

Multimodal Neural Machine Translation for English to Hindi

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

Department of Computer Science and Engineering

National Institute of Technology Silchar

Assam, India

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

Abstract

Machine translation (MT) focuses on the automatic translation of text from one natural language to another natural language. Neural machine translation (NMT) achieves state-of-the-art results in the task of machine translation because of utilizing advanced deep learning techniques and handles issues like long-term dependency, and context-analysis. Nevertheless, NMT still suffers low translation quality for low resource languages. To encounter this challenge, the multi-modal concept comes in. The multi-modal concept combines textual and visual features to improve the translation quality of low resource languages. Moreover, the utilization of monolingual data in the pre-training step can improve the performance of the system for low resource language translations. Workshop on Asian Translation 2020 (WAT2020) organized a translation task for multimodal translation in English to Hindi. We have participated in the same in two-track submission, namely text-only and multi-modal translation with team name CNLP-NITS. The evaluated results are declared at the WAT2020 translation task, which reports that our multimodal NMT system attained higher scores than our text-only NMT on both challenge and evaluation test set. For the challenge test data, our multi-modal neural machine translation system achieves Bilingual Evaluation Understudy (BLEU) score of 33.57, Rank-based Intuitive Bilingual Evaluation Score (RIBES) 0.754141, Adequacy-Fluency Metrics (AMFM) score 0.787320 and for evaluation test data, BLEU, RIBES, and, AMFM score of 40.51, 0.803208, and 0.820980 for English to Hindi translation respectively.

1 Introduction

Multi-modal NMT aims to draw information from the input data from different modalities like text, image, and audio. By combining information from

more than one modality, it attempts to amend the quality of low resource language translation. The work undertaken by (Shah et al., 2016) merges the visual features of images from the corresponding input data with textual features of the input bibtex to translate sentences, which outperforms text-only translation. For text-only based NMT, encoder-decoder architecture is a widely used technique in the MT community. Because 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). However, in the case of very long sentences, the basic encoder-decoder architecture is unable to encode all the information. To resolve this issue, the attention mechanism is proposed which pays attention to all source words locally as well as globally (Bahdanau et al., 2015; Luong et al., 2015). For Indian language translation, attention-based NMT yields remarkable performance (Pathak and Pakray, 2018; Pathak et al., 2018; Laskar et al., 2019b,a). Besides, without modifying the system architecture, NMT performance can be improved using monolingual data (Sennrich et al., 2016; Zhang and Zong, 2016), which is very effective in the case of low resource language translation. This paper investigates English to Hindi translation using the multi-modal concept with monolingual data to improve the translation quality at the WAT2020 translation task.

2 Related Works

The literature survey finds out very limited existing works on English-Hindi language pair translation using multi-modal NMT (Dutta Chowdhury et al., 2018; Sanayai Meetei et al., 2019; Laskar et al., 2019c). The work by (Dutta Chowdhury et al., 2018) uses synthetic data, following multi-modal

Type	Name	Instances/Items	Tokens (English / Hindi)
Train	Text Data (English - Hindi)	28,927	143,164 / 145,448
	Image Data	28,927	
Test (Evaluation Set)	Text Data (English - Hindi)	1,595	7,853 / 7,852
	Image Data	1,595	
Test (Challenge Set)	Text Data (English - Hindi)	1,400	8,186 / 8,639
	Image Data	1,400	
Validation	Text Data (English - Hindi)	998	4,922 / 4,978
	Image Data	998	

Table 1: Parallel Data Statistics (Nakazawa et al., 2020; Parida et al., 2019).

NMT settings of (Calixto and Liu, 2017), achieves BLEU score 24.2 for Hindi to English translation. Moreover, in the WAT2019 multi-modal translation task of English to Hindi, (Sanayai Meetei et al., 2019) based on recurrent neural network (RNN) (Calixto and Liu, 2017) achieves BLEU score of 12.58, 28.45 for the challenge and evaluation test respectively. And, on the same task of WAT2019, we have achieved the highest BLEU score of 20.37, 40.55 for the challenge and evaluation test respectively (Laskar et al., 2019c). We have used RNN encoder and doubly-attentive RNN decoder based model (Calixto and Liu, 2017; Calixto et al., 2017). The loophole in the existing works of English to Hindi translation using multi-modal NMT is that they have not used monolingual data to improve the performance of multi-modal NMT (Sennrich et al., 2016). In this paper, we have used monolingual corpus in the pre-training step to enhance the performance of the multi-modal NMT for English to Hindi translation respectively.

3 Dataset Description

Hindi Visual Genome 1.1 consists of parallel text and image data, which is provided by the WAT2020 organizers (Nakazawa et al., 2020; Parida et al., 2019). The original train parallel text data consists of 28,930 sentences and 28,928 images. We have removed three duplicate sentences (id:2328549, 2385507, 2391240) from the parallel data. Also, we have removed one image (id:2326837) from the image dataset since it is not available in train parallel text data. Therefore, the total number of parallel sentences and image become 28,927 in Hindi Visual Genome 1.1 dataset as summarized in Table 1. Additionally, we have used English-Hindi parallel corpus and monolingual data of Hindi from

Monolingual Data	Sentences	Tokens
English	107,597,494	1,832,008,594
Hindi	44,949,045	743,723,731

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

IITB¹ (Kunchukuttan et al., 2018) and English monolingual data from WMT16² as shown in Table 2.

4 System Description

We have used OpenNMT-py (Klein et al., 2017) to setup our multi-modal NMT and text-only NMT systems. The key process of the operations include data preprocessing, system training to generate an optimum trained model, and then obtained trained model is used in the testing/translation process to predict translation on the given unseen data.

4.1 Data Preprocessing

For multi-modal translation, pre-trained CNN with VGG19 is used for the extraction of global and local features from the provided image dataset. The pre-trained CNN with VGG19 is publicly available in OpenNMT-py. In the text-only and multi-modal task, we have used GloVe (Pennington et al., 2014) to pretrain on monolingual data of English-Hindi and generated global vectors of word embedding. The OpenNMT-py tool is used to create a vocabulary size of 5004 for both source and target sentences. We have not used any word-segmentation technique.

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

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

Our System	Test Set	BLEU	RIBES	AMFM
Text-only NMT	Challenge	27.75	0.714980	0.750320
	Evaluation	38.84	0.793416	0.804250
Multi-modal NMT	Challenge	33.57	0.754141	0.787320
	Evaluation	40.51	0.803208	0.820980

Table 3: Our system’s results on English to Hindi multi-modal translation Task.


Image id: 2853	
	
Multi-modal Translation Track Source Language: English Target Language: Hindi	
Source Sentence	a bunch of books on book stand
Predicted Sentence	किताब स्टैंड पर पुस्तकों का एक गुच्छा
Reference Sentence	पुस्तक स्टैंड पर पुस्तकों का एक गुच्छा
Google Translation	पुस्तक स्टैंड पर पुस्तकों का एक गुच्छा
Text-only Translation Track	
Predicted Sentence:	बुक स्टैंड पर पुस्तकों का एक गुच्छा

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


Image id: 2417756	
	
Multi-modal Translation Track Source Language: English Target Language: Hindi	
Source Sentence	March 7th is the date on the calendar
Predicted Sentence	the चालक कैलेंडर पर है
Reference Sentence	कैलेंडर पर 7 मार्च की तारीख है
Google Translation	7 मार्च को कैलेंडर की तारीख है
Text-only Translation Track	
Predicted Sentence:	HALO निकालने पर बनी है

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

4.2 Training

The training process for each track is carried out separately. For multi-modal translation, the obtained pretrained vectors, extracted visual features from data preprocessing are fine-tuned with the parallel text data during the training process. 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). BRNN uses two distinct RNN, one for the forward direction and another for backward, and two different attention mechanisms are incorporated across the source words and visual features at a single RNN decoder. Two layer LSTM networks having 500 nodes in each layer are used in both encoder and decoder. Our multi-modal NMT is trained on a single GPU up to 40 epochs with 0.3 drop out, batch size 40 and the best model is obtained at epoch 10. For text-only translation, we have not used visual features and only used pretrained vectors of monolingual data to fine-tune with parallel corpus in the training process. The text-only NMT is trained up to 20,000 epoch since learning curve raises up to 18,000 and then drops. We have selected best trained model at epoch 18,000. The difference between (Laskar et al., 2019c) and this paper, is that in this work, our multi-modal NMT adopts BRNN at encoder type unlike RNN in (Laskar et al., 2019c) and utilizes pretrain word embeddings of monolingual corpus.

4.3 Testing

In this process, the obtained trained models of both multi-modal and text-only NMT system, are used to translate the given test data in each track separately.

5 Result and Analysis

The WAT2020 translation task organizer declared the evaluation result³ of multi-modal translation

³<http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html>

task for English to Hindi and our system’s results are presented in Table 3. Our team name is CNLP-NITS and participated in text-only and multi-modal submission track of the same task. In text-only translation submission track, a total of four teams participated for both challenges and evaluation test data and for multi-modal translation submission track, only our team participated. The submitted predicted translations are evaluated via standard evaluation metrics namely, BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and AMFM (Banchs et al., 2015). From the Table 2, it is observed that our multi-modal NMT system obtained higher scores on the ground of BLEU, RIBES, AMFM than our text-only NMT system. This reasons about combination of visual and textual features in multi-modal NMT shows better performance than only textual features based NMT. Moreover, our systems used pretrained word embedding of monolingual data and adopted BRNN encoder that reasons about outperform previous work (Laskar et al., 2019c) at WAT2019. Figure 1 and 2 present best and worst performance our systems outputs, where included Google translation for comparative analysis.

6 Conclusion and Future Work

This work participates in two different translation tracks at WAT2020 multi-modal translation task of English to Hindi namely: multi-modal and text-only. In this competition our multi-modal NMT achieves higher BLEU, RIBES and AMFM scores than text-only NMT. From the best of our knowledge, our multi-modal NMT achieves best score on English to Hindi multi-modal translation. In future work, more experiments, analysis will be carried out to enhance the performance of multi-modal NMT.

Acknowledgement

We would like to thank Center for Natural Language Processing (CNLP) and Department of Computer Science and Engineering at National Institute of Technology, Silchar, India for providing the requisite support and infrastructure to execute this work. We also thank the WAT2020 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 Bhat-tacharyya. 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, 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, 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, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. 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.
- 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.