

Improved English to Hindi Multimodal Neural Machine Translation

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji,
Darsh Kaushik, Partha Pakray, Sivaji Bandyopadhyay

Department of Computer Science and Engineering
National Institute of Technology Silchar

Assam, India

{sahinur_rs, abduallah_ug, darsh_ug, partha}@cse.nits.ac.in,
sivaji.cse.ju@gmail.com

Abstract

Machine translation performs automatic translation from one natural language to another. Neural machine translation attains a state-of-the-art approach in machine translation, but it requires adequate training data, which is a severe problem for low-resource language pairs translation. The concept of multimodal is introduced in neural machine translation (NMT) by merging textual features with visual features to improve low-resource pair translation. WAT2021 (Workshop on Asian Translation 2021) organizes a shared task of multimodal translation for English to Hindi. We have participated the same with team name CNLP-NITS-PP in two submissions: multimodal and text-only translation. This work investigates phrase pairs injection via data augmentation approach and attains improvement over our previous work at WAT2020 on the same task in both text-only and multimodal translation. We have achieved second rank on the challenge test set for English to Hindi multimodal translation where Bilingual Evaluation Understudy (BLEU) score of 39.28, Rank-based Intuitive Bilingual Evaluation Score (RIBES) 0.792097, and Adequacy-Fluency Metrics (AMFM) score 0.830230 respectively.

1 Introduction

Multimodal NMT (MNMT) intends to draw insights from the input data through different modalities like text, image, and audio. Combining information from more than one modality attempts to amend the quality of low resource language translation. (Shah et al., 2016) show, combining the visual features of images with corresponding textual features of the input bitext to translate sentences outperform text-only translation. Encoder-decoder architecture is a widely used technique in the MT community for text-only-based NMT as it handles

various issues like variable-length phrases using sequence to sequence learning, the problem of long-term dependency using Long Short Term Memory (LSTM) (Sutskever et al., 2014). Nevertheless, the basic encoder-decoder architecture cannot encode all the information when it comes to very long sentences. The attention mechanism is proposed to handle such issues, which pays attention to all source words locally and globally (Bahdanau et al., 2015; Luong et al., 2015). The attention-based NMT yields substantial performance for Indian language translation (Pathak and Pakray, 2018; Pathak et al., 2018; Laskar et al., 2019a,b, 2020a, 2021b,a). Moreover, NMT performance can be enhanced by utilizing monolingual data (Sennrich et al., 2016; Zhang and Zong, 2016; Laskar et al., 2020b) and phrase pair injection (Sen et al., 2020), effective in low resource language pair translation. This paper aims English to Hindi translation using the multimodal concept by taking advantage of monolingual data and phrase pair injections to improve the translation quality at the WAT2021 translation task.

2 Related Works

For the English-Hindi language pair, the literature survey revealed minor existing works on translation using multimodal NMT (Dutta Chowdhury et al., 2018; Sanayai Meetei et al., 2019; Laskar et al., 2019c). (Dutta Chowdhury et al., 2018) uses synthetic data, following multimodal NMT settings (Calixto and Liu, 2017), and attains a BLEU score of 24.2 for Hindi to English translation. However, in the WAT 2019 multimodal translation task of English to Hindi, we achieved the highest BLEU score of 20.37 for the challenge test set (Laskar et al., 2019c). This score was improved later in the task of WAT2020 (Laskar et al., 2020c) to obtain the BLEU score of 33.57 on the challenge

Type	Name	Items/Instances	Tokens in millions (En / Hi)
Train	Text Data (En - Hi)	28,927	0.143164 / 0.145448
	Image Data	28,927	
Test (Evaluation Set)	Text Data (En - Hi)	1,595	0.007853 / 0.007852
	Image Data	1,595	
Test (Challenge Set)	Text Data (En - Hi)	1,400	0.008186 / 0.008639
	Image Data	1,400	
Validation	Text Data (En - Hi)	998	0.004922 / 0.004978
	Image Data	998	

Table 1: Parallel Data Statistics (Nakazawa et al., 2021; Parida et al., 2019).

Monolingual Sentences	Tokens in millions	
Data		
En	107,597,494	1832.008594
Hi	44,949,045	743.723731

Table 2: Monolingual Data Statistics collected from IITB and WMT16.

test set. In (Laskar et al., 2020c), we have used bidirectional RNN (BRNN) at encoder type, and doubly-attentive RNN at decoder type following default settings of (Calixto and Liu, 2017; Calixto et al., 2017) and utilizes pre-train word embeddings of the monolingual corpus and additional parallel data of IITB. This work attempts to utilize phrase pairs (Sen et al., 2020) to enhance the translational performance of the WAT2021: English to Hindi multimodal translation task.

3 Dataset Description

We have used the Hindi Visual Genome 1.1 dataset provided by WAT2021 organizers (Nakazawa et al., 2021; Parida et al., 2019). The train data contains English-Hindi 28,930 parallel sentences and 28,928 images. After removing duplicate sentences having ID numbers 2391240, 2385507, 2328549 from parallel data and one image having ID number 2326837 (since corresponding text not present in parallel data), the parallel and image train data reduced to 28,927. Moreover, English-Hindi (En-Hi) parallel and Hindi monolingual corpus¹ (Kunchukuttan et al., 2018) and also, English monolingual data available at WMT16² are used. Table 1 and 2 depict the data statistics.

¹http://www.cfilt.iitb.ac.in/iitb_parallel/

²<http://www.statmt.org/wmt16/translation-task.html>

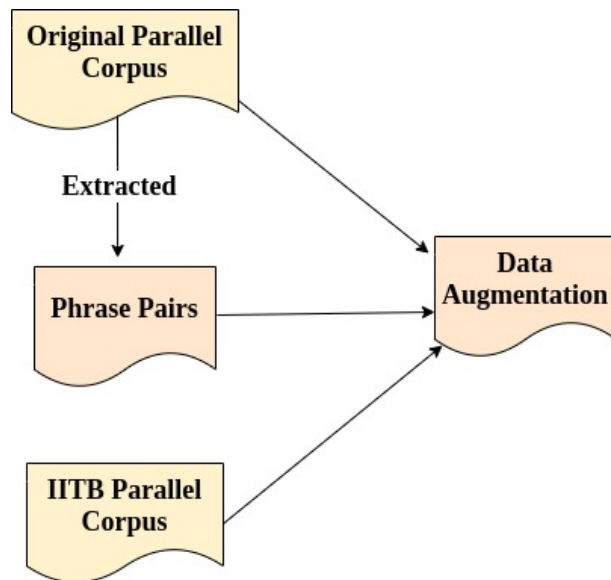


Figure 1: Data augmentation for English-to-Hindi multimodal NMT.

4 System Description

To build multimodal and text-only NMT models, OpenNMT-py (Klein et al., 2017) tool is used. There are four operations which include data augmentation, preprocessing, training and testing. Our multi-model NMT gets advantages from both image and textual features with phrase pairs and word embeddings.

4.1 Data Augmentation

In (Sen et al., 2020), authors used SMT-based phrase pairs to augment with the original parallel data to improve low-resource language pairs translation. In SMT³, Giza++ word alignment tool is used to extract phrase pair. Inspired by the work (Sen et al., 2020), we have extracted phrase

³<http://www.statmt.org/moses/>


Image id: 1159278	
	
Multi-modal Translation Track Source Language: English Target Language: Hindi	
Source Sentence	The top white cross.
Predicted Sentence	ऊपर सफेद क्रॉस (Upor shafed cross)
Reference Sentence	शीर्ष पर सफेद क्रॉस (Shirse par shafed cross)
Google Translation	शीर्ष सफेद क्रॉस (Shirse shafed cross)
Text-only Translation Track	
Predicted Sentence:	ऊपर सफेद <unk> (Upor shafed <unk>)

Figure 2: Examples of our best predicted output on challenge test data.

pairs using Giza++⁴. Then after removing duplicates and blank lines, the obtained phrase pairs are augmented to the original parallel data. The data statistics of extracted phrase pairs is given in Table 3. Additionally, IITB parallel data is directly augmented with the original parallel to expand the train data. The diagram of data augmentation is presented in Figure 1.

4.2 Data Preprocessing

To extract visual features from image data, we have used publicly available⁵ pre-trained CNN with VGG19. The visual features are extracted independently for train, validation, and test data. To get the advantage of monolingual data on both multimodal and text-only, GloVe (Pennington et al., 2014) is used to generate vectors of word embeddings. For tokenization of text data, the OpenNMT-py tool is utilized and obtained a vocabulary size of 50004 for source-target sentences. We have not used any word-segmentation technique.

4.3 Training

The multimodal and text-only based NMT are trained independently. During the multimodal training process, extracted visual features, pre-trained

⁴<https://github.com/ayushidalmia/Phrase-Based-Model>

⁵<https://github.com/iacercalixto/MultimodalNMT>


Image id: 2351980	
	
Multi-modal Translation Track Source Language: English Target Language: Hindi	
Source Sentence	Dirt on the players pants.
Predicted Sentence	खिलाड़ी <unk> पर <unk> (khillari <unk> par <unk>)
Reference Sentence	खिलाड़ियों की पेट पर मिट्टी। (khillariyo ki pant par mitti)
Google Translation	खिलाड़ियों की पेट पर गंदगी। (khillaiyo ki pant par ghandagi)
Text-only Translation Track	
Predicted Sentence:	खिलाड़ी <unk> पर <unk> (khillari <unk> par <unk>)

Figure 3: Examples of our worst predicted output on challenge test data.

word vectors are fine-tuned with the augmented parallel data. The bidirectional RNN (BRNN) at encoder type and doubly-attentive RNN at decoder type following default settings of (Calixto and Liu, 2017; Calixto et al., 2017) are used for multimodal NMT. Two different RNNs are used in BRNN, one for backward and another for forwards directions, and two distinct attention mechanisms are utilized over source words and image features at a single decoder. The multimodal NMT is trained up to 40 epochs with 0.3 drop-outs and batch size 32 on a single GPU. During the training process of text-only NMT, we have used only textual data i.e., pre-trained word vectors are fine-tuned with the augmented parallel data, and the model is trained up to 100000 steps using BRNN encoder and RNN decoder following default settings of OpenNMT-py. The primary difference between our previous work (Laskar et al., 2020c) and this work is that the present work uses phrase pairs in augmented parallel data.

4.4 Testing

The obtained trained NMT models of both multimodal and text-only are tested on both test data: evaluation and challenge set independently. During testing, the only difference between text-only and multimodal NMT is that multimodal NMT uses

Number of Phrase Pairs	Tokens in millions	
	En	Hi
158,131	0.392966	0.410696

Table 3: Data Statistics of extracted phrase pairs.

Our System	Test Set	BLEU	RIBES	AMFM
Text-only NMT	Challenge	37.16	0.770621	0.798670
	Evaluation	37.01	0.795302	0.812190
Multi-modal NMT	Challenge	39.28	0.792097	0.830230
	Evaluation	39.46	0.802055	0.823270

Table 4: Our system’s results on English to Hindi multimodal translation Task.

visual features of image test data.

5 Result and Analysis

The WAT2021 shared task organizer published the evaluation result⁶ of multimodal translation task for English to Hindi and our team stood second position in multimodal submission for challenge test set. Our team name is CNLP-NITS-PP, and we have participated in the multimodal and text-only submission tracks of the same task. In both multimodal and text-only translation submission tracks, a total of three teams participated in both evaluation and challenges test data. The results are evaluated using automatic metrics: BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and AMFM (Banchs et al., 2015). The results of our system is reported in Table 4, and it is noticed that the multimodal NMT obtains higher than text-only NMT. It is because the combination of textual and visual features outperforms text-only NMT. Furthermore, our system’s results are improved as compared to our previous work on the same task (Laskar et al., 2020c). It shows the BLEU, RIBES, AMFM scores of present work show (+5.71, +9.41), (+0.037956, +0.055641), (+0.04291, +0.04835) increments on the challenge test set for multimodal and text-only NMT, where it is realised that phrase pairs augmentation improves translational performance. The sample examples of best and worst outputs, along with Google translation and transliteration of Hindi words, are presented in Figure 2 and 3. In Figure 2 and 3, highlighted the region in the image for the given caption by a red colour rectangular box.

⁶<http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html>

6 Conclusion and Future Work

In this work, we have participated in a shared task at WAT2021 multimodal translation task of English to Hindi, where translation submitted at tracks: multimodal and text-only. This work investigates phrase pairs through data augmentation approach in both multimodal and text-only NMT, which shows better performance than our previous work on the same task (Laskar et al., 2020c). In future work, we will investigate a multilingual approach to improve the performance of multimodal NMT.

Acknowledgement

We want to thank the Center for Natural Language Processing (CNLP), the Artificial Intelligence (AI) Lab, and the Department of Computer Science and Engineering at the National Institute of Technology, Silchar, India, for providing the requisite support and infrastructure to execute this work. We also thank the WAT2021 Translation task organizers.

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*.
- Rafael E. Banchs, Luis F. D’Haro, and Haizhou Li. 2015. *Adequacy-fluency metrics: Evaluating mt in the continuous space model framework*. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(3):472–482.
- 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.
- 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.
- Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. **The IIT Bombay English-Hindi parallel corpus**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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, 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. **Hindi-Marathi cross lingual model**. In *Proceedings of the Fifth Conference on Machine Translation*, pages 396–401, Online. Association for Computational Linguistics.
- Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020c. **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. 2021a. **Neural machine translation: Assamese–bengali**. In *Modeling, Simulation and Optimization: Proceedings of CoMSO 2020*, pages 571–579. Springer Singapore.
- 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: 13CS 2020, NEHU, Shillong, India*, volume 170, page 35. Springer.
- 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.
- 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, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. **Overview of the 8th workshop on Asian translation**. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. Association for Computational Linguistics.
- 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, Ondřej Bojar, and Satya Ranjan Dash. 2019. **Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation**. *Computación y Sistemas*, 23(4):1499–1505.
- 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.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- 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.
- 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.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- 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.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to sequence learning with neural networks](#). In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS’14*, pages 3104–3112, Cambridge, MA, USA. MIT Press.
- Jiajun Zhang and Chengqing Zong. 2016. [Exploiting source-side monolingual data in neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.