ISSN 2007-9737

A Domain Specific Parallel Corpus and Enhanced English-Assamese Neural Machine Translation

Sahinur Rahman Laskar[1], Riyanka Manna[2], Partha Pakray[1], Sivaji Bandyopadhyay[1]

[1] National Institute of Technology Silchar,
Department of Computer Science and Engineering,
India

[2] Adamas University,
Department of Computer Science and Engineering,
India

{sahinurlaskar.nits, riyankamanna16, parthapakray, sivaji.cse.ju}@gmail.com

Abstract. Machine translation deals with automatic translation from one natural language to another. Neural machine translation is a widely accepted technique of the corpus-based machine translation approach. However, an adequate amount of training data is required, and there is a need for the domain-wise parallel corpus to improve translational performance that shows translational coverages in various domains. In this work, a domain-specific parallel corpus is prepared that includes different domain coverages, namely, Agriculture, Government Office, Judiciary, Social Media, Tourism, COVID-19, Sports, and Literature domains for low-resource English-Assamese pair translation. Moreover, we have tackled data scarcity and word-order divergence problems via data augmentation and prior alignment concept. Also, we have contributed Assamese pretrained LM, Assamese word-embeddings by utilizing Assamese monolingual data, and a bilingual dictionary-based post-processing step to enhance transformer-based neural machine translation. We have achieved state-of-the-art results for both forward (English-to-Assamese) and backward (Assamese-to-English) directions of translation.

Keywords. English-Assamese, low-resource, neural machine translation, parallel corpus, data augmentation, prior alignment, language model.

1 Introduction

Machine translation (MT) is a sub-field of natural language processing (NLP) that helps to bridge gaps in communication via automatic translation without human assistance. With the advancement of deep learning techniques, machine translation technique, namely, neural machine translation (NMT) shows remarkable translation accuracy [3, 26]. The NMT is a corpus-based approach of MT, which requires large amount of bilingual corpus for training a NMT model to achieve a good translation performance.

However, the adequate amount of training data is a challenging issue for low-resource settings [19]. Generally, low-resource pairs are considered if the training amount of parallel data is less than 1 million [16]. For instance, English–Mizo (En-Mz) [33, 21, 15], English–Assamese (En-As) [23, 24], English–Khasi (En-Kha) [22] are the examples of low-resource pairs.

The majority languages of the worldwide can be considered as “low-resource” based on the availability resources [29, 36]. Furthermore, the precise definition of “low-resource language pair” is a research question itself since the morphological rich low-resource languages in addition to the presence of varieties of inflected words, require more bitext data to achieve equivalent translation performance of languages that have less inflected words [7].

Moreover, NMT shows weakness in case of out-domain data [19], which demand to develop domain specific parallel corpus to improve low-resource pair translation.

doi: 10.13053/CyS-26-4-4423

In this paper, we have investigated a low-resource pair “En-As” to improve NMT for both directions, En-to-As and As-to-En translation. From the linguistic aspects, En and As are very different to each other, for instance, unlike En [23], as follows subject-object-verb (SOV), morphological rich language and adopts Assamese-Bengali script [28] originated from the Gupta script [8]. Our contributions are summarized as follows:

— We have created a domain specific En-As parallel corpus, which covers various domains, namely, social media, agriculture, Government office, judiciary, sports, tourism, COVID-19 and literature.

— We have addressed data scarcity and word-order divergence problems to enhance NMT for En-As language pair translation. By utilizing monolingual As data, synthetic En-As parallel sentences are prepared and extracted phrase pairs from the original parallel sentences (train set).

  To tackle the data scarcity issue, the extracted phrase pairs are augmented to the original parallel data and leveraging synthetic parallel data in the training model via two steps process: pretrain on the train data with synthetic parallel data and then fine-tuned on the train data without synthetic parallel data.

  Moreover, we have utilized pretrained multilingual contextual embeddings-based alignment technique to extract alignment information and that is used as prior alignment information during the training phase to tackle the word-order divergence issue.

— We have contributed an Assamese pretrained language model (AsLM) and word-embeddings vectors (AsGloVe) that shall be used in various downstream NLP tasks of Assamese language. The AsLM and AsGloVe are used for the improvement of En-As NMT.

— We have contributed a bilingual dictionary of En-As that is used in the post-processing step to tackle out-of-vocabulary issue and enhance En-to-As and As-to-En translations.

— We have achieved state-of-the-art results for low-resource En–As MT translational performance in terms of automatic and manual evaluation.

The rest of the paper is structured as follows: Section 2 discuss background concept and the related works. The domain specific parallel corpus and dataset description is presented in Section 3. Section 4 reported the baseline system results. Section 5 and 6 describe the proposed approach and reported results with analysis. Lastly, Section 7 conclude the paper with future scopes.

2 NMT Background and Related Work

Statistical machine translation (SMT) and NMT are two well-studied corpus-based MT techniques in the MT. To enhance low-resource pair translation, researchers have started experimenting with NMT recently. In this section, we have discussed the fundamentals of NMT and also emphasizes earlier research on English-Assamese MT.

2.1 NMT

The corpus-based (also known as data-driven) approach of NMT introduces RNN-based encoder-decoder architectures, where seq-2-seq learning is achievable by tackling variable length phrases of source-target sentences [3, 26].

To learn the long-term features of the source and target words for encoding and decoding, long short-term memory (LSTM) has demonstrated remarkable performance in this case. When encountering too lengthy sentences, it is unable to encode all the necessary information.

For that reason, the attention mechanism has been introduced in NMT [3, 26] that enables the decoder to take into account various segments of the source sequence during various decoding steps.

In the encoder-decoder based NMT, the encoder is responsible for the encoding of input sequences $sr_1, sr_2 \ldots sr_n$ and generates a vector U.

ISSN 2007-9737

Whereas, the decoder decodes the output $tr_1, tr_2 \ldots tr_m$ using computation of condition probability, as given in Eq. (1):

$$P(tr \mid sr) = \sum_{i=1}^{m} P\left(tr_i \mid tr_{<1},\ U\right). \tag{1}$$

Using Eq. (2), the value of at_o correlates to the frequency of time steps in the input sentence. Therein, in the source side (h_i) and target side (h_o), the series of hidden states are computed, which are finally correlated to produce the attention vector at_o:

$$at_o = \frac{\exp\left(\text{score}\left(h_o,\ h_i'\right)\right)}{\sum_{i'} \exp\left(\text{score}\left(h_o,\ h_{i'}\right)\right)}. \tag{2}$$

The general estimate of score function defined in Eq. (3) is considered in this work for the preliminary experiments of the baseline system:

$$\text{score}\left(h_o,\ h_i'\right) = h_o\ W_a\ h_i'. \tag{3}$$

Then, the context vector c_l is computed by using the hidden states average input weights with the attention vector. The attentional hidden vector is computed using Eq. (4) by the concatenation of h_o and c_t:

$$h_o' = \tanh\left(W_c\left[c_t,\ h_o\right]\right). \tag{4}$$

Finally, the softmax layer is included to the vector h_o' using Eq.(5) to obtain the predicted target sequence:

$$P\left(t_j \mid t_{<1}, U\right) = \text{softmax}\left(W_s\ h_o'\right). \tag{5}$$

The disadvantages of RNN-based NMT in terms of parallelization and long-term dependencies are tackled by introducing transformer-based NMT [42]. The primary idea behind the transformer model is to make use of the self-attention mechanism, an attention mechanism found inside the encoder.

Each token position is encoded by the transformer model, and self-attention is employed to connect two different tokens that aid in parallelization to quicken learning. The self-attention, also known as multi-head attention, computes attention several times.

The encoder-decoder architecture of transformer-based NMT contains six identical attention layers that are placed on stack of each other. The position of the input sequence is encoded and embedded to combine the sequence of tokens prior to feeding the sequence into the network.

The encoder consists of a point-wise connected feed-forward network layer and multiple headed attention layer. Whereas, the decoder comprises three layers and two of these layers are identical to the encoder.

The another multi-head attention layer is the third layer of the decoder that focuses to attend the output sequence a headed by the encoder. Here, the attention is calculated by considering the dot product of the input and utilizing a softmax function to get the weight of each token at a given position using Eq. (6):

$$\text{Attn}(Q,\ K,\ V) = \text{softmax}\left(\frac{QT^k}{\sqrt{d_k}}\right)V. \tag{6}$$

To compute the attention, input vectors such as query (Q), key (K) with dimension d_k, and value (V) are used. The advantage of using multi-head (MHD) attention in the transformer model over single-head attention is that it allows you to deal with different word representations through multiple positions. As shown in Eq. (7) and (8), the number of parallel attention heads accounts for $h = 8$:

$$\text{MHD}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\ W^O, \tag{7}$$

$$\text{head}\ _i = \text{Attn}\left(QW_i^Q, KW_i^K, VW_i^v\right), \tag{8}$$

where the parameter matrices $W_i^Q \in R^{d_{\text{model}} \times d_k}$, $W_i^K \in R^{d_{\text{model}} \times d_k}$, and $W_i^V \in R^{d_{\text{model}} \times d_r}$.

2.2 Related Work on English–Assamese MT

In literature review of the MT for English-Assamese pair, it is noted that the researchers are working on the dataset preparation to overcome the dataset's scarcity for such a low-resource pair [4, 14, 23, 37]. The authors of [4] build a phrase-based SMT translation system via preparation of a small En-As parallel corpus of 14,371 sentences.

doi: 10.13053/CyS-26-4-4423

Table 1. Data statistics for domain wise parallel sentences

Domain	Parallel Sentences
Agriculture	2,150
COVID-19	5,500
Government Office	9,500
Judiciary	4,500
Social Media	3,220
Sports	8,600
Tourism	4,750
Literature	19,300
Total	**57,520**

Table 2. Train, validation and test data statistics

Type	Sent	Tokens	
		En	As
Train Set-1 [23]	203,315	2,414,172	1,986,270
Train Set-2 [37]	138,353	1,715,435	1,377,336
Train Set-3	46,016	560,972	446,500
Total	**387,684**	**4,690,579**	**3,810,106**
Validation Set-1 [23]	4,500	74,561	59,677
Validation Set-2 [37]	1,000	19,922	16,824
Validation Set-3	5,752	75,652	65,612
Total	**11,252**	**170,135**	**142,113**
Test Set-1 [23]	2,500	41,985	34,643
Test Set-2	5,752	75,348	65,576

Table 3. En/As Monolingual data statistics

Type	Sentences	Tokens
As	2,810,197	47,740,981
En	3,387,704	58,847,760

In our previous work [23], a parallel corpus, namely, EnAsCorp1.0 [23] is developed, and it contains 210,315 parallel sentences. And, the same has been used to build baseline models for En-As pair translation using the phrase-based SMT and RNN-based NMT.

Then in the previous work [24], we have explored different NMT models (RNN and transformer) with data augmentation approach and attains better results on the same test set [23] for En-As pair translation.

Moreover, a parallel corpora, namely, Samanantar [37] that contains 11 Indian languages with English, and it includes 141,353 English-Assamese parallel sentences. Also, they [37] implemented transformer-based NMT model for the En-to-Indic and Indic-to-En.

It is noted that all the prior works that have been conducted on this English-Assamese MT are not domain specific. In this work, we have prepared domain specific English-Assamese parallel corpus and utilized parallel corpus of EnAsCorp1.0 and Samanantar to enhance the translational performance for both forward and backward directions of translation. We have addressed data scarcity and word order divergence issues via data augmentation and guided alignment concept.

3 Domain Specific Parallel Corpus Preparation and Dataset Description

In this section, we briefly discuss dataset preparation. First, we have collected Assamese monolingual data from the available online sources. For agriculture and social media domains, we have collected from Assamese monolingual sentences[1] from [34].

The Assamese monolingual sentences of sports[2], literature[3] domains are extracted from the News and Xahityo online sources. For extraction, we used the technique of web scraping, which is an automatic method to obtain large amounts of data from websites.

We employed Scrapy[4] for this purpose. Scrapy is a free and open-source web-crawling framework written in Python. While scrapping, we faced several challenges, which were mainly because of different web page structure in different websites and dynamic web content. Then, Assamese monolingual sentences are translated into English sentences using Bing translator[5].

[1] https://github.com/anononymus/assamese-redup

[2] https://www.asomiyapratidin.in/

[3] https://xahitya.org/

[4] https://github.com/scrapy/scrapy

[5] https://www.bing.com/translator

Table 4. Baseline system results on test set-1 in terms of automatic evaluation scores

Translation	Model	BLEU	TER	RIBES	METEOR	F-measure
En-to-As	PBSMT (Baseline-1)	4.85	103.2	0.2598	0.0768	0.1745
	RNN (Baseline-2)	6.78	93.4	0.2847	0.0996	0.2074
	Transformer (Baseline-3)	6.92	93.1	0.2878	0.1043	0.2106
As-to-En	PBSMT (Baseline-1)	8.58	90.5	0.2938	0.1070	0.2095
	RNN (Baseline-2)	12.52	88.6	0.4262	0.1421	0.2871
	Transformer (Baseline-3)	12.84	88.1	0.4284	0.1477	0.2876

Table 5. Baseline system results on test set-2 in terms of automatic evaluation scores

Translation	Model	BLEU	TER	RIBES	METEOR	F-measure
En-to-As	PBSMT (Baseline-1)	3.62	105.6	0.1676	0.0472	0.1356
	RNN (Baseline-2)	4.26	98.3	0.1706	0.0647	0.1994
	Transformer (Baseline-3)	4.66	98.2	0.1732	0.0686	0.2006
As-to-En	PBSMT (Baseline-1)	4.02	100.8	0.1710	0.0526	0.1487
	RNN (Baseline-2)	6.28	96.6	0.2064	0.1062	0.2008
	Transformer (Baseline-3)	6.49	96.5	0.2098	0.1084	0.2096

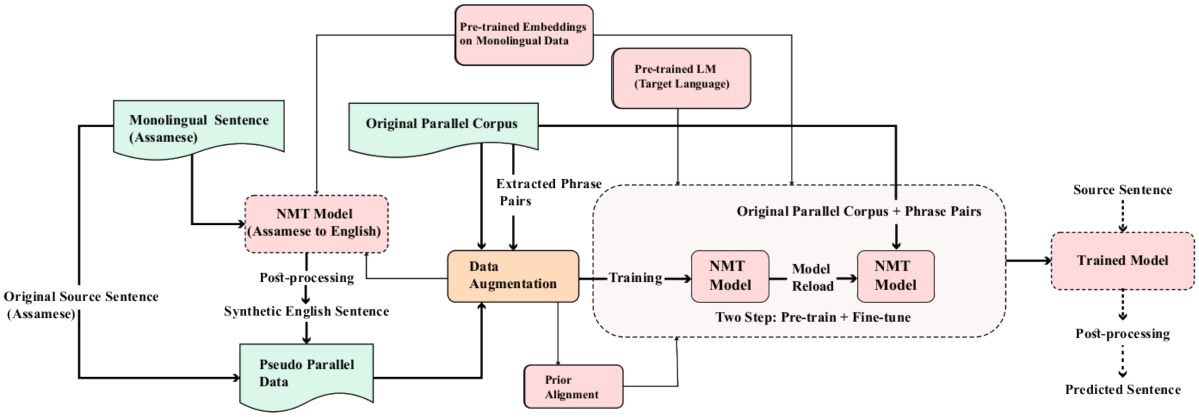
Pre-trained Embeddings on Monolingual Data
Pre-trained LM (Target Language)
Monolingual Sentence (Assamese)
Original Parallel Corpus
Extracted Phrase Pairs
Original Parallel Corpus + Phrase Pairs
Source Sentence
NMT Model (Assamese to English)
Data Augmentation
Training
NMT Model
Model Reload
NMT Model
Trained Model
Post-processing
Synthetic English Sentence
Original Source Sentence (Assamese)
Two Step: Pre-train + Fine-tune
Pseudo Parallel Data
Prior Alignment
Post-processing
Predicted Sentence

Fig. 1. Proposed approach for English-Assamese NMT

Similarly, English side sentences are extracted from News[6] via scraping for the domain of COVID-19 and tourism domains. And, we have collected English sentences of Government office and judiciary domains from IIT Bombay English-Hindi parallel corpus[7]. Then, utilize Bing translator to generate corresponding Assamese sentences. We have considered maximum sentence length 50 words.

Further, we have manually corrected and verified the parallel sentences. For manual verification, we have hired three linguistic experts who possess linguistic knowledge of both English and Assamese, and it took about 70 days.

The statistics of domain-wise parallel sentences are summarized in Table 1. The domain-wise parallel data is split into train, validation, and test data by considering 90%, 10%, 10% from each domain (Agriculture / COVID-19 / Government Office / Judiciary / Social media / Sports / Tourism / Literature) for train, validation and test set.

[6] https://theprint.in/

[7] https://www.cfilt.iitb.ac.in/iitb_parallel/

doi: 10.13053/CyS-26-4-4423

Table 6. Train data statistics before and after phrase pairs augmentation, OPC:"original parallel corpus", PP: "phrase pairs"

Type	Sentences
OPC	387,684
PP	10,103,84
OPC + PP	13,980,68

Table 7. BLEU scores results of transformer-based NMT on test set-1, M1 (Without domain-specific parallel data (Train Set-3)): Train Set-1+Train Set-2; M2 (baseline-3): With domain-specific parallel data (Train Set-3) + Train Set-1+Train Set-2; M3: M1+PSP (Post-processing); M4: M2+PSP (Post-processing)

Translation	M1	M2	M3	M4
En-to-As	6.74	6.92	6.87	7.04
As-to-En	12.54	12.84	12.78	12.96

Table 8. BLEU scores results of transformer-based NMT on test set-2, M1 (Without domain-specific parallel data (Train Set-3)): Train Set-1+Train Set-2; M2 (baseline-3): With domain-specific parallel data (Train Set-3) + Train Set-1+Train Set-2; M3: M1+PSP (Post-processing); M4: M2+PSP (Post-processing)

Translation	M1	M2	M3	M4
En-to-As	1.16	4.66	1.21	4.84
As-to-En	2.24	6.49	2.32	6.68

We have named these sets: train set-3, validation-3 and test set-2. The data set statistics, that are used in this work, are summarized in Table 2.

In Table 2, we have merged parallel corpora, namely, EnAsCorp1.0 [23] and Samanantar [37].

Furthermore, we have used monolingual data of Assamese/English from [23] and Assamese/English side monolingual sentences from train set-3.

The data statistics of monolingual data are presented in Table 3. It is mainly used for the preparation of pretrained word embeddings and LM.

4 Baseline System

In our previous work, we have prepared EnAsCorp1.0 [23], wherein, parallel En-As corpus and monolingual sentences of As are collected. The same dataset was used to implement baseline systems by considering two models, namely, phrase-based SMT (baseline-1) and RNN-based NMT (baseline-2).

In this work, we have considered domain-wise En-As parallel corpus (as mentioned in Section 3) in addition to EnAsCorp1.0 [23] and Samanantar [37], data statistics are shown in Table 2. Moreover, custom pretrained word embeddings using GloVe [35] is utilized in NMT models.

For baseline systems, transformer-based NMT [42] (baseline-3) is also considered in addition to RNN-based NMT (baseline-2) and phrase-based SMT (baseline-1).

The reason behind choosing transformer-based NMT in baseline systems is that it outperforms RNN-based NMT and PBSMT (as reported in Table 4, 5) and performs fair comparisons with improved transformer-based NMT (as discussed in Section 5).

To evaluate quantitative results, standard evaluation metrics [32], namely, BLEU (bilingual evaluation under study), TER (translation error rate) [41], RIBES (rank-based intuitive bilingual evaluation score) [11], METEOR (metric for evaluation of translation with explicit ordering) [25], and F-measure scores are considered.

5 Enhanced English-Assamese NMT

In the previous section, we have reported baseline system results, and it is noticed that transformer-based NMT achieves best results for both directions of translation. Therefore, we have chosen transformer-based NMT for further investigation.

In this section, we have briefly described the improved transformer-based NMT for low-resource En-As pair by investigating different approaches like data augmentation, prior alignment, pretrained LM and post-processing step. Figure 1 depicts the proposed approach for En-As NMT.

Table 9. BLEU scores results of transformer-based NMT on test set-1, M5:M2+PP (Phrase pairs); M6: M5+PSP (Post-processing)

Translation	M2	M5	M6
En-to-As	6.92	8.46	9.12
As-to-En	12.84	14.34	15.06

Table 10. BLEU scores results of transformer-based NMT on test set-2, M5:M2+PP (Phrase pairs); M6: M5+PSP (Post-processing)

Translation	M2	M5	M6
En-to-As	4.66	7.86	8.17
As-to-En	6.49	10.34	10.84

Table 11. BLEU scores results of transformer-based NMT on test set-1, M7: M5+SP (synthetic parallel data (pretrain + fine-tune)); M8: M7+PSP (Post-processing)

Translation	M2	M5	M7	M8
En-to-As	6.92	8.46	9.52	10.12
As-to-En	12.84	14.34	15.66	16.04

Table 12. BLEU scores results of transformer-based NMT on test set-2, M7: M5+SP (synthetic parallel data (pretrain + fine-tune)); M8: M7+PSP (Post-processing)

Translation	M2	M5	M7	M8
En-to-As	4.66	7.86	8.64	9.06
As-to-En	6.49	10.34	11.74	12.10

5.1 Data Augmentation

We have tackled data scarcity problem via data augmentation in two-ways: augmenting phrase-pairs and utilizing synthetic parallel data without modifying the NMT model architecture.

Following the strategy [38], phrase-based SMT is trained on original parallel data using Moses[8] toolkit and extracted phrase pairs from the generated phrase table.

However, in our previous work [24], it is noticed that the extracted phrase pairs contain wrong alignment phrases [20].

Therefore, we have extracted phrase pairs by considering different translation probabilities (Set$_{p\geq0.5}$ / Set$_{p=1}$ / Set$_{\text{all}}$) of target phrases given source phrases following [38, 24] and observed that the translation accuracy with augmentation of extracted phrase pairs having translation probability Set $_{p\geq0.5}$ are higher.

Therefore, we have considered phrase pairs with translation probability Set$_{p\geq0.5}$ and the data statistics are reported in Table 6.

Further, to expand the parallel corpus, monolingual data is used to generate synthetic parallel data following BT strategy [39, 24]. However, it is observed that the translational accuracy with augmented data is lower than the without augmented one.

Therefore, following our previous work [24] a two-step solution is used [1]. First, pretrain the NMT model with synthetic data and “original parallel corpus + phrase pairs” and then fine-tune or reload it on the “original parallel corpus + phrase pairs”.

As a result of this , the final model initializes the parameters from the pretrained model that gains the training performance when the “original parallel corpus + phrase pairs” is utilized. We have used As-to-En transformer-based NMT model to generate synthetic parallel data using Assamese monolingual sentences since it gives higher translation accuracy, as shown in Table 4, 5.

To examine the effect of augmented synthetic parallel data, we have performed a series of experiments like our previous work [24] on the ratio of parallel and synthetic corpora. It is noticed that 1:3 + phrase pairs attain higher translation accuracy for As-to-En and similar observation is found in case of En-to-As with 1:4 + phrase pairs and therefore, we have reported these results in Section 6.

5.2 Prior Alignment and Pretrained LM

The word order or token position of English is different from Assamese [24] that leads to word-order divergence issue. In this work, we have attempted to extract token alignment information from the En-As bi-text data and feeded into NMT to enhance En-to-As and As-to-En directions of

[8] http://www.statmt.org/moses/

Table 13. BLEU scores results of transformer-based NMT on test set-1, M11:M7 with PA1 (BA); M12: M7 with PA1 (UA); M13: M7 with PA1 (RA); M14: M7 with PA2 (BA); M15: M7 with PA2 (UA); M16: M7 with PA2 (RA), where PA1:Prior Alignment (FastAlign), PA2:Prior Alignment (SimAlign), UA: Unidirectional Alignment, BA: Bidirectional Alignment (grow-diagonal heuristics), RA: Reverse Direction Alignment

Translation	Model	BLEU
En-to-As	M7	9.52
	M11	10.12
	M12	10.46
	M13	13.12
	M14	11.24
	M15	12.43
	M16	14.54
As-to-En	M7	15.66
	M11	16.42
	M12	17.32
	M14	17.44
	M15	18.32

Table 14. BLEU scores results of transformer-based NMT on test set-2, M11:M5 with PA1 (BA); M12: M5 with PA1 (UA); M13: M5 with PA1 (RA); M14: M5 with PA2 (BA); M15: M5 with PA2 (UA); M16: M5 with PA2 (RA), where PA1:Prior Alignment (FastAlign), PA2:Prior Alignment (SimAlign), UA: Unidirectional Alignment, BA: Bidirectional Alignment (grow-diagonal heuristics), RA: Reverse Direction Alignment

Translation	Model	BLEU
En-to-As	M7	8.64
	M11	8.86
	M12	8.94
	M13	9.08
	M14	8.98
	M15	9.04
	M16	9.16
As-to-En	M7	11.74
	M11	11.82
	M12	11.96
	M14	12.10
	M15	12.28

translation. In [30], FastAlign tool is used to extract the token alignment information from the parallel data and adopted the guided alignment concept in the transformer-based NMT [10].

In [30, 10], the optimization criteria for training the baseline transformer model [42] is presented in Eq. 10, where T denotes the number of target tokens, p represents the output probability distribution, and $r_{i,j}$ indicates $j-th$ the token in the dictionary is the true value at the $i-th$ position in the target sentence.

The modified optimization criteria is represented in Eq. 11, where a pair of source-target sentences of length K and T, respectively, and a prior alignment set:

$$A \subseteq (j-i)) : j = 1, \ ..., \ k, \ \ i = 1, \ ..., \ T. \qquad (9)$$

It takes randomly the output of just a head from the fifth decoder layer and project it into T target token probability distribution over K corresponding source token.

It compares the probability distributions q_{ij} with the reference probability generated from prior alignments through cross-entropy. The symbol $a_{i,j}$ represents the $i-th$ target token is properly aligned with the $j-th$ source token.

Both $L1$ and $L2$ are combined in Eq. 12 [10, 30] which is the sum of cross-entropy for tokens and alignment weights of source-target sentences:

$$L_1 = -\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{m}(r_{i,j} \times \log(p_{i,j})), \qquad (10)$$

$$L_2 = -\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{K}(a_{i,j} \times \log(q_{i,j})), \qquad (11)$$

$$L = L_1 + \lambda L_2, \qquad (12)$$

where, λ is a weighted cross-entropy for alignments (a hyperparameter), the authors [30] considered 0.05. For comparative analysis, we also considered FastAlign to extract alignment information and, however, we have considered weighted alignment $\lambda = 0.03$ since it yields the lowest training cost.

In this work, we have proposed to use SimAlign[9] [12] tool to extract the token alignment information. The SimAlign is a word alignment tool that uses static and pretrained multilingual language model (mBERT) based contextualized embeddings.

[9] https://github.com/cisnlp/simalign

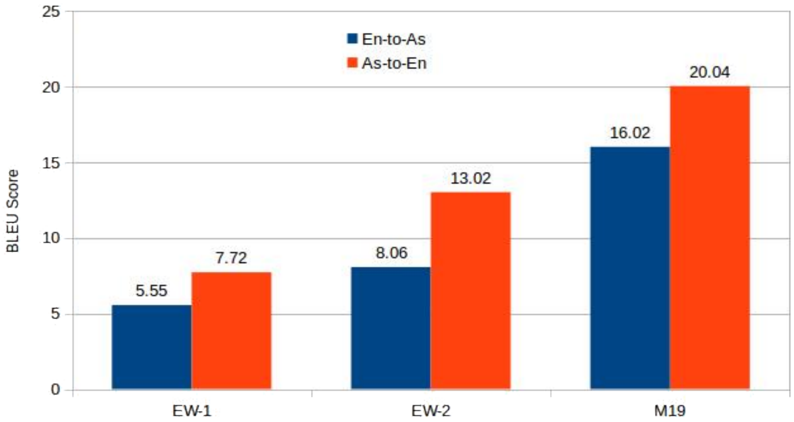

Fig. 2. BLEU score comparison among existing works, EW-1 [23], EW-2 [24] and M19:our best model for En-As pair translation on test set-1

It uses sub-word (BPE) level processing in three various methods: Argmax, Itermax and Match to obtain the alignment information. The basic difference among these three methods is Argmax finds a local optimum and Itermax uses greedy algorithm, whereas Match finds a global optimum via maximum-weight maximal matching technique [12].

Although, we have extracted alignment information using these three methods, reported only Match-based SimAlign, since it shows higher translational performance in the NMT. We have used the two-step process to construct the alignments. First, extract the alignment information from both the forward and backward direction, i.e., En-to-As and As-to-En.

Then, combine the bidirectional alignments using the grow-diagonal heuristics of [17]. For the comparative analysis, we also considered extracted unidirectional (En-to-As or As-to-En) alignment information (as reported in Section 6.2).

It is noticed that the backward direction, i.e., As-to-En translation attains a higher score than that of En-to-As translation. Therefore, we have proposed to use the backward/reverse direction (As-to-En) of alignment information in the forward direction (En-to-As) translation using a simple two-step solution.

First, we reverse the extracted alignment information of the backward direction and then sort them to obtain the alignment information of forward direction.

Table 15. BLEU scores results of transformer-based NMT on test set-1, M17:M16 (En-to-As) / M15 (As-to-En)+ PSP (Post-processing)

Translation	M7	M16/M15	M17
En-to-As	9.52	14.54	15.10
As-to-En	15.66	18.32	19.16

Table 16. BLEU scores results of transformer-based NMT on test set-2, M17:M16 (En-to-As) / M15 (As-to-En)+ PSP (Post-processing)

Translation	M7	M16/M15	M17
En-to-As	8.64	9.16	9.65
As-to-En	11.74	12.28	13.18

The Marian [13] toolkit is employed to uses the source-target prior alignment information in the training process of transformer-based NMT.

Moreover, the pretrained language model (LM) [5] could be used to improve low-resource NMT. We have used the Marian[10] toolkit that allows to use the pretrained language model (LM) in the training process of NMT.

We have utilized the monolingual data of the target language to train and generate an LM using the transformer model, and the weight matrices are loaded from the pretrained LM by initializing the decoder of an encoder-decoder architecture of transformer-based NMT. We have named AsLM for the custom pretrained Assamese LM.

[10] https://marian-nmt.github.io/docs/

doi: 10.13053/CyS-26-4-4423

Table 17. BLEU scores results of transformer-based NMT on test set-1, M18:M16 (En-to-As) / M15 (As-to-En)+PLM (Pretrained LM); M19: M18+ PSP (Post-processing)

Translation	M16/M15	M18	M19
En-to-As	14.54	15.46	16.02
As-to-En	18.32	19.62	20.04

Table 18. BLEU scores results of transformer-based NMT on test set-2, M18:M16 (En-to-As) / M15 (As-to-En)+PLM (Pretrained LM); M19: M18+ PSP (Post-processing)

Translation	M16/M15	M18	M19
En-to-As	9.16	9.42	10.52
As-to-En	12.28	12.56	13.93

5.3 Post-processing

The post-processing step is used to handle out-of-vocabulary issue. It arises due to the named-entities, compounds, technical terms and misspelled words [2]. The OOV is of two types: Completely Out-of-Vocabulary (COOV) and Sense Out-of-Vocabulary (SOOV).

If the words are not present in the training data, then it is known as COOV, on the other hand SOOV are those words which are present in the training data with different usage or sense from the test set words. NMT generates <unk> (unknown) tokens against OOV.

Furthermore, NMT shows weakness in case of rare word translation since fixed-size vocabulary, which forces producing <unk> [27]. The authors [40] introduced byte pair encoding (BPE) to handle the OOV issue. Likewise, we have used BPE and proposed to use a post-processing step.

The post-processing step contains two key components: Bilingual Dictionary and Transliteration Module **Bilingual Dictionary:** We have prepared a bilingual English - Assamese dictionary since there is lack of available dictionary data for En-As pair.

In our previous work [22], we have collected 200,151 a number of En-As parallel sentences from an online dictionary, namely, Glosbe.

Moreover, we have extracted 10, 103, 84 phrase pairs from the train set (as discussed in Section 5.1). We have used both (Glosbe and phrase pairs) to filter out single and double parallel words.

In the prepared dictionary, the total number of parallel single/double words are 464,586, wherein 87,024 from Glosbe and rest are from phrase pairs. We have filtered parallel noun phrases from the phrase pairs using two steps: first, extracted noun phrases from the English side of phrase pairs using NLTK[11] tool and then mapped those sentences in the phrase pairs to collect corresponding Assamese noun phrases.

The bilingual dictionary is used to replace the <unk> tokens with the appropriate target words concerning source words. **Transliteration Module:** We have used this module to source words which are not present in the bilingual dictionary.

It is mainly used to handle the unseen tokens that produce <unk>. We have used indic-trans[12] [6] to convert the source word into the target word script in the predicted sentence for both En-to-As and As-to-En transliteration.

6 Experiment and Result

In this section, we briefly present experimental setup and reported quantitative results with error analysis.

6.1 Experimental Setup

We have employed two setups in the baseline system experiments, namely, phrase-based SMT (PBSMT) and NMT. For PBSMT, the Moses[13] [18] toolkit is used, wherein, GIZA++ [31] and IRSTLM [9] are used to extract phrase pairs to build the translational model and language model, following default settings of Moses.

The NMT experiments are carried out using the publicly available Marian [13] toolkit in three basic operations, data preprocessing, training and testing.

[11] https://www.nltk.org/

[12] https://github.com/libindic/indic-trans

[13] http://www.statmt.org/moses/

Table 19. BLEU scores on different sentence group distribution for Test Set-1, SG: Sentence Group, NM-1: M7, NM-2: M16 (En-to-As) / M15 (As-to-En), NM-3: NM-2 + PLM

SG	Length	No. of Sentences	NM-1	NM-2	NM-3
1	1-15	1344	En-to-As: 15.98 As-to-En: 18.98	En-to-As: 17.72 As-to-En: 23.32	En-to-As: 17.96 As-to-En: 23.96
2	16-30	944	En-to-As: 10.52 As-to-En: 12.16	En-to-As: 11.22 As-to-En: 17.38	En-to-As: 11.36 As-to-En: 17.87
3	31-45	179	En-to-As: 9.32 As-to-En: 11.47	En-to-As: 10.52 As-to-En: 15.26	En-to-As: 11.69 As-to-En: 15.57
4	46-80	33	En-to-As: 4.40 As-to-En: 7.34	En-to-As: 7.48 As-to-En: 9.54	En-to-As: 9.29 As-to-En: 11.38

Table 20. Comparative quantitative results on test set-1 in terms of automatic evaluation scores

Translation	Model	BLEU	TER	RIBES	METEOR	F-measure
En-to-As	M2 (baseline-3)	6.92	93.1	0.2878	0.1043	0.2106
	M19 (best)	16.02	79.4	0.4226	0.2712	0.6346
As-to-En	M2 (baseline-3)	12.84	88.1	0.4284	0.1477	0.2876
	M19 (best)	20.04	74.5	0.4738	0.3846	0.7584

Table 21. Comparative quantitative results on test set-2 in terms of automatic evaluation scores

Translation	Model	BLEU	TER	RIBES	METEOR	F-measure
En-to-As	M2 (baseline-3)	7.66	91.3	0.3032	0.1286	0.2306
	M19 (best)	10.52	88.4	0.4027	0.1406	0.2798
As-to-En	M2 (baseline-3)	10.49	89.5	0.4098	0.1384	0.2796
	M19 (best)	13.93	82.4	0.3826	0.2394	0.2847

In the data preprocessing step, the word-segmentation technique, namely, byte pair encoding (BPE) [40] with $32k$ merge operations is utilized. The vocabulary size of English and Assamese are $32,404$ and, $31,920$ at sub-word level (BPE).

Moreover, we have used GloVe [35] word embeddings as subword level, wherein, the pretraining is performed up to 100 iterations with embedding vector size 200. We have named AsGloVe for custom Assamese word embeddings on Assamese side monolingual data.

For RNN-based NMT, we have investigated RNN and bidirectional RNN in our previous work [23, 24] and it is observed that the bidirectional RNN-based NMT shows better translational accuracy.

Thus, we have considered bidirectional RNN-based NMT in baseline-2, where, 0.3 drop-out in two-layer LSTM-based encoder-decoder architecture is used [26].

The default configuration of six layers with eight attention heads, drop-out of 0.1, and Adam optimizer with a learning rate of 0.001 are used in the training process of NMT and LM.

A single NVIDIA Quadro P2000 GPU is utilized to train the models with early stopping criteria, i.e., the model training is halted if it does not converge on the validation set for more than 10 epochs.

6.2 Result and Error Analysis

We have used automatic evaluation metrics, namely, BLEU, TER, RIBES, METEOR,

Table 22. Human evaluation scores on test set-1, AD: Adequacy, FL: Fluency, OR: Overall Ratings

Translation	Model	AD	FL	OR
En-to-As	M2 (baseline-3)	2.56	3.26	2.91
	M19 (best)	4.92	5.84	5.38
As-to-En	M2 (baseline-3)	2.96	3.76	3.36
	M19 (best)	5.12	6.26	5.69

Table 23. Human evaluation scores on test set-2, AD: Adequacy, FL: Fluency, OR: Overall Ratings

Translation	Model	AD	FL	OR
En-to-As	M2 (baseline-3)	1.36	2.02	1.69
	M19 (best)	2.04	3.12	2.58
As-to-En	M2 (baseline-3)	1.84	2.38	2.11
	M19 (best)	2.12	3.18	2.65

F-measure, and human evaluation scores to evaluate the quantitative results of predicted translations.

Table 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, reports the comparative BLEU score results of exploring the transformer-based NMT in different configurations i.e, with or without domain-specific parallel data, phrase pairs augmentation, synthetic parallel data, prior alignment, pretrained LM and along with the post-processing step on test set-1 and test set-2.

Furthermore, we have reported statistical significance in Table 19, wherein, BLEU scores are evaluated on the test set-1 in four groups of sentence length. It is noticed that translation accuracy decreases as the increase in sentence length (number of words).

The effect of LM is realized in the sentences of group 4 (length: 46 - 80). Table 20 and 21 presents comparative results in terms of different automatic evaluation metrics of our best model (enhanced transformer-based NMT) over the baseline transformer model (baseline-3).

Figure 2 presents comparative results of our best model over the existing works [23, 24] in terms of BLEU scores. All facets of translation accuracy cannot be evaluated using the automatic evaluation measures. Thus, the human evaluation (HE) or manual evaluation metric is taken into account. It consists of two aspects: adequacy and fluency.

The adequacy factor measures how well the predicted translation, which corresponds to the reference sentence, is contextually represented. Whereas, fluency is a different criterion that determines whether the predicted translation is well-formed or not.

The overall rating[14] of HE is calculated by the average score of adequacy and fluency. For example, if a reference sentence is: *“He is coming to the park”* and the predicted sentence is: *“He is a good boy.”* Here, the predicted sentence, inadequate with respect to the reference sentence.

But, the predicted sentence is fluent since it is a well-formed or grammatical well-structured sentence. We have hired three human evaluators who possess linguistic knowledge of both the languages, i.e., English and Assamese, and considered the assessment criteria on a scale of 1-5 on randomly selected 100 sample sentences following [33]. Table 22, 23 report the manual evaluation results of transformer model (baseline) and the best model, wherein, the average scores of three human evaluators are presented.

From the quantitative results, it is observed that our best model (M19) attains higher translation accuracy than the baseline models.

[14] https://nlp.amrita.edu/mtil_cen/\#results

Also, it is observed that As-to-En translation attains higher translational performance than En-to-As.

It is because of the presence of more number of tokens in En side as compared to As (as mentioned in Table 2) and as a result, more number of En tokens are encoded by the encoder and the decoder can produce a better translation for As-to-En translation.

From Table 8, it is observed the NMT performance lowers in M1 (without domain-specific parallel train set) [19], therefore, by contributing domain-specific parallel corpus in this work, NMT translational performance improves for both directions of translation covering various domains.

To closely analyse the effect of domain-specific parallel data, the sample predicted sentences of best model and with Google[15] and Bing[16] translation are discussed using the following notations:

— SS: Source sentence.

— TT: Reference / Target sentence.

— PT1: Predicted sentence by the best model (En-to-As).

— PT2: Predicted sentence by the best model (As-to-En).

— BT: Bing translation.

— GT: Google translation.

1. (a) Example-1 (Agriculture): En-to-As

SS: *More than 50 percent of these bamboos are found across the North East including Assam.*

TT: ইয়াৰে ৫০ শতাংশৰো অধিক বাঁহ অসমকে ধৰি সমগ্ৰ উত্তৰ পূত্ৰ্বাঞ্চলত পোৱা যায় । (*yare 50 shatangshu adhik banh asomake dhari samagra uttar purtwanchalat poua jai*)

PT1: প্ৰায় ৫০টাতকৈ অধিক বাঁহবিলাক ধৰি উত্তৰপূৰ্বাঞ্চলত পোৱা যায় । (*praiy 50tatki adhik banhhbilak dhari uttarapurbanchalat poua jai*)

BT: ইয়াৰে ৫০ শতাংশতকৈও অধিক বাঁহ অসমকে ধৰি উত্তৰ-পূৰ্বাঞ্চলত পোৱা যায় । (*yare 50 shatanshtkio adhik banh asamake dhari uttar-purbanchalat poua jai*)

GT: অসমকে ধৰি সমগ্ৰ উত্তৰ পূৰ্বাঞ্চলত ৫০ শতাংশতকৈ অধিক বাঁহ পোৱা যায় (*asamake dhari samagr uttar purbanchalat 50 shatanshtaki adhik banh poua jai*)

1. (b) Example-1 (Agriculture): As-to-En

SS ইয়াৰে ৫০ শতাংশৰো অধিক বাঁহ অসমকে ধৰি সমগ্ৰ উত্তৰ পূত্ৰ্বাঞ্চলত পোৱা যায় ।(*yare 50 shatangshu adhik banh asomake dhari samagra uttar purtwanchalat poua jai*)

TT: *More than 50 percent of these bamboos are found across the North East including Assam.*

PT2: *More than 50 per cent bamboo available in the state of North East India.*

BT: *More than 50 per cent of these bamboos are found in the entire north-eastern region including Assam.*

GT: *More than 50 per cent of this bamboo is found in the entire North East including Assam.*

Discussion: In the above examples, both directions of predicted translation of PT1 and PT2 are fluent like BT and GT. However, predicted translations are partial adequate unlike BT and GT, since PT2 misses “including Assamese”, and plural form of “bamboo”. Whereas, PT1 misses “অসমকে”

2. (a) Example-2 (Social Media): En-to-As

SS: *The moon hangs in the sky like a huge plate.*

TT: আকাশত প্ৰকাণ্ড থালিখনৰ দৰেই জোনবাইজনী ওলমি আছে। (*akashat prakando thalikhanar darei jonvaijani ulomi aache.*)

PT1: আকাশত ওলমি থকা চন্দ্ৰ। (*akashat ulomi thaka chandra*)

BT: চন্দ্ৰটো এটা ডাঙৰ প্লেটৰ দৰে আকাশত ওলমি আছে। (*chandrato eta dangor plater dore aakasot ulomi ase*)

GT: বিশাল প্লেটৰ দৰে আকাশত ওলমি আছে চন্দ্ৰ। (*vishal plator dore akashat ulomi ase chandra*)

[15] https://translate.google.co.in/ (Accessed on 30.09.2022)

[16] https://www.bing.com/translator(Accessed on 30.09.2022)

2. (b) Example-2 (Social Media): As-to-En

SS: আকাশত প্ৰকাণ্ড থালিখনৰ দৰেই জোনবাইজনী ওলমি আছে। (*akashat prakando thalikhanar darei jonvaijani ulomi aache.*)

TT: *The moon hangs in the sky like a huge plate.*

PT2: *Jonbil is hanging in the sky.*

BT: *The zonbaijani is hanging in the sky like a huge thali.*

GT: *The moon hangs in the sky like a huge plate.*

Discussion: Here, both PT1 and PT2 generate partially adequate translation, but sentences are fluent like BT and GT. Also, unlike GT, BT unable to produce correct word (“zonbaijani”, “thali”) for As-to-En translation.

3. (a) Example-3 (Judiciary): En-to-As

SS: *The respondent asserted that after show cause notice dated 15th June 2001 was replied by the petitioner by letter dated 8th July.*

TT: উত্তৰদাতাই দৃঢ়তাৰে কৈছিল যে 15 জুন 2001 তাৰিখৰ কাৰণ দৰ্শোৱাৰ জাননীৰ পিছত আবেদনকাৰীয়ে 8 জুলাই তাৰিখৰ পত্ৰৰ দ্বাৰা ইয়াৰ উত্তৰ দিছিল ।

PT1: উত্তৰদাতাই স্বীকাৰ কৰিছিল যে 8 জুন 2001 তাৰিখৰ কাৰণ দৰ্শোৱাৰ জাননী জাৰী কৰা হৈছিল ।

BT: উত্তৰদাতাই দৃঢ়তাৰে কৈছিল যে 15 জুন 2001 তাৰিখৰ কাৰণ দৰ্শোৱাৰ জাননীৰ পিছত আবেদনকাৰীয়ে 8 জুলাই তাৰিখৰ পত্ৰৰ দ্বাৰা উত্তৰ দিছিল ।

GT: প্ৰতিবাদীয়ে দৃঢ়তাৰে কয় যে ২০০১ চনৰ ১৫ জুন তাৰিখৰ কাৰণ দেখুৱাৰ পিছত আবেদনকাৰীয়ে ৮ জুলাই তাৰিখৰ পত্ৰযোগে উত্তৰ দিছিল ।

3. (b) Example-3 (Judiciary): As-to-En

SS: উত্তৰদাতাই দৃঢ়তাৰে কৈছিল যে 15 জুন 2001 তাৰিখৰ কাৰণ দৰ্শোৱাৰ জাননীৰ পিছত আবেদনকাৰীয়ে 8 জুলাই তাৰিখৰ পত্ৰৰ দ্বাৰা ইয়াৰ উত্তৰ দিছিল ।

TT: *The respondent asserted that after show cause notice dated 15th June 2001 was replied by the petitioner by letter dated 8th July .*

PT2: *The respondent asserted that after the issuance of the show cause notice dated 15 June 2001 the petitioner submitted its reply by the letter dated 8th July .*

BT: *The respondent asserted that after the show cause notice dated June 15, 2001, the petitioner had replied to it by letter dated July 8 .*

GT: *The respondent asserted that after the show cause notice dated 15 June 2001, the petitioner replied to it by letter dated 8 July .*

Discussion: In the above examples, PT1 and PT2 show weakness in adequacy since both unable to produce correct translation of last sub-phrase “by the petitioner by letter dated 8th July” / পিছত আবেদনকাৰীয়ে 8 জুলাই তাৰিখৰ পত্ৰৰ দ্বাৰা ইয়াৰ উত্তৰ দিছিল unlike BT and GT. However, fluency is fine in all the predicted translations.

4. (a) Example-4 (Government Office): En-to-As

SS: *Official receiver or assignee in insolvency proceedings*

TT: দেউলিয়া প্ৰক্ৰিয়াত অফিচিয়েল ৰিচিভাৰ বা আৱণ্টনকাৰী

PT1: চৰকাৰী অনুসন্ধানৰ কাৰ্যক্ৰমসমূহৰ ৰিচিভাৰ

BT: দেউলিয়া প্ৰক্ৰিয়াত অফিচিয়েল ৰিচিভাৰ বা আৱণ্টনকাৰী

GT: ইনছলভেন্সি প্ৰক্ৰিয়াত অফিচিয়েল ৰিচিভাৰ বা এচাইনী

4. (b) Example-4 (Government Office): As-to-En

SS: দেউলিয়া প্ৰক্ৰিয়াত অফিচিয়েল ৰিচিভাৰ বা আৱণ্টনকাৰী

TT: *Official receiver or assignee in insolvency proceedings*.

PT2: *Official resource in bankruptcy proceeding or the allocation*.

BT: *Official receiver or allottee in insolvency proceedings*.

GT: *The official receiver or allocator in bankruptcy proceedings*.

Discussion: Here, PT1 and PT2 produce inadequate as well as not fluent translations, unlike BT ang GT.

5. (a) Example-5 (Tourism): En-to-As

SS: *Taj Mahal ticket to increase by Rs 200*.

TT: তাজমহলৰ টিকট ২০০ টকা বৃদ্ধি হ'ব

PT1: ২০০ টকা খৰচ বৃদ্ধি কৰাৰ বাবে তাজমহলৰ টিকটৰ

BT: তাজমহলৰ টিকট ২০০ টকা বৃদ্ধি হ'ব

GT: ২০০ টকা বৃদ্ধি হ’ব তাজমহলৰ টিকট

5. (b) Example-5 (Tourism): As-to-En

SS: তাজমহলৰ টিকট ২০০ টকা বৃদ্ধি হ'ব

TT: *Taj Mahal ticket to increase by Rs 200*.

PT2: *200 Rs will increase in Taj Mahal*.

BT: *Taj Mahal tickets to be increased by Rs 200*.

GT: *Tickets for the Taj Mahal will be increased by Rs*.

Discussion: Here, PT2 missed the word “ticket”, that leads to inadequate translation Unlike BT. Whereas, GT unable produce “200” in output. However, PT1 produce correct translation like BT and GT in terms of both adequacy and fluency factors of translation.

6. (a) Example-6 (COVID-19): En-to-As

SS: *The fresh order comes amid concerns in the government about the Covid19 lockdown disrupting the supply chain of essential goods.*

TT: কভিড১৯ লকডাউনে অত্যাৱশ্যকীয় সামগ্ৰীৰ যোগান শৃংখলা ব্যাহত কৰাৰ বিষয়ে চৰকাৰত উদ্বেগৰ মাজতে নতুন নিৰ্দেশটো আহিছে ।

PT1: চৰকাৰে কঠোৰ সামগ্ৰীৰ যোগান ব্যাহত কৰাৰ বাবে চৰকাৰৰ কোভিড১৯ লকডাউনৰ বিষয়ে উদ্বেগ প্ৰকাশ কৰে ।

BT: কভিড১৯ লকডাউনে অত্যাৱশ্যকীয় সামগ্ৰীৰ যোগান শৃংখলা ব্যাহত কৰাৰ বিষয়ে চৰকাৰত উদ্বেগৰ মাজতে নতুন নিৰ্দেশটো আহিছে।

GT: Covid19 লকডাউনে অত্যাৱশ্যকীয় সামগ্ৰীৰ যোগান শৃংখলত ব্যাঘাত জন্মাবলৈ চৰকাৰত উদ্বেগ প্ৰকাশ কৰাৰ সময়তে এই সতেজ নিৰ্দেশ ।

6. (b) Example-6 (COVID-19): As-to-En

SS: কভিড১৯ লকডাউনে অত্যাৱশ্যকীয় সামগ্ৰীৰ যোগান শৃংখলা ব্যাহত কৰাৰ বিষয়ে চৰকাৰত উদ্বেগৰ মাজতে নতুন নিৰ্দেশটো আহিছে ।

TT: *The fresh order comes amid concerns in the government about the Covid19 lockdown disrupting the supply chain of essential goods.*

PT2: *The new orders have come when the Covid19 lockdown avoids an essential commotion.*

BT: *The new order comes amid concerns in the government about the Covid-19 lockdown disrupting the supply chain of essential commodities.*

GT: *The new directive comes amid concerns in the government that the lockdown has disrupted the supply chain of essential commodities.*

Discussion: Both PT1 and PT2 yield fluent translation like BT and GT. But partially adequate translation in PT1 and PT2, unlike BT and GT.

7. (a) Example-7 (Sports): En-to-As

SS: *Indian boxers to start practice for Olympics from June 10.*

TT: ভাৰতীয় বক্সাৰসকলে ১০ জুনৰ পৰা অলিম্পিকৰ বাবে আৰম্ভ কৰিব অনুশীলন

PT1: ভাৰতীয় বক্সাৰসকলে ১০ জুনৰ পৰা অলিম্পিকৰ বাবে আৰম্ভ কৰিব

BT: ভাৰতীয় বক্সাৰসকলে ১০ জুনৰ পৰা অলিম্পিকৰ বাবে অনুশীলন আৰম্ভ কৰিব

GT: ১০ জুনৰ পৰা অলিম্পিকৰ বাবে অনুশীলন আৰম্ভ কৰিব ভাৰতীয় বক্সাৰসকলে

7. (b) Example-7 (Sports): As-to-En

SS: ভাৰতীয় বক্সাৰসকলে ১০ জুনৰ পৰা অলিম্পিকৰ বাবে আৰম্ভ কৰিব অনুশীলন

TT: *Indian boxers to start practice for Olympics from June 10.*

PT2: *Indian boxers should start on Olympics from June 10.*

BT: *Indian boxers to start training for Olympics from June 10.*

GT: *Indian boxers will start training for the Olympics from June.*

Discussion: Both PT1 and PT2 missed the word “practice” or অনুশীলন in output, that lead to partially adequate unlike GT and BT. However, all the sentences are fluent.

8. (a) Example-8 (Literature): En-to-As

SS: *A practice called Mizwah has been prevalent among Jewish people.*

TT: ইহুদি ধৰ্মাৱলম্বী লোকসকলৰ মাজত মিজ্ৱাহ নামৰ এটা প্ৰথা প্ৰচলিত হৈ আহিছে ।

PT1: মিজুলুই ইহুদী লোকসকলৰ মাজত মিলুই প্ৰচলিত কৰিছে।

BT: ইহুদী লোকসকলৰ মাজত মিজৱাহ নামৰ এটা প্ৰথা প্ৰচলিত হৈ আহিছে।

GT: ইহুদী লোকসকলৰ মাজত মিজৱা নামৰ এটা প্ৰথা প্ৰচলিত হৈ আহিছে

8. (b) Example-8 (Literature): As-to-En

SS: ইহুদি ধৰ্মাৱলম্বী লোকসকলৰ মাজত মিজ্ৱাহ নামৰ এটা প্ৰথা প্ৰচলিত হৈ আহিছে ।

TT: *A practice called Mizwah has been prevalent among Jewish people.*

PT2: *A practice of worship is prevalent among Jewish people.*

BT: *There has been a custom called Mizwah among the Jewish people.*

GT: *There is a custom called mizvah among the Jewish people.*

Discussion: Like BT and GT, both PT1 and PT2 generate fluent translation. However, inadequate translation in case of PT1 and PT2 unlike BT and GT.

7 Conclusion and Future Work

In this work, we have contributed domain-wise parallel corpus into previous developed dataset, EnAsCorp1.0 [23], we have improved NMT to cover different domains, such as, Agriculture, Social Media, Judiciary, Government Office, COVID-19, Sports, Tourism, Literature for En-As pair translation.

By data augmentation via phrase pairs in addition to the original parallel corpus, more token alignment information is passed into the training model. Also, utilization of synthetic parallel sentences via pretrain and fine-tune steps, we have handled the data scarcity issues for En-As pair translation. It improves translational performance for both directions of translation.

By injecting prior alignment information with pretrained multilingual contextual embeddings-based alignment technique i.e., SimAlign in the transformer-based NMT attains higher translation accuracy than the FastAlign-based prior alignment information or without alignment information.

Moreover, the backward direction, i.e., As-to-En achieves better translational performance than the forward direction En-to-As. Therefore, we have proposed to use reverse order (As-to-En) alignment information in the forward direction (En-to-As) and it shows enhancement in the forward direction of translation i.e., En-to-As.

With custom pretrained LM, translation accuracy is higher in the long-type sentences (as mentioned in Table 19). However, it is inadequate since contextual meaning is different from the source sentence, but fluency is better in the case of the best model for both directions of translation. The domain-wise parallel data will be increased in future work, and attempt to apply the multilingual transfer learning-based approach for further research.

Acknowledgments

The authors are thankful to the Department of Computer Science and Engineering and Center for Natural Language Processing (CNLP), Artificial Intelligence (AI) Lab at the National Institute of Technology, Silchar for providing the requisite support and infrastructure to execute this work.

References

1. **Abdulmumin, I., Galadanci, B. S., Garba, A. (2019).** Tag-less back-translation. DOI: 10.48550/ARXIV.1912.10514.

2. **Aminian, M., Ghoneim, M., Diab, M. (2014).** Handling OOV words in dialectal Arabic to English machine translation. Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants, Association for Computational Linguistics, pp. 99–108.

3. **Bahdanau, D., Cho, K., Bengio, Y. (2015).** Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations. Conference Track Proceedings, pp. 1–15. DOI: 10.48550/ARXIV.1409.0473.

4. **Barman, A., Sarmah, J., Sarma, S. (2014).** Assamese wordnet based quality enhancement of bilingual machine translation system. Proceedings of the Seventh Global Wordnet Conference, pp. 256–261.

5. **Baziotis, C., Haddow, B., Birch, A. (2020).** Language model prior for low-resource neural machine translation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp. 7622–7634. DOI: 10.48550/ARXIV.2004.14928.

6. **Bhat, I. A., Mujadia, V., Tammewar, A., Bhat, R. A., Shrivastava, M. (2014).** Iiit-h system submission for fire2014 shared task on transliterated search. Proceedings of the Forum for Information Retrieval Evaluation, Association for Computing Machinery, pp. 48—53. DOI: 10.1145/2824864.2824872.

7. **Denkowski, M., Neubig, G. (2017).** Stronger baselines for trustable results in neural machine translation. Proceedings of the First Workshop on Neural Machine Translation, Association for Computational Linguistics, pp. 18–27. DOI: 10.48550/ARXIV.1706.09733.

8. **Dutta, H. (2019).** Assamese Orthography: An Introduction and Some Applications for Literacy Development. Springer International Publishing, pp. 181–194. DOI: 10.1007/978-3-030-05977-4_10.

9. **Federico, M., Bertoldi, N., Cettolo, M. (2008).** IRSTLM: an open source toolkit for handling large scale language models. INTERSPEECH, ISCA, pp. 1618–1621. DOI: 10.21437/interspeech.2008-271.

10. **Garg, S., Peitz, S., Nallasamy, U., Paulik, M. (2019).** Jointly learning to align and translate with transformer models. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, pp. 4452–4461. DOI: 10.18653/v1/D19-1453.

11. **Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H. (2010).** Automatic evaluation of translation quality for distant language pairs. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 944–952.

12. **Jalili Sabet, M., Dufter, P., Yvon, F., Schütze, H. (2020).** SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Association for Computational Linguistics, pp. 1627–1643. DOI: 10.48550/ARXIV.2004.08728.

13. **Junczys Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., et al. (2018).** Marian: Fast neural machine translation in C++. Proceedings of ACL 2018, System Demonstrations, pp. 116–121. DOI: 10.48550/ARXIV.1804.00344.

14. **Kanchan Baruah, K., Das, P., Hannan, A., Sarma, S. K. (2014).** Assamese-English bilingual machine translation. International Journal on Natural Language Computing (IJNLC).

15. **Khenglawt, V., Laskar, S. R., Pakray, P., Manna, R., Khan, A. K. (2022).** Machine translation for low-resource English-Mizo pair encountering tonal words. Computación y Sistemas, Vol. 26, No. 3.

16. **Kocmi, T. (2020).** Exploring benefits of transfer learning in neural machine translation. DOI: 10 .48550/ARXIV.2001.01622.

17. **Koehn, P., Axelrod, A., Birch, A., Callison-Burch, C., Osborne, M., Talbot, D. (2005).** Edinburgh system description for the 2005 IWSLT speech translation evaluation. 2005 International Workshop on Spoken Language Translation, ISCA, pp. 68–75.

18. **Koehn, P., Hoang, H., Birch, A., Burch, C. C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007).** Moses: Open source toolkit for statistical machine translation. ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, The Association for Computational Linguistics, pp. 177—180.

19. **Koehn, P., Knowles, R. (2017).** Six challenges for neural machine translation. Proceedings of the First Workshop on Neural Machine Translation, Association for Computational Linguistics, pp. 28–39. DOI: 10.18653/v1/W17-3204.

20. **Koehn, P., Och, F. J., Marcu, D. (2003).** Statistical phrase-based translation. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 127–133.

21. **Lalrempuii, C., Soni, B., Pakray, P. (2021).** An improved English-to-Mizo neural machine translation. ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 20, No. 4, pp. 1–21. DOI: 10.1145/3445974.

22. **Laskar, S. R., Faiz Ur Rahman Khilji Darsh Kaushik, A., Pakray, P., Bandyopadhyay, S. (2021).** EnKhCorp1.0: An English-Khasi corpus. Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021), Association for Machine Translation in the Americas, pp. 89–95.

23. **Laskar, S. R., Khilji, A. F. U. R., Pakray, P., Bandyopadhyay, S. (2020).** EnAsCorp1.0: English-Assamese corpus. Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, Association for Computational Linguistics, pp. 62–68.

24. **Laskar, S. R., Ur Rahman Khilji, A. F., Pakray, P., Bandyopadhyay, S. (2022).** Improved neural machine translation for low-resource English–Assamese pair. Journal of Intelligent and Fuzzy Systems, Vol. 42, No. 5, pp. 4727–4738.

25. **Lavie, A., Denkowski, M. J. (2009).** The meteor metric for automatic evaluation of machine translation. Machine Translation, Vol. 23, No. 3, pp. 105—115. DOI: 10.1007/s10590-009-9059-4.

26. **Luong, T., Pham, H., Manning, C. D. (2015).** Effective approaches to attention-based neural machine translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 1412–1421.

27. **Luong, T., Sutskever, I., Le, Q., Vinyals, O., Zaremba, W. (2015).** Addressing the rare word problem in neural machine translation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Vol. 1, pp. 11–19. DOI: 10.48550/ARXIV.1410.8206.

28. **Mahanta, S. (2012).** Assamese. Journal of the International Phonetic Association, Vol. 42, No. 2, pp. 217–224. DOI: 10.1017/s0025100 312000096.

29. **Megerdoomian, K., Parvaz, D. (2008).** Low-density language bootstrapping: the case of tajiki Persian. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08),

European Language Resources Association (ELRA), pp. 3293–3298.

30. **Nguyen, T., Nguyen, L., Tran, P., Nguyen, H. (2021).** Improving transformer-based neural machine translation with prior alignments. Complexity, Vol. 2021, pp. 1–10. DOI: 10.1155/2021/5515407.

31. **Och, F. J., Ney, H. (2003).** A systematic comparison of various statistical alignment models. Computational Linguistics, Vol. 29, No. 1, pp. 19–51. DOI: 10.1162/089120103321337421.

32. **Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002).** BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135.

33. **Pathak, A., Pakray, P., Bentham, J. (2018).** English–Mizo machine translation using neural and statistical approaches. Neural Computing and Applications, Vol. 30, pp. 1–17.

34. **Pathak, D., Nandi, S., Sarmah, P. (2022).** Reduplication in Assamese: Identification and modeling. Transactions on Asian and Low-Resource Language Information Processing, Vol. 21, No. 5, pp. 1–18. DOI: 10.1145/3510419.

35. **Pennington, J., Socher, R., Manning, C. D. (2014).** Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP. A meeting of SIGDAT, a Special Interest Group of the ACL, Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/d14-1162.

36. **Probst, K., Brown, R., Carbonell, J., Lavie, A., Levin, L. S., Peterson, E. (2001).** Design and implementation of controlled elicitation for machine translation of low-density languages. pp. 3293–3298.

37. **Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., Sharma, A., Sahoo, S., Diddee, H., Kakwani, D., Kumar, N. (2021).** Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. Transactions of the Association for Computational Linguistics, Vol. 10, pp. 145–162.

38. **Sen, S., Hasanuzzaman, M., Ekbal, A., Bhattacharyya, P., Way, A. (2020).** Neural machine translation of low-resource languages using SMT phrase pair injection. Natural Language Engineering, Vol. 27, No. 3, pp. 271–292. DOI: 10.1017/s1351324920000303.

39. **Sennrich, R., Haddow, B., Birch, A. (2016).** Improving neural machine translation models with monolingual data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 86–96. DOI: 10.48550/ARXIV.1511.06709.

40. **Sennrich, R., Haddow, B., Birch, A. (2016).** Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1715–1725. DOI: 10.48550/ARXIV.1508.07909.

41. **Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006).** A study of translation edit rate with targeted human annotation. Proceedings of Association for Machine Translation in the Americas, pp. 223–231.

42. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017).** Attention is all you need. In Advances in Neural Information Processing Systems. pp. 5998–6008.

Article received on 09/07/2022; accepted on 07/10/2022.
Corresponding author is Partha Pakray.