Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 979–988

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

English-Assamese Multimodal Neural Machine Translation using Transliteration-based Phrase Augmentation Approach

Sahinur Rahman Laskar[a,*], Bishwaraj Paul[a], Partha Pakray[a], Sivaji Bandyopadhyay[a]

[a]Department of Computer Science and Engineering, National Institute of Technology Silchar, 788010 Assam, India

Abstract

Neural machine translation (NMT) is a popular machine translation method due to its contextual analyzing ability and end-to-end process flexibility. However, NMT suffers poor translation quality in low-resource contexts, particularly for diverse language pairs. To overcome this issue, multimodal concept has been introduced in NMT, wherein leverage information from different modalities like image or speech in addition to text to enhance automatic translation quality. In this paper, we have investigated multimodal NMT for a low-resource language diverse pair, English-Assamese, by addressing data scarcity and word-order divergence issues. To tackle such issues, a transliteration-based phrase augmentation approach is proposed, that leverages the sub-word level tokens sharing among source-target sequences in the training process via transliteration and provides more word alignment information by the addition of phrase pairs. Also, the relevant image features corresponding to the phrase pairs are augmented by considering a filtering step. With the proposed approach, state-of-the-art multimodal NMT results are attained for both directions of English-Assamese pair translation.

Keywords: English-Assamese; MNMT; Multimodal; Phrase Pairs; Transliteration

1. Introduction

Deep neural networks (DNN) have achieved state-of-the-art results for a variety of tasks nowadays, including natural language processing (NLP), computer vision, and speech processing. And, also inspiring researchers to develop a system that benefits from the integration of multiple modalities [2]. Machine translation (MT) is a subfield of NLP where one natural language is automatically translated to another natural language. The DNN-based MT, also known as NMT is a well-known approach of MT which shows remark-

[*] Corresponding author. Tel.: +91-7002460508
E-mail address: sahinur_rs@cse.nits.ac.in

10.1016/j.procs.2023.01.078

able translation performance because of its contextual analyzing potential. Moreover, multimodal NMT (MNMT) aims to extract information from more than one modality, like the image is used along with textual data to train an NMT model. Fig. 1 presents an example of English–Assamese (En-As) multimodal translation where En and As are the source and target language. The researchers have been proposed different approaches in MNMT, which include processing of global image features [5], processing the object tags obtained from the images [8] and cross-lingual visual pre-training [3]. In this paper, our goal is to improve MNMT performance for a low-resource En-As language pair [16]. The linguistic characteristic of As is very different from En, which includes morphological richness, word-order of As is subject-object-verb (SOV) unlike En [14]. The contributions are summarized below:

- We have proposed a transliteration-based phrase augmentation approach to enhance the translational performance of MNMT for the En-As language pair. We have tackled the data scarcity and word-order divergence problem by the augmentation of phrase pairs to provide more token alignment information and also augmented relevant image features of phrase pairs by considering a filtering step. Also, investigate the transliteration-based approach that allows sub-word level vocabulary sharing among the source-target sentences in the training process. Our proposed method shows improvement over the prior work of En-As MNMT [16].
- We have explored different MNMT models (BRNN and Transformer) and performed quantitative comparative analysis for both directions of translation, En-to-As, and As-to-En. We have achieved state-of-the-art MNMT results for the low-resource En-As pair.

The rest of the paper is organized as follows: Section 2 briefly presents related works, and Section 3 discusses the fundamental concept of MNMT and describes the proposed approach. The experimental results and analysis are presented in Section 4. Lastly, the paper is concluded in Section 5.

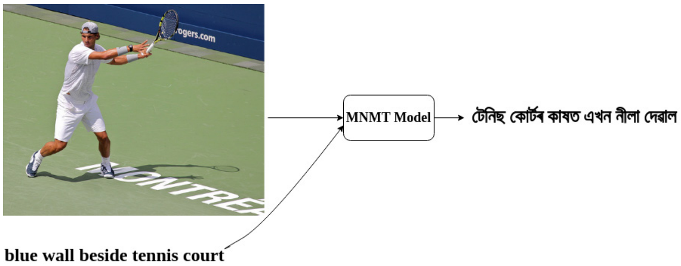
MNMT Model
টৈনিছ কোৰ্টৰ কাষত এখন নীলা দেৱাল
blue wall beside tennis court

Fig. 1. Example of multimodal translation

2. Related Work

In this section, we have focused on earlier MNMT works on Indian languages. In the area of multimodal translation for En–As language pair, the first work was done by [16]. That work developed a multimodal translation model using RNN-based MNMT [4, 5]. Their significant contribution was the development of the multimodal dataset for English–Assamese translation called as Assamese Visual Genome 1.0 consists of 28,927 train images and sentences along with 1400 test and 998 valid set. Their models consisted of RNN and BRNN-based attentive neural networks out of which BRNN performed better with a BLEU score of 23.84 and 22.91 for En-to-As and As-to-En respectively. However, the multimodal translation systems have been explored in other language pairs, namely, English–Hindi [6, 15, 13, 7]. The authors [6] used synthetic data and deployed an RNN-based model of MNMT [4, 5]. The Hindi Visual Genome [21] dataset is used in [13, 7] and attained BLEU scores of 39.28 and 37.50 on the challenge test set. Although they used the same MNMT architecture of [4, 5], [13] utilized a data augmentation approach to improve the translational results. Similar to Hindi Visual Genome, Bengali Visual Genome [24] was proposed, and the same had been used to develop the English-to-Bengali MNMT system [22]. They followed ViTA [9] that utilizes a multilingual seq2seq denoising auto-encoder called as mBART [18], and achieved a BLEU score of 26.8 in English-to-Bengali translation. Recently, MNMT model architecture is improved by introducing a co-attention network, wherein, text and visual context vector is filtered to acquire the text-aware visual context vector [29]. They used Multi30k dataset for investing English-German language pair MNMT task. Moreover, the authors [12] investigate MNMT by incorporating pre-trained language model. This paper investigates MNMT for En-As pair with a proposed transliteration-based phrase augmentation approach to enhance the translational performance for forward (En-to-As) and backward (As-to-En) direction.

3. MNMT

We have explored two different model architectures, namely, RNN [4, 5]and transformer [31] to build En-As MNMT system. Both model architectures consist of encoder-decoder architecture. In RNN-based MNMT, each of the RNNs within encoder, single-layer feed-forward neural networks are used to initialize the hidden states. To initialize the encoder/decoder hidden states of RNN, [4] uses global image features. The visual features are calculated using Eq. (1).

$$d = PM_I^2 . \left(PM_I^1 .q + c_I^1 \right) + c_I^2 \tag{1}$$

Where, c and PM denote the bias and projection matrix, respectively.
Using Eq. (2) and (3), the encoder hidden states are then computed from the vector d.

$$\overrightarrow{s}_{init} = tanh(PM_{fw} d + c_{fw}) \; (\; forward \; rnn) \tag{2}$$

$$\overleftarrow{s}_{init} = tanh(PM_{bw} d + c_{bw}) \; (\; backward \; rnn) \tag{3}$$

Where, the forward and backward directions are represented by the subscript fw and bw of PM.
At the decoder side, a doubly-attentive RNN [5] is adopted. There are three computations that include computation of the hidden states, attention and the final hidden states. A single layer feed-forward network calculates the alignment $e_{t,i}^{ip}$ between each source or input vector s_i and the expected or output token $\hat{w}_t$ at t (time step), using Eq. (4) and (5).

$$e_{t,i}^{ip} = \left(v_a^{ip} \right)^T tanh \left(U_a^{ip} \acute{s}_t + m_a^{ip} s_i \right) \tag{4}$$

$$a_{t,i}^{ip} = \frac{exp\left(e_{t,i}^{ip} \right)}{\sum_{j=1}^{N} exp\left(e_{t,j}^{ip} \right)} \tag{5}$$

Where, $a^{ip}_{t,i}$ represents the alignments between the target token $\hat{w}_t$ and each input vector s_i at t (time step), and the model parameters v^{ip}_a, U^{ip}_a and m^{ip}_a. The weighted sum over the input vectors is calculated via a context vector v_t, wherein, each vector is represented by the attention weight $a^{ip}_{t,i}$; $v_t = \sum_{i=1}^{N} a^{src}_{t,i} s_i$

In transformer-based MNMT [31], multimodal self-attention is incorporated in transformer model architecture [30]. The authors [31] considered two modalities, namely, *text* and *image*, denoted by $X^{text} \in R^{n\times d}$ and $X^{img}W^{img} \in R^{p\times d}$. The multimodal self-attention induces hidden representations of image from text with the help of image-aware attention and computed using Eq. 6.

$$C_i = \sum_{j=1}^{m} A_{i,j} \left(X^{text}_j W^V \right) \tag{6}$$

$$S = \text{softmax}\left(\frac{(x_i W^Q)(X^{text}_j W^K)^T}{\sqrt{d}} \right) \tag{7}$$

Where, $X^{text}_j \in R^{n\times d}$ is a text entry modality and $A_{i,j}$ is a weight coefficient calculated via a softmax function Eq. 7, $C \in R^{(n+p)\times d}$ denote hidden representation of words and the image, W^V, W^Q, W^K are parameter matrices of layer-specific.

We have proposed a transliteration-based phrase augmentation approach and employed RNN-based MNMT [4, 5]and transformer-based MNMT [31] without modifying model architectures. The following steps are considered:

- Phrase Pairs Extraction: We have extracted phrase pairs from the original parallel corpus following [25, 13]. In [25, 13] uses SMT[1], wherein, Giza++ word alignment tool is utilized for phrase pairs extraction and then, directly appended with the original parallel corpus to enhance the performance of a low-resource pair translation. Here, the objective is to yield more token alignment information into the training model and to increase training amount of data. Like [13], the blank lines, duplicates are removed from the extracted phrase pairs and then performed augmentation with the original parallel corpus. The details of the dataset is reported in Table 1.
- Transliteration: Then, we have performed transliteration on source language text to transliterate into the target language script (TSL). For example, As-to-En translation. Here, the source language As is transliterated into the En script. Similarly, the source language, En for En-to-As, is transliterated into Assamese script. The indic-trans[2] [1] is utilized for both directions of transliteration. Then, applied jointly learn byte pair encoding (sub-word level with 20k merge operations) [26] for the source-target sentences in the data preprocessing step. The vocabulary size of En-As is, 7156 (En) and 8813 (As). The motive of the transliteration method is to allow sub-word-level lexical sharing between source and target sentences while also providing a reduced sub-word vocabulary to be shared during the training phase.
- Pre-trained Embeddings: Using GloVe [23] embeddings technique, we have pre-trained on monolingual data of target and transliterated-source languages. In the training process, we have used the obtained GloVe vectors at sub-word-level.
- Image Feature Extraction: Like [16, 15], a pretrained VGG-19 CNN model [27] is utilized to extract the local and global features. The authors of [16, 15], did not consider co-ordinate information of the

[1] http://www.statmt.org/moses/
[2] https://github.com/libindic/indic-trans

images. In this work, we have considered the co-ordinate or bounded box region information (X, Y, width, height) of the images which is available in the Hindi Visual Genome 1.1 [20]. Moreover, we have augmented image features of extracted phrase pairs. To select relevant images of the corresponding phrase pairs, each phrase is searched in the original parallel corpus. And, if it is available, the accompanying image and its co-ordinate information are taken into account. However, there is a difficulty when the same phrase subset appears in many sentences. This problem is addressed by using a filtering step solution.

- First, we identified the matching English segments from the corpus that contain the English phrase of each extracted En-As phrase pair as a sub-string for each phrase pair that was taken from the corpus (filter-1).
- The phrase is skipped as it is invalid if the length of the generated data-frame, or the number of matching English segments for the English part of the sentence, is 0. When length is 1, just one English segment matches it, thus it is directly chosen.
- The resulting English segments are then filtered again (filter-2) to see if the corresponding Assamese phrase of the phrase pairs also has subset in the Assamese segments, if length is greater than 1, i.e. more than 1 English segments have the English phrase as sub-string.
  - The final segment from matching English segments is chosen at random from the filter-1 data frame if the result of filter-2 is 0, meaning there are no Assamese segments that match in the Assamese phrase as a sub-string.
  - If Assamese segment matching is 1, then that single segment is chosen.
  - If there are more than one Assamese phrase matches, a matching segment is chosen at random using a seed value.

The pictorial view of the English–Assamese MNMT system using transliteration-based phrase augmentation approach is depicted in Fig. 2.

4. Experimental Results and Analysis

In this work, the En-As multimodal "Assamese Visual Genome 1.0 (AVG)" [16] dataset is used in experiments and the data statistics are presented in Table 1.

Table 1. Data statistics of AVG [16] and augmented phrase pairs

Type	Name	Items/Instances	Tokens (En / As)
Train	Text (En–As) Image	28,927 28,927	143,164 / 145,448
Train (Phrase Pairs)	Text (En–As) Image	73,299 73,299	223,553 / 259,756
Validation	Text (En–As) Image	998 998	4,922 / 4,978
Test	Text (En–As) Image	1,400 1,400	8,186 / 8,639

The RNN-based MNMT[3] [4, 5] is build by using the fork of the OpenNMT-py [11] and the transformer-based MNMT[4] is build following default settings of [31]. The MNMT model is trained independently for each direction of translation. For RNN-based MNMT, we have used 2-layer, 16 batch size, 0.3 drop-outs, Adam optimizer and 0.001 learning rate. Similar hyperparameters are used in the training process of transformer-based MNMT, with a drop-outs of 0.1. A single GPU (NVIDIA Quadro P2000) is used to train the model

[3] https://github.com/iacercalixto/MultimodalNMT

[4] https://github.com/QAQ-v/MMT

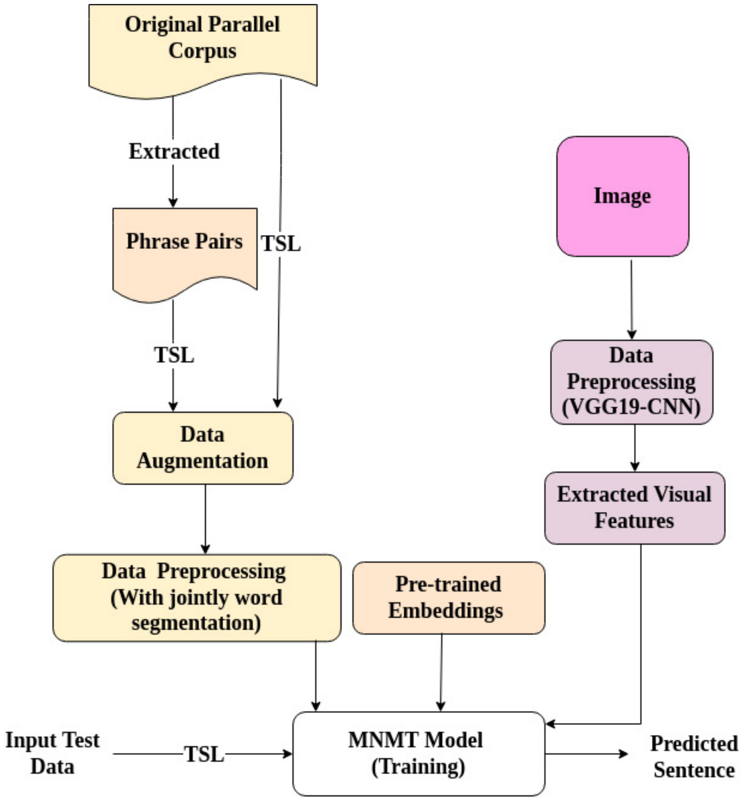

Fig. 2. En–As MNMT, TSL: Transliterated Source Language

for 40 epochs. During the testing process, the best model acquired from the training phase is utilized on the test data. The source language sentences of test data are transliterated (TSL) and then applied to the trained model to generate the predicted target sentences.

The generated target sentences are evaluated in terms of standard metrics, BLEU (bilingual evaluation under study) [19], TER (translation error rate) [28], RIBES (rank-based intuitive bilingual evaluation score) [10], METEOR (metric for evaluation of translation with explicit ordering) [17] and F-measure scores. More the values of score in all the evaluation metric indicate higher prediction accuracy, except TER. The quantitative results of automatic evaluation scores are reported in Table 2. From Table 2, it is observed that both RNN and transformer-based MNMT model with our approach, i.e., transliteration-based phrase augmentation approach attain higher scores than baseline MNMT model [16]. Our best MNMT model, transformer (transliteration-based phrase-augmentation) attain (+0.58, +1.86 (BLEU)), (+2.52, +2.49 (TER)), (+0.037989, +0.025294 (RIBES)), (+0.038054, +0.042707(METEOR)), (+0.041095, +0.039502 (F-measure)) increment in comparison to baseline MNMT [16] for forward (En-to-As) and backward (As-to-En) translation. The quantitative comparative results, in terms of BLEU scores, are presented in Fig. 3.

In Fig. 4 of forward translation, the MNMT with our best model predicts the correct sentence corresponding to the reference sentence unlike MNMT baseline model, which unable to predict “তাৰিখ" and “সময়" in translation. For, As-to-En translation in Fig. 4, the MNMT with our best model predicted sentence that contains “image” which represent the same contextual meaning of “photo”. On the other hand, MNMT baseline model predicts completely wrong translation. In Fig. 5, both the baseline and our best MNMT predict partially correct word “সময়" instead of “সময়ৰ" for En-to-As translation. Here, it misses suffix letter “ৰ" . While in case of As-to-En translation, the MNMT with our model predicts correctly “stamp” unlike baseline MNMT model. But both MNMT models predict wrong verb word “has” instead of “is” that result in wrong translation.

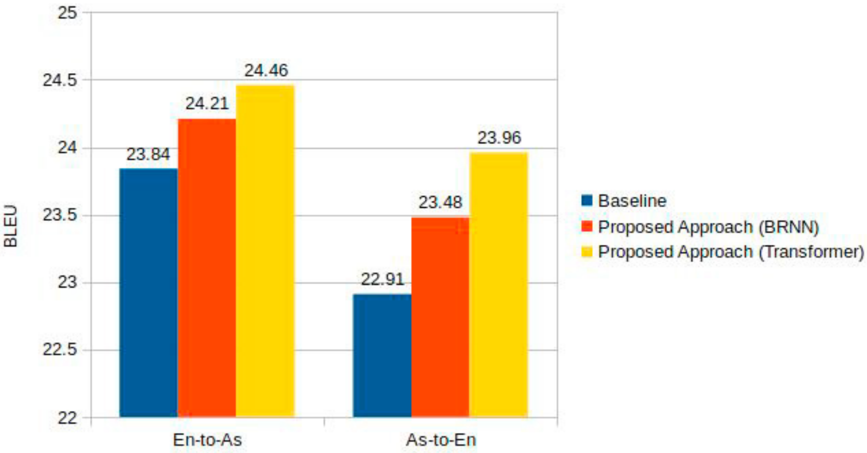

Fig. 3. BLEU score comparison among baseline (existing work) [16], proposed approach in BRNN and Transformer-based MNMT models (best model)

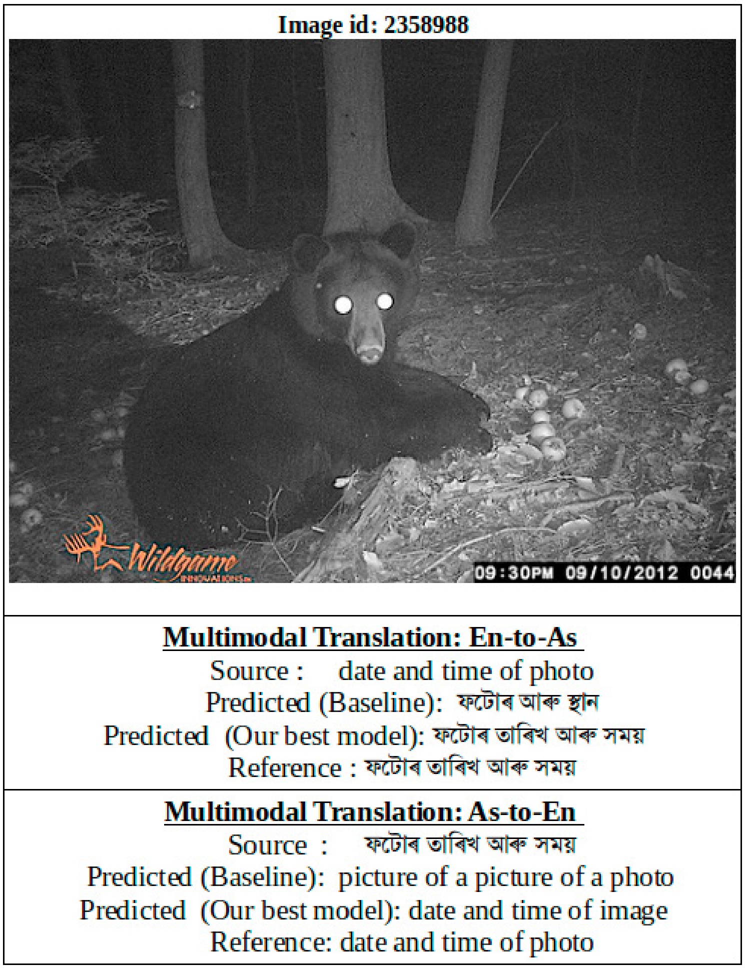

Image id: 2358988

Multimodal Translation: En-to-As

Source : date and time of photo

Predicted (Baseline): ফটোৰ আৰু স্থান

Predicted (Our best model): ফটোৰ তাৰিখ আৰু সময়

Reference : ফটোৰ তাৰিখ আৰু সময়

Multimodal Translation: As-to-En

Source : ফটোৰ তাৰিখ আৰু সময়

Predicted (Baseline): picture of a picture of a photo

Predicted (Our best model): date and time of image

Reference: date and time of photo

Fig. 4. Example-1: predicted output of En-As MNMT

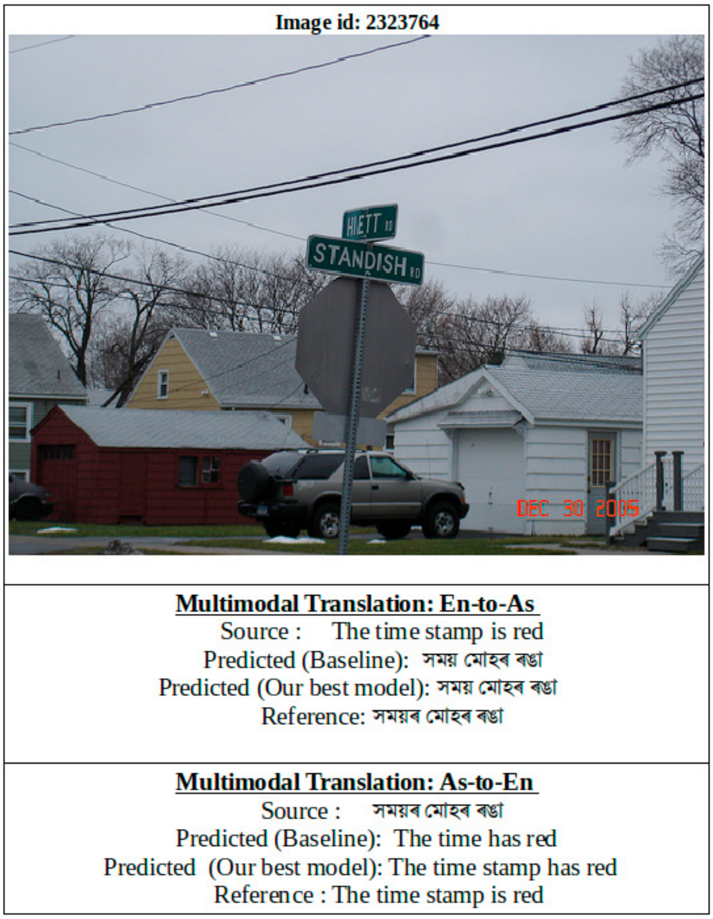

Image id: 2323764

Multimodal Translation: En-to-As
Source : The time stamp is red
Predicted (Baseline): সময় মোহৰ ৰঙা
Predicted (Our best model): সময় মোহৰ ৰঙা
Reference: সময়ৰ মোহৰ ৰঙা

Multimodal Translation: As-to-En
Source : সময়ৰ মোহৰ ৰঙা
Predicted (Baseline): The time has red
Predicted (Our best model): The time stamp has red
Reference : The time stamp is red

Fig. 5. Example-2: predicted output of En-As MNMT

Table 2. Automatic evaluation scores of En-As MNMT system

Translation	MNMT Model	BLEU	TER	RIBES	METEOR	F-measure
En-to-As	BRNN (Baseline [16])	23.84	54.80	0.585941	0.366624	0.720262
	BRNN (Transliteration)	24.14	53.67	0.606854	0.385867	0.746534
	BRNN (Phrase-augmentation)	23.96	54.24	0.588764	0.368764	0.726786
	BRNN (Transliteration-based Phrase-augmentation)	24.21	52.84	0.621784	0.401256	0.760463
	Transformer (Transliteration)	24.32	52.68	0.622567	0.402686	0.760894
	Transformer (Phrase-augmentation)	24.04	52.76	0.621674	0.400532	0.760248
	Transformer (Transliteration-based Phrase-augmentation)	24.46	52.28	0.623930	0.404678	0.761357
As-to-En	BRNN (Baseline [16])	22.91	55.73	0.678821	0.281901	0.595173
	BRNN (Transliteration)	23.26	54.78	0.694856	0.312864	0.626389
	BRNN (Phrase-augmentation)	23.18	54.94	0.692648	0.310753	0.624862
	BRNN (Transliteration-based Phrase-augmentation)	23.48	54.46	0.698764	0.322783	0.631849
	Transformer (Transliteration)	23.78	53.94	0.701278	0.322874	0.632849
	Transformer (Phrase-augmentation)	23.65	53.78	0.703675	0.323784	0.633735
	Transformer (Transliteration-based Phrase-augmentation)	23.96	53.24	0.704115	0.324608	0.634675

5. Conclusion and Future Work

In this work, we have investigated MNMT for En-As pair translation and the proposed transliteration-based phrase augmentation approach attains improvement in the translational performance over baseline MNMT [16] for both forward and backward directions of translation. The proposed approach allows the model to share sub-word-level tokens information through transliteration of the source language into the target language script and gets more alignment information via phrase pairs augmentation in order to tackle the word-order divergence issue. By the augmentation of phrase segment image features in addition to phrase pairs bi-text data, we have expanded the train data by handling data scarcity problems for the improvement of the low-resource MNMT task in terms of quantitative results of standard evaluation metrics. The multimodal dataset size will be increased in the future work and incorporate multilingual transfer learning approach in MNMT model for further research.

Acknowledgements

Authors would like to thank Artificial Intelligence (AI) Lab, Center for Natural Language Processing (CNLP) and the Department of Computer Science and Engineering, National Institute of Technology, Silchar for providing the infrastructure to execute this work.

References

[1] Bhat, I.A., Mujadia, V., Tammewar, A., Bhat, R.A., Shrivastava, M., 2014. Iiit-h system submission for fire2014 shared task on transliterated search, in: Proceedings of the Forum for Information Retrieval Evaluation, Association for Computing Machinery, New York, NY, USA. p. 48–53.

[2] Caglayan, O., Barrault, L., Bougares, F., 2016. Multimodal attention for neural machine translation. CoRR abs/1609.03976.

[3] Caglayan, O., Kuyu, M., Amac, M.S., Madhyastha, P., Erdem, E., Erdem, A., Specia, L., 2021. Cross-lingual visual pre-training for multimodal machine translation, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, Association for Computational Linguistics. pp. 1317–1324.

[4] Calixto, I., Liu, Q., 2017. Incorporating global visual features into attention-based neural machine translation., in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark. pp. 992–1003.

[5] Calixto, I., Liu, Q., Campbell, N., 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada. pp. 1913–1924.

[6] Dutta Chowdhury, K., Hasanuzzaman, M., Liu, Q., 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data, in: Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, Association for Computational Linguistics, Melbourne. pp. 33–42.

[7] Gain, B., Bandyopadhyay, D., Ekbal, A., 2021. IITP at WAT 2021: System description for English-Hindi multimodal translation task, in: Proceedings of the 8th Workshop on Asian Translation (WAT2021), Association for Computational Linguistics, Online. pp. 161–165.

[8] Gupta, K., Gautam, D., Mamidi, R., 2021a. Vita: Visual-linguistic translation by aligning object tags, in: Proceedings of the 8th Workshop on Asian Translation, WAT@ACL/IJCNLP 2021, Online, August 5-6, 2021, Association for Computational Linguistics. pp. 166–173.

[9] Gupta, K., Gautam, D., Mamidi, R., 2021b. ViTA: Visual-linguistic translation by aligning object tags, in: Proceedings of the 8th Workshop on Asian Translation (WAT2021), Association for Computational Linguistics, Online. pp. 166–173.

[10] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H., 2010. Automatic evaluation of translation quality for distant language pairs, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Cambridge, MA. pp. 944–952.

[11] Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A., 2017. Opennmt: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations, Association for Computational Linguistics, Vancouver, Canada. pp. 67–72.

[12] Kong, Y., Fan, K., 2021. Probing multi-modal machine translation with pre-trained language model, in: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, Association for Computational Linguistics. pp. 3689–3699.

[13] Laskar, S.R., Khilji, A.F.U.R., Kaushik, D., Pakray, P., Bandyopadhyay, S., 2021a. Improved English to Hindi multimodal neural machine translation, in: Proceedings of the 8th Workshop on Asian Translation (WAT2021), Association for Computational Linguistics, Online. pp. 155–160.

[14] Laskar, S.R., Khilji, A.F.U.R., Pakray, P., Bandyopadhyay, S., 2020a. EnAsCorp1.0: English-Assamese corpus, in: Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, Association for Computational Linguistics, Suzhou, China. pp. 62–68.

[15] Laskar, S.R., Khilji, A.F.U.R., Pakray, P., Bandyopadhyay, S., 2020b. Multimodal neural machine translation for English to Hindi, in: Proceedings of the 7th Workshop on Asian Translation, Association for Computational Linguistics, Suzhou, China. pp. 109–113.

[16] Laskar, S.R., Paul, B., Paudwal, S., Gautam, P., Biswas, N., Pakray, P., 2021b. Multimodal neural machine translation for english-assamese pair, in: 2021 International Conference on Computational Performance Evaluation (ComPE), pp. 387–392.

[17] Lavie, A., Denkowski, M.J., 2009. The meteor metric for automatic evaluation of machine translation. Machine Translation 23, 105–115.

[18] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L., 2020. Multilingual denoising pre-training for neural machine translation. Trans. Assoc. Comput. Linguistics 8, 726–742.

[19] Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 311–318.

[20] Parida, S., Bojar, O., 2020. Hindi visual genome 1.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[21] Parida, S., Bojar, O., Dash, S.R., 2019. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. Computación y Sistemas 23.

[22] Parida, S., Panda, S., Biswal, S.P., Kotwal, K., Sen, A., Dash, S.R., Motlicek, P., 2021. Multimodal neural machine translation system for english to bengali, in: Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021), pp. 31–39.

[23] Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, Doha, Qatar. pp. 1532–1543.

[24] Sen, A., Parida, S., Kotwal, K., Panda, S., Bojar, O., Dash, S.R., 2022. Bengali visual genome: A multimodal dataset for machine translation and image captioning, in: Satapathy, S.C., Peer, P., Tang, J., Bhateja, V., Ghosh, A. (Eds.), Intelligent Data Engineering and Analytics, Springer Nature Singapore, Singapore. pp. 63–70.

[25] 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 , 1–22.

[26] Sennrich, R., Haddow, B., Birch, A., 2016. Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany. pp. 1715–1725.

[27] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[28] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2006. A study of translation edit rate with targeted human annotation, in: In Proceedings of Association for Machine Translation in the Americas, pp. 223–231.

[29] Su, J., Chen, J., Jiang, H., Zhou, C., Lin, H., Ge, Y., Wu, Q., Lai, Y., 2021. Multi-modal neural machine translation with deep semantic interactions. Inf. Sci. 554, 47–60.

[30] 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 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008.

[31] Yao, S., Wan, X., 2020. Multimodal transformer for multimodal machine translation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 4346–4350.