State-of-the-Art Translation of Text-to-Gloss using mBART: A Case Study of Bangla

Anonymous ACL submission

Abstract

Gloss is a written approximation that bridges Sign Language (SL) and its corresponding spoken language. Despite a deaf and hard-of-hearing population of 1.7 million, Bangla Sign Language (BdSL) remains largely understudied, with no prior work on Bangla text-to-gloss translation and no publicly accessible datasets. To address this gap, we construct a dataset of Bangla sentences and their gloss representations, adapting rule-based glossing methods from German and American Sign Languages to fit BdSL. We further augment the dataset using GPT-4o, alongside back-translation and text generation techniques. We fine-tune pretrained mBART-large-50 (hereafter, mBART) and mBERT-multiclass-uncased models, and train traditional baselines including RNN, GRU, and a novel seq-to-seq model with multi-head attention. Fine-tuning mBART achieves the best performance (sacreBLEU = 79.53). We hypothesize that mBART's training on shuffled and masked text aligns well with the inherently non-linear structure of gloss. Testing this on the PHOENIX-14T benchmark confirms our hypothesis, where mBART achieves State-of-the-Art results across six metrics, including sacreBLEU = 63.89 and COMET = 0.624. Our work introduces the first Bangla text-to-gloss framework and highlights the effectiveness of rule-based synthetic data in tackling low-resource sign language translation. Our study presents a novel approach to Bangla text-to-gloss translation using mBART and demonstrates the value of rule-based synthetic data in addressing low-resource sign language translation challenges.

1 Introduction

Sign language is the most natural mode of communication for deaf people. However, in a predominantly hearing society, deaf people often resort to lip reading, text-based communication, or media-facilitated interpersonal communication (Barnett, 2002) to interact with others. Sign language translation (SLT) is an important research area that aims to improve communication between signers and non-signers while allowing each party to use their preferred language. In addition, the deaf community has expressed their ease in communicating using sign rather than textual communications (Middleton et al., 2010).

Bangla, a language spoken by 272.7 million individuals (Sultana et al., 2025), also has a substantial deaf community that employs the Bangla Sign Language daily. According to the government defined disability categories, among the people of Bangladesh hearing disability is at 0.29%.(of Statistics) However, BdSL is a low-resource language. Although a recent work conducted end-to-end Sign Language Translation (SLT) (Zeeon et al., 2024), research has shown that the use of glosses as intermediaries improves the translation process (Gómez et al., 2021). As for Sign Language, gloss is considered as the written approximation of SL which uses words as "labels" for each sign along with various grammatical notes. A sign can have multiple meanings depending on the context of the sentence. For example, the English sentence "I am going to the store." is translated into gloss as "I GO STORE". Gloss thus acts as an intermediary between the SL and it's corresponding language. Additionally, the task of translating text into corresponding signs can greatly benefit from intermediary gloss form. Moryossef and Goldberg (Moryossef and Goldberg, 2021) represent SLT using a simple graph (Figure-1)

that describes 20 distinct tasks conceptually defined by this graph. Among them, BdSL only covers the video (sign) to text task (Zeeon et al., 2024; Sams et al., 2023; Sonare et al., 2021; Alam et al., 2021; Hoque et al., 2016; Harini et al., 2020). Few works have focused on the text-to-sign and text-to-pose tasks (Shahriar et al., 2017; Sarkar et al., 2009; Hoque et al., 2016). To the best of our knowledge,

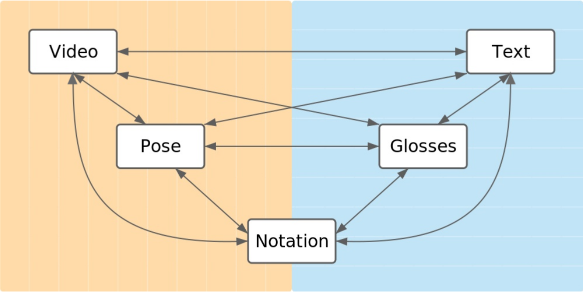

Figure 1: 20 Tasks of STL (Moryossef and Goldberg, 2021)

there are no previous works on Bangla text-to-gloss or vice versa. Research has shown that the use of glosses improves the translation process (Gómez et al., 2021), although challenges such as dataset limitations and under-articulated signing remain significant (Babbitt and Mansueto, 2012; Walsh et al., 2024). To address this gap, we present our contributions in this paper.

- State Of The Art model for Bangla-to-Gloss Translation
- A dataset for training the model
- Rule based algorithm for generating the dataset

We first construct a Bangla text-to-gloss dataset. Since annotating gloss for a large volume of sentences is costly, we leverage GPT-4o for this task. GPT has been widely used to generate synthetic data in the low-resource domain in recent years and has shown sufficient reliability (Zehady et al., 2024; del Barrio et al., 2024). Hence, we first annotated the gloss in 159 Bangla sentences that are used in day-to-day life with the help of a BdSL expert who taught deaf children at Bangladesh Deaf School. We used these data as reference for GPT-4o to generate future data. The prompt used for generating the data was: "Use this xlsx file as the source of truth, and based on the translation patterns in this file, translate the following bangla sentences to their gloss forms" We then evaluated these 159 text-to-gloss translations with two human annotators. The agreement score between the annotators was 93.96%, with Cohen's Kappa 0.7494. This result indicates substantial agreement between the annotators. Based on majority voting with a third annotator as tiebreaker, we prepare the ground truth and find that the accuracy of synthetic data is 86.57%. Therefore, we generate gloss form for 2062 Bangla sentences, randomly sampled from the Bangla POS-tagged corpus. This synthetic data from GPT-4o is our first source of data. We also utilize rule-based gloss generation approaches. A recent paper (Moryossef et al., 2021) proposes rule sets to generate gloss form from text. We adapt it for Bangla. We then use these heuristics to generate gloss form for 6509 Bangla sentences, sampled from Bangla book-review corpus (ManyThings.org). Our final data source is the widely used STL corpus - PHOENIX-Weather-2014T dataset (Camgoz et al., 2020). The text-to-gloss task has been shown to improve the performance of the multilingual corpus. By combining these datasets, we prepare our final dataset for Bangla text-to-gloss translation. We then train and fine-tuned different architectures and observe significantly high performance with the mBART model. Our approach can greatly reduce the cost of annotating huge amounts of gloss data with experts. We show that a small amount of high-quality/annotated data can be augmented with rule-generated gloss data to improve overall performance of BdSL text-to-gloss.

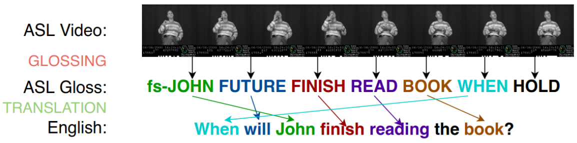
ASL Video:
GLOSSING
ASL Gloss:
TRANSLATION
English:
fs-JOHN FUTURE FINISH READ BOOK WHEN HOLD
When will John finish reading the book?

Figure 2: ASL video with gloss annotation and English translation (Moryossef et al., 2021)

2 Literature Review

We organize the literature review into four subsections. First, we define the concept of gloss. Next, we examine existing research on text-to-gloss translation. This is followed by a discussion of the limited body of work specific to Bangla text-to-gloss translation. Finally, we review recent advances in the use of large language models (LLMs) to generate synthetic data for low-resource languages.

2.1 What is gloss?

In linguistics, a gloss represents a word-by-word or morpheme-by-morpheme explanation of text from one language into another, often with grammatical and syntactic annotations. This is particularly helpful for languages with complex morphology, as it allows a detailed analysis of word formation and sentence structure. Glossing typically uses a three-line system: the first line contains the source language, the second provides a morpheme-by-morpheme breakdown, and the third line gives the natural translation (for Evolutionary Anthropology and of Leipzig, 2015; Lehmann, 1982; Croft, 2003). An example can be seen in Figure-2.

2.2 Text-to-Gloss

Text-to-gloss is a necessary task for sign language translation (SLT). Research has shown that considering gloss form can improve the performance of SLT compared to end-to-end neural networks. Additionally, for tasks like text-to-sign, text-to-gloss translation is a necessary step. Furthermore, as we can see in Figure-1, many SLT tasks depend on gloss form. Although the Hamburg Notation System (HamNoSys) can provide phonetic level notations via data-driven learning, the literature has shown that such an unsupervised approach can improve performance with supervised gloss notations. Despite such a necessity, text-to-gloss is a low-resource domain.

While our task is translation from a language to its corresponding sign language gloss (Text2Gloss), previous work on sign language translation has primarily focused on the opposite translation direction (Gloss2Text) as an intermediate step in translating from sign language videos to text (Sign2Text). Previous findings by Camgoz et al. showed that using a gloss as a mid-level representation in sign-to-text modeling improves performance on the Sign2Text task. For this intermediate Gloss2Text model, the authors trained an RNN-based encoder-decoder model with Gated Recurrent Units (GRUs) and report results on the PHOENIX-Weather-2014T dataset (Camgoz et al., 2020). Yin and Read also report results for a Gloss2Text model that uses the basic transformer architecture of Vaswani et al., tested on the PHOENIX-Weather-2014T dataset and the ASLG-PC12 dataset (Yin and Read, 2020; Vaswani et al., 2017). These two sets of results, while from the opposite translation direction, nonetheless offer a baseline for our Text2Gloss model.

For the Text2Gloss task, Stoll et al. used an RNN-based encoder-decoder with GRUs. They evaluated their model on the PHOENIX-Weather-2014T dataset and achieved performance comparable to that of Camgoz et al. for the opposite translation direction. This RNN-based Text2Gloss model provides a useful baseline for our Text2Gloss transformer model (Stoll et al., 2018). Babbit and Mansueto et al. use the POS tag and multilingual dataset to improve the performance of the text2gloss model (Babbitt and Mansueto, 2012).

2.3 Bangla text-to-gloss

Bangla, a morphologically rich language, lacks extensive research in gloss annotation. Although corpora such as the Bangla POS-tagged corpus or Bangla dependency treebanks exist, they focus primarily on syntactic and lexical tagging rather than morpheme-level glossing. There is no prior work on the Bangla text-to-gloss task to the best of our knowledge.

In fact, there is a general scarcity of text-to-gloss dataset across languages. Moryossef et al. (Moryossef et al., 2021) summarizes publicly available text-to-gloss datasets in an attempt to demonstrate the small size of corpus (Table-1).

Language Pair	Gloss Text Pairs	(Gloss/ Spoken)
Signum DGS-German (von Agris and Kraiss, 2007)	780	565 / 1,051
NCSLGR ASL-English (SignStream, 2007)	1,875	2,484 / 3,104
RWTH-PHOENIX -Weather-2014T (Camgoz et al., 2018)	8,257	1,870 / 4,839
French SL-French (Limsi, 2019)	2,904	2,266 / 5,028

Table 1: Some publicly available SLT corpora with gloss annotations and spoken language translation.

2.4 LLM for synthetic data generation

Incorporating synthetic data with real-world data has been shown to improve model adaptability and contextual understanding, particularly in domain-specific applications. This hybrid approach often outperforms models trained solely on real or synthetic data, as it provides a diverse and enriched training set (Zhezherau and Yanockin, 2024). Recent advances in leveraging large language models (LLMs) such as GPT-4 have demonstrated their efficacy in generating high-quality synthetic data for specialized tasks, including conversational semantic frame analysis. The research by Matta et al. (Matta et al., 2024) highlights the cost-efficiency of using LLM-generated data when combined with human-labeled data. The study shows that as budget constraints become more stringent, incorporating synthetic data significantly improves model performance compared to relying solely on human-labeled datasets. This indicates that LLMs can play a pivotal role in resource-limited settings, offering scalable data generation while maintaining relevance to the application domain.

However, the use of LLMs for synthetic data generation introduces potential biases, especially when the same model is used for both data generation and evaluation. Smaller models tend to exhibit biases towards their generated data, while larger models show more reliability (Maheshwari et al., 2024).

3 Methodology

3.1 Constructing Bangla T2G dataset

We construct a Bangla text-to-gloss (T2G) dataset which, to the best of our knowledge, is the first for Bangla language. We consider five data sources for Bangla T2G - synthetic data generated using LLM, gloss generation using grammatical rules, hand annotation, data augmentation and finally multilingual(German and English) data. With the help of BdSL experts, who are currently employed in sectors dedicate to disability studies and disability helping, we collected 159 hand annotated data. So we discuss the data collection process from the rest of the sources in this section.

3.1.1 Synthetic data generation using LLM

LLM has been widely used for generating synthetic data in low-resource domain in recent works and has shown sufficient reliability (Zehady et al., 2024; del Barrio et al., 2024). So we first annotate the gloss in 159 Bangla sentences with the BdSL experts. All annotators were informed about the goals and use of the dataset, and provided with a consent form prior to the annotation process. Verbal approval was obtained from each participant. Then we utilize this dataset to generate gloss form for 120 Bangla sentences with GPT-4o. We then evaluate these 120 text-to-gloss translations with three human annotators. The agreement score between the annotators was 93.96%, with Cohen Kappa 0.7494 indicating a substantial agreement between the annotators. Based on majority voting, we prepare the ground truth and find that the accuracy of synthetic data is 86.57%. Subsequently, we generated gloss representations for an additional 2062 Bangla sentences using GPT-4o, with the sentences randomly sampled from the Bangla POS-tagged corpus. (Dash, 2013). This data was further verified by BdSL experts.

3.1.2 Generating gloss from text using grammatical rules

Moryossef et al. proposes general rules and language specific rules for generating gloss form (Moryossef et al., 2021). For a given sentence S, we use general rules (Algorithm 1) to generate gloss for Bangla.

Algorithm 1 General Gloss Generation

```
Input: Set of tokens $S$
Output: Processed set of tokens $S'$
$S \leftarrow \{t \in S \mid POS(t) \in \{\text{noun, verb, adjective, adverb, numeral}\}\}$
foreach $t \in S$ do
    Discard $t$ with probability $p = 0.2$
$S \leftarrow \text{Lemmatize}(S)$
$S' \leftarrow$ Apply random permutation $\sigma$ such that $\forall i \in \{1, |S|\}, |\sigma(i) - i| \leq 4$
```

As for language-specific rules, we adopt their German-DGS Rules (Moryossef et al., 2021) for Bangla by considering Subject-Object-Verb position and negation (Algorithm 2).

Algorithm 2 Language Specific Gloss Generation

```
Input: Set of tokens $S$ with subject-object-verb
       triplets, POS tags, named entities (NER),
       dependency tags (DEP).
Output: Processed set of tokens $S'$.
foreach $(s, o, v) \in S$ do
    Swap the positions of $o$ and $v$ in $S$
$S \leftarrow \{t \in S \mid POS(t) \in \{\text{noun, verb, adjective, adverb, numeral}\}\}$
foreach $t \in S$ do
    if $POS(t) = adverb$ then
        Move $t$ to the end of $s$
    if $NER(t) = location$ then
        Move $t$ to the start of $s$
    if $DEP(t) = negation$ then
        Move $t$ to the end of $s$
    if $t$ is a compound noun $c_1c_2 \ldots c_n$ then
        Replace $t$ by $c_1$
$S \leftarrow \text{Lemmatize}(S)$
return $S'$
```

We first applied the generic rules on 158,065 Bangla samples from a dataset of book reviews

(Kabir et al., 2023). We then applied the language-specific rules on 6509 simple conversational Bangla samples from a Bangla-English machine translation dataset (ManyThings.org). We observed that simple conversational glosses were more meaningful than book-review glosses. A manual inspection showed that the book-reviews contained many complex words, grammars, and complex sentence structure. BdSL dictionary only has around 1200 words (ban, 1997), far little compared to Bangla dictionary (100,000 words) (Wikipedia contributors, n.d.). For that reason, it is hard to map 100,000 words into 1,200 vocabulary using a simple rule-based approach.

3.1.3 Incorporating multilingual Corpora

Most existing T2G models are trained or benchmarked on the RWTH-PHOENIX-Weather 2014T dataset (PHOENIX14-T) (Babbitt and Mansueto, 2012; Ouargani and Khattabi, 2023). This dataset comprises 8,247 sentences, divided into 7,096 training samples, 519 development samples, and 642 testing samples. It features a vocabulary of 1,085 signs and is used for both Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT) tasks. The dataset is derived from German public TV broadcasts that cover daily weather forecasts and news (Camgöz et al., 2018). As Moryossef et al. have shown in their work, considering high-quality multilingual dataset can improve T2G performance (Moryossef et al., 2021). So we consider PHOENIX14-T dataset in our experiment since BdSL text-to-gloss is a low-resource task and may benefit from high-quality multilingual data.

3.1.4 Data Augmentation

From the three data sources discussed so far, GPT-4o and hand-annotation guarantees high-quality data. The first was due to quantitative evaluation (acc=86.57%) and the latter due to expert supervision. We therefore had $2062 + 159 = 2221$ high quality Bangla text-to-gloss data at hand which we augment using back-translation and text generation (Sarker, 2021). Through these 2 techniques, we generated 4232 new samples.

3.1.5 Final Datasets

We combine the data sources above (GPT-4o, rule-based, multilingual, data augmentation, hand annotation) to construct 3 datasets during our experiments (Table-**??**). We observed that rule-based gloss generation works better on the conversational dataset, so we included it in our final dataset *bangla-gloss*.

3.2 Model Training for T2G Translation

We experimented with five models whose specifications and hyper-parameter choices are described in this section.

- BERT (Bidirectional Encoder Representations from Transformers): We use BERT model, particularly BERT multilingual base model (cased), as this supports Masked Language Modeling(MLM), which is essential for translation task. Furthermore, this model is intended to be fine-tuned for downstream tasks, such as ours. BERT's bidirectional attention mechanism enables it to capture context from both preceding and succeeding tokens that helps it understand linguistic nuances. We trained the model on our dataset. We tokenized the input text using WordPiece tokenizer and apply preprocessing steps to standardize gloss annotations. We use the AdamW optimizer with a learning rate of 0.00005 and train the model for 25 epochs with a batch size of 16.

- RNN (Recurrent Neural Network): RNNs are fast and efficient for predicting the next word. Due to the fast training time, we trained the RNN on 12910 samples from *bangla-gloss* dataset. However, since we are training from scratch, the low amount of data resulted in lower performance. We plan to fine-tune pre-trained RWKV as future work (Hochreiter and Schmidhuber, 1997).

- mBART (multilingual Bidirectional and Auto-Regressive Transformers): The choice of BART model came from an interesting observation: BART models are trained on shuffled and masked words, and as we saw in Algorithm 1, gloss involves shuffling making it suitable for gloss translation tasks. So we fine-tuned the mBART model using two datasets: *bangla-gloss* and *multilingual-gloss*. The mBART serves as a good base language model to fine-tune on (Liu et al., 2020). The learning rate was set to 0.0.00002 after some hyper-parameter finetuning.

- GRU (Gated Recurrent Unit): We also applied Gated Recurrent Unit since it has a larger context. In author's opinion, GRU-like models are the most suitable for text-to-gloss translation due to their limited context size because gloss sentences are smaller (Amin et al., 2021). However, due to limited data, training from scratch did not result in good performance (Cho et al., 2014).

- Sequence-to-Sequence model with Multi-head Attention Mechanism: Transformer can consider long context while RNN works great for short context. Gloss forms are small sentences. With that in mind, we decided to find an intersection between the context length of Transformer and the efficiency of RNN, inspired by the work of (Vaswani, 2017). We built a traditional Seq-to-Seq architecture with LSTM-based decoder and encoders. This is the commonly used LSTM variant for Seq-to-Seq tasks (i.e. translation of text-to-gloss). The traditional architecture has a single attention head. We explored multi-head attention with Seq-to-Seq model in this experiment. We trained it on 176857 samples from *rule-based-multilingual-gloss* dataset. To make training faster on such a large dataset, we used 16 bit floating point precision. The learning rate was 0.001.

We train *bangla-gloss* dataset on RNN, GRU. We fine-tuned bert-uncased multilingual model on *bangla-gloss*. We trained Seq-to-Seq with multi-head attention on the large *rule-based-multilingual-gloss*. Finally, we finetuned mBART on *bangla-gloss* and *multilingual gloss* dataset during experiment because the observed performance was high in mBART with only 3 epochs of finetuning.

3.3 Evaluation Scheme

To evaluate the text-to-gloss translation models, we use six metrics - sacreBLEU, BLEU-1, BLEU-2, BLEU-3, BLEU-4 and COMET based on (Babbitt and Mansueto, 2012; Moryossef et al., 2021). SacreBLEU is a standardized version of BLEU designed to ensure reproducibility by applying a consistent tokenization scheme and calculating n-gram precision with a brevity penalty to account for differences in sequence length. BLEU-1 through BLEU-4 assess n-gram overlap between generated and reference glosses, with BLEU-1 considering unigram precision and subsequent BLEU scores incorporating higher-order n-grams up to four. These metrics provide insight into the surface-level similarity of generated outputs with reference glosses, with higher-order BLEU scores emphasizing longer, contiguous n-gram matches. COMET, on the other hand, is a neural evaluation metric that predicts the quality of translations by comparing generated outputs against reference glosses using a combination of semantic embeddings and pre-trained language models.

4 Result

The performance of the models across various datasets and evaluation metrics is summarized in Table 2. The six evaluation metrics include sacreBLEU, BLEU-1, BLEU-2, BLEU-3, BLEU-4, and COMET, which provide a comprehensive assessment of translation quality. The mBART model, fine-tuned on specific datasets, consistently outperformed other configurations, demonstrating the advantages of pretrained multilingual transformers for gloss translation tasks.

In the *bangla-gloss* dataset, the mBART model achieved the highest sacreBLEU score of 79.53 and excelled in all BLEU metrics, with a BLEU-4 score of 44.71. The COMET score of 0.965 further validates the high quality of these translations, showcasing the effectiveness of domain-specific fine-tuning. For the *multilingual-gloss* dataset, the same model achieved a sacreBLEU score of 21.36 and a BLEU-4 score of 42.62, highlighting its ability to generalize across multiple languages, although with a slight reduction in performance compared to the Bangla-specific dataset.

The sequence-to-sequence model with an attention mechanism, trained on the *rule-based-multilingual-gloss* dataset, produced the lowest sacreBLEU score (6.63) and BLEU metrics. This outcome indicates the limitations of a simpler architecture in handling complex gloss translation tasks, particularly in comparison to the mBART model. Note that we skip RNN and GRU performance due to their very low performance in our experiment.

For the *PHOENIX-14T* dataset, the mBART model achieved a sacreBLEU score of 63.89 and a BLEU-4 score of 20.68, outperforming the results reported by Mansueto et al. (2024), Stoll et al. (2018), and Moryossef et al. (2021). Although the BLEU and sacreBLEU scores are competitive, the COMET score of 0.624 suggests challenges in capturing nuanced semantic equivalence in this dataset. These results emphasize the effectiveness of fine-

Dataset	Model	Training Size	Test Size	sacre-BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	COMET
rule-based-multi-lingual-gloss	Seq-to-Seq model with Attention	131,656	32914	6.63	20.75	7.52	4.62	3.77	-
multi-lingual-gloss	mBART	13320	857	21.36	83.27	70.55	55.66	42.62	0.908
multi-lingual-gloss	mBART	13320	519	26.86	59	40.84	29.08	21.96	0.953
bangla-gloss	mBART	7705	857	79.53	85.1	73.46	57.4	44.71	0.965
PHOENIX 14-T	mBART	8257	519	**63.89**	**55.14**	**38.07**	**27.13**	**20.68**	**0.624**
PHOENIX 14-T	Mansueto et al. (2024) (Babbitt and Mansueto, 2012)	8257	519	-	60.3	21.6	8.7	5.1	-
PHOENIX 14-T	Stoll et al. (2018) (Babbitt and Mansueto, 2012)	8257	519	-	50.67	32.25	21.54	15.26	-
PHOENIX 14-T	(Moryossef et al., 2021)	8257	519	23.35	-	-	-	-	0.1365
bangla-gloss	bert-base-uncased	7705	857	3.95	15.91	8.68	5.58	4.42	-

Table 2: Performance Comparison Across Datasets and Models

tuning large pre-trained models on domain-specific datasets for gloss translation.

5 Findings

Based on these results, we try to answer three research questions.

RQ1: Does mBART improve text-to-gloss performane compared to mBERT since it is inherently trained to handle shuffling?

mBART is trained on shuffled and masked words, while mBERT is trained on masked sentences in a bidirectional fashion. As we saw in Algorithm 1, gloss form has shuffling property. Since mBART is inherently trained on shuffling, it is better suited for gloss tasks, something the gloss literature has not considered before to the best of our knowledge. In our experiment, we finetune mBART and bert-base-uncased multilingual model on the same *bangla-gloss* dataset. We observe significantly better result using mBART.

RQ2: Does mBART outperform state-of-the-art text-to-gloss models?

As discussed in RQ1, mBART has inherent properties that make it very suitable for text-to-gloss translation task. However, it has not been considered in literature before. So we benchmark mBART against 3 works from 2024 (Babbitt and Mansueto, 2012) , 2021 (Moryossef et al., 2021) and 2018 (Babbitt and Mansueto, 2012). We observe that our finetuned mBART consistently outperforms their proposed models in all six metrics on the PHOENIX-14T benchmark. This provides strong support for the notion that mBART outperforms state-of-the-art text-to-gloss models.

RQ3: Can rule based gloss generation help Bangla text-to-gloss task?

As we can see in Table 2, the performance of Bangla text-to-gloss translation is very high (sacreBLEU=79.53). We evaluated it with expert-annotated dataset as well and the sacreBLEU was similarly high. This finding supports previous literature where (Moryossef et al., 2021) similarly showed that rule-based gloss generation can improve text-to-gloss performance for German sign language. Additionally, we observe that finetuned language models significantly perform better than training a large model from scratch (from BLEU 20.75 to 85.10). We recommend considering rule-based gloss generated from simple conversational dataset. Because it is a low-cost solution to improve performance of text-to-gloss models. We also recommend finetuning over training from scratch since gloss is inherently a low-resource domain.

6 Discussion & Future Work

In this work we achieved state-of-the-art performance on the text-to-gloss task for BdSL. We finetune mBART model on the Phoenix Weather 2014T benchmark and achieve a sacreBLEU score of 63.89 and COMET score of 0.624. We performed a preliminary literature review to identify the reasons behind achieving such a SOTA performance where a surprisingly lack of mention of fine-tuning base language models for text-to-gloss translation was observed. Additionally, we also observe that synthetic data generated using Bangla gloss generation rules can indeed effectively improve text-to-gloss performance. This finding supports the conclusion proposed in (Moryossef et al., 2021). In the future, we aim to explore more intuitive and native gloss generation rules for Bangla. Additionally, we want to incorporate the insights from the linguists into the Bangla gloss generation algorithm to generate better glosses.

Limitations

Our dataset primarily consists of hand-annotated examples; however, the volume of such annotations remains limited. This constraint affects the coverage and diversity of linguistic constructions, particularly for complex or less frequently used sentence structures. Additionally, our gloss generation relies on a rule-based approach inspired by American and German Sign Language. Although this provides a useful starting point, Bangla exhibits unique syntactic and morphological characteristics that are reflected in Bangla Sign Language (BdSL). As such, a set of glossing rules tailored specifically to BdSL would likely produce more accurate and linguistically faithful representations. Finally, the annotated dataset predominantly includes simple sentences, which can limit the model's ability to translate more complex or context-rich sentences.

References

1997. *Bangla Ishara Vasha Ovidhan*. Accessed: December 29, 2024.

Md Shahinur Alam, Mahib Tanvir, Dip Kumar Saha, and Sajal K Das. 2021. Two dimensional convolutional neural network approach for real-time bangla sign language characters recognition and translation. *SN Computer Science*, 2:1–13.

Mohamed Amin, Hesahm Hefny, and Mohammed Ammar. 2021. Sign language gloss translation using deep learning models. *International Journal of Advanced Computer Science and Applications*, 12(11).

Luke Babbitt and Jenna Mansueto. 2012. Text2gloss: Translation into sign language gloss with transformers. In *Proceedings of the Stanford CS224N Custom Project*.

Steven Barnett. 2002. Communication with deaf and hard-of-hearing people: a guide for medical education. *Academic Medicine*, 77(7):694–700.

Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Parallel sign language translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Rwth-phoenix-weather 2014 t: Parallel corpus of sign language video, gloss and translation. *CVPR, Salt Lake City, UT*, 3:6.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fouad Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1724–1734. Citeseer.

William Croft. 2003. *Typology and universals*, 2nd edition. Cambridge University Press, Cambridge.

Niladri Sekhar Dash. 2013. Part-of-speech (pos) tagging in bengali written text corpus. *International Journal on Linguistics and Language Technology*, 1(1):53–96.

David Alonso del Barrio, Max Tiel, and Daniel Gatica-Perez. 2024. Human interest or conflict? leveraging llms for automated framing analysis in tv shows.

Max Planck Institute for Evolutionary Anthropology and University of Leipzig. 2015. Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Last updated: May 31, 2015.

Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th workshop on building and using comparable corpora (BUCC 2021)*, pages 18–27.

R. Harini, R. Janani, S. Keerthana, S. Madhubala, and S. Venkatasubramanian. 2020. Sign language translation. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 883–886.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Md. Tazimul Hoque, Md. Rifat-Ut-Tauwab, Md. Fasihul Kabir, Farhana Sarker, Mohammad Nurul Huda, and Khandaker Abdullah-Al-Mamun. 2016. Automated bangla sign language translation system: Prospects, limitations and applications. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 856–862.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. Banglabook: A large-scale bangla dataset for sentiment analysis from book reviews. *arXiv preprint arXiv:2305.06595*.

Christian Lehmann. 1982. Directions for interlinear morphemic translations. *Folia Linguistica*, 16:199–224.

Limsi. 2019. Dicta-sign-lsf-v2. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.

Yuxian Liu, Shijie Yang, Junyang Li, Xu Du, Haitao Yang, and Xipeng Liu. 2020. Multilingual denoising pre-training for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 3008–3018. Association for Computational Linguistics.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *arXiv preprint arXiv:2409.11968*.

ManyThings.org. Bangla-english sentence pairs. http://www.manythings.org/anki/ben-eng.zip. Accessed: 2024-12-26.

Shiho Matta, Yin Jou Huang, Fei Cheng, Hirokazu Kiyomaru, and Yugo Murawaki. 2024. Investigating cost-efficiency of llm-generated training data for conversational semantic frame analysis. *arXiv preprint arXiv:2410.06550*.

Anna Middleton, Graham H Turner, Maria Bitner-Glindzicz, Peter Lewis, Martin Richards, Angus Clarke, and Dafydd Stephens. 2010. Preferences for communication in clinic from deaf people: A cross-sectional study. *Journal of Evaluation in Clinical Practice*, 16(4):811–817.

Amit Moryossef and Yoav Goldberg. 2021. Sign Language Processing. https://sign-language-processing.github.io/.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.

Bangladesh Bureau of Statistics. National survey on persons with disabilities (nspd) 2021. https://bbs.portal.gov.bd/sites/default/files/files/bbs.portal.gov.bd/page/b343a8b4_956b_45ca_872f_4cf9b2f1a6e0/2022-06-13-15-24-ca6f018ab83c88a4db8ff51386439794.pdf. Accessed: 25-6-2025.

Younes Ouargani and Noussaima El Khattabi. 2023. Advancing text-to-gloss neural translation using a novel hyper-parameter optimization technique. *arXiv preprint arXiv:2309.02162*.

Ataher Sams, Ahsan Habib Akash, and S. M. Mahbubur Rahman. 2023. Signbd-word: Video-based bangla word-level sign language and pose translation. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Biswajit Sarkar, Kaushik Datta, CD Datta, Debranjan Sarkar, Shashanka J Dutta, Indranil Das Roy, Amalesh Paul, Joshim Uddin Molla, and Anirban Paul. 2009. A translator for bangla text to sign language. In *2009 Annual IEEE India Conference*, pages 1–4. IEEE.

Sagor Sarker. 2021. bnaug. https://github.com/sagorbrur/bnaug.

Rhythm Shahriar, AGM Zaman, Tanvir Ahmed, Saqib Mahtab Khan, and HM Maruf. 2017. A communication platform between bangla and sign language. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, pages 1–4. IEEE.

NCSLGR SignStream. 2007. *Volumes 2–7*.

Babita Sonare, Aditya Padgal, Yash Gaikwad, and Aniket Patil. 2021. Video-based sign language translation system using machine learning. In *2021 2nd International Conference for Emerging Technology (INCET)*, pages 1–4.

Stefan Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2018. Text2gloss: Sequence learning for text representation in sign language translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nusrat Sultana, Rumana Yasmin, Bijon Mallik, and Mohammad Shorif Uddin. 2025. Onubad: A comprehensive dataset for automated conversion of bangla regional dialects into standard bengali dialect. *Data in Brief*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, and 1 others. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*.

Ulrich von Agris and K. Kraiss. 2007. Towards a video corpus for signer-independent continuous sign language recognition. In *Gesture in Human-Computer Interaction and Simulation*.

Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, and Richard Bowden. 2024. A data-driven representation for sign language production. In *Proceedings of the 18th International Conference on Automatic Face and Gesture Recognition (FG 2024)*. Institute of Electrical and Electronics Engineers (IEEE).

Wikipedia contributors. n.d. Bengali vocabulary. Accessed: December 29, 2024.

Xiang Yin and Jonathan Read. 2020. Sign language translation with transformer networks. In *Proceedings of the Workshop on Human-Machine Communication*.

Iftekhar E Mahbub Zeeon, Mir Mahathir Mohammad, and Muhammad Abdullah Adnan. 2024. Btvsl: A novel sentence-level annotated dataset for bangla sign language translation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE.

Abdullah Khan Zehady, Safi Al Mamun, Naymul Islam, and Santu Karmaker. 2024. Bongllama: Llama for bangla language. *arXiv preprint arXiv:2410.21200*.

Alexey Zhezherau and Alexei Yanockin. 2024. Hybrid training approaches for llms: Leveraging real and synthetic data to enhance model performance in domain-specific applications. *arXiv preprint arXiv:2410.09168*.

A Combined datasets across languages for gloss experiments

The combined datasets across languages for gloss experiments are shown in Table 3

Dataset	**rule-based-multi-lingual-gloss**	**multi-lingual-gloss**	**bangla-gloss**
Annotated	173	173	173
Rule Based	158065 (von Agris and Kraiss, 2007), 6509 (Wikipedia contributors, n.d.)	0	6509
GPT-4o-mini	1996	1996	1996
Augmented	0	4232	4232
English	1875 (Limsi, 2019)	0	0
German	8257 (Zhezherau and Yanockin, 2024)	8257	0
Size	176857	14658	12910

Table 3: Combined datasets across languages for gloss experiments.