English to Hindi Multi-modal Neural Machine Translation and Hindi Image Captioning

Sahinur Rahman Laskar, Rohit Pratap Singh, Partha Pakray and **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering
National Institute of Technology Silchar
Assam, India
{sahinurlaskar.nits,rohitkako,parthapakray,sivaji.cse.ju}@gmail.com

Abstract

With the widespread use of Machine Translation (MT) techniques, attempt to minimize communication gap among people from diverse linguistic backgrounds. We have participated in Workshop on Asian Translation 2019 (WAT2019) multi-modal translation task. There are three types of submission track namely, multi-modal translation, Hindi-only image captioning and text-only translation for English to Hindi translation. The main challenge is to provide a precise MT output. The multi-modal concept incorporates textual and visual features in the translation task. In this work, multi-modal translation track relies on pre-trained convolutional neural networks (CNN) with Visual Geometry Group having 19 layered (VGG19) to extract image features and attention-based Neural Machine Translation (NMT) system for translation. The merge-model of recurrent neural network (RNN) and CNN is used for the Hindi-only image captioning. The text-only translation track is based on the transformer model of the NMT system. The official results evaluated at WAT2019 translation task, which shows that our multi-modal NMT system achieved Bilingual Evaluation Understudy (BLEU) score 20.37, Rank-based Intuitive Bilingual Evaluation Score (RIBES) 0.642838, Adequacy-Fluency Metrics (AMFM) score 0.668260 for challenge test data and BLEU score 40.55, RIBES 0.760080, AMFM score 0.770860 for evaluation test data in English to Hindi multi-modal translation respectively.

1 Introduction

The multi-modal translation is an emerging task of the MT community, where visual features of image combine with textual features of parallel source-target text to translate sentences (Shah et al., 2016). Interestingly, multi-modal concept improved the translation quality of generating the captions of the images (Dash et al., 2019) as well as significant improvement over text-only NMT system (Huang et al., 2016). In text-only NMT system, the encoder-decoder framework of NMT is a widely accepted technique used in the task of MT. Because it handles sequence to sequence learning problem for variable length source and target sentences and also, handles long term dependency problem using long short term memory (LSTM) (Sutskever et al., 2014). The demerits of basic encoder-decoder model is that it fails to encode all necessary information into the context vector when the sentence is too long. Hence, to handle such problem attention-based encoder-decoder model is introduced, which allows the decoder to focus on different parts of the source sequence at different decoding steps (Bahdanau et al., 2015). (Luong et al., 2015) enhanced the attention model that merges global, accompanying to all source words and local, only pay attention to a part of source words. The attention-based NMT system shows a promising outcome in various languages (Pathak and Pakray, 2018; Pathak et al., 2018; Laskar et al., 2019). Current work has been investigated for English to Hindi translation. There are three different tracks, namely, multi-modal translation, Hindi-only image captioning and text-only translation using NMT system and participated in WAT2019 multi-modal translation task.

2 Related Works

Literature survey mainly focused on multimodal based NMT works, where multimodal informa-

Proceedings of the 6th Workshop on Asian Translation, pages 62–67
Hong Kong, China, November 4, 2019.

tion (text and image) integrating into the attention-based encoder-decoder architecture. (Huang et al., 2016), proposed a model using attention based NMT, where regional and global visual features are attached in parallel with multiple encoding threads and each thread is followed by the text sequence. They obtained BLEU score 36.5, which outperformed the text-only baseline model BLEU score 34.5. (Calixto and Liu, 2017) used bi-directional recurrent neural network (RNN) with gated recurrent unit (GRU) in the encoding phase instead of single-layer unidirectional LSTM in (Huang et al., 2016) and also, used image features separately either as a word in the source sentence or directly for encoder or decoder initialization unlike word only in (Huang et al., 2016), achieved BLEU score 38.5, 43.9 in English to German and German to English translation respectively. (Calixto et al., 2017), introduced two independent attention mechanisms over source language words and visual features in a single decoder RNN, which significantly improve over the models used in (Huang et al., 2016), obtained BLEU score 39.0, 43.2 in English to German and German to English translation respectively. (Dutta Chowdhury et al., 2018), investigated multimodel NMT following settings of (Calixto and Liu, 2017) for Hindi to English translation and acquired BLEU score 24.2.

3 System Description

The primary steps of the system operations are data preprocessing, system training and system testing and the same have been illustrated in following subsections. The multimodal NMT toolkit (Calixto and Liu, 2017; Calixto et al., 2017) is employed to build the multimodal NMT system for multimodal translation task, which are based on the pytorch port of OpenNMT (Klein et al., 2017). For text-only translation task, OpenNMT is deployed to build the NMT system and in the case of Hindi-only image captioning track, publicly available VGG16 and LSTM in Keras library, are used to build the system (Simonyan and Zisserman, 2015; Tanti et al., 2018). We have used Hindi visual genome dataset in each track of WAT2019 multi-modal translation task provided by the organizer (Nakazawa et al., 2019). We have not used image coordinates (Width, Height) provided in the dataset to indicate the rectangular region in the image described by the caption. Because, we have used global features of the images.

3.1 Data Preprocessing

The data preprocessing steps of each track are carried out separately. In the multi-modal translation track, firstly, image features for training, validation and test data are extracted from the image data set as mentioned in Table 1. We have used publicly available pre-trained CNN with VGG19 via batch normalization for extraction of both global and local visual features from the image dataset as shown in Table 1. Secondly, primary functions of preprocessing step, tokenization, lowercasing and applying byte pair encoding (BPE) model of source and target sentences. For this purpose, OpenNMT toolkit is used to make a dictionary of vocabulary size of dimension 8300, 7984 for English-Hindi parallel sentence pairs, which indexes the words during the training process. In the text-only translation track, we have considered only source-target corresponding sentences as shown in Table 1 to build the dictionary, vocabulary size of dimension 8300, 7984 using the OpenNMT toolkit. In the Hindi-only image captioning track, image features are extracted via pre-trained CNN with VGG16 from the image data set as shown in Table 1. The image extracted features are 1-dimensional 4,096 element vector. The text input sequences, maximum description length of 22 words, are cleaned to get the vocabulary size of 5605.

3.2 System Training

After preprocessing of data, the system training process is performed in each track separately in Multiple Graphics Processing Units (GPU) environment to boost the performance of training. In the multi-modal translation track, the source (English) and target (Hindi) sentences are fed into encoder-decoder RNN. The multi-modal NMT system is trained using doubly-attentive decoder following settings of (Calixto et al., 2017), where the multi-modal NMT incorporates two different attention mechanism across the source-language words and visual features in a single decoder RNN. Both encoder and decoder consists of a two-layer network of LSTM nodes, which contains 500 units in each layer. The multi-modal NMT system is trained up to 100 epoch. The default settings drop out of 0.3, batch size 40 and layer normalization are used for a stable training run. In the

Nature of corpus	Name of Corpus	Number of instances/items
Training	Englsih-Hindi (Text data)	28,929
	Image data	28,929
Test (Evaluation Set)	English to Hindi (Text data)	1595
	Image data	1595
Test (Challenge Set)	English to Hindi (Text data)	1400
	Image data	1400
Validation	English-Hindi (Text data)	998
	Image data	998

Table 1: Corpus Statistics (Nakazawa et al., 2019).

training process of text-only translation track, the NMT system is trained up to 25,000 epoch to build the train models by transformer model of NMT system. For a small dataset in text-only translation, it is not required up to 25,000 epoch. But in this dataset, we need to trained up to 25,000 because of learning curve grows up to 24,000 then falls. Hence, we have chosen predicted translation at an optimum point on 24,000 epoch. In the training process of Hindi-only image captioning track, we have used merge-model following settings of (Tanti et al., 2018). The preprocessed image feature vector of 4096 elements are processed by a dense layer to provide 256 elements for representation of the image. Afterward, the input text sequence of 22 words length are fed into a word embedding layer to convert it into vector form which is followed by LSTM based RNN layer contains 256 nodes. Both the fixed-length vectors (Image and text) generated are merged together and processed by a dense layer to build the train models up to 20 epoch.

3.3 System Testing

System training is followed by the system testing process in each track separately. This process is required for predicting translations of test instances/items as shown in Table 1.

4 Result and Analysis

The official evaluation results of the competition for English to Hindi multi-modal translation task are reported by the organizer [1]. Automatic evaluation metrics namely, BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and AMFM (Banchs et al., 2015) are used to measure performance of predicted translations. We have participated in all the track of the multi-modal translation task and our team name is 683. In multi-modal translation track, a total of three teams, including our team participated for both challenge and evaluation test data in English to Hindi translation. We have acquired BLEU, RIBES, AMFM score 20.37, 0.642838, 0.668260 for challenge test set and BLEU, RIBES, AMFM score 40.55, 0.760080, 0.770860 for evaluation test set respectively, higher than other teams as shown in Table 2. However, we have attained lower BLEU, RIBES and AMFM scores than other teams in text-only and Hindi-only image captioning translation track as shown in Table 3 and 4 respectively. Moreover, from Table 2, 3 and 4, it is observed that when translating English to Hindi our multi-modal translation outperforms our text only translation as well as our Hindi-only image captioning. To further analyze the best and worst performance of multi-modal translation in comparison to text-only and Hindi-only image captioning, sample predicted sentences on challenge test data, reference target sentences and Google translation on same test data are considered in Table 5, 6. In Table 5, our multi-modal NMT system provides perfect prediction like reference target sentence, Google translation and close to text-only translation but wrong translation in Hindi-only image captioning. However, in Table 6, prediction of source word "court" is inappropriate like Google translation, text-only translation and wrong translation in Hindi-only image captioning.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

System	BLEU	
	Challenge Test Set	Evaluation Test Set
System-1 (Our system)	**20.37**	**40.55**
System-2	12.58	28.45
System-3	11.77	28.27
System-4	10.19	27.39
	RIBES	
	Challenge Test Set	**Evaluation Test Set**
System-1 (Our system)	**0.642838**	**0.760080**
System-2	0.507192	0.692880
System-3	0.487897	0.676444
System-4	0.482373	0.634567
	AMFM	
	Challenge Test Set	**Evaluation Test Set**
System-1 (Our system)	**0.668260**	**0.770860**
System-2	0.659840	0.722110
System-3	0.632060	0.707520
System-4	0.559990	0.682060

Table 2: BLEU, RIBES and AMFM scores result of participated teams for multi-modal translation track.

System	BLEU	
	Challenge Test Set	Evaluation Test Set
System-1	30.94	41.32
System-2	30.34	-
System-3	-	38.95
System-4 (Our system)	**15.85**	**38.19**
System-5 (Our system)	**14.69**	**25.34**
System-6	5.56	20.13
	RIBES	
	Challenge Test Set	**Evaluation Test Set**
System-1	0.734435	0.770754
System-2	0.726998	-
System-3	-	0.749535
System-4 (Our system)	**0.550964**	**0.744158**
System-5 (Our system)	**0.550568**	**0.636152**
System-6	0.373560	0.574366
	AMFM	
	Challenge Test Set	**Evaluation Test Set**
System-1	0.775890	0.784950
System-2	0.773260	-
System-3 (Our system)	**0.632910**	**0.763940**
System-4	-	0.762180
System-5 (Our system)	**0.578930**	**0.656370**
System-6	0.461110	0.615290

Table 3: BLEU, RIBES and AMFM scores result of participated teams for text-only translation track.

System	Challenge Test Set	
	RIBES	AMFM
System-1	0.080028	0.385960
System-2 (Our system)	**0.034482**	**0.335390**

Table 4: RIBES, AMFM scores result of participated teams for Hindi-only image captioning track.

5 Conclusion and Future Work

Current work participates in three different translation tracks at WAT2019 namely, multi-modal, text-only and Hindi-only image captioning for

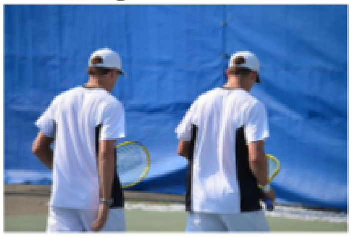
English to Hindi translation. In the current competition, our multi-modal NMT system obtained higher BLEU scores than other participants in case of challenge as well as evaluation test data. The multi-modal NMT system is based on a doubly-attentive decoder to predict sentences, which shows better performance than text-only as well as Hindi-only image captioning. The combination of textual as well as visual features reasons about multi-modal translation outperforms text-only translation as well as Hindi-only image captioning tasks. However, close analysis of predicted sentences on the given test data remarks that more experiment and analysis are needed in future work to enhance the performance of multi-modal NMT system.

Image id: 2417491

Multi-modal translation track Source Language: English Target Language: Hindi	
Source Test Sentence	wooden sign with white letters on second bus
Predicted Target Sentence	दूसरी बस पर सफेद अक्षरों के साथ लकड़ी के चिन्ह
Reference Target Sentence	दूसरी बस में सफेद अक्षरों के साथ लकड़ी का चिन्ह
Google Translation	दूसरी बस में सफेद अक्षरों के साथ लकड़ी का चिन्ह
Text-only translation track Predicted Target Sentence: दूसरे बस बस पर सफेद अक्षर के साथ लकड़ी संकेत	
Hindi-only image captioning track Predicted Caption: एक सड़क पर एक बस	

Table 5: Best performance examples in English to Hindi multi-modal translation.

Image id: 2407547



Multi-modal translation track Source Language: English Target Language: Hindi	
Source Test Sentence	there are two players in the court
Predicted Target Sentence	अदालत में दो खिलाड़ी हैं
Reference Target Sentence	क़ोर्ट में दो खिलाड़ी हैं
Google Translation	अदालत में दो खिलाड़ी हैं
Text-only translation track Predicted Target Sentence: अदालत में दो खिलाड़ी हैं	
Hindi-only image captioning track Predicted Caption: एक व्यक्ति के हाथ में एक सफेद और सफेद टेनिस खिलाड़ी	

Table 6: Worst performance example in English to Hindi multi-modal translation.

Acknowledgement

Authors would like to thank WAT2019 Translation task organizers for organizing this competition and also, thank Centre for Natural Language Processing (CNLP) and Department of Computer Science and Engineering at 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*.

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.

Sandeep Kumar Dash, Saurav Saha, Partha Pakray, and Alexander Gelbukh. 2019. Generating image captions through multimodal embedding. *Journal of Intelligent and Fuzzy Systems*, 36(5):4787–4796.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. pages 33–42.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. 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, Partha Pakray, and Sivaji Bandyopadhyay. 2019. 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.

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, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. 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.

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.

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.

Karen Simonyan and Andrew Zisserman. 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*.

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.

Marc Tanti, Albert Gatt, and Kenneth P. Camilleri. 2018. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489.