Shironaam: Bengali News Headline Generation using Auxiliary Information

Abu Ubaida Akash∗§ Mir Tafseer Nayeem∗◊ Faisal Tareque Shohan§ Tanvir Islam‡

∗Ahsanullah University of Science and Technology ○University of Alberta
§University of Hawaii at Manoa

akash.ubaida@gmail.com, mnayeem@ualberta.ca
faisaltareque@hotmail.com, tislam@hawaii.edu

Abstract

Automatic headline generation systems have the potential to assist editors in finding interesting headlines to attract visitors or readers. However, the performance of headline generation systems remains challenging due to the unavailability of sufficient parallel data for low-resource languages like Bengali and the lack of ideal approaches to develop a system for headline generation using pre-trained language models, especially for long news articles. To address these challenges, we present Shironaam, a large-scale dataset in Bengali containing over 240K news article-headline pairings with auxiliary data such as image captions, topic words, and category information. Unlike other headline generation models, this paper uses this auxiliary information to better model this task. Furthermore, we utilize the contextualized language models to design encoder-decoder model for Bengali news headline generation and follow a simple yet cost-effective coarse-to-fine approach using topic-words to retrieve important sentences considering the fixed length requirement of the pre-trained language models. Finally, we conduct extensive experiments on our dataset containing news articles of 13 different categories to demonstrate the effectiveness of incorporating auxiliary information and evaluate our system on a wide range of metrics. The experimental results demonstrate that our methods bring significant improvements (i.e., 3 to 10 percentage points across all evaluation metrics) over the baselines1. Also to illustrate the utility and robustness, we report experimental results in few-shot and non-few-shot settings.

1 Introduction

News headlines can significantly affect the number of visitors and play a crucial part in the life-cycle of a news article (Murao et al., 2019). Therefore, representative and interesting headlines are arguably essential to any news document to grab the attention of potential readers (Mishra et al., 2021; Ao et al., 2021). Nowadays, online and printed news releases significantly increase the article’s visibility, support, and context by using multimedia content. As a picture is worth a thousand words, digital assets such as images and videos are the go-to candidates for the thumbnails used in different social media, blogs, and many other platforms. The captions that go with the images or videos are equally significant as the actual content. Captions describing the images can clarify and enhance the image, optimize news articles for search engines, and improve the accessibility of the news for people with vision impairments2.

Headline generation, given a news article, is a special case of abstractive summarization (Yamada et al., 2021), which involves sentence compression, syntactic reorganization, sentence fusion, and lexical paraphrasing (See et al., 2017; Gehrmann et al., 2018; Zhong et al., 2019; Nayeem et al., 2019; Nayeem and Chali, 2017b). Unlike text summaries, which often feature many or single long sentences to summarize a document's important concepts (Nayeem and Chali, 2017a), news headlines frequently have a single short catchy statement to grab the readers attention and entice them to read the story. Even though Bengali is the seventh most spoken language with approximately 337 million speakers worldwide3 (Chakraborty et al., 2021; Chowdhury et al., 2021), generating quality headlines for a low-resource language such as Bengali is more challenging due to the unavailability of large-scale human-annotated dataset (Haque et al., 2016; Nayeem et al., 2018; Joshi et al., 2019).

2 Equal contribution.

3 In this paper, we limit our focus to only captions to improve the news headlines. Using multimodal information for this task is left as possible future work.

1 Code, dataset, and model checkpoints: https://github.com/dialect-ai/BenHeadGen

2 https://w.wiki/57
Contextualized language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020) help improving several downstream tasks in NLP, such as summarization, question answering, and text classification. Unfortunately, these models suffer from a limitation as they can handle input sequences up to a certain limit (Sun et al., 2019). As a result, this limitation burdens some NLP tasks, especially where the input is necessarily long (Kitaev et al., 2020), such as transcript analysis of the phone calls, document topic prediction, news headline generation, etc. The most natural way to address this problem is to trim the input sequences to a maximum length. However, trimming the long input document is tricky, especially for headline generation. The news articles usually maintain coherence and relevant parts may be located at the bottom of the document, which may prevent models from generalizing well to positions beyond the cutoff point. In this paper, we utilize topic words to retrieve important sentences as a context for the BERT model by following a simple yet cost-effective coarse-to-fine approach.

We present Shironaam, a large-scale abstractive Bengali news article dataset that includes over 240K professionally annotated headline-article pairings as well as auxiliary information such as image captions, topic words, and category information. Each sample can be represented as a tuple of (article, image caption, topic-words, category, headline). To the best of our knowledge, Shironaam is the first Bengali news article dataset incorporating auxiliary information and a benchmark for the news headline generation task. This corpus has the potential to authorize and encourage research on such a low-resource language, bringing technological advancements to a previously underserved community. Rather than the one-to-one mapping (i.e., input is an article, and output is a headline) used in the earlier works (Takase et al., 2016; Zhang et al., 2018; Murao et al., 2019; Colmenares et al., 2019; Song et al., 2020; Li et al., 2021), we treat the headline generation task as a three-to-one mapping with the inputs being an image caption, a list of topic words, and an article where the output is a headline. Based on the transformer architecture, we utilize pre-trained language models for generating headlines and present a new concept of fusing image caption parallelly (Liu et al., 2020a) with the input article to support the three-to-one mapping and to encode long documents. We design and compare numerous input mechanism alternatives as part of the suggested strategy. Extensive experiments on our proposed dataset reveal that the suggested method is capable of generating high-quality news headlines (see Section F in the Appendix) and brings significant improvements over the state-of-the-art baselines across all evaluation metrics (Table 5).

Our main contributions can be summarized as follows:

- We provide Shironaam, a large-scale news headline generation dataset of a low-resource language i.e., Bengali containing over 240K news headline-article pairings with auxiliary information such as image captions, topic words, and category information (Table 2). Also, this dataset can potentially be used for other tasks such as document categorization, news clustering, keyword identification, etc.

- We present a new concept of incorporating auxiliary information to model input with articles to improve the quality of headlines. We train an encoder-decoder model for this task almost from scratch, which utilizes pre-trained language model (Figure 1).

- We develop BenSim, an independent module for measuring the semantic similarity among Bengali sentences. We make use of the BenSim module and utilize topic words to encode long articles by following a simple yet effective approach (Figure 1(c)).

- To illustrate the utility and robustness, we also evaluate the performance with few-shot settings where the domains don’t have enough training samples (Table 6).

2 The Shironaam Corpus

In this section, we present the first-of-its-kind corpus (we name it Shironaam) for news headline generation in Bengali like low-resource language. This includes auxiliary information in addition to the usual headline-article pairs. We explain the curation process involving raw data crawling, pre-processing, and cleaning.
we remove the date/time and the embedded items using regular expressions. To preserve only the Bengali texts, we construct a vocabulary of Bengali unit characters and perform character level matching in the article bodies and headlines. The English numbers, however, are retained since they are used frequently in regular Bengali texts.

The image captions sometimes include extra/irrelevant information (e.g., ছবি [Picture], সংগ্রহীত [Collected], ফাইল ছবি [File Image], রয়েটার্স [Reuters], ইন্টারনেট [Internet], পৰ্তীকী ছিব [Symbolic Image], etc.)\(^6\), which are common in any news article. Thus, we identify these repetitive words using a simple frequency-based approach over all the samples and remove them from the image captions. Furthermore, we discard the samples whose captions are smaller than four words in length; from our manual inspection, we observed that these words often describe the named entities present in the image, such as name, location, date/time, etc.

Different newspapers use different names to categorize their contents. Consequently, each domain is represented with different category names in all the news dailies. For extracting the categories, we map them with their corresponding representative domains and label each domain with its relevant names. For instance, national, whole-country, city-news, country, capital, city-roundup, south-city, etc. are distinct categorical terms, but they can be grouped easily under the national domain. Table 1 shows the distribution of the final domains in the Shironaam corpus. We use sbnltk\(^7\) for tokenizing the documents into sentences. Finally, we discard the samples where any of the information (i.e., headline, article, or image caption) is missing.

### 2.1 Raw Data Crawling

We crawl around 900,000 raw data samples from seven famous Bengali newspapers (names in Section C in the Appendix) concentrating on certain criteria, such as headline, article, image caption, topic words, and various categories. Since each of the newspapers mentioned above has its own professional authors and distinct writing style, we consider multiple sources to prevent the bias of a particular annotation style. To ensure content diversity, we also cover various domains from all the news dailies. The majority of the news samples are extracted from HTML bodies of the corresponding publications, while some are rendered using JavaScript. However, two of them (see in Appendix Section C) do not provide the archives on their websites; therefore, we collect the samples through their APIs.

### 2.2 Dataset Preprocessing

The overall crawled corpus contains a lot of noise, such as irrelevant details about the publisher and the date/time of the news in multiple formats, embedded advertisements, phrases from different languages (especially English), reference URLs, inconsistent bold sections, emoticons, extrinsic symbols, and various Unicode representations. Thus,

<table>
<thead>
<tr>
<th>Category</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
<th>Total</th>
<th>Jaccard (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entertainment</td>
<td>16,104</td>
<td>365</td>
<td>1095</td>
<td>17,565</td>
<td>13.56</td>
</tr>
<tr>
<td>National</td>
<td>117,566</td>
<td>2,664</td>
<td>7,994</td>
<td>128,226</td>
<td>24.60</td>
</tr>
<tr>
<td>Nature</td>
<td>467</td>
<td>10</td>
<td>31</td>
<td>510</td>
<td>23.66</td>
</tr>
<tr>
<td>International</td>
<td>30,558</td>
<td>692</td>
<td>2,078</td>
<td>33,329</td>
<td>17.83</td>
</tr>
<tr>
<td>Sports</td>
<td>17,635</td>
<td>399</td>
<td>1,199</td>
<td>19,235</td>
<td>17.82</td>
</tr>
<tr>
<td>Economy</td>
<td>6,447</td>
<td>146</td>
<td>438</td>
<td>7,032</td>
<td>39.37</td>
</tr>
<tr>
<td>Life-Health</td>
<td>6,356</td>
<td>144</td>
<td>432</td>
<td>6,933</td>
<td>17.83</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>1,599</td>
<td>36</td>
<td>108</td>
<td>1,744</td>
<td>11.71</td>
</tr>
<tr>
<td>Opinion</td>
<td>3,501</td>
<td>79</td>
<td>238</td>
<td>3,819</td>
<td>38.41</td>
</tr>
<tr>
<td>Politics</td>
<td>15,018</td>
<td>340</td>
<td>1,021</td>
<td>16,380</td>
<td>23.02</td>
</tr>
<tr>
<td>Edu-Career</td>
<td>4,008</td>
<td>90</td>
<td>272</td>
<td>4,372</td>
<td>53.58</td>
</tr>
<tr>
<td>Science-Tech</td>
<td>1,046</td>
<td>23</td>
<td>71</td>
<td>1,141</td>
<td>22.95</td>
</tr>
<tr>
<td>Religion</td>
<td>269</td>
<td>6</td>
<td>18</td>
<td>294</td>
<td>71.59</td>
</tr>
</tbody>
</table>

Table 1: Our headline generation dataset (Shironaam) distribution over 13 different domains. Jaccard scores\(^6\) represent the similarities of each domain in between the image captions and headlines.

<table>
<thead>
<tr>
<th>Features</th>
<th>IndicNLG-BN</th>
<th>Shironaam (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Headline</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Category</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Topic words</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Image Caption</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>#Examples</td>
<td>142,731</td>
<td>240,580</td>
</tr>
</tbody>
</table>

Table 2: Feature-level comparison between IndicNLG-BN (2022) and Shironaam dataset (ours).

### 2.3 Dataset Statistics

After preprocessing the raw corpus, we have 240,580 news samples as a tuple of (headline, article, image caption, topic words, category). To en-

\(^6\)The square brackets contain the English translations.

\(^7\)https://pypi.org/project/sbnltk
Table 4: Quantitative statistics compared to IndicNLG-BN (2022) and our proposed dataset Shironaam.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Article</th>
<th>Headline</th>
<th>Image Caption</th>
<th>Topic Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average number of words</td>
<td>252.01</td>
<td>6.53</td>
<td>6.80</td>
</tr>
<tr>
<td>Shironaam</td>
<td>199.83</td>
<td>10.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IndicNLG</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
- Shironaam: Average number of sentences
- IndicNLG: Average number of sentences
- Shironaam: Vocabulary size
- IndicNLG: Vocabulary size

Therefore, the Shironaam corpus comprises a diverse range of headline styles and provides the largest collection of Bengali news articles. Moreover, it is the first benchmarking dataset in such a low-resource language that includes auxiliary information in addition to the headline-article pairs. We hope it will motivate further study and serve as a baseline for future works on this task for this low-resource language.

3 News Headline Generation

3.1 Task

We establish a new concept of incorporating auxiliary information in order to generate high-quality headlines in Bengali, a low-resource language. In the context of this generation task, we assume that a) we have enough data with auxiliary information to train a headline generation model in Bengali language (can be referred to Shironaam corpus); b) the auxiliary information refers to the image captions and topic words used in tagging documents; c) we have access to a module that filters a document based on the contextual similarity with a list of topic words (we refer BensSim in Section 3.3). The task can be formalized as follows. Given article \( A \), image caption \( C \), and a set of topic words \( T \) as input, our goal is to generate high-quality headline \( \mathcal{H} \) for the corresponding news article.

3.2 Approach

To carry out the idea, we need several benchmarks to compare with and evaluate our proposed hypothesis. But, no SOTA benchmark is available for this task in Bengali language, except the IndicBART (Dabre et al., 2022). So, we set multiple baselines (Section 4.2), both of extractive and abstractive types, that take article \( A \) as input and generate corresponding headline \( \mathcal{H} \) as output. We follow LEAD-1 and EXT-ORACLE approaches among the extractive types, whereas from the abstractive types, we initialize an encoder-decoder model for Bengali language with a pre-trained encoder-only checkpoint to skip the costly pre-training (Rothe et al., 2020). To train the encoder-decoder model (BED), we use BanglaBERT (Bhattacharjee et al., 2022a) as the encoder checkpoint. Additionally, we utilize other pre-trained models (i.e., BanglaT5, Bedrock, etc.) for necessary background.
(Bhattacharjee et al., 2022b), IndicBART (Dabre et al., 2022) based on transformer architecture (Vaswani et al., 2017). After comparing all the baselines (see Section 4.2), we select the best performing one for further ablations. Experimental results (in Table 5) reveal that BED model outperforms other baselines, even though the fine-tuned BanglaT5 (Bhattacharjee et al., 2022b) scores competitively.

3.3 BERT-based Encoder-Decoder (BED)
A BED model consists of an encoder that has been initialized with BERT, termed as BERTenc, coupled with a decoder that has also been initialized with BERT, which we call BERTdec. The initializing point for each weight’s calculation is a public BERT checkpoint. The only variable initialized at random is the encoder-decoder attention (Rothe et al., 2020).

**Article Only** In Figure 1(a), we implement the basic version i.e., BED (base) model, which takes word tokens of an article as a sequence of inputs $A_{1:n}$ and describes a conditional distribution of target vectors $H_{1:l}$ of variable length $l$, in our case, generated words for headline:

$$p_{\theta_{BERTenc},\theta_{BERTdec}}(H_{1:l}|A_{1:n}).$$

The input sequence $A_{1:n}$ is sent to the BERTenc component, which then converts it into a sequence of hidden states, $\overline{A}_{1:n}$. The mapping can be defined as:

$$f_{\theta_{BERTenc}}: A_{1:n} \rightarrow \overline{A}_{1:n}. \quad (2)$$

The BERTdec component will simulate the conditional probability distribution of the target vector sequence $H_{1:l}$, assuming that the sequence of encoded hidden states $\overline{A}_{1:n}$ has been provided:

$$p_{\theta_{BERTdec}}(H_{1:l}|\overline{A}_{1:n}). \quad (3)$$

Bayes’ rule lets us turn this distribution into a product of the conditional probability distribution of the target vector $h_i$, given the encoded hidden states $\overline{A}_{1:n}$ and all the previous target vectors $H_{0:i−1}$:

$$p_{\theta_{BERTdec}}(H_{1:l}|\overline{A}_{1:n}) = \prod_{i=1}^{n} p_{\theta_{BERTdec}}(h_{i}|H_{0:i−1}, \overline{A}_{1:n}). \quad (4)$$

All preceding target vectors $H_{0:i−1}$ and the encoded hidden state sequence $\overline{A}_{1:n}$ are mapped to the logit vector $V_i$ by the BERTdec. The next step is to run the softmax operation on the logit vector $V_i$. This helps to define the conditional distribution $p_{\theta_{BERTdec}}(h_{i}|H_{0:i−1}, \overline{A}_{1:n})$ by making sure that the distribution of the target vector $h_i$ depends on the distributions of all previous target vectors $h_0, \ldots, h_{i−1}$:

$$p_{\theta_{BERTdec}}(h_{i}|H_{0:i−1}, \overline{A}_{1:n}) = \text{Softmax}(V_i). \quad (5)$$

The first target vector $h_0$ is going to be represented by a unique BOS vector that is referred to as the “beginning-of-sentence”. After the conditional distribution $p_{\theta_{BERTdec}}(h_{i}|H_{0:i−1}, \overline{A}_{1:n})$ has been set, the output can be made in an auto-regressive way. This makes it possible to define a mapping between an input sequence $A_{1:n}$ and an output sequence $H_{1:l}$ at the time of inference.

**Fusing Article and Image Caption** In order to explore more ways to improve the quality
of the generated headlines, we employ BED (u/Article + Caption) model, which incorporates image caption $C_{1:m}$ with the corresponding article $A_{1:n}$ as in Figure 1(b), where $m << n$, and passes them through the BERTenc using parallel-fusion (Liu et al., 2020a) mechanism:

$$K_{1:r} = C_{1:m} \oplus A_{1:n}, \quad (6)$$

$$f_{\theta_{BERTenc}} : C_{1:m}, A_{1:n} \rightarrow \overline{K}_{1:r}, \quad (7)$$

Here, $K_{1:r}$ denotes the model input sequence, where $r$ represents the new input sequence length, and $\oplus$ is concatenation operator separated by a special token. The sequence of hidden states $\overline{K}_{1:r}$ are then processed through the BERTdec likewise the Shironaam(base) model and the headline is generated as output:

$$p_{\theta_{BERTenc}, \theta_{BERTdec}}(H_{1:l}|C_{1:m}, A_{1:n}). \quad (8)$$

However, the image caption may not always serve the full context if the news article becomes too long for BERTenc. Moreover, the image caption length is generally much smaller than the news article length. Thus, the impact of using image caption as a context is less sensitive for lengthier articles.

**Bengali Sentence Similarity (BenSim)** Since many of the news articles’ lengths exceed the input sequence limit that BERTenc can process, we therefore, utilize the sequence length by ensuring all the relevant sentences are present in the limited input sequences. To ensure the extraction of relevant sentences, we develop BenSim module\footnote{https://github.com/dialect-ai/BenSim}, a tool for measuring semantic similarity between Bengali sentences utilizing BERT embeddings. It takes news article $A_{1:n}$ and corresponding topic words $T_{1:k}$ as input for getting most of the contextual sentences and employs pre-trained bangla-bert-base (Sarker, 2020) model on both of the input sequences to generate the contextualized encoded representations. After performing mean pooling operation, cosine similarity (Singhal, 2001) is then applied to the encoded sequences to get the similarity score. After measuring the similarities between the topic words and input sentences, a filtered article $A'_{1:r}$ is returned as output, which is then fused parallelly with the image caption $C_{1:m}$ and sent to the model input. Finally, the BED (u/FilteredArticle + Caption) model produces a headline after processing the fused input.

### 4 Experiments and Benchmarks

In this section, we set a new benchmark for Bengali news headline generation using Shironaam corpus and compare it with the other state-of-the-art baselines. After a clean comparison, we perform two ablation experiments on the superior base model. Then, we analyze the performance gap between the baselines and the ablations and further evaluate the best model on news domains with a few samples (few-shot). Finally, after proper analysis, we seek to find out the answers to the following research questions:

- **RQ#1**: Can we use auxiliary information (e.g., image caption and topic words) to improve the performance of the headline generation?
- **RQ#2**: Which domain(s) benefit from the auxiliary information in few-shot and non-few-shot settings?

#### 4.1 Implementation Details

We utilize the encoder-decoder paradigm\footnote{Encoder-Decoder models documentation} of HuggingFace, where pre-trained BanglaBERT (Bhattachjee et al., 2022a)\footnote{BanglaBERT usage (HuggingFace)} is used to initialize both of the weights of encoder and decoder. Before proceeding to tokenization, we perform sentence normalization, introduced in Hasan et al. (2020). For tokenization, we use the pre-trained tokenizer\footnote{BanglaBERT usage (HuggingFace)} that comes with the model. All the hyper-parameters used for training and decoding are presented in Section B in the Appendix.

**Evaluation Metrics** We compare the performance with the following baselines across several evaluation metrics presented in Section D in the Appendix.

#### 4.2 Baselines

**LEAD-1** LEAD-1 is a commonly used baseline for setting the lower bound of news headline generation task (Kumar et al., 2022; Narayanan et al., 2018). It also indicates the degree of positional biasness of article body sentences in generating headlines. We pick the article’s first sentence as the system headline and compare it with the original headline to generate the LEAD-1 scores.


<table>
<thead>
<tr>
<th>Models</th>
<th>ROUGE</th>
<th>BLEU</th>
<th>BERT</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1 R-2 R-L</td>
<td>Score Brevity</td>
<td>Score</td>
<td>Score</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Penalty Length</td>
<td></td>
<td>Length</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td>Ratio</td>
<td></td>
<td>Ratio</td>
</tr>
<tr>
<td>LEAD-1 (Extractive)</td>
<td>30.50 13.86 28.00</td>
<td>5.65 97.71 2.48</td>
<td>74.63 29.90</td>
<td></td>
</tr>
<tr>
<td>EXT-ORACLE (Extractive)</td>
<td>39.92 22.89 37.28</td>
<td>9.17 97.16 2.30</td>
<td>77.16 39.65</td>
<td></td>
</tr>
<tr>
<td>IndicBART (mBART)</td>
<td>28.76 12.65 27.11</td>
<td>15.03 99.91 1.14</td>
<td>74.95 20.39</td>
<td></td>
</tr>
<tr>
<td>BanglaT5 (mT5)</td>
<td>44.13 23.03 42.12</td>
<td>13.05 91.33 1.15</td>
<td>80.13 34.65</td>
<td></td>
</tr>
</tbody>
</table>

| Our Ablations                |                |                  |       |                |
|------------------------------|                |                  |       |                |
| BED Base (BERT2BERT)         | 44.22 24.18 42.28 | 22.06 94.47 0.94 | 80.33 34.16 |
| -w/ Article + Caption        | 51.62 33.62 49.94 | 31.39 96.02 0.96 | 82.93 42.57 |
| -w/ FilteredArticle + Caption| 52.19 34.27 50.31 | 31.80 98.57 0.99 | 83.10 43.52 |

Table 5: Performance on Shironaam (test) corpus compared to the baselines (Section 4.2) and the results of our ablation study (see Appendix Section E for validation scores) across various automatic evaluation metrics, where **bold-faced** scores indicate superior performance.

**EXT-ORACLE** On the other hand, EXT-ORACLE can be considered as the upper bound of generating headlines by an extractive approach (Kumar et al., 2022; Narayan et al., 2018). We implement this baseline on the Shironaam (test) corpus by aligning a sentence from the input article with the reference headline based on the ROUGE-2 metric.

**IndicBART** Kumar et al. (2022) releases a multilingual model, which is fine-tuned on IndicBART (Dabre et al., 2022) checkpoint for the headline generation task focusing on Indic languages including Bengali. IndicBART is a sequence-to-sequence multilingual pre-trained model (Dabre et al., 2022) based on the mBART (Liu et al., 2020b) architecture.

**BanglaT5** We fine-tune BanglaT5 (Bhattacharjee et al., 2022b), a sequence-to-sequence transformer model based on mT5 (Xue et al., 2021) architecture for Bengali language, on the Shironaam (train) corpus for the headline generation task. For a fair comparison, we maintain the same hyper-parameters.

**BED (base) Model** We implement the model (article only), illustrated in Figure 1(a) on the Shironaam (train) corpus to make the baseline. We utilize 220,500 news samples from the train set to train the BED model, which takes the article only as input and generates a headline as output. The evaluation result on the Shironaam (test) set is shown in Table 5, which is a new benchmark for the Bengali headline generation task. In the following experiments, we utilize the auxiliary information with the same hyper-parameter settings to generate better-quality headlines.

### 4.3 Ablation Experiments

**BED (w/ Article + Caption) Model** As per the demonstration in Figure 1(b), the image caption is incorporated with the input article. This leads to a much improved result across all evaluation metrics compared to article only model (a.k.a., BED (base)) as shown in Table 5.

**BED (w/ FilteredArticle + Caption) Model** Since the utilization of image caption in model input gives better results, therefore we further enrich the inputs by incorporating topic words. We use topic words in filtering the longer articles through BenSim rather than using them directly to the input, as shown in Figure 1(c). First, we set a threshold value (40 in our case) for BenSim to extract the number of top semantically similar sentences. BenSim maintains the relative appearance order of the sentences in the original article to construct the corresponding filtered article. To fix the number of sentences in a filtered article, we consider the maximum use of the number of tokens BED model can afford *i.e.* 512. Fusing filtered article with image caption achieves the best results across several evaluation metrics as shown in Table 5 (also see Appendix Section F for the generation quality).

### 4.4 Discussions

**Result Analysis** Table 5 shows that the LEAD-1 baseline performs inadequately on the Shironaam (test) corpus. More specifically, the ROUGE-2 and BLEU scores and the length ratio indicate that the original headlines are more abstractive in nature, and the first sentence of an article does not contain sufficient information for generating a headline. Unlike LEAD-1, comparatively higher ROUGE scores are obtained by using...
EXT-ORACLE, but at the same time, BLEU score gain is lower. This trade-off indicates that the reference headlines consist of the subset of words present in the sentences selected by EXT-ORACLE. However, because of the concise nature of news headlines, this approach does not fit well but can be considered a strong baseline for other models. Among the abstractive types, IndicBART performs poorly on the Shironaam (test) corpus and even is unable to beat the weak LEAD–1 baseline, let alone EXT-ORACLE. On the other hand, the fine-tuned BanglaT5 yields a good score for this task. Although the generated results are slightly lengthier than the reference ones, they can be considered a strong baseline. The BED (base) model provides the best performance in terms of ROUGE, BLEU, and BERT scores. So, we consider it the strongest baseline and look for further ablations.

To this end, Table 5 shows that the best baseline is outperformed by our proposed technique of modeling input using auxiliary data. We want to emphasize that we use image caption and topic words purely as auxiliary data. While collecting the data from various news portals, we observe that it is very common to include images to help support and communicate the story and image captions are a crucial part of it that only describe the referred image. Although image captions are mostly correlated with the corresponding article in terms of context, we argue that they are not headlines. First, there is not much overlap in terms of Jaccard similarity measured between image captions and headlines (as from Table 1 we have approximately 29% overlap across different categories). Second, headlines differ from image captions in terms of styles and content.

Image captions usually give the model some signal on which parts of the document model need to attend more. Hence, as a result of combining image caption with article, BED (w/ Article + Caption) model improves the performance by about 3 (BERT score) to 10 (ROUGE-2 score) percentage points and produces more human-like headlines. Moreover, it often begins generating sentient headlines that are more abstract and profound than the reference ones. The BED (w/ FilteredArticle + Caption) model performs slightly better than the previous ablation. Since, there are fewer lengthier articles in the Shironaam corpus, the variations in the scores of the two ablation models are rather small. We observe that when we include filtered articles led by relevant topic words in the model’s input, it begins to learn faster than the model without topic words (demonstrated in Figure 2). Therefore, the differences between the scores of two ablation models will increase with the number of lengthier articles.

So, following the question RQ#1, we may conclude from the preceding discussion that auxiliary information definitely aids in creating better headlines. Although we achieve superior performance compared to the state-of-the-art baselines across several evaluation metrics, these quantitative measures can not determine the generation quality. Therefore, we present generated samples from our model categorized into several abstractive types (see Section F Appendix). We leave the human evaluation of our generated samples as one of the future works.

**Domain Specific Analysis** We evaluate our proposed BED (w/ FilteredArticle + Caption) [denoted as BED (FA+C)] model on individual domains by comparing with a base model and to answer RQ#2. We also observe the performance of the presented model on the domains with fewer samples (few-shot). We employ two baselines here: BED (base) and BanglaT5 (Bhattacharjee et al., 2022b) [denoted as BNT5]. Although, the BNT5 has demonstrated competitive performance, Table 6 shows that BED (base) model performs better on the maximum number of domains. To calculate the exact performance gap, we maintain a uniform baseline i.e., BED (base) to compare with the proposed model. For the few-shot observation, based on the number of training samples, we split the domains into two folds (see Table 6). The Few-Shot domains contain less than 6500 train samples, whereas rest of the domains are considered as Non-Few-Shot.
Table 6 demonstrates that our proposed model improves the scores by a satisfactory margin of almost all the domains except Entertainment and Miscellaneous. These two categories get comparatively lower scores. The majority of headlines in the Entertainment domain are casual and clickbait-style and do not maintain the identical nature of a particular domain. We argue that the discrepancy, in this case, decreases the scores. The Miscellaneous domain is comprised of different sorts of randomness containing articles of various domains. Therefore, it is anticipated that this genre will get a lower score. Table 6 shows that our proposed model maintains consistent performance when there are few samples to train.

5 Related Works

Headline generation is an under-explored subtask of abstractive summarization, particularly in languages with limited resources. For the English language, an attention-based neural network has been proposed by Rush et al. (2015) for abstractive sentence summarization. The authors propose a model that utilizes a recurrent neural network (RNN) and an attention mechanism to summarize input sentences into a compact summary. Takase et al. (2016) build an AMR encoder for headline creation based on an encoder-decoder architecture. Using a dual-attention seq2seq model, Zhang et al. (2018) proposes a way for question headline development. In limited resource settings, Tilk and Alumäe (2017) pretrain a neural encoder and decoder model to enhance headline generation outputs. A sentence encoder, a gate network for sentence selection, and a headline decoder are the three stages of Zhou et al. (2017)’s headline generation approach. Tan et al. (2017) proposes a coarse-to-fine strategy that extracts the most important sentences before generating the headlines based on the context. For headline generation, Kumar et al. (2022) have released the IndicNLG, a collection of multilingual datasets. However, they do not provide any additional attributes besides the headline-article pairs. In summary, the majority of the past works for generating headlines primarily used the article content to generate headlines.

6 Conclusion and Future Work

In this paper, we contributed a large-scale dataset (a.k.a., Shironaam) with auxiliary information such as image captions, topic words, and category for Bengali news headline generation. We employ contextualized language models to incorporate such auxiliary information and proposed a simple yet effective solution to encode long articles using topic words. Experimental results demonstrate the superiority of our approach across different domains and settings. We anticipate that our efforts will motivate the community to expand the scope of headline generation tasks beyond English, particularly for a low-resource language like Bengali. Our future work will look into incorporating auxiliary information to support more languages and build a language-agnostic model.
Limitations

Our model relies on auxiliary information such as image captions and topic words to achieve superior performance. However, it is quite common to include images and extra information (e.g., topic words) to increase the article’s visibility, support, and context. Also, our base model without auxiliary information demonstrates improved performance compared to the well-established and state-of-the-art baselines. Another limitation we observed that our model did not perform as well as for the Miscellaneous and Entertainment categories compared to the other 11 different categories because of the clickbaity nature of these categories. Finally, our headline generation model works only for Bengali, a widely spoken but low-resource language. Still, this idea of using auxiliary information to improve headline generation performance can easily be extendable for many languages.

Ethics Statement

We considered some ethical aspects while scraping the data. We requested data at a reasonable rate without any intention of a DDoS attack. Moreover, for each website, we read the instructions listed in robots.txt to check whether we can crawl the intended content. We tried to minimize offensive texts in the data by explicitly crawling the sites where such contents are minimal. Further, we removed the Personal Identifying Information (PII) such as name, phone number, email address etc from the corpus.

Acknowledgements

We thank all the anonymous reviewers for their valuable feedback and constructive suggestions for improving this work. Mir Taseer Nayeem is supported by a Huawei Doctoral Scholarship.

References


A Preliminaries

Abstractive text summarization (Rush et al., 2015; See et al., 2017; Zhang et al., 2020) was considerably more challenging before the development of sequence-to-sequence (seq2seq) models (Cho et al., 2014; Sutskever et al., 2014) and recent advances in transformer-based models (Vaswani et al., 2017; Devlin et al., 2019) due to a lack of sufficient datasets. Many text-summarizing applications are still hindered by the lack of suitable datasets, particularly for low-resource languages (Joshi et al., 2019). After being presented in Vaswani et al. (2017), models based on transformer architectures have been proven to perform better on sequence-to-sequence tasks than decoder-only language models e.g. Raffel et al. (2020). In its most basic form, an encoder-decoder model comprises a stand-alone encoder, like BERT (Devlin et al., 2019), and a stand-alone decoder model, like GPT2 (Radford et al., 2019). It has been demonstrated that the huge pre-trained encoder-decoder models may considerably improve performance on a range of sequence-to-sequence tasks Lewis et al. (2020); Raffel et al. (2020). On the other hand, pre-training encoder-decoder models are very expensive to build since the models require a lot of computational resources.

Rothe et al. (2020) introduces the encoder-decoder model using pre-trained encoder and/or decoder-only checkpoints (such as BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019)) to avoid the time-consuming pre-training process. According to Rothe et al. (2020), these encoder-decoder models can do well as large pre-trained encoder-decoder models like T5 (Raffel et al., 2020) and Pegasus (Zhang et al., 2020) on different sequence-to-sequence tasks at a fraction of the training cost.

B Hyper-parameters, Training, and Decoding

All the BED models (Figure 1) are trained almost from scratch by maintaining uniform hyper-parameters and trained for 110,250 global steps with the learning rate 5e−5, and batch size 12. We save the best checkpoint by ensuring the lowest validation loss. We use AdamW (Loshchilov and Hutter, 2019) for optimizing the loss with default linear warmup. The maximum lengths of encoder and decoder are limited to 512 and 32 tokens, respectively. Each of the BED models is trained on a single NVIDIA Tesla P100 GPU and trained for approximately 33 hours, which takes almost 5 hours 30 minutes per epoch. The total number of trainable parameters is 249,044,480.

Decoding When validating and testing, we use beam search algorithm (Sutskever et al., 2014) with 4 beams to generate headlines. The maximum and minimum lengths used in generating the headlines are 16 and 4, respectively. We use ‘early stopping’ to stop the beam search when at least 4 sentences are finished per batch. The ‘no-repeat n-gram size’ is set to 2, where the exponential penalty to the length is 1.2. Regarding vocabulary
size, we use the fixed 32,000 vocabularies from the encoder.

C Data Sources

<table>
<thead>
<tr>
<th>Newspaper</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prothom Alo</td>
<td><a href="http://www.prothomalo.com">www.prothomalo.com</a></td>
</tr>
<tr>
<td>Naya Diganta</td>
<td><a href="http://www.dailynayadiganta.com">www.dailynayadiganta.com</a></td>
</tr>
<tr>
<td>Ajker Patrika</td>
<td><a href="http://www.ajkerpatrika.com">www.ajkerpatrika.com</a></td>
</tr>
<tr>
<td>Bangladesh Protidin</td>
<td><a href="http://www.bd-pratidin.com">www.bd-pratidin.com</a></td>
</tr>
<tr>
<td>Samakal</td>
<td><a href="http://www.samakal.com">www.samakal.com</a></td>
</tr>
<tr>
<td>Bhorer Kagoj</td>
<td><a href="http://www.bhorerkagoj.com">www.bhorerkagoj.com</a></td>
</tr>
<tr>
<td>Dhaka Tribune</td>
<td><a href="http://www.dhakatribune.com">www.dhakatribune.com</a></td>
</tr>
</tbody>
</table>

Table 7: List of Bengali newspapers to form the Shiron-aam corpus with their corresponding URLs. Samples from the italic-faced newspapers were crawled through their APIs.

D Evaluation Metrics

We evaluate the predicted headlines with some automatic metrics used for generation tasks. The generation quality is measured with the ROUGE (Lin, 2004) F1 score\(^{13}\). ROUGE-1 and ROUGE-2 measure informativeness, where fluency is measured by the longest common subsequence (ROUGE-L). We include BLEU (Papineni et al., 2002) score which indicates the similarity between reference and predicted sentences by comparing the overlap within tokens\(^{14}\). Brevity penalty and length ratio are shown to justify the BLEU score. The contextual similarity between the generated and reference headline is measured using F1 BERT score (Zhang* et al., 2020)\(^{15}\), where the correlation between them is reported by METEOR score (Banerjee and Lavie, 2005)\(^{16}\). We use the available open-source implementations for the above metrics.

\(^{13}\)ROUGE (multilingual)
\(^{14}\)BLEU (HuggingFace)
\(^{15}\)BERTScore (HuggingFace)
\(^{16}\)METEOR (HuggingFace)
### E Validation Results

<table>
<thead>
<tr>
<th>Train Loss</th>
<th>Valid Loss</th>
<th>ROUGE</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a) 1.0892</td>
<td>2.4332</td>
<td>44.51</td>
<td>23.56</td>
<td>42.38</td>
</tr>
<tr>
<td>b) 1.5083</td>
<td>2.1227</td>
<td>49.59</td>
<td>30.53</td>
<td>47.79</td>
</tr>
<tr>
<td>c) 1.2199</td>
<td>2.0836</td>
<td>49.77</td>
<td>31.42</td>
<td>48.05</td>
</tr>
</tbody>
</table>

Table 8: All the scores are reported for BED model with ablations on Shironaam (valid) corpus. The labels indicate the ablations of BED model: a) Base, b) Article + Caption, c) FilteredArticle + Caption. Only the Train Loss is measured on the training set and kept for comparison with the Valid Loss.

### F Generated Headlines

Generated headlines on Shironaam (test) corpus across all the categories are presented in Table 9.

<table>
<thead>
<tr>
<th>Category</th>
<th>Headline</th>
</tr>
</thead>
</table>
| Economy  |GH চিনা আমদানিতে শুল্ক কমােলা সরকার  
ET The government reduced the duty on sugar import  
RH চিনা আমদানিতে শুল্ক কমােলা  
ET Import duty on sugar reduced |
| Edu-Career |GH সহকারী জজের লিখিত পরীক্ষার সূচি পুনর্নির্দেশ  
ET Release of written test schedule for the appointment of Assistant Judge  
RH সহকারী জজের লিখিত পরীক্ষার সূচি পুনর্নির্দেশ  
ET Release of written test schedule for the appointment of Assistant Judge |
| Entertainment |GH "আমি বিয়ে করব না, দেখ কে আমাকে বিয়ে করে"  
ET "I will not marry, let’s see who marries me"  
RH "আমি বিয়ে করব না, কে আমাকে বিয়ে করে দেখি...",  
ET "I will not marry, who will marry me let’s see..." |
| International |GH জাপােন পযর্টকবাহী জাহাজ ডুেব িনেখাঁজ ২৬  
ET Tourist ship sinks in Japan and goes missing 26  
RH জাপােন পযর্টকবাহী জাহাজ ডুেব িনেখাঁজ ২৬  
ET Tourist ship sinks in Japan and goes missing 26 |
| Life-Health |GH নতুন মৃতুয্ ৩৭, শনাক্ত ৩০৪৫  
ET New deaths 37, detections 3045  
RH আকৰ্ান্ত ছাড়ােলা ৫৫ হাজার  
ET Number of infected has crossed 55 thousand |
| Miscellaneous |GH চার অক্ষের সন্তানের নাম!  
ET Four-letters in children’s name!  
RH সন্তানের নাম চার অক্ষের!  
ET Children’s names are in four-letters! |
| National |GH চোরা প্রতিকে না পেয়ে কেদে ফেলেন চেয়ারমান প্রাথমিক  
ET Chairman candidate cried after not getting the horse symbol  
RH চোরা প্রতিকে না পেয়ে কেদে ফেলেন সেই চেয়ারমান প্রাথমিক  
ET That chairman candidate cried after not getting the horse symbol |
<p>| Nature |GH লেখার নিবন্ধে বৃশ্চিক পূর্বায়ন |</p>
<table>
<thead>
<tr>
<th>Category</th>
<th>Bengali Headline</th>
<th>English Headline</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rain</td>
<td>রাতে বাড়বে তাপমাত্রা, ৩ বিভাগে বৃষ্টির আভাস</td>
<td>The temperature will increase at night, there is a chance of rain in 3 divisions</td>
<td></td>
</tr>
<tr>
<td>Opinion</td>
<td>ই-কমাস বাজেট চাই</td>
<td>Want e-commerce friendly budgeting</td>
<td>Deleted</td>
</tr>
<tr>
<td>Politics</td>
<td>বিদ্যুৎ-গ্যাসের মূলবিন্দুর সিদ্ধান্ত পণ্ডিত</td>
<td>The decision to increase the price of electricity and gas is an anti-people move: Public Forum</td>
<td>Inserted</td>
</tr>
<tr>
<td>Religion</td>
<td>পাবলিক শেঁব বরাত পালিত</td>
<td>Devoted Muslims engaged in prayer</td>
<td>Sentient</td>
</tr>
<tr>
<td>Science-Tech</td>
<td>মহাকাশ সায়াটেলাইটের সংখ্যা বাড়াচ্ছে ওয়ানওেব</td>
<td>OneWeb is increasing number of satellites into space</td>
<td>Inserted</td>
</tr>
<tr>
<td>Sports</td>
<td>মেসি এখন পিএসজির জায়গাতে</td>
<td>Messi is now in PSG jersey</td>
<td>Paraphrased</td>
</tr>
</tbody>
</table>

Table 9: High quality headlines generated on Shironaam (test) corpus across all the categories. Here, “Type” means the how the generated headlines are different from the references. We categorize the differences into 5 types: **Inserted** (only one/some word(s) is/are added to reference headline), **Matched** (generated exactly the same), **Swapped** (the only difference is made by swapping one/some word(s) within the reference headline), **Deleted** (the output is about similar to the reference with one/some word(s) less), **Sentient** (generated headline is completely different but a potential competitor against the reference one), **Paraphrased** (paraphrased version of the reference headline). The colored words (i.e. teal for Inserted, cyan for Swapped, brown for Sentient, magenta for Deleted, and violet for Paraphrased) indicate the exact positions where the generated ones are different from the references and no color refers to no change. The generated and reference Bengali headlines, and their corresponding English version are denoted by GH, RH, and ET respectively.