Summarization Models

The evolution of summarization models reflects the broader progress in deep learning for NLP. A seminal work by Rush et al. (2015) introduced a neural attention model for abstractive sentence summarization, marking one of the first major applications of deep learning in summarization. Their encoder–decoder model with attention laid the groundwork for subsequent innovations. Later, See et al. (2017) refined this approach by incorporating pointer-generator networks, which directly copy words from the source text to address the OOV problem and reduce repetition. The introduction of the Transformer architecture (Vaswani et al., 2017) and large pre-trained language models like BERT (Devlin et al., 2018) ushered in a new era of summarization. Transformers enable parallel processing of sequences and capture long-range dependencies more effectively than RNNs, leading to significant performance gains. Research has also explored reinforcement learning to better align training objectives with evaluation metrics (Paulus et al., 2017) and adversarial training to improve the realism of generated summaries (Liu et al., 2018). Despite these advances, challenges remain – particularly in maintaining factual accuracy and handling ambiguous or rare terms.

2.4 Deep Learning Architectures for Text Summarization.

The core deep learning architectures applied to text summarization include recurrent neural networks and Transformer-based models. Each architecture offers distinct advantages and has influenced the design of summarization systems over time. Figure 1 illustrates a conceptual framework of a typical encoder–decoder summarization model with an attention mechanism, which is common to many of these architectures.

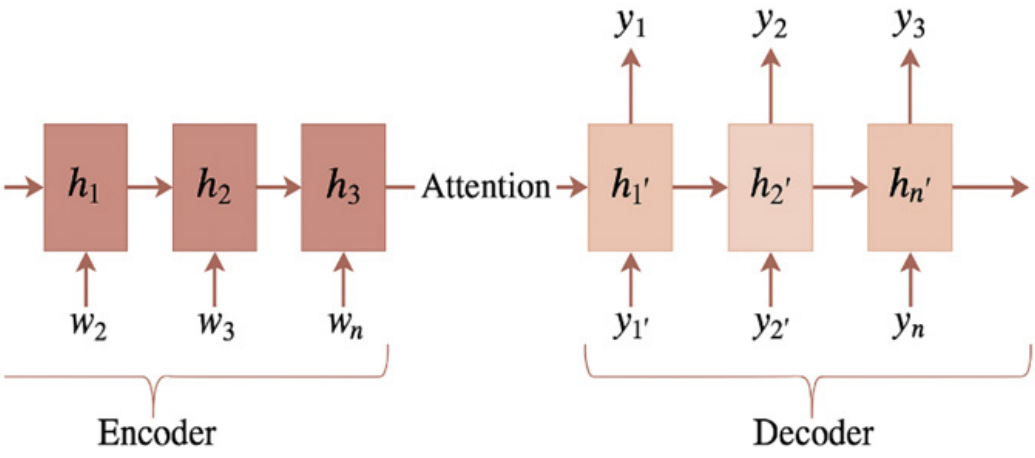

Figure 1: Conceptual framework of a deep learning-based text summarization model (encoder–decoder with attention).

2.4.1 Recurrent Neural Networks (RNNs): RNNs are among the earliest deep learning models applied to sequence-to-sequence tasks like summarization. Their ability to handle variable-length input sequences makes them a natural choice for processing text. In an RNN-based encoder–decoder architecture, the encoder processes the input text word-by-word and produces a sequence of hidden states that capture the context of the document. The final hidden state (or a combination of states) serves as a context vector that initializes the decoder, which then generates the summary sequentially, one word at a time. RNNs, however, suffer from limitations such as the vanishing gradient problem, which hampers learning long-term dependencies. For summarizing lengthy documents, vanilla RNNs often struggle to retain context over many time steps. This makes it challenging for them to produce coherent summaries of long texts without architectural enhancements.

2.4.2 Long Short-Term Memory Networks (LSTMs): LSTMs were introduced to address the shortcomings of traditional RNNs, particularly the vanishing gradient issue (Hochreiter & Schmidhuber, 1997). LSTMs incorporate gating mechanisms – including input, forget, and output gates – to regulate information flow through the network. These gates allow the model to maintain or forget information over long sequences, enabling the retention of important context across many time steps. In text summarization, LSTM-based encoder–decoder models have demonstrated improved performance, especially on tasks requiring understanding of extended context. For instance, summarization systems using LSTM encoders and decoders can generate more coherent and contextually relevant summaries than standard RNNs. The gating mechanisms help the model to focus on salient parts of the input text and ignore less relevant information, which contributes to the accuracy and fluency of the resulting summaries.

2.4.3 Gated Recurrent Units (GRUs): GRUs are a simplified variant of LSTMs that combine the input and forget gates into a single update gate, thus reducing the number of parameters and computational complexity (Cho et al., 2014). GRUs often achieve performance comparable to LSTMs while training faster due to their simpler structure. In abstractive summarization, GRU-based encoder–decoder models have been used effectively, especially when computational resources or training time are limited. They capture sequential dependencies in text sufficiently well to generate meaningful summaries. GRUs have been applied in real-time summarization scenarios – such as summarizing news as it streams or condensing social media content – where speed is crucial. Despite their relative simplicity, GRUs can maintain the context needed for summarization tasks, making them a practical choice when balancing performance and efficiency.

2.4.4 Transformer-Based Models and Attention Mechanisms: The introduction of the Transformer architecture (Vaswani et al., 2017) marked a paradigm shift in deep learning for NLP. Unlike RNNs, Transformers dispense with sequential processing in favor of a self-attention mechanism that allows them to consider all words in the input simultaneously. This design enables efficient parallelization and better capture of long-range dependencies in text. Attention mechanisms, which can be viewed as the core of Transformers, are also used within RNN/LSTM models (often as additive or multiplicative attention) to improve performance. In summarization models, attention dynamically focuses the decoder on different parts of the input text at each generation step, effectively telling the model which words to attend to when producing the next word of the summary. This results in more relevant and contextually appropriate word choices. Transformer-based models, such as those built on the original “Attention

is All You Need" framework, have achieved state-of-the-art results in summarization. Moreover, pre-trained Transformer models like BERTSUM (a BERT-based summarizer) and variants of GPT have been fine-tuned for summarization tasks, yielding significant improvements on benchmark datasets. These models leverage vast amounts of pre-training to better understand language, and when adapted to summarization, they produce summaries that are fluent and information-rich. Hybrid approaches have also emerged, combining Transformers with the pointer-generator concept to allow direct copying from the source text, which helps handle OOV words and maintain factual consistency. Overall, Transformer architectures and attention mechanisms have greatly advanced the capability of summarization systems to produce high-quality abstractive summaries.

3. Results and Discussion

3.1 Datasets and Evaluation

This section presents the datasets commonly used for developing and evaluating text summarization models, as well as the metrics used to evaluate summary quality. A comparison of model performance on standard metrics is provided, followed by a comparison of dataset characteristics. These elements align with the second objective of examining datasets and metrics in summarization research.

3.1.1 Datasets for Text Summarization

The availability of benchmark datasets has been crucial for training and evaluating deep learning summarization models. Among the most widely used datasets are:

- Gigaword: A large-scale corpus with millions of news articles and headline-length summaries (Napoles et al., 2012). Gigaword is commonly used for training models to generate very short summaries or headlines. It focuses on concise summaries, typically one sentence (about 15–20 tokens on average)
- CNN/Daily Mail: A collection of news articles paired with multi-sentence summaries (Hermann et al., 2015; See et al., 2017). This dataset contains on the order of 300k articles, each with an associated summary of several sentences (approximately 50–70 tokens). CNN/Daily Mail is used to train models for more detailed, paragraph-length summarization.
- DUC 2003/2004: Datasets from the Document Understanding Conference containing a smaller set of documents (e.g., ~500 in DUC 2004) with human-written multi-sentence summaries, often used for evaluation in both extractive and abstractive tasks. Summaries in DUC are longer (100–150 tokens on average) and the dataset covers a variety of domains.

Each dataset serves different summarization needs: Gigaword is ideal for generating very short summaries (like headlines), CNN/Daily Mail supports training for moderate-length news summaries, and DUC focuses on multi-document or longer abstractive summarization. Dataset selection depends on the task requirements – for instance, Gigaword is useful when brevity is paramount, whereas CNN/Daily Mail and DUC enable models to handle more context and produce longer summaries. Table 2 provides a comparison of these datasets. Notably, these benchmarks have driven progress in the field, but they also come with biases and idiosyncrasies (e.g., lead bias in news). Ensuring models generalize beyond these specifics is an ongoing challenge.

3.1.2 Preprocessing Techniques

Data preprocessing is a critical step in training effective summarization models. Common preprocessing techniques include tokenization (splitting text into words or subwords), lowercasing, and the handling of rare words. For example, many approaches replace low-frequency words with an <UNK> (unknown) token or use subword segmentation methods like Byte-Pair Encoding (BPE) to mitigate OOV issues. Consistent tokenization (such as using the Penn Treebank tokenizer) and normalization (removing or standardizing punctuation, numbers, etc.) are often applied to both inputs and reference summaries to ensure alignment in vocabulary. In some cases, sentence splitting is performed for multi-sentence summaries, and stopword removal may be considered for extractive methods (though typically not for abstractive methods, as it could remove important functional words). Preprocessing also involves constructing a vocabulary or using a pre-trained embedding vocabulary. When using pre-trained models (like BERT or GPT), the text is usually preprocessed according to that model's requirements (including special tokens and subword units). Overall, careful preprocessing helps improve model training stability and performance by reducing noise and focusing the model on salient patterns.

3.1.3 Evaluation Metrics

Evaluating the quality of generated summaries is challenging because it involves assessing both informativeness and linguistic quality. Several automated evaluation metrics have become standard in summarization research:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): The most widely used metric family in summarization. ROUGE measures overlap between the system-generated summary and a reference summary in terms of n-grams. Common variants include ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence). Higher ROUGE scores indicate that the generated summary has more words or sequences in common with the reference, which is interpreted as capturing more of the reference content. ROUGE is easy to compute and correlates reasonably with informativeness, but it has limitations (discussed later).
- BLEU (Bilingual Evaluation Understudy): Originally developed for machine translation, BLEU calculates the precision of n-grams of various lengths between the generated text and reference. It is sometimes reported for summarization, but it tends to penalize abstractive summaries that use synonyms or rephrasings not present in the reference (since BLEU expects exact matches). Thus, BLEU is generally less favored than ROUGE for summarization evaluation.
- METEOR: An improved MT metric that accounts for synonymy and stemming. METEOR aligns generated and reference summaries based on not only exact matches but also stemmed forms and synonyms, attempting to capture semantic similarity. METEOR often shows better correlation with human judgments than BLEU for summarization, particularly for abstractive summaries where wording can differ. These metrics provide quantitative benchmarks: for example, state-of-the-art models might achieve ROUGE-1 and ROUGE-2 scores in the 40s and high teens, respectively, on datasets like CNN/Daily Mail. Table 1 shows an example comparison of different model types by these metrics. However, it is well-known that automated metrics are imperfect proxies for summary quality. They do not capture readability, coherence, or factual correctness directly. As a result, researchers often perform human evaluation (ranking or rating summaries) to complement automated metrics, especially for high-quality systems that may have similar ROUGE scores.

3.1.4 Summarization Model Performance Comparison

Table 1 compares the performance of representative summarization models – RNN-based, LSTM-based, Transformer, and a BERT-based model (BERTSUM) – using standard evaluation metrics (ROUGE, BLEU, METEOR).

Table 1: Performance of different summarization models on standard metrics.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
RNN-based	36.5	14.2	33.1	18.9	20.3
LSTM-based	38.2	16.5	35.0	20.1	22.5
Transformer	41.8	18.6	39.4	24.3	25.7
BERTSUM (Transf.)	43.9	20.3	40.1	26.5	27.8

Table 1: Performance metrics are aggregated from multiple studies, including Rush et al. (2015), See et al. (2017), and Liu & Lapata (2019). These results are based on evaluations primarily conducted using the CNN/Daily Mail dataset.

As shown in Table 1, the evaluation results indicate that Transformer-based models achieve substantially higher scores across most metrics compared to earlier RNN/LSTM models. In particular, the BERTSUM model (which leverages a pre-trained Transformer encoder) achieves the highest ROUGE, BLEU, and METEOR scores of the group. This suggests that Transformer architectures – especially when combined with pre-training – generate more informative and fluent summaries. For instance, BERTSUM's ROUGE-1 of ~43.9 and ROUGE-2 of ~20.3 surpass those of the LSTM model by several points, reflecting its stronger ability to retain important information and phrase it effectively. The trend underscores the evolution noted earlier: Transformer-based summarization models have an advantage in capturing the gist of documents and phrasing summaries in a way that overlaps well with reference summaries. While these metrics-driven improvements are clear, it is important to also consider qualitative factors. Often, transformer-based summaries are judged by humans to be more coherent and less repetitive than those from RNN-based models, aligning with the metric gains.

3.1.5 Dataset Comparison

Table 2 compares the key characteristics of three common summarization datasets: Gigaword, CNN/Daily Mail, and DUC 2004.

Table 2: Comparison of summarization datasets.

Dataset	Number of Documents	Summary Type	Avg. Summary Length	Domain	METEOR
Gigaword	~4,000,000	Single sentence	15–20 tokens	News (wires)	20.3
CNN/Daily Mail	~300,000	Multi-sentence	50–70 tokens	News (articles)	22.5
DUC 2004	500	Multi-sentence	100–150 tokens	Mixed (news, etc.)	25.7

Table 2: Dataset information sourced from Napoles et al. (2012), Hermann et al. (2015), and DUC Workshop Proceedings (2004).

Table 2 shows that Gigaword is orders of magnitude larger than CNN/Daily Mail and DUC in the number of documents, but its summaries are single sentences (headlines) and much shorter. CNN/Daily Mail provides multi-sentence summaries of moderate length and focuses on the news domain, which makes it suitable for training models to produce paragraph-length news summaries. DUC 2004 has very few documents by comparison, but the summaries are relatively long and often used for evaluating summarization systems in a controlled setting (especially for multi-document summarization). Each dataset's size and summary length influence the kind of model that can be trained: Gigaword supports training of high-capacity models due to its size, whereas DUC's limited size means models often must be pre-trained or trained on other data before fine-tuning. Dataset selection, therefore, depends on the summarization task. For example, training a model on Gigaword would be appropriate for headline generation, while fine-tuning a pre-trained model on DUC 2004 might be useful for evaluating multi-document summarization techniques. Overall, these datasets complement each other, and using multiple datasets can help a model generalize better.Dataset selection also impacts evaluation. Models tend to achieve higher ROUGE scores on datasets with formulaic summaries (like news) due to easily predictable content (e.g., lead bias), whereas more diverse datasets can lower scores but may test a model's generality. As summarization research progresses, new datasets (like XSum, Newsroom, etc.) have been introduced to address some limitations of these classic datasets by providing different domains or requiring more abstraction.

3.2 Technical Limitations of Traditional Methods

Traditional deep learning architectures such as Recurrent Neural Networks (RNNs), including their variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), have contributed significantly to early progress in text summarization. However, they suffer from inherent limitations that impact performance and scalability, particularly when handling long or complex texts. RNNs, by design, process sequences one time step at a time, making their computation inherently sequential. This characteristic impedes parallel processing during training, resulting in slower convergence and higher computational costs. Moreover, standard RNNs are prone to the vanishing and exploding gradient problem, which affects their ability to learn long-term dependencies. This issue becomes especially problematic in summarization tasks involving lengthy documents, where the model struggles to retain contextual information from earlier parts of the input (Hochreiter & Schmidhuber, 1997). While LSTM and GRU architectures address some of these concerns by incorporating gating mechanisms to better manage information flow, they still maintain sequential processing and do not scale well to large datasets or long-range dependencies (Chopra et al., 2016; See et al., 2017).

In contrast, Transformer architectures (Vaswani et al., 2017) introduce a paradigm shift through their self-attention mechanism, which allows each token in the input sequence to attend to all others simultaneously. This enables parallelization during training, significantly improving speed and efficiency. Additionally, Transformers capture long-range dependencies more effectively, as they do not rely on incremental state propagation. These advantages translate into better summary quality, especially in abstractive tasks that require a deeper understanding of the overall document structure and semantics. The transition from RNN-based models to Transformers has thus addressed many bottlenecks of earlier approaches, enabling state-of-the-art performance on benchmark summarization tasks (Liu & Lapata, 2019).

4. Conclusions

This review highlights the transformative impact of deep learning on text summarization, tracing the evolution from early RNN-based architectures to state-of-the-art Transformer models. Traditional sequential models such as RNNs, LSTMs, and GRUs improved the capacity to retain contextual information across sequences, but they were hindered by technical limitations like vanishing gradients and inefficient training. The emergence of Transformer architectures, particularly those leveraging self-attention mechanisms and pre-training (e.g., BERTSUM), has significantly advanced the field by enabling parallelization and capturing long-range dependencies more effectively. Empirical results from benchmark datasets such as CNN/Daily Mail demonstrate that Transformer-based models consistently outperform earlier approaches in terms of ROUGE, BLEU, and METEOR scores.

In addition to architectural innovations, this study underscores the importance of understanding the limitations of current datasets and evaluation practices. While datasets like Gigaword, CNN/Daily Mail, and DUC have enabled substantial progress, they often exhibit biases—such as formulaic writing styles or lead bias in news articles—that can inflate evaluation scores and reduce generalizability. Similarly, standard metrics like ROUGE and BLEU, though widely used, rely on n-gram overlap and fail to capture semantic accuracy, coherence, or factual correctness. These shortcomings point to the need for more nuanced, human-aligned evaluation methods.

Looking forward, future research should prioritize the development of hybrid models that combine extractive and abstractive strategies, leveraging the strengths of both to generate more coherent and factually grounded summaries. Additionally, enhancing factual consistency remains a critical challenge; this calls for the integration of external knowledge sources and the adoption of advanced evaluation metrics that account for semantic and factual alignment. Addressing these issues will be key to building more robust, domain-adaptive summarization systems capable of supporting real-world applications in news, academia, healthcare, and beyond.

References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K.(2017). Text summarization techniques: A brief survey. arXiv preprint arXiv:1707.02268.

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 93–98).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Liu, Y., & Lapata, M. (2019). Text summarization with pre-trained encoders. arXiv preprint arXiv:1908.08345.

Liu, Y., Shen, X., & Song, L. (2018). Generative adversarial network for abstractive text summarization. arXiv preprint arXiv:1805.00000.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Napoles, C., Gormley, M. R., & Van Durme, B. (2012). Annotated Gigaword. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction (pp. 95–100).

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer- generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017) (pp. 107–117).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).