Leveraging Pivoting Techniques for Summarization in Low-Resource Languages: Insights from Bangla

Saifullah Bin Yusuf[1] and S. M. Mahbubur Rahman[1,2]

[1]Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka 1205, Bangladesh.
[2]Department of Electrical and Electronic Engineering
University of Liberal Arts Bangladesh Dhaka 1207, Bangladesh.

Contributing authors: saifsearcher@gmail.com;
mahbubur@eee.buet.ac.bd; mahbubur.rahman@ulab.edu.bd;

Abstract

This paper accesses the performance of using pivoted datasets for Bangla text summarizing as compared to human-generated datasets by a fine-tuned transformer model such as mT5-small, BanglaT5 or mBART. The pivoted dataset is created by translating Bangla text to English, summarizing it using a pre-trained T5-small model and then translating back to Bangla. The goal is to assess, whether the pivoted data can effectively substitute human-generated data for a low-resource language like Bangla. The XLsum dataset that contains Bangla-and-English article-summary pairs from BBC, is used to create synthetic datasets through the pivoting technique. The models are fine-tuned on both original and synthetic datasets by maintaining consistent training parameters. Summarization performance is evaluated using the BLEU, METEOR, and chrF++ scores. Experimental results reveal that pivoted datasets achieve around 85% of the performance of human-generated datasets on average in METEOR and chrF++ scores. In terms of BLEU scores, the performance is comparable when human generated data is combined with the pivoted data in different models. Our findings suggest that while human-generated data provides a better performance, the pivoted datasets can be viably used for summarizing tasks in a low-resource language, like Bangla.

Keywords: Low-resource language, pivoting technique, summarization, transformer models

1 Introduction

Bangla, although spoken by over 230 million people, is a low-resource language lacking sufficient datasets and NLP tools, hindering advancements in key areas like machine translation, sentiment analysis, and text summarization. This research addresses the scarcity of high-quality Bangla text summarization datasets, critical for applications such as summarizing news articles, video descriptions, and literature. Manually creating such datasets is both time-consuming and resource-intensive. If proven viable, the pivoting technique can efficiently synthesize datasets, significantly reducing time and effort. This study synthesizes new Bangla datasets, evaluates their performance alongside human-generated ones, and explores the effectiveness of combining real and synthetic data for Bangla text summarization. Thus, the contributions of this study include the creation of a new dataset comprising pivoted 'text paragraph and summary' pairs, which can be used for future research purposes. The study further compares summaries generated using the human-generated and pivoted summaries from the same text dataset, evaluating them based on BLEU[1], METEOR[2] and chrF++[3] scores, as well as through human assessment. Finally, the research examines the comparative performance of these approaches in different news categories including sports, politics, international news, and science and technology.

This paper is organized as follows: Section II discusses the related works previously conducted in this area; Section III presents the proposed methodology and details of the experiment; the results obtained are discussed in Section IV; finally, Section V provides the conclusion. The codes and datasets for this work are made available at the GitHub repository [4] for the sake of reproducibility and further research.

2 Background

This section provides the related works of the research study, scope, and our specific contributions presented in this research.

2.1 Related Work

Bangla text summarization remains an under-explored field, primarily constrained by the limited availability of datasets and resources, as discussed in [5], and [6]. Most research in dominant language, English benefits from large datasets and advanced models, employing both extractive and abstraction methods. Extractive approaches focus on selecting key sentences from the text, while abstraction methods aim to generate new sentences that convey the document's essence [7]. However, these methodologies often struggle in Bangla due to the lack of high-quality training data. A notable review in [5] analyzed fourteen Bangla text summarization approaches, highlighting their strengths, limitations, and areas for improvement. This study emphasized the need for more comprehensive datasets and refined methods to enhance summarization performance. Similarly, [6] trained its model with only a few hundred text-summary pairs, a stark contrast to English models like [8], which use hundreds of thousands of pairs, reflecting the significant data disparity. This data scarcity remains a critical bottleneck for developing robust Bangla NLP models.

Despite these challenges, some research studies have made notable progress such as the Bengali Abstraction News Summarization (BANS) model [7] that utilizes a sequence-to-sequence LSTM network with attention mechanisms leveraging a dataset of over 19,000 articles from bangla.bdnews24.com. The attention mechanisms in both the encoder and decoder of this method play a vital role in enhancing the output. In extractive summarization, the authors of [9] rank sentences based on term frequency, sentence position, and keyword presence, creating a system that efficiently reduces document length while preserving critical information to produce concise summaries. Unsupervised approaches have also shown promise in Bangla summarization. In [10], a method is introduced that uses sentence embeddings and clustering techniques, where pre-trained embeddings represent sentences in a high-dimensional space. By clustering similar sentences, the approach identified the most representative ones for summarization. Synthetic data generation has also been explored in [11], wherein a pivoting technique is used to synthesize Bangla paraphrasing datasets by employing English as an intermediary. While this work focused on paraphrasing, it highlights potential strategies for addressing resource limitations in Bangla summarization. Large-scale multilingual generative models like mT5 and mBART offer additional avenues for Bangla NLP. However, their effectiveness is limited by sufficient fine-tuning.

2.2 Scope

The literature reveals that there remains a scope to addresses the pressing challenges in Bangla text 'summarization' by evaluating both human-generated and synthetic datasets. The pivoting technique is examined as a potential solution to bridge the resource gap by leveraging English as an intermediary language similar to 'paraphrasing' done in [12]. In addition, the research can assess how transformer models such as the Google mT5-small [13], BanglaT5 [14], and mBART [15] perform when fine-tuned on human-generated datasets versus synthetic pivoted datasets. By exploring these approaches, the study can evaluate the potential of the pivoted dataset for low resource language 'summarization'.

2.3 Specific Contributions

The specific contributions of the paper are as follows:

- Introducing a novel application of the pivoting technique for dataset synthesis in Bangla text 'summarization'
- Evaluation by transformer models such as Google mT5-small, BanglaT5 and mBART in low-resource settings, highlighting category-wise performance
- Revealing that pivoted datasets achieve up to 85% of the performance of human-generated datasets using METEOR and chrF++ metrics

3 Experiments

This section describes how the human-generated and pivoted datasets are available for summarization. Next, the setup of the experiment and evaluation metrics are presented.

3.1 Datasets

The paper uses the BBC XLsum Bangla text summarization dataset [16], that is human generated. It has 8102 pairs of summary for training and 1012 each for validation and testing. The pivoting technique take advantage of English as an intermediate language through translation, summarizing, and back-translation. In particular, this pivoted dataset has been synthesized by translating full set of Bangla paragraphs in the BBC XLsum dataset into English using the Bangla-EnglishNMT model [3]. These translations were summarized using the mT5-small model, which is fine-tuned on 10,000 English text-summary pairs from the same BBC XLsum dataset. The summarized English text was back-translated into Bangla using the same NMT model, generating the final pivoted dataset.

3.2 Setup

The experiments is carried out by using three fine-tune transformer models, viz., the Google mT5-small [13], BanglaT5 [14], and mBART [15]. The performance of three distinct datasets, namely, the original human-generated dataset, the synthesized pivoted dataset, and a generic dataset combining these are evaluated. These datasets undergo a pre-processing that is appending 'summarize:' prefix to the text paragraphs to indicate the summarization task. Tokenization and padding techniques are applied uniformly across all datasets to ensure a standardized input format.

As per The transformer models are initialized with pre-trained weights and fine-tuned for the summarization task using the AdamW optimizer with a decay rate of 0.01. Fine-tuning is performed on each of the three datasets by adjusting hyper-parameters like batch size and training epochs to optimize the performance. The training, validation and testing parts of the experiments on processed dataset are kept same as that of the BBC XLsum.

3.3 Evaluation Metrics

The performance of the transformer models were evaluated bu using the BLEU[1], METEOR[2] and chrF++[3] scores that are popular for assessing the quality of the generated summaries. The higher the score of these metrics the better in the summary. This analysis offered insights into how the combined dataset enhance performance as compared to the individual human-generated or the pivoted dataset.

4 Results

Table 1 shows the performance comparison of the transformer models Google mT5-small, BanglaT5, and mBART on Bangla text summarization in terms of BLEU, METEOR, and chrF++ scores. It cab be seen from this table that human-generated dataset consistently outperforms pivoted synthetic dataset across all metrics. In addition, the BanglaT5 method outperforms the Google mT5-small and mBART across all metrics, since this method is pre-trained for Bangla text. In this comparison, the Google mT5-small ranks the second slightly ahead of the mBART method. In the case, METEOR and chrF++ scores, it is seen that for the method mBART, the combined

Table 1 Summary of Model Performances Using the Three Datasets

Dataset	Model	chrF++	METEOR	BLEU
Human-generated dataset	BanglaT5	38.496	0.235	5.93
	Google mT5-small	33.059	0.173	4.61
	mBART	31.167	0.167	4.32
Pivoted-synthetic dataset	BanglaT5	34.86	0.202	3.01
	Google mT5-small	32.05	0.176	2.76
	mBART	29.94	0.171	2.43
Combined dataset	BanglaT5	36.90	0.235	4.85
	Google mT5-small	32.45	0.179	3.77
	mBART	32.27	0.182	3.28

Distribution Bar 1: Percentage of Instances of A Category of Summary Deemed Better

Distribution Bar 2: Distribution of the Preferred Summary for Each Text News Article

Fig. 1 Subjective evaluation of summarization task

dataset improves the performance from that of the human-generated dataset. In the case of Google mT5-small method, similar observation can be made for METEOR score. When comparing all three methods in terms of the scores considered the combined dataset improves the performance from the pivoted one, but it lags slightly behind human-generated one. Thus, the pivoted dataset has the capability of providing a satisfactory summarization tasks, even some times the performance of in combination with human-generated dataset can surpass that of the human-generated ones.

An online survey [17] was conducted to compare the perceived quality of summaries generated using human-generated and pivoted datasets. A total of 36 native Bangladeshi respondents evaluated 30 news articles, choosing the better summary between those produced by the experimental models. Fig. 1 shows the results of this survey comparing the real-life efficacy of pivot-based summaries relative to human-generated ones. It turned out that for 12 out of the 30 given news articles, the summary resulting from the pivoted dataset is deemed better. In the survey there were $36 \times 30 = 1080$ votes, in which 46% of the votes were favored for the summarization tasks of the pivoted-dataset.

Tables 2 and 3 shows the comparing performance of experimental models on the human-generated and pivoted datasets for different topics of news articles in terms

Table 2 Topic-Based METEOR Scores

Topic	Human Data	Pivoted Data	Deficit
Sports	0.221	0.198	10.4%
Politics	0.244	0.204	16.4%
Science and Tech	0.219	0.187	14.6%
International	0.240	0.207	13.75%
Others	0.265	0.221	16.6%

Table 3 Topic-Based chrF++ Scores

Topic	Human Data	Pivoted Data	Deficit
Sports	36.77	34.36	6.55%
Politics	39.6	35.10	11.36%
Science and Tech	36.49	34.01	6.8%
International	39.17	35.21	10.1%
Others	41.79	35.88	14.1%

text to be summarized: পাঁচ বছরের কারাদন্ড দেয়া হয়েছে বিএনপি চেয়ারপার্সন খালেদা জিয়াকে জাতীয় নির্বাচনের আগে দলের চেয়ারপার্সনকে কারাগারে পাঠানোর কী প্রভাব পড়বে বিএনপির নির্বাচনী কৌশলে? চলতি বছরের শেষ দিকে এ নির্বাচন হওয়ার কথা রয়েছে। রাজনৈতিক বিশ্লেষক ও নিউজ টুডে'র সম্পাদক রিয়াজউদ্দিন আহমেদ মনে করেন এই রায়ে খুব বেশি অপ্রস্তুত অবস্থায় পড়বে না বিএনপি। কারণ তার মতে দলটি যথেষ্ট সময় পেয়েছে এই বিষয়ে পূর্বপ্রস্তুতি নেয়ার। ... মি. আহমেদের ধারণা খালেদা জিয়ার নির্বাচন করার পক্ষেই পরবর্তী পদক্ষেপ নেবে উচ্চ আদালত। তার মতে, রাজনৈতিক বিবেচনায় সরকার যদি মনে করে খালেদা জিয়াকে নির্বাচনে রাখবে না তাহলে সেটি সরকারের জন্য খুব একটা লাভজনক হবে না। তবে খালেদা ... ওদিকে দলের যুগ্ম মহাসচিব মাহবুব উদ্দিন খোকন বলেছেন তারা আদালতে রায়ের কপির জন্য আবেদন করেছেন এবং সেটি পেলে রবিবার বা সোমবারে এ রায়ের বিরুদ্ধ আপীল করবেন।

target summary: জিয়া অরফানেজ ট্রাস্ট দুর্নীতি মামলায় বিএনপি চেয়ারপার্সন খালেদা জিয়াকে ৫ বছরের কারাদণ্ড দিয়েছে বিশেষ আদালত।

summary after training with human-generated dataset: বাংলাদেশে বিরোধীদল বিএনপির চেয়ারপার্সন খালেদা জিয়াকে পাঁচ বছরের কারাদন্ড দিয়েছে হাইকোর্ট।

summary after training with pivoted synthetic dataset: বাংলাদেশের প্রধানমন্ত্রী খালেদা জিয়ার পাঁচ বছরের কারাদণ্ডের রায় বিএনপির জন্য একটি বড় সঙ্কট।

Fig. 2 Example of text and summaries from the two datasets

of the METEOR and chrF++ scores, respectively. It is seen from these tables that scores of the pivoted datasets is around 85% of that of human-generated datasets. It is also seen that there is less deviation of performance in the 'sports' category, and the highest deviation is in the 'politics' and 'others' categories. This could be due the fact that these two topics have limited number of news articles in the BBC XLsum dataset. An example of summarization using human-generated and pivoted dataset is shown in Fig.2. The difference of these two summary is indistinguishable.

5 Discussions

The transformer models are trained four times using 2000, 4000, 6000, and 8102 entries from the datasets. This allows us to observe a progressive improvement in the evaluation scores, BLEU and METEOR, for both the human-generated and pivoted data as shown in Fig.3. As expected, the quality of the summarization task, whether in BLEU or METEOR, improves with the number of training data. This indicates that a further increase in the size of the dataset is likely to yield even better summarization quality for any score. The bar charts with normalized scores shown in Fig.4illustrate some key findings:

- The chrF++ metric reveals the closest performance of the pivoted dataset compared to the human-generated dataset.
- The METEOR score exhibits better comparability, with the pivoted dataset performing closely with the human-generated dataset.
- The BLEU score indicates reduced performance for the pivoted dataset.

The observed differences in performance can be attributed to the different approaches of the metrics in assessing the similarity between reference summaries and generated summaries. The BLEU evaluates similarity based on 'word n-grams', whereas chrF++ assesses it through 'character n-grams'. Due to this distinction, difference of performance in terms of these two scores is acceptable for a morphologically very rich language like Bangla. In other words, there are numerous variations of the same word depending on its grammatical role (e.g., possessive, subject, or object) in Bangla. Consequently, 'character n-gram' analysis provides a more accurate reflection of similarity for Bangla.

6 Conclusion

This study has investigated three fine-tuned transformer models, namely, the BanglaT5, Google mT5-small, and mBART for Bangla text summarization using human-generated and pivoted datasets. The models were evaluated using three metrics, viz., chrF++, METEOR and BLEU offering insights into their performance. The Bangla database used for the experimentation is obtained from BBC XLsum. The training, validation, and testing sets are same as the partitions recommended in this database. In the experiments, it has been found that the BanglaT5 consistently outperforms other models, leveraging its Bangla-specific pre-training. Using the human-generated dataset, it has achieved the highest performance in terms of BLEU, METEOR, and chrF++ scores. The Google mT5-small has marginally performed better than mBART, but falls short of BanglaT5. Training on human-generated data has consistently outperformed pivoted data, but training on the latter provides a significantly close performance to the former. For example, METEOR score for the human-generated dataset has been found to be 0.244 for 'politics' and 0.265 for 'other' topics compared to 0.204 and 0.221 for the pivoted dataset, with performance ranging from 83% to 93%. Similarly, the chrF++ score for 'politics' has been found to be 39.6 for human-generated data and 35.10 for pivoted data, i.e., the latter performing 88% of the former. Training by using the combined datasets for summarization has

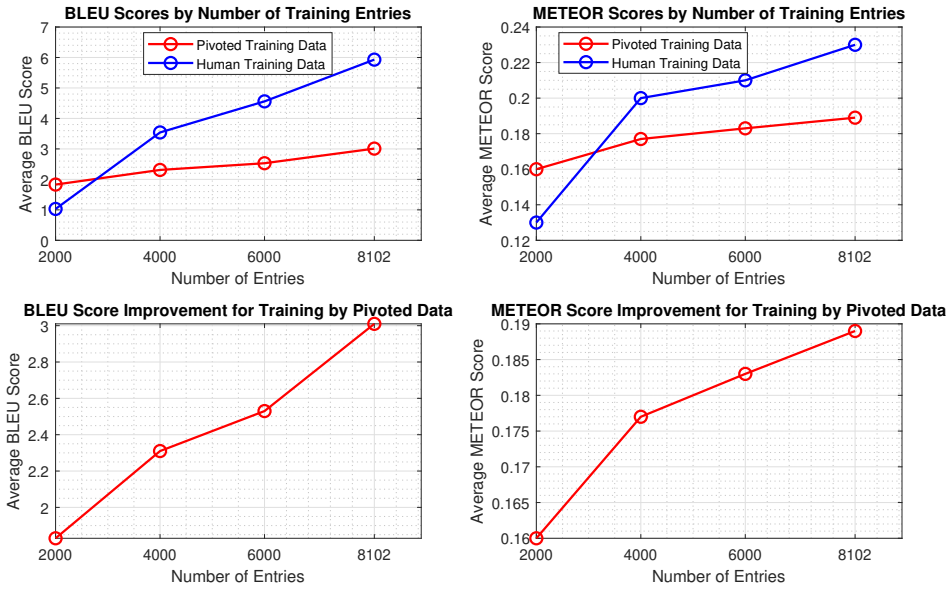

Fig. 3 BLEU and METEOR scores with increasing training data

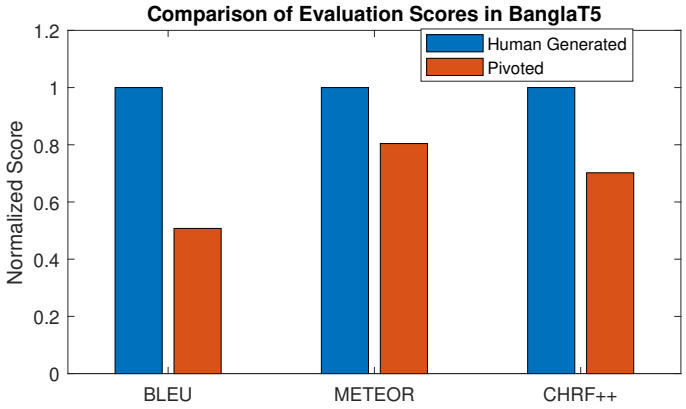

Fig. 4 Normalized scores in terms of BLEU, METEOR, and chrF++ summarization

revealed mixed performance. For example, METEOR scores for Google mT5-small and mBART have been enhanced due to the combination, for BanglaT5 the scores achieved at best the same as the human-generated ones. In other cases, the performance of the combined dataset lags behind that obtained from the human-generated data. In conclusion, while pivoted datasets are a viable alternative in resource-constrained scenarios, they can achieve summarization performance approximately 85% of that of the

human-generated datasets for fine-tuned models. These findings highlight the importance of data quality in Bangla NLP. The fine-tuned models demonstrated strong capabilities in Bangla text summarization, offering a foundation for future research to explore advanced data augmentation and model designs to enhance performance in low-resource settings.

Acknowledgment

The authors would like to thank the American Institute of Business Intelligence (AIBI), McLean, VA for their support in conducting this research work.

References

[1] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proc. 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, p. 311 (2001)

[2] Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, Ann Arbor, Michigan, pp. 65–72 (2005)

[3] Popović, M.: chrF++: Words Helping Character n-grams. In: Proc. 2nd ACL Conf. Machine Translation, Stroudsburg, PA, USA, pp. 612–618 (2017)

[4] Github repository for "Leveraging Pivoting Techniques for Summarization in Low-Resource Languages: Insights from Bangla". https://github.com/saifullahBUET144/Bengali_pivot_summary

[5] Haque, M.M., Pervin, S., Hossain, A., Begum, Z.: Approaches and Trends of Automatic Bangla Text Summarization. Int. Journal of Technology Diffusion **11**, 67–83 (2020)

[6] Chowdhury, R.R., Nayeem, M.T., Mim, T.T., Chowdhury, M.S.R., Jannat, T.: Unsupervised Abstractive Summarization of Bengali Text Documents. In: Proc. 16th Conf. European Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 2612–2619 (2021)

[7] Bhattacharjee, P., Mallick, A., Islam, M.S., Marium-E-Jannat: Bengali Abstractive News Summarization: A Neural Attention Approach. In: Proc. Int. Conf. Trends in Computational and Cognitive Engineering, Dhaka, Bangladesh, pp. 41–51 (2021)

[8] See, A., Liu, P.J., Manning, C.D.: Get To The Point: Summarization with Pointer-Generator Networks. In: Proc. 55th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1073–1083 (2017)

[9] Sarkar, K.: Bengali Text Summarization by Sentence Extraction. ArXiv (2012)

[10] Chowdhury, S.R., Sarkar, K., Maji, A.: Unsupervised Bengali Text Summarization Using Sentence Embedding and Spectral Clustering. In: Proc. 19th ACL Int. Conf. on Natural Language Processing, New Delhi, India, pp. 337–346 (2022)

[11] Akil, A., Sultana, N., Bhattacharjee, A., Shahriyar, R.: BanglaParaphrase: A High-Quality Bangla Paraphrase Dataset. In: Proc. 2nd Conf. Asia-Pacific Chapter of the Association for Computational Linguistics and 12th Int. Joint Conf. Natural Language Processing, Stroudsburg, PA, USA, pp. 261–272 (2022)

[12] Mallinson, J., Sennrich, R., Lapata, M.: Paraphrasing Revisited with Neural Machine Translation. In: Proc. 15th Conf. European Chapter of the Association for Computational Linguistics, Valencia, Spain, pp. 881–893 (2017)

[13] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: Proc. Conf. North American Chap. Assoc. Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA, pp. 483–498 (2021)

[14] Bhattacharjee, A., Hasan, T., Ahmad, W.U., Shahriyar, R.: BanglaNLG and BanglaT5: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla. In: Find. Assoc. Computational Linguistics, Stroudsburg, PA, USA, pp. 726–735 (2023)

[15] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual Denoising Pre-training for Neural Machine Translation. Trans. Assoc. Computational Linguistics **8**, 726–742 (2020)

[16] Hasan, T., Bhattacharjee, A., Islam, M.S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M.S., Shahriyar, R.: XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In: Find. Assoc. Computational Linguistics, Stroudsburg, PA, USA, pp. 4693–4703 (2021)

[17] Google Form: "Machine Generated Bengali Summary Evaluation". https://docs.google.com/forms/d/e/1FAIpQLScbV_OAfmdDRVgnZtf_Vi7soM-fVsRbiJXq_7e2G44285FOZQ/viewform