

An Enhanced Text Compression Approach Using Transformer-based Language Models

Chowdhury Mofizur Rahman¹, Mahbub E Sobhani², Anika Tasnim Rodela² and Swakkhar Shatabda³

¹Department of Computer Science and Engineering, State University of Bangladesh

²Department of Computer Science and Engineering, United International University

³Department of Computer Science and Engineering, BRAC University

Email: cmr.cse@sub.edu.bd, msobhani2410011@mscse.uiu.ac.bd,

anika@cse.uiu.ac.bd, swakkhar.shatabda@bracu.ac.bd

Abstract—Text compression shrinks textual data while keeping crucial information, eradicating constraints on storage, bandwidth, and computational efficacy. The integration of lossless compression techniques with transformer-based text decompression has received negligible attention, despite the increasing volume of English text data in communication. The primary barrier in advancing text compression and restoration involves optimizing transformer-based approaches with efficient pre-processing and integrating lossless compression algorithms, that remained unresolved in the prior attempts. Here, we propose a transformer-based method named RejuvenateFormer for text decompression, addressing prior issues by harnessing a new pre-processing technique and a lossless compression method. Our meticulous pre-processing technique incorporating the Lempel-Ziv-Welch algorithm achieves compression ratios of 12.57, 13.38, and 11.42 on the BookCorpus, EN-DE, and EN-FR corpora, thus showing state-of-the-art compression ratios compared to other deep learning and traditional approaches. Furthermore, the RejuvenateFormer achieves a BLEU score of 27.31, 25.78, and 50.45 on the EN-DE, EN-FR, and BookCorpus corpora, showcasing its comprehensive efficacy. In contrast, the pre-trained T5-Small exhibits better performance over prior state-of-the-art models.

Keywords—Text Compression, Transformer, Lossless Compression, Lossy Compression, Compression Ratio, Deep learning.

I. INTRODUCTION

Text compression, a fundamental aspect of data compression, refers to the process of reducing the size of textual data yet preserving its important information content. In simple terms, it indicates encoding text to represent it substantially smaller in size, thereby optimizing storage space, transmission bandwidth, and computational resources. The increasing amount of textual data generated and shared in a broad spectrum of domains, such as communication networks, digital libraries, web content, and document management systems, makes text compression indispensable. Worldwide, 18.7 billion texts are transmitted each day, not including messages exchanged between apps [1]. By 2022, 95% of businesses shifted from paper to digital documentation, increasing operational efficiency and minimizing security risks. Because of the scalability and security, cloud-based document management systems are utilized by 98% of tech companies

[2]. Due to the exponential growth of digital content, robust solutions for compressing text without losing its integrity or readability are important. Text compression techniques address this issue by ensuring efficient storage, retrieval, and transmission of textual data, thereby enhancing information systems' scalability, performance, and cost-effectiveness. Additionally, text compression has become vital for enabling efficient data transfer and interoperability across heterogeneous systems and platforms.

A range of approaches can be found in text compression techniques to reduce the volume of text data while maintaining its important information content. Lossless compression techniques, such as the Lempel-Ziv-Welch (LZW) [3] algorithm and Huffman coding [4], leverage patterns and redundancies in the text to produce compact representations while achieving compression without losing any data. Conversely, lossy compression techniques, such as statistical and dictionary-based approaches, trade off some quality in exchange for greater compression ratios. Recently, deep learning techniques have become efficient means for attaining advanced compression efficiency. These models may automatically find effective representations and encoding schemes by training on large corpora of text data. This enables greater compression performance over conventional techniques. The impact of transformer-based models—like BERT [5], LLaMA [6], and ALBERT [7]—has been emphasized in earlier research. These models have proven to be successful in preserving contextual information for both high and low-resource languages during text decompression. Architectural changes exhibiting good performance, such as the BBFNMT [8], are noteworthy. Not to mention, following the trend of refining domain-specific pre-trained models and integrating the LoRA [9] technique has also produced excellent outcomes [10]. Cross-lingual augmentation techniques improve transformer models' performance for languages based on various resource availability—a topic that was not addressed in the research of Mao et al. [11]. The study of Huang et al. [12] utilized the operational process of arithmetic coding and combined it with GPT for lossless text compression which was more of a complex approach with additional computational latency. For a comprehensive assessment of compression ratios, at least three general-purpose

compressors must be considered. It is also crucial to keep in mind that iterative assessment of meaning preservation may take time, and non-autoregressive decoding may not restore the original text flawlessly [13].

We addressed several constraints associated with text compression, especially concerning the lack of a futuristic approach that integrates lossless compressors and transformer-based methods. In this study, we propose a simple yet effective pre-processing technique to improve the compression ratio and restore the original text by leveraging a transformer-based method called RejuvenateFormer, wherein we optimize complexity and computational efficiency through a tailored architecture consisting of six encoder and decoder layers. The contribution of our study is summarized below:

- A simple strategic pre-processing technique tailored for the English language has been introduced, aimed to improve text compression ratios.
- After thorough pre-processing, various general-purpose lossless compression methods are utilized across multiple corpora to evaluate the effectiveness of our technique in terms of compression ratio. Notably, our method demonstrated state-of-the-art results, with the Lempel-Ziv-Welch (LZW) algorithm showcasing exceptional performance.
- The pre-trained T5-Small showcases adequate performance over RejuvenateFormer and BBFNMT; underscoring our approach of strategic pre-processing incorporating transformer-based encoder-decoder models has the potential to attain benchmark results; this paved the way for large-scale text restoration experiments.
- An exploration into the impact of the size of the training corpus on the efficacy of RejuvenateFormer in restoring missing characters within English text is conducted, shedding light on its potential contributions in the realm of decompression.

The following sections of the paper are organized as follows: an exhaustive review of the most relevant prior research on text compression. Next, we explain our text pre-processing technique. The subsequent section elucidates our proposed architecture. Afterward, we present a detailed account of the experimental analysis, and, finally, we offer concluding remarks and discuss future research avenues.

II. RELATED WORK

In recent years, we have experienced an upsurge of digital documents in a broad spectrum of domains, including government, business, healthcare, and academia. Such rapid changes have led to an extreme need for reliable compression algorithms that can reduce data sizes without risking the content's quality and integrity. This literature review presents an in-depth evaluation of document compression techniques for lossless and lossy methods, ranging from traditional to trendy techniques while highlighting the methodologies, guiding principles, benefits, and limitations.

Transformer-based techniques, utilizing GPT, LSTM, and a single-layer transformer, have been widely utilized for text

compression. Valmeekam et al. [6] propose a novel study that leverages large language models (LLMs) such as LLaMA-7B [14] to present a novel approach for lossless text compression. Using a window of prior tokens, the method utilizes the LLM to determine the next token (based on probability ranking) in a text sequence. The rank or probability of the actual token is then encoded using a lossless compression scheme, such as zlib, token-by-token, or arithmetic coding. Subsequently, harnessing the decoder-only transformer Schmidt et al. [15] developed PathPiece, a new tokenizer that splits a document's content into the minimum number of tokens necessary for specific vocabulary, to test the hypothesis that fewer tokens improve downstream performance. In this study, the authors trained 64 language models utilizing distinct tokenization and parameter sizes ranging from 350M to 2.4B.

One of the LSTM-based techniques that Goyal et al. [16] proposed is a technique that uses recurrent neural networks, for prediction and arithmetic coder algorithms for lossless compression that achieves around ideal compression for synthetic datasets, outperforming Gzip on real datasets. In contrast, the LSTM-based model can perform significantly well on classification tasks [17].

Among lossy text compression, Mingxiao et al. [5] presents an innovative text compression strategy that offers superior compression performance and semantic stability by masking fewer significant words and restoring them through a transformer-based model. The masked text is converted into a bit string using LZW encoding, and then sentence-BERT evaluates semantic importance. Masked words can be restored using a transformer-based demasking module, lowering the cross-entropy between ground truth and soft decisions. Whereas, Zuchao et al. [7] proposed an approach whose goal was to extract the core information of the input text by improving the text encoding of transformer models with text compression using two different approaches. They are Explicit Text Compression (ETC) and Implicit Text Compression (ITC). The basic idea of ITC is to produce compressed text features without generating an explicit text sequence through the use of a non-autoregressive decoder.

Amid BERT-based methods, Kale et al. [18] used neural networks for mutual information estimation and information-theoretic compression in addition to a benchmark language model for initial text representation. In terms of the predictive accuracy of downstream models trained on the compressed data, their studies show notable improvements in keeping maximal targeted information and minimal sensitive information over detrimental compression ratios. Furthermore, Zhang et al. [19] offer the DC-Graph model to overcome the issues with reconstructing and augmenting information in lengthy texts incorporating different optimized modules with pre-trained BERT. Their proposed architecture outperforms Recurrent Chunking Mechanisms and BERT.

Anisimov et al. [20] created a basic monotonic mapping from the set of non-negative integers to the codeword set, which is utilized to construct a fast byte-aligned decoding algorithm showcasing a better compression ratio when compared

to known codes of the same type.

Khan et al. [21] proposed a method that combines Named Entity Recognition (NER) and Graph Neural Networks (GNNs) for text summarization, focusing on important document structures and relevant entities. For summarizing large text, this integration ensures both relevance and efficiency.

Aslanyurek et al. [22] introduced WSDC (Word-based Static Dictionary Compression) and used iterative clustering to build static dictionaries for high compression ratios in short texts. DSWF (Dictionary Selection by Word Frequency) is proposed to select the most suitable dictionary for compressing the source text. WSDC shows superior performance than other methods for short texts under 200 and 1000 bytes(except Zstd).

III. DATASET

A. Data Sourcing

The source of our data is from two publicly available standard large-scale corpora: WMT14 and BookCorpus. Within the WMT14, EN-DE, and EN-FR language pairs datasets comprise 1.5 million and 1.6 million sentence pairs respectively and the BookCorpus dataset contains 7.8 million English sentences.

B. Data Preprocessing

Beneath the WMT14 dataset, we consider English-to-German and English-to-French language pairs. We only extracted the column containing English text from both the dataset. In the BookCorpus dataset, it was not necessary to do this procedure as it has a single column containing English text. Then, we identified 5 vowel letters, in both capital and small letters, that appear in the English text. These vowel letters are denoted as $VC = \{VC_1, VC_2, \dots, VC_{10}\}$. Next, we take every sentence indicated as $S = \{S_1, S_2, \dots, S_N\}$, where N indicates the number of characters in the sentence. We traverse through each of the characters $S_i \in S$ and remove any character that is present in the character set VC . By following the above process, the dataset has been constructed in source-target pairs. In this source-target pair, the source sentence has no vowel letters, while the target sentences are corrected versions with missing vowel letters.

IV. PROPOSED METHOD

A. Problem Formulation & Overview

Consider input sentences $S = \{S_1, S_2, \dots, S_N\}$, where N is the number of character. The vowel remover ($VR(\cdot)$) takes the input sentence and removes the vowels. The Lempel-Ziv-Welch compressor ($LZWC(\cdot)$) takes the sentence and the compressed representation is stored in a hard disk, whereupon calculations regarding length are computed. The Lempel-Ziv-Welch decompressor ($LZWD(\cdot)$) decompresses the compressed representation. Subsequently, consider pair of tokens, $SRC_I = \{src_1, src_2, \dots, src_n\}$ and $TGT_I = \{tgt_1, tgt_2, \dots, tgt_k\}$, where SRC_I represents the input sequence that has no vowel, and TGT_I represents the target sequence with vowel restored. Erroneous input sequence SRC_I is tokenized using a subword tokenizer ($T(\cdot)$) and then fed

into the encoder ($E(\cdot)$) and generates context vectors of the sentence, denoted as $CV = \{CV_1, CV_2, \dots, CV_{512}\}$. The decoder ($D(\cdot)$) utilized the context vector CV , alongside the previously decoder generated tokens to generate the correct sentence autoregressively. The pre-processing can integrate seamlessly with any transformers-based encoder-decoder models. The entire procedure can be mathematically abbreviated as follows:

$$SRC_I = LZWD(LZWC(VR(S))) \quad (1)$$

$$T\hat{G}T = D((E(T([SRC_I]), W^E), D_{out}^{t-1}), W^D) \quad (2)$$

B. RejuvenateFormer

In this section, we discuss the details of RejuvenateFormer.

1) **Lempel-Ziv-Welch:** Lempel-Ziv-Welch (LZW) compression forms a dictionary of symbols, replaces repeated substrings with dictionary indices, and updates the dictionary during compression. During decompression, it initializes, reads, retrieves, and updates the dictionary to reconstruct the original data.

2) **Vowel Removal and Compression:** In our text processing pipeline, the function ($VR(\cdot)$) removes vowels from input sentences, producing a text devoid of vowels. Subsequently, the Lempel-Ziv-Welch compressor ($LZWC(\cdot)$) generates a compressed representation of the text, which is then used to calculate the length optimization ratio. Lastly, the compressed representation is decompressed using the ($LZWD(\cdot)$), restoring the original text without vowels, which is then fed into our transformer model to recover the vowels.

3) **Encoder:** We transform each sentence into a sequence of tokens $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ and assign discrete values to each subword. Here n is the sequence length. We confirmed the same input dimension to every input sequence by incorporating padding to $\mathbf{X}_i$. Subsequently, each token x_i went through an embedding layer $\mathbf{E}$ to transform the discrete values to continuous vectors which represent trainable matrix $\mathbf{E}x_i = \text{Embedding}(x_i)$. Through backpropagation during training, these metrics are fine-tuned to minimize the loss. To preserve the positional order of the tokens in a sentence we incorporated positional encoding $\mathbf{PE}$ for each token x_i , denoted as $\mathbf{PE}x_i$ with the embedding. This composite embedding, represented as $\mathbf{Z}x_i = \mathbf{E}x_i + \mathbf{PE}x_i$ is then fed into the stack of $\mathbf{K}$ identical encoder layers which are built upon two main components: a position-wise-feed-forward network and a multi-head-self-attention layer. Furthermore, the self-attention mechanism calculates attention score for each token by looking into all other tokens in a sequence $\mathbf{X}$ using three learnable vectors, query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) for capturing the contextual dependencies in a sequence of tokens. The mathematical expression of self-attention is defined as follows [23]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

The position-wise-feed-forward mechanism introduces non-linearity to the encoder through two linear transformations by incorporating ReLU, which is a non-linear activation function

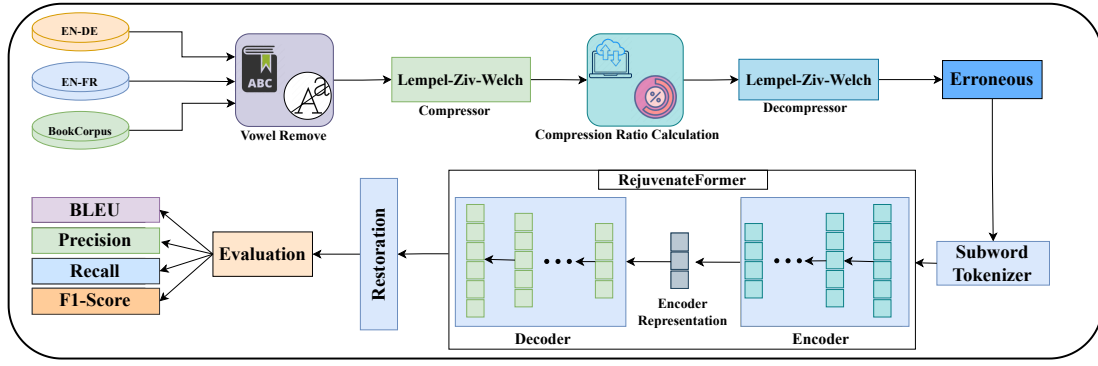


Fig. 1. **(Top)** Each corpus undergoes vowel removal, and the compression ratio is calculated using the compressed representation. The text is then reverted to its earliest form without vowels. **(Bottom)** After tokenization, the RejuvenateFormer is trained on each corpus, to proficiently generate expected outcomes.

for enhancing the obtained representation. The output of each encoder layer is propagated sequentially to K identical layers which output rich contextualized representation of input sequence X covering local and global dependencies.

4) **Decoder:** The target sequence, $Y = \{y_1, y_2, \dots, y_m\}$, where m is the sequence length, undergoes through an embedding layer $Ey_i = \text{Embed}(y_i)$ for converting tokens with discrete values into continuous vectors. For preserving positional information of tokens in a sequence, positional encoding PEy_i is blended with embedding Ey_i and the resulting embedding is $Zy_i = Ey_i + PEy_i$. This combined embedding is passed through L identical decoder layers. The decoder layers are composed of two main components: a position-wise-feed-forward layer and a masked multi-head-self-attention layer. Masking is incorporated in the decoder so that it can not generate the current token by attending the future tokens. Following this distinction, the computational process of masked-multi-head-self-attention follows the same equation (3) as the encoder's multi-head-self-attention. Furthermore, the position-wise-feed-forward network enhances the learned representation by following the same procedure as the encoder's position-wise-feed-forward network. The acquired learned representation undergoes through L identical decoder layers and produces the fine-grained target sequence.

5) **Hyperparameters:** The hidden dimension is kept as 512 through all the encoder and decoder layers for maintaining consistency. To maintain the model's depth and capacity the number of neurons is kept at 2048 for the feed-forward layer and a 0.1 dropout ratio is applied to prevent overfitting. To maintain the efficient computation and non-linearity over the network we have incorporated ReLU. The model went through 50 epochs with 5×10^{-5} learning rate incorporating AdamW optimizer. We incorporated the categorical cross-entropy loss function for the optimization process, which leads the model towards desired translations.

V. EXPERIMENTAL ANALYSIS

A. Datasets

Because of limited computational resources, we selected 100K sentence pairs from each of the three datasets.

- **WMT14 [24]** The WMT14 dataset was used as shared tasks of the Ninth Workshop on Statistical Machine Translation. Within five language pairs, we used English-to-French (EN-FR) and English-to-German (EN-DE) datasets, which are both large-scale standard corpus.

- **English-French.** It is composed of 1.6M source-target pairs after our rigorous pre-processing. The corpus was divided into training, validation (newstest2013), and test (newstest2014) sets. As a result, the training, validation, and test sets comprise 100000 (100K), 3000 (3K), and 3003 (3k) source-target pairs respectively.
- **English-German.** The dataset contains a total of 1.5M source-target pairs after our thorough pre-processing. We separated the corpus into training (100K), validation, and test sets, keeping newstest2013 as the validation set and newstest2014 as the data in the test set. Accordingly, the resulting training, validation, and test set comprise 100000 (100K), 3000 (3K), and 3003 (3k) source-target pairs respectively.

- **BookCorpus [25]** An extensive set of free novel books, mostly authored by unpublished authors are included in the BookCorpus dataset. Following a meticulous text pre-processing phase, we found a total of 7.8M source-target pairs within the corpus. Consequently, we partitioned the corpus into 100000 (100K) pairs in the training and 5000 pairs in the test set to create distinct training and test sets.

B. Baselines

- **T5-Small [26]** Text-to-Text Transformer (T5) is a language model developed by Google. T5-Small is the smaller variant of T5 consisting of ($\approx 70M$) parameters where the base version is of (220M) parameters. The T5-Small was created aiming to maintain good performance with a smaller number of parameters.

C. Performance Evaluation

We evaluated the performance of our approach with compression ratio (**Original Length/Compressed Length**). We

TABLE I
THE COMPARISON OF THE QUANTITATIVE PERFORMANCE OF DIFFERENT EXISTING METHODS ACROSS DIFFERENT DATASETS.

Method	#Params.	WMT14 EN-DE				WMT14 EN-FR				BookCorpus			
		BLEU	PR	RE	F1	BLEU	PR	RE	F1	BLEU	PR	RE	F1
BBFNMT [8]	230.3M	29.37	—	—	—	42.52	—	—	—	—	—	—	—
T5-Small	60.51M	41.44	0.93169	0.91312	0.92211	34.57	0.9224	0.9039	0.9129	63.61	0.9703	0.9539	0.9618
RejuvenateFormer	63.23M	27.31	0.8717	0.9139	0.8921	25.78	0.8691	0.9131	0.8903	50.45	0.9384	0.9527	0.9454

also evaluated the performance of our model in restoring vowels by calculating BERTScore (6) and BLEU (7) scores.

- **BERTScore.** [27] The BERTScore calculates the semantic similarity of two pieces of text by calculating the cosine similarity of their embedding tokens. This metric outputs precision, recall, and f1 score, and their equations are as follows:

$$Recall_{BERT} = \frac{1}{N} \times \sum_1^N \left(\frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \right) \quad (4)$$

$$Precision_{BERT} = \frac{1}{N} \times \sum_1^N \left(\frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \right) \quad (5)$$

$$F1_{BERT} = \frac{1}{N} \times \sum_1^N \left(2 \times \frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}} \right) \quad (6)$$

where $\mathbf{x}_i^\top \hat{\mathbf{x}}_j$ is the cosine similarity between two pieces of text and N is the total number of sentence pairs.

- **BLEU.** [28] The BLEU metric estimates the quality of candidate text by assigning precision scores to n-grams and comparing them with one or more reference texts. Scores range from 0 and 100, where a higher score denotes better results. The mathematical formula for BLEU is as follows:

$$BLEU = BP \times e^{\sum_{n=1}^N (w_n \cdot \log p_n)} \quad (7)$$

Here, The Brevity Penalty (BP) punishes shorter predictions. N is the maximum n-gram length. w_n are weights for n-gram precision, and $\log p_n$ is the logarithm of n-gram precision in the candidate text.

D. Experimental Results

1) **Quantitative Results:** We evaluated the results of our pre-processing technique on all three corpora. We tested the significance of removing vowels in compressing text using various general-purpose lossless compressors such as LZMA, LZW, GZIP [29], ZLIB [30], and Arithmetic Coding (AC) [31]. Our rigorous experiment found that LZW compresses text 13.38 $\times$, 12.57 $\times$, and 11.42 $\times$ on EN-DE, BookCorpus, and EN-FR corpus respectively, and establishes itself as the state-of-the-art in compression ratio by outperforming GPT-based approach [12], TRACE [11], and other compressors. Table II demonstrates that removing vowels significantly improves the compression ratio and lowers the length incorporating compressed representation. In contrast, we presented the quantitative performance of different Transformer-based

methods in Table I. Our proposed model, RejuvenateFormer, shows favorable performance across all three corpora in small-scale experiments being 3.6 $\times$ small in terms of parameter count. In contrast, the pre-trained T5-Small performed significantly well over BBFNMT and RejuvenateFormer across all

TABLE II
THE JUXTAPOSITION OF THE COMPRESSION RATIO OF DIFFERENT EXISTING METHODS ACROSS VARIOUS CORPORA.

Dataset	Corpus Size	Vowel Remove (Ours)					GPT	TRACE
		LZW	LZMA	GZIP	GLIB	AC		
BookCorpus	6.7M	12.57	4.583	3.603	3.575	2.907	10.55	4.49
EN-DE	1.5M	13.38	5.257	3.959	3.927	2.947	—	—
EN-FR	1.6M	11.42	4.656	3.643	3.626	2.953	—	—

evaluation metrics and corpora, except for the EN-FR dataset with only 40K sentence pairs and 30 epochs. This highlights the effectiveness of larger contextual models when integrated with our proposed text pre-processing strategy. However, we did not consider the scores for each sentence separately; instead, we calculated the overall BLEU and BERTScore by evaluating all the sentences in a corpus.

2) **Qualitative Results:** We evaluated the qualitative performance of T5-Small and RejuvenateFormer. Our proposed model showed superior performance in both training and inference times compared to T5-Small in a variety of sentence conditions. We found that while T5-Small was accurate in restoring short sentences with a low number of missing vowels, it struggled with longer sentences that had more missing vowels. However, due to its training on a larger corpus, T5-Small outperformed RejuvenateFormer in terms of the overall score, benefiting from a wider linguistic context. In contrast, our RejuvenateFormer model demonstrated its competitive performance by accurately restoring sentences with a higher number of missing vowels in different sentence structures.

3) **Ablation Study:** Table III shows the impact of corpus

TABLE III
THE INFLUENCE OF CORPUS SIZE (EN-DE) ON THE PERFORMANCE OF OUR PROPOSED METHOD.

Method	Corpus Size	Inference			
		BLEU	PR	RE	F1
RejuvenateFormer	30K	23.07	0.8714	0.9012	0.8859
RejuvenateFormer	50K	25.12	0.8720	0.9071	0.8890
RejuvenateFormer	100K	27.31	0.8717	0.9139	0.8921

size on model performance. In this extensive study, we used three large-scale datasets but the correlation between the size of the corpus and the effectiveness of the model performance has been shown for the WMT14 EN-DE language pair dataset. Due to computational constraints, we limited our experiment to a smaller number of sentence pairs. Surprisingly, the corpus with 100K instances outperformed those with 30K and 50K instances. Conversely, the corpus with 30K instances showed the least significant results, while the 50K instances corpus delivered moderate outcomes. This tendency was seen in all three datasets, signifying that larger corpus sizes lead to enhanced performance [32].

VI. CONCLUSION

This study identified the primary obstacle in text compression and proposed a comprehensive architecture. We introduced a new pre-processing technique that leverages the compatibility of the Lempel-Ziv-Welch algorithm to reduce the length of text, resulting in a state-of-the-art compression ratio on EN-DE, EN-FR, and BookCorpus dataset. Concurrently, we validated the efficacy of our proposed technique on these three large-scale corpora, confirming its compatibility with transformers-based encoder-decoder architecture like pre-trained T5-small and establishing it as a promising method for text restoration tasks. Notably, we introduced a transformer-based text restoration method, meticulously designed to address complex linguistic patterns by harnessing attention mechanisms and positional encoding resulting in promising evaluation scores even in small-scale experiments. In our future study, we will incorporate our transformer-based method into large-scale training and experiments, and evaluate the efficiency of knowledge distillation.

REFERENCES

- [1] V. S. S. H. Office, "Worldwide texting statistics." <https://shso.vermont.gov/sites/ghsp/files/documents/Worldwide> [Online; accessed 2024-07-01].
- [2] pdfreaderpro, "100 document management statistics for the digital era." <https://www.pdfreaderpro.com/blog/document-management-statistics>, 8 2023. [Online; accessed 2024-07-01].
- [3] M. R. Nelson, "Lzw data compression," *Dr. Dobbs' Journal*, vol. 14, no. 10, pp. 29–36, 1989.
- [4] D. E. Knuth, "Dynamic huffman coding," *Journal of algorithms*, vol. 6, no. 2, pp. 163–180, 1985.
- [5] M. Li, R. Jin, L. Xiang, K. Shen, and S. Cui, "Crossword: A semantic approach to data compression via masking," *arXiv preprint arXiv:2304.01106*, 2023.
- [6] C. S. K. Valmeekam, K. Narayanan, D. Kalathil, J.-F. Chamberland, and S. Shakkottai, "Llmzip: Lossless text compression using large language models," *arXiv preprint arXiv:2306.04050*, 2023.
- [7] Z. Li, Z. Zhang, H. Zhao, R. Wang, K. Chen, M. Utiyama, and E. Sumita, "Text compression-aided transformer encoding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3840–3857, 2021.
- [8] Z. Li, R. Wang, K. Chen, M. Utiyama, E. Sumita, Z. Zhang, and H. Zhao, "Explicit sentence compression for neural machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8311–8318, 2020.
- [9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [10] T. Ge, J. Hu, X. Wang, S.-Q. Chen, and F. Wei, "In-context autoencoder for context compression in a large language model," *arXiv preprint arXiv:2307.06945*, 2023.
- [11] Y. Mao, Y. Cui, T.-W. Kuo, and C. J. Xue, "Trace: A fast transformer-based general-purpose lossless compressor," in *Proceedings of the ACM Web Conference 2022*, pp. 1829–1838, 2022.
- [12] C. Huang, Y. Xie, Z. Jiang, J. Lin, and M. Li, "Approximating human-like few-shot learning with gpt-based compression," *arXiv preprint arXiv:2308.06942*, 2023.
- [13] T. Ge, H. Xia, X. Sun, S.-Q. Chen, and F. Wei, "Lossless acceleration for seq2seq generation with aggressive decoding," *arXiv preprint arXiv:2205.10350*, 2022.
- [14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.
- [15] C. W. Schmidt, V. Reddy, H. Zhang, A. Alameddine, O. Uzan, Y. Pinter, and C. Tanner, "Tokenization is more than compression," *arXiv preprint arXiv:2402.18376*, 2024.
- [16] M. Goyal, K. Tatwawadi, S. Chandak, and I. Ochoa, "Deepzip: Lossless data compression using recurrent neural networks," *arXiv preprint arXiv:1811.08162*, 2018.
- [17] A. T. Rodela, H.-H. Nguyen, D. M. Farid, and M. N. Huda, "Bangla social media cyberbullying detection using deep learning," in *International Conference on Intelligent Systems and Data Science*, pp. 170–184, Springer, 2023.
- [18] H. K. Kale, H. Esfahanizadeh, N. Elias, O. Baser, M. Medard, and S. Vishwanath, "Texshape: Information theoretic sentence embedding for language models," *arXiv preprint arXiv:2402.05132*, 2024.
- [19] G. Zhang, X. Li, and H. Zhang, "Dc-graph: A chunk optimization model based on document classification and graph learning," *Artificial Intelligence Review*, 2024.
- [20] A. Anisimov, I. Zavadskyi, and T. Chudakov, "Natural-language text compression using reverse multi-delimiter codes," *Cybernetics and Systems Analysis*, pp. 1–12, 2024.
- [21] I. Z. Khan, A. A. Sheikh, and U. Sinha, "Graph neural network and ner-based text summarization," *arXiv preprint arXiv:2402.05126*, 2024.
- [22] M. Aslanyürek and A. Mesut, "A new method for short text compression," *IEEE Access*, 2023.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, E. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, et al., "Findings of the 2014 workshop on statistical machine translation," in *Proceedings of the ninth workshop on statistical machine translation*, pp. 12–58, 2014.
- [25] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.
- [26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [27] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.
- [28] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, (Belgium, Brussels), pp. 186–191, Association for Computational Linguistics, Oct. 2018.
- [29] P. Deutsch, "Gzip file format specification version 4.3," tech. rep., Association for Computing Machinery, 1996.
- [30] P. Deutsch and J.-L. Gailly, "Zlib compressed data format specification version 3.3," tech. rep., Association for Computing Machinery, 1996.
- [31] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, 1987.
- [32] M. H. Bijoy, M. F. A. Faria, M. E. Sobhani, T. Ferdoush, and S. Shatabda, "Advancing bangla punctuation restoration by a monolingual transformer-based method and a large-scale corpus," in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pp. 18–25, 2023.