

English to Bengali Multimodal Neural Machine Translation using Transliteration-based Phrase Pairs Augmentation

Sahinur Rahman Laskar¹, Pankaj Dadure²,
Riyanka Manna³, Partha Pakray¹, Sivaji Bandyopadhyay¹

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

²School of Computer Science, University of Petroleum & Energy Studies, Dehradun, India

³Department of Computer Science and Engineering, Adamas University, Kolkata, India

{sahinurlaskar.nits, krdadure, riyankamanna16}@gmail.com

{parthapakray, sivaji.cse.ju}@gmail.com

Abstract

Automatic translation of one natural language to another is a popular task of natural language processing. Although the deep learning-based technique known as neural machine translation (NMT) is a widely accepted machine translation approach, it needs an adequate amount of training data, which is a challenging issue for low-resource pair translation. Moreover, the multimodal concept utilizes text and visual features to improve low-resource pair translation. WAT2022 (Workshop on Asian Translation 2022) organizes (hosted by the COLING 2022) English to Bengali multimodal translation task where we have participated as a team named CNLP-NITS-PP in two tracks: 1) text-only and 2) multimodal translation. Herein, we have proposed a transliteration-based phrase pairs augmentation approach which shows improvement in the multimodal translation task and achieved benchmark results on Bengali Visual Genome 1.0 dataset. We have attained the best results on the challenge and evaluation test set for English to Bengali multimodal translation with BLEU scores of 28.70, 43.90 and RIBES scores of 0.688931, 0.780669, respectively.

1 Introduction

In recent years, multimodal approaches have shown remarkable contributions in various NLP applications such as machine translation, caption generation, etc. Especially in machine translation, multiple input modalities, like text, image, or audio/speech, integrate with NMT, known as multimodal NMT (MNMT), attempts to improve low-resource pair translation by merging visual features in addition to textual features (Shah et al., 2016). The attention-based encoder-decoder architecture of NMT handles various issues of long-term dependency and variable-length phrases via sequence-to-sequence learning and attains a state-of-the-art technique of machine translation (MT) (Bahdanau

et al., 2015; Luong et al., 2015). Also, NMT shows remarkable performance for low-resource Indian languages (Pathak and Pakray, 2018; Pathak et al., 2018; Laskar et al., 2019a,b, 2020a, 2021b, 2022b). Further, to handle the data scarcity problem, the authors (Sen et al., 2020) augmenting phrase pairs and the source language transliteration-based (Laskar et al., 2022a) approach to enhance text-only based for low-resource pair translation. This paper aims to investigate English to Bengali multimodal translation task in WAT2022 with a proposed transliteration-based phrase pairs augmentation approach.

The rest of the paper is organized as follows: Section 2 presents the review of related works. The system description is briefly discussed in Section 3. Section 4 reports the results and Section 5 concludes the paper with future scope.

2 Related Work

In the literature survey, there is minimal existing work, particularly on the English to Bengali multimodal translation task (Parida et al., 2021). In (Parida et al., 2021), they used Bengali Visual Genome 1.0 (Sen et al., 2022b) adopted ViTA (Gupta et al., 2021) approach where they extracted object tags from the image and utilized mBART model (Liu et al., 2020) for encoding English sentences with the object tags and decoding to generate the Bengali translation. The obtained BLEU scores were 43.5 and 26.8 on the evaluation and challenge test sets, respectively. Moreover, the related existing works are available on English to Hindi multimodal translation task (Dutta Chowdhury et al., 2018; Sanayai Meetei et al., 2019; Laskar et al., 2019c, 2020b, 2021a). The authors (Laskar et al., 2020b, 2021a) used Hindi Visual Genome 1.1 and adopts RNN-based MNMT model (Calixto and Liu, 2017; Calixto et al., 2017) with advantages pre-trained word embeddings on monolingual corpus, achieved BLEU scores of 39.28, 39.46 on

challenge and evaluation test set respectively. This work investigates transliteration-based phrase pairs augmentation to improve the multimodal translation task of English to Bengali.

3 System Description

We have carried out four operations: transliteration-based phrase pairs augmentation, data preprocessing, model training, and testing. The OpenNMT-py (Klein et al., 2017) tool is utilized to build multimodal and text-only models separately.

3.1 Dataset Description

The dataset namely, Bengali Visual Genome 1.0¹ (Sen et al., 2022b,a) is used in this task, which is provided by WAT2022 organizer (Nakazawa et al., 2022). In this dataset, the duplicates (text and image) are present in the train set, which have image ID numbers 2328549, 2391240, and 2385507. Therefore, we have removed those duplicates, and thus train set contains 28,927 images and the same number of corresponding English-Bengali parallel sentences. The validation and test (evaluation and challenge) set contains 998, 1,595, and 1,400 images and parallel text data.

3.2 Transliteration-based Phrase Pairs Augmentation

In this phase, firstly, we have expanded the training amount of data via augmentation of phrase pairs to the train set. To improve low-resource pair translation, (Sen et al., 2020) utilized SMT-based phrase pairs to increase training data via augmentation strategy. We have also followed same (Sen et al., 2020) and utilized Giza++ (Och and Ney, 2003) to extract phrase pairs (Laskar et al., 2021a) from the English-Bengali parallel train set. Before augmentation to the parallel train set, duplicates and blank lines are removed. The statistics of extracted phrase pairs are shown in Table 1.

Secondly, English source sentences are transliterated using indic-trans² (Bhat et al., 2014) in to Bengali script following (Laskar et al., 2022a). The goal of the transliteration approach is to allow subword-level lexical sharing between source and target sentences that will be shared during the training process.

¹<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3722>

²<https://github.com/libindic/indic-trans>

3.3 Data Preprocessing, Model Training, and Testing

The image/visual features are independently extracted from the image data using pre-trained CNN-VGG19³ for train, validation, and test data. During feature extraction, the coordinate or bounded box region information (X, Y, width, height) of the images is considered, which is available in the Bengali Visual Genome 1.0 (Sen et al., 2022b). Moreover, we have augmented image features of extracted phrase pairs. To select relevant images of the corresponding phrase pairs, we have searched each phrase in the original parallel corpus (train set), if it is found, then the corresponding image and its coordinate information are considered. But there is a problem if multiple sentences contain the same phrase subset. To tackle this issue, a filtering step solution is considered.

- First, for every phrase pair extracted from the corpus, we found the matching English segments from the corpus which have the English phrase of the En-Bn phrase pair as a sub-string (filter-1).
- If the length of the resulting data frame, i.e., the number of matching English segments for the English part of the phrase is 0, then the phrase is skipped and considered invalid. If the length is 1, since only one English segment matches it, that segment is directly selected.
- On the other hand, if the length is more than 1, i.e., more than 1 English segments have the English phrase as a sub-string, the resulting English segments are again filtered (filter-2) to check if the corresponding Bengali phrase of the phrase pairs also has subset in the Bengali segments.
 - If after filter-2, the result is 0, i.e., there are no matching Bengali segments that have the Bengali phrase as a sub-string, then from the filter-1 data-frame, i.e., the final segment from matching English segments is randomly selected.
 - If the number of matches after Bengali segment matching is 1, then that single segment is selected.

³<https://github.com/iacercalixto/MultimodalNMT>

Number of Phrase Pairs	Tokens	
	En	Bn
127,897	442,657	364,644

Table 1: Statistics of extracted phrase pairs.

- If the number of Bengali phrase matches is more than 1, then a matching segment is randomly selected with a seed value.

For tokenization and preprocessing of text data, the OpenNMT-py toolkit is utilized. We have trained separately for multimodal and text-only NMT using the OpenNMT-py toolkit. During multimodal NMT training, the bidirectional RNN (BRNN) at the encoder and doubly-attentive RNN at the decoder are used by following default settings of (Calixto and Liu, 2017; Calixto et al., 2017). We have trained on a single GPU with early stopping criteria i.e., the model training is halted if does not converge on the validation set for more than 10 epochs. We have used a batch size of 32 during the training process. The optimum trained models of multimodal and text-only NMT are applied to the evaluation and challenge test set. The primary difference in the testing phase is that multimodal NMT uses visual features of image test data. The source English sentences of test data are transliterated and then applied to the trained model to generate the predicted target Bengali sentences.

4 Result and Analysis

The WAT2022 shared task organizer (Nakazawa et al., 2022) published the evaluation result⁴ of the multimodal translation task for English to Bengali, where our team achieves the first position in multimodal submission for both challenge and evaluation test set. Herein, we have participated with a team named CNLP-NITS-PP in the multimodal and text-only submission tracks, where a total of three teams participated. The automatic evaluation metrics, BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) are used for evaluation of results. Table 2 presents the results of our system. The quantitative results show that the multimodal NMT outperforms text-only NMT due to the use of visual and textual features. Furthermore, we have attained benchmark results on the evaluation and challenge test set, which is higher compared

⁴<http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html>

to (Parida et al., 2021). It shows +0.40 and +1.9 increment in terms of BLEU score, which realized that our approach i.e., transliteration-based phrase pairs augmentation improves the translational performance of multimodal NMT. Moreover, Figure 1 and 2 present best and worst outputs along with transliteration of Bengali words and Google translation. In Figure 1, the predicted sentences for both multimodal and text-only represent the same contextual meaning. Here, the only difference is that *prachir* (“wall”) word in the case of the multimodal predicted sentence whereas *dewal* word in the case of the text-only predicted sentence and Google translation. These two words represent the same meaning corresponding to the reference sentence. However, both multimodal and text-only predicted wrong translations.

5 Conclusion and Future Work

In this work, we have proposed a transliteration-based phrase pairs augmentation approach which has been introduced in the WAT2022 multimodal translation task of English to Bengali. The multimodal NMT attains a higher score than the text-only NMT model and other existing works. Furthermore, the designed multilingual-based approach will be investigated to improve the translational performance of low-resource multimodal NMT.

Acknowledgements

We want to thank the Department of Computer Science and Engineering, 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

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. [Neural machine translation by jointly learning to align and translate](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Our System	Test Set	BLEU	RIBES
Text-only NMT	Challenge	26.70	0.680655
	Evaluation	40.90	0.752543
Multi-modal NMT	Challenge	28.70	0.688931
	Evaluation	43.90	0.780669

Table 2: Our system’s results (official) on English to Bengali multimodal translation task.

Image id: 2373836	
	
Multi-modal Translation Track Source Language: English Target Language: Bengali	
Source	a blue wall beside tennis court
Predicted	টেনিস কোর্টের পাশে একটি নীল প্রাচীর (tenis courter pashey ekti nil prachir)
Reference	টেনিস কোর্টের পাশে একটি নীল প্রাচীর (tenis courter pashey ekti nil prachir)
Google Translation	টেনিস কোর্টের পাশে একটি নীল দেয়াল (tenis courter pashey ekti nil dewal)
Text-only Translation Track	
Predicted:	টেনিস কোর্ট পাশে একটি নীল দেয়াল (tenis courter pashey ekti nil dewal)

Figure 1: Sample output of best predicted on challenge test data.


Image id: 2417756	
	
Multi-modal Translation Track Source Language: English Target Language: Bengali	
Source	March 7th is the date on the calendar
Predicted	টটিকা বৃড়ি টেরনে তারিখ রয়েছে (tatka juri traine tarikh royeche)
Reference	৭ই মার্চ ক্যালেন্ডারে তারিখ (7e march kelendare tarikh)
Google Translation	7 ই মার্চ ক্যালেন্ডারে তারিখ (7e march kelendare tarikh)
Text-only Translation Track	
Predicted:	টেমস টেবিলটি স্নানের তারিখ তারিখ (tems tableti snaner tarikh tarikh)

Figure 2: Sample output of worst predicted on challenge test data.

- Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tamemwar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2014. Iit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation, FIRE '14*, page 48–53, New York, NY, USA. Association for Computing Machinery.
- Iacer Calixto and Qun Liu. 2017. **Incorporating global visual features into attention-based neural machine translation**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.
- Iacer Calixto, Qun Liu, and Nick Campbell. 2017. **Doubly-attentive decoder for multi-modal neural machine translation**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1913–1924. Association for Computational Linguistics.
- Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. **Multimodal neural machine translation for low-resource language pairs using synthetic data**. In " ", pages 33–42.
- Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. ViTA: Visual-linguistic translation by aligning object tags. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 166–173, Online. Association for Computational Linguistics.
- Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. **Automatic evaluation of translation quality for distant language pairs**. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. **Opennmt: Open-source toolkit for neural machine translation**. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Abinash Dutta, Partha Pakray, and Sivaji Bandyopadhyay. 2019a. Neural machine translation: English to hindi. In *2019 IEEE Conference on Information and Communication Technology*, pages 1–6.
- Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021a. Improved English to Hindi multimodal neural machine translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 155–160, Online. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020a. **EnAsCorp1.0: English-Assamese corpus**. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020b. **Multimodal neural machine translation for English to Hindi**. In *Proceedings of the 7th Workshop on Asian Translation*, pages 109–113, Suzhou, China. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2019b. **Neural machine translation: Hindi-Nepali**. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 202–207, Florence, Italy. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2021b. Neural machine translation for low resource assamese–english. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, volume 170, page 35. Springer.
- Sahinur Rahman Laskar, Bishwaraj Paul, Partha Pakray, and Sivaji Bandyopadhyay. 2022a. Improving english-assamese neural machine translation using transliteration-based approach. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications, FICTA 2022*. In press.
- Sahinur Rahman Laskar, Rohit Pratap Singh, Partha Pakray, and Sivaji Bandyopadhyay. 2019c. **English to Hindi multi-modal neural machine translation and Hindi image captioning**. In *Proceedings of the 6th Workshop on Asian Translation*, pages 62–67, Hong Kong, China. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2022b. Improved neural machine translation for low-resource english–assamese pair. *Journal of Intelligent and Fuzzy Systems*, 42(5):4727–4738.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. **Effective approaches to attention-based neural machine translation**. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida,

- Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation (WAT2022)*, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Shantipriya Parida, Subhadarshi Panda, Satya Prakash Biswal, Ketan Kotwal, Arghyadeep Sen, Satya Ranjan Dash, and Petr Motlicek. 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)*, pages 31–39, Online (Virtual Mode). INCOMA Ltd.
- Amarnath Pathak and Partha Pakray. 2018. [Neural machine translation for indian languages](#). *Journal of Intelligent Systems*, pages 1–13.
- Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2018. [English–mizo machine translation using neural and statistical approaches](#). *Neural Computing and Applications*, 30:1–17.
- Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. [WAT2019: English-Hindi translation on Hindi visual genome dataset](#). In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188, Hong Kong, China. Association for Computational Linguistics.
- Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022a. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70. Springer.
- Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022b. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70, Singapore. Springer Nature Singapore.
- Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2020. [Neural machine translation of low-resource languages using smt phrase pair injection](#). *Natural Language Engineering*, page 1–22.
- Kashif Shah, Josiah Wang, and Lucia Specia. 2016. [SHEF-multimodal: Grounding machine translation on images](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 660–665, Berlin, Germany. Association for Computational Linguistics.