Multiple Captions Embellished Multilingual Multi-Modal Neural Machine Translation

Salam Michael Singh[1], **Loitongbam Sanayai Meetei**[1], **Thoudam Doren Singh**[1], and **Sivaji Bandyopadhyay**[1]

[1]Centre for Natural Language Processing (CNLP) & Dept. of CSE, NIT Silchar, India
{salammichaelcse,loisanayai,thoudam.doren,sivaji.cse.ju}@gmail.com

Abstract

Neural machine translation based on bilingual text with limited training data suffers from lexical diversity, which lowers the rare word translation accuracy and reduces the generalizability of the translation system. In this work, we utilise the multiple captions from the `Multi-30K` dataset to increase the lexical diversity aided with the cross-lingual transfer of information among the languages in a multilingual setup. In this multilingual and multimodal setting, the inclusion of the visual features boosts the translation quality by a significant margin. Empirical study affirms that our proposed multimodal approach achieves substantial gain in terms of the automatic score and shows robustness in handling the rare word translation in the pretext of English to/from Hindi and Telugu translation tasks.

1 Introduction

The machine translation (MT) systems by (Koehn et al., 2003; Sutskever et al., 2014; Gehring et al., 2017; Vaswani et al., 2017) has been the de-facto standard which are based on parallel dataset. But, in recent times, use of monolingual data (Singh and Singh, 2020) or incorporating multiple languages in a jointly trained single multilingual model (Johnson et al., 2017; Fan et al., 2020) has improved the translation quality of low resource languages. Compared to training separate bi-lingual models with the same parameters, the ability to handle translation between multiple language pairs provides an inherent advantage of having relatively compact model parameters. Typically, in such models, the encoder and decoders are shared among all the languages and attention. The sharing of the encoder is crucial to learn the initial multilingual cross-lingual information (Sachan and Neubig, 2018) however, a single shared decoder is often insufficient in handling the translation of multiple languages. This decoder degeneracy is addressed by partial sharing of decoder and attention parameters (Sachan and Neubig, 2018) or through a language-agnostic universal models (Bapna and Firat, 2019).

Image features along with the text data has been used in sequence generation tasks such as the image caption generation (Singh et al., 2021a,b) and multimodal machine translation (MMT) which incorporates visual features into ordinary NMT systems for low resource languages. With the introduction of MMT datasets such as `Multi-30k` (Elliott et al., 2016) and Hindi Visual Genome Parida et al. (2019), MT researchers (Huang et al., 2016; Caglayan et al., 2016, 2019; Meetei et al., 2019) have highlighted improvement in translation quality by incorporating image features in the MT systems.

In this work, we adopt a single shared multilingual machine translation system between English and under resourced languages viz., Hindi and Telugu, aided by linguistic information in the form of multiple captions. The inclusion of multiple captions during training makes the system implicitly robust to lexical and syntactic diversity. In addition to the multiple captions, we infuse our multicaption multilingual model with the visual information in a multimodal (Calixto et al., 2017a; Elliott and Kádár, 2017; Yao and Wan, 2020) setting. English and Hindi belong to the Indo-European language family, while Telugu is a Dravidian language. All three languages use different scripts; Roman for English, Devanagari for Hindi and Telugu is written in Telugu script, an abugida writing system from the Brahmic family of scripts.

2 Related Works

Callison-Burch et al. (2006) used paraphrase in a phrase-based statistical machine translation model

Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021),
in conjuction with RANLP-2021, September 7, 2021. pages 2–11
https://doi.org/10.26615/978-954-452-073-1_002

and found that their method improves over the single parallel corpora PB-SMT baseline in terms of the overall word coverage and the translation quality. Paraphrase has also been leveraged as a data augmentation technique to improve dialog generation (Gao et al., 2020) and question generation (Jia et al., 2020) tasks. More similar to our proposed work Zhou et al. (2019) decomposed the paraphrase as a foreign language in a multilingual scenario. Similar to the findings of Callison-Burch et al. (2006), Zhou et al. (2019) found that their method improves the word coverage with diverse lexical choices.

However, these works are purely uni-modal, and on the other hand, a visually informed multimodal system involves the extraction of the global semantic features from the image and initialize either the encoder or decoder to fuse the visual context along with the textual input (Calixto et al., 2017b). In cases where the textual context is restricted, Caglayan et al. (2019) showed that visual features could help to generate better translations. Similar to the proposed work, (Chakravarthi et al., 2019) trained a multimodal machine translation system in the pretext of Tamil, Kannada and Malayalam by generating a synthetic dataset from Flickr30k (Plummer et al., 2015). They showed that transliteration of the Dravidian languages into Latin script and the multilingual setup improves the multimodal system over the bilingual multimodal baseline.

3 Methodology

The proposed multi-caption enabled multi-modal multilingual machine translation system employs two major steps: first, we create the training corpus and then train the multi-caption multilingual system fused with the visual features.

3.1 Corpus Creation

The creation of a training corpus for the experimentation is the first step. Multilingual machine translation with visual features for English (*en*) to/from {Hindi (*hi*), Telugu (*te*)} is the experiment's premise. `Multi-30K` (Elliott et al., 2016), on the other hand, lacks the *hi* and *te* data. As a result, for training, validation, and test data, a publicly available machine translation model (Ramesh et al., 2021) generates the *hi* and *te* translations corresponding to the English captions[1] (`caption-1` and `caption-2`).

Initially, all the `caption-1` instances are suffixed with the *prefix.lang1* while *prefix.lang2* for the `caption-2` where *prefix* $\in$ (train, validation, test) and *lang* $\in$ (*en*, *hi*, *te*). Furthermore, during the many-to-one (**m2o**) training, all the *en* instances of the train and validation are merged into a single compound target language while all the non-English instances are merged as the source language. Here, we do not append any artificial target token at the source side to denote the target language as English is the sole target language. On the other hand, for the one-to-many (**o2m**) training all the *en* instances of the train and validation are merged into as the source language while all the non-English instances are merged as the target language. In this case, an artificial target language token <__*tgt*__*lang*> is appended at the beginning of the source sentence to denote the target language *lang* $\in$ (*hi1*, *te1*, *hi2*, *te2*).

3.2 Neural Machine Translation (NMT)

NMT is an encoder-decoder based sequence-to-sequence approach to machine translation which jointly models the conditional probability $p(\mathbf{y}|\mathbf{x})$ to translate a target sequence, $\mathbf{y} = \{y_1, \ldots, y_m\}$ given a source sequence, $\mathbf{x} = \{x_1, \ldots, x_n\}$ as:

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{j=1}^{m} p(y_j|\mathbf{y}_{<j}, \mathbf{x};\theta),$$

where θ is the set of learnable model parameters. Furthermore, the model objective is to maximize the log-likelihood $\mathcal{L}$ w.r.t θ by the following equation:

$$\mathcal{L}_\theta = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} log\; p(\mathbf{y}|\mathbf{x};\theta), \qquad (1)$$

where $\mathcal{D}$ is the parallel corpus. RNN (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015), CNN (Gehring et al., 2017) and Transformers (Vaswani et al., 2017) are the popular choice of encoder-decoder models. In this work, we use RNN with cross attention between the encoder and the decoder.

3.3 Multilingual NMT (MNMT)

Multilingual NMT facilitates the translation between multiple languages via pivot based (Dabre et al., 2015), transfer learning (Zoph et al., 2016) or through a jointly trained single NMT model (Johnson et al., 2017). In this work, we utilise the jointly

[1] Further details are provided in the Dataset section.

trained single multilingual NMT model. Additionally, this single MNMT can be further divided into three types according to the mapping of the source and the target languages:

1. **Many-to-one** (**m2o**). In this setting, the model is trained to translate multiple source languages into a single target language.
2. **One-to-many** (**o2m**). This MNMT model translates from a single source language to multiple target languages.
3. **Many-to-many** (**m2m**). Here, translation between many source and many target languages is possible.

Moreover, as there are several target languages in the **o2m** and **m2m**, a target language tag is appended at the beginning of the source sentence to specify the predicted target language. Given K sentence pairs and L language pairs the training objective of an MNMT model is to maximise the log-likelihood over the whole parallel pairs $\{\mathbf{x}^{(l,k)}, \mathbf{y}^{(l,k)}\}_{k\in(1,\dots,K_l)}^{l\in(1,\dots,L)}$ as:

$$\mathcal{L}_\theta = \frac{1}{K}\sum_{l=1}^{L}\sum_{k=1}^{K_l} log\ p(\mathbf{y}^{(l,k)}|\mathbf{x}^{(l,k)};\theta), \qquad (2)$$

where the total parallel sentences $K = \sum_{l=1}^{L} K_l$.

3.4 Multimodal Machine Translation (MMT)

We follow the Calixto et al. (2017a) MMT model, an expansion of the attention-based NMT framework (Bahdanau et al., 2014), where visual features are incorporated. Spatial features extracted from the image using pre-trained CNNs are incorporated with an attention mechanism in the decoder. When generating the target words, the model learns to independently input textual and visual context using separate attention mechanisms in a single decoder RNN. The decoder RNN is conditioned on the image, source sentence, previous hidden state, and previously emitted words through an independent attention mechanism. Given a new multimodal hidden state $\mathbf{s}_j$, the previous emitted word $\mathbf{y}_{<j}$, context vector $\mathbf{c}_j$ from encoder source sentence and context vector $\boldsymbol{i}_j$ from image features, the decoder computes a new probability to generate a target word using Equation 3.

$$\begin{aligned} p(y_j = k|\mathbf{y}_{<j}, C, A) = softmax(\boldsymbol{L}_o tanh(\\ \boldsymbol{L}_s\mathbf{s}_j + \boldsymbol{L}_w\boldsymbol{E}_y[y_{j-1}]\boldsymbol{L}_{cs}\mathbf{c}_j + \boldsymbol{L}_{ci}\boldsymbol{i}_j)) \end{aligned} \qquad (3)$$

where $\boldsymbol{L}_o$, $\boldsymbol{L}_s$, $\boldsymbol{L}_w$, $\boldsymbol{L}_{cs}$, and $\boldsymbol{L}_{ci}$ are projection matrices.

3.5 Multilingual Multimodal MT (M2MT) with Multiple Captions (MC)

Using the multilingual multi-caption augmented corpus discussed in Section 3.1, a multi-modal machine translation model is trained in **o2m** and **m2o** fashion. Both the visual information fused **o2m** and **m2o** models are trained by maximizing the following log-likelihood:

$$\mathcal{L}_\theta = \frac{1}{K}\sum_{l=1}^{L}\sum_{k=1}^{K_l} log\ p(\mathbf{y}^{(l,k)}|\mathbf{x}^{(l,k)}, \boldsymbol{i}^k;\theta), \qquad (4)$$

where $\boldsymbol{i}^k$ is the image feature corresponding to the k^{th} caption.

4 Experimental setup

4.1 Dataset

In this work, we use the `Multi-30K` (Elliott et al., 2016) corpus. Specifically, the first two caption descriptions (`caption-1` and `caption-2`) of *task2*[2] are used and evaluated upon the `test_2016` test set. Since our work is the multilingual machine translation for *en* to/from {*hi*,*te*}, we use a publicly available pre-trained machine translation model (Ramesh et al., 2021) to translate the English data of the `Multi-30K` to *hi* and *te*, which is further used as the parallel corpus for our experimentation.

During the evaluation process, we use the `caption-1` instance of the test set only.

4.1.1 Text Preprocessing

The text preprocessing step initially tokenizes the raw texts. English side data is tokenized using the *moses-scripts*[3] while the *hi* and *te* data are normalized and tokenized using the `IndicNLP` toolkit[4]. Additionally, the *te* side data is transliterated into Devanagari script (the same script as the *hi*). Finally, sentencepiece (Kudo and Richardson, 2018) BPE (Sennrich et al., 2016) of 10,000 subword merge operations is learnt across all the training data jointly from both the captions to increase the vocabulary coverage. The transliteration of *te* into the *hi* script further maximises the common

[2] https://github.com/multi30k/dataset/tree/master/data/task2

[3] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

[4] https://github.com/anoopkunchukuttan/indic_nlp_library

word/subword overlaps even with smaller vocabulary size. Furthermore, using a common script prevents subword vocabulary fragmentation between the *hi* and *te* and improves lexical sharing among the languages (Ramesh et al., 2021). We use the same vocabulary throughout the experimentation.

4.2 Training Settings

NMT System: We use attention (Luong et al., 2015) based LSTM (Hochreiter and Schmidhuber, 1997) having 2 layers encoder-decoder network with 500 word embedding vector and hidden vector size. Stochastic Gradient Descent is used to optimise the training, which is trained for a maximum of 200,000 update steps, validated after every 10,000 update steps, and halted after 10 consecutive stalls in the validation perplexity. During the decoding time, the <UNK> tokens are replaced by a source word with the highest alignment score, and the translation is generated using beam search (Tillmann and Ney, 2003) with a beam size of 5. Finally, the training is done on `OpenNMT-Py` (Klein et al., 2017).

MMT System: The MMT system is trained by incorporating the image features extracted using a VGG19-CNN pre-trained model and the processed text with a batch size of 64 until validation perplexity doesn't improve for 10 consecutive epochs. A bidirectional LSTM encoder with 2 layers and a dropout with a probability of 0.2 in both source and target word embeddings is used.

4.3 Baselines

We compare our approach with the following baselines:

1. **Single caption bilingual models (NMT)**. The first baseline is the four bilingual models, i.e. *en-hi*, *en-te*, *hi-en* and *te-en* trained only on `caption-1`.

2. **Multiple caption bilingual models (NMT+MC)**. The second baseline is the four bilingual models trained on both `caption-1` and `caption-2`.

3. **Single caption text only multilingual models (MNMT)**. The third baseline is the two multilingual models (**m2o** and **o2m**) trained on `caption-1` only.

4. **Multiple caption text only multilingual models (MNMT+MC)**. Multilingual models (**m2o** and **o2m**) jointly trained on both the captions.

5. **Single caption multilingual multimodal models (M2MT)**. Multilingual models (**m2o** and **o2m**) trained on `caption-1` only along with the visual features.

4.4 Evaluation

The systems are evaluated using both the automatic and human evaluation approach.

4.4.1 Automatic Evaluation

To report the automatic evaluation score, we use the BLEU (Papineni et al., 2002) and chrF both computed using the SacreBLEU (Post, 2018) implementation.

1. **BLEU**: The BLEU score is reported over the geometric mean of the 4-gram precision or BLEU-4, ranging from 0-100, with 100 being the highest. The hypothesis is detokenized and then retokenized using SacreBLEU's inbuilt mteval-13a tokenizer[5] for the Indic to English evaluation. Meanwhile, for the English to Indic translation evaluation, the hypothesis is detokenized and then retokenized using the IndicNLP tokenizer and then evaluated without using any tokenizer in SacreBLEU[6]. Furthermore, the Telugu translation is transliterated back to the Telugu script for the assessment.

2. **chrF**: The processing step for reporting the evaluation score of the chrF[7] is similar to that of the BLEU. Additionally, the chrF score scales from 0-1, where the perfect translation gets a score of 1.

4.4.2 Human Evaluation

Human evaluation is carried out by considering the fluency and adequacy of the translated output. In this pretext, a human translator fluent in English and Hindi is assigned to separately rate each sentence from 1-5 for the fluency and the adequacy criteria. Finally, the sentence wise scores are averaged to get the corpus level score for both the criteria.

[5] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1

[6] BLEU+case.mixed+numrefs.1+smooth.exp+tok.none+version.1.5.1

[7] chrF2+numchars.6+space.false+version.1.5.1

Systems	*en-hi*		*hi-en*		*en-te*		*te-en*	
	BLEU-4	chrF	BLEU-4	chrF	BLEU-4	chrF	BLEU-4	chrF
NMT-full	35.3	0.52	38.9	0.57	27	0.56	38.8	0.58
MNMT-full	48.6	0.62	47	0.65	31.2	0.59	41.1	0.6
M2MT-full	54.4	0.7	55.1	0.72	37.8	0.66	48.3	0.67

Table 1: Overall scores of the systems

Systems	*en-hi*	*hi-en*	*en-te*	*te-en*
NMT	33.2	34.5	24.9	31.3
NMT + MC	35.3	38.9	27	38.8
MNMT	45.5	43.6	29	36.4
MNMT + MC	48.6	47	31.2	41.1
MNMT + V (M2MT)	51.9	53.3	35	46.1
MNMT + V + MC (M2MT + MC)	**54.4**	**55.1**	**37.8**	**48.3**

Table 2: Ablation study of the systems based on the BLEU-4 scores.

5 Results and Analysis

This section presents the quantitative results obtained by all the models and their analysis.

5.1 Overall Result of the Automatic Evaluation

Table 1 compares the overall scores of the proposed multimodal system (M2MT-full) to the bilingual (NMT-full) and multilingual (MNMT-full) baseline variants in terms of BLEU-4 and chrF scores. M2MT-full significantly improves in BLEU and chrF scores over the NMT-full and MNMT-full baselines, with +19.1, +5.8 BLEU and +0.18, +0.08 chrF improvements for the *en-hi* translation, +16.2, +8.1 BLEU and +0.15, +0.07 chrf improvements for the *hi-en* translation. Similarly, +10.8, +6.6 BLEU and +0.1, +0.07 chrf improvements for the *en-te* translation. And, +9.5, +7.2 BLEU and +0.09, +0.07 chrf improvements for the *te-en* translation.

5.2 Ablation Study

Table 2 summarises the ablation study and quantifies the contribution of the components, namely multilinguality, multiple captions (MC), and visual features (V).

1. **Effect of the multiple languages**: In Table 2, the multilingual system (MNMT) achieves a substantial gain over the bilingual system (NMT) in BLEU scores for all translation directions. The *en-hi* translation benefits the most from the MNMT with a +12.3 BLEU improvement over the NMT followed by *hi-en* translation with +9.1 increment. Hence for this dataset, the multilinguality is effective for all the translation directions and more prominent when *hi* language is involved. This substantial gain in BLEU score in the multilingual setting can be credited to the cross-lingual information shared between Hindi and Telugu. Furthermore, the transliteration of Telugu into the Hindi script maximizes the subword overlapping during the vocabulary buildup, thus increasing the word coverage of the model as a whole. We illustrate the word coverage in the form of rare word translation accuracy in Section 5.4.

2. **Effect of the multiple captions**: We utilise the multiple captions (**MC**) provided in the `Multi-30K` dataset in our models. In doing so, a substantial gain in the BLEU score is observed when multiple captions are added over the single caption models for all the translation directions. We hypothesize that the multiple captions increase the models' generalizability by imposing a lexical enhancement as it introduces diverse word forms for the same context. Further, the increase in lexical choice makes the system more robust in handling the rare word translation.

3. **Effect of the visual features**: The addition of the visual features to the MNMT system further outperforms all the text only models even without the inclusion of the multiple captions. Furthermore, the translation accuracy of the

Systems	Adequacy	Fluency
NMT	1.85	2.95
NMT + MC	2.1	3
MNMT	2.65	2.95
MNMT + MC	3	2.85
M2MT	3.15	3.05
M2MT + MC	3.2	3.55

Table 3: Human evaluation of the systems for English to Hindi translation.

words based on their frequency in the training corpus, which we discuss in Section 5.4, reveals that visual features are beneficial for handling the rare word translation.

5.2.1 Human Evaluation Results

Apart from the automatic scores reported in Table 1 and Table 2, human evaluation result for the English to Hindi translation of all the systems is also reported in Table 3. The evaluation is conducted based the fluency and adequacy criteria. As such, both the human evaluation and the automatic scores correlates well suggesting the effectiveness of the proposed method. However, the highest adequacy and fluency scores (3.2 and 3.55 of **M2MT+MC**) is far lesser than the upper limit of the scores (a score of 5). On the other hand, the automatic scores are relatively higher. This is due to training the systems on the synthetic data as only the English side training data is real, as a result the errors in the training data are accumulated to the translated outputs. However, the human evaluator catches these errors and hence the low adequacy and fluency scores.

5.3 Qualitative Analysis

We also present the qualitative analysis of the models based on the translation outputs. In doing so, we use certain abbreviations for the readability purpose of the non-English sentences: TT is the English transliteration, Gloss denotes the word-to-word English translation, and ET denotes the English translation for the same. Moreover, texts are also colour coded based on the similarity/dissimilarity of the translation with the reference. Blue colour depicts the exact match between the reference and the translation. Meanwhile, cyan and red are errors with red denoting wrong translation and the outputs with good translation but absent in reference is denoted by cyan. The first source sentence (**English Source**) in Figure 4 presents the English to Hindi translations while the second source sentence (**Hindi Source**) illustrates the Hindi to English translations of the models.

In Figure 4 for the English to Hindi translation, both the bilingual variants wrongly generates the phrases "peeth karate hain" (with their backs) by the **NMT** and "kaimare ke saath daudate hain" (run with the camera) by the **NMT+MC** which is highlighted in red color, thus changing the actual meaning of the source sentence. Additionally, the **M2MT** system generates "yaard" (yard) instead of the "maidaan" (field) as in the reference, which is technically correct but penalised by the automatic score. Finally, **M2MT+MC** is the only system to generate the correct determiner "ek" (a). Apart from these, all the system outputs follow the subject-object-verb (SOV) word order of the target language (Hindi), irrespective of preserving the actual context of the reference. However, there are cases of interchanging the present continuous tense to simple present tense form such as generation of "daudate hain" (run) instead of "daud rahe hain" (running) as in the reference. In this regard, **MNMT**, **M2MT** and **M2MT+MC** systems produce the correct generation.

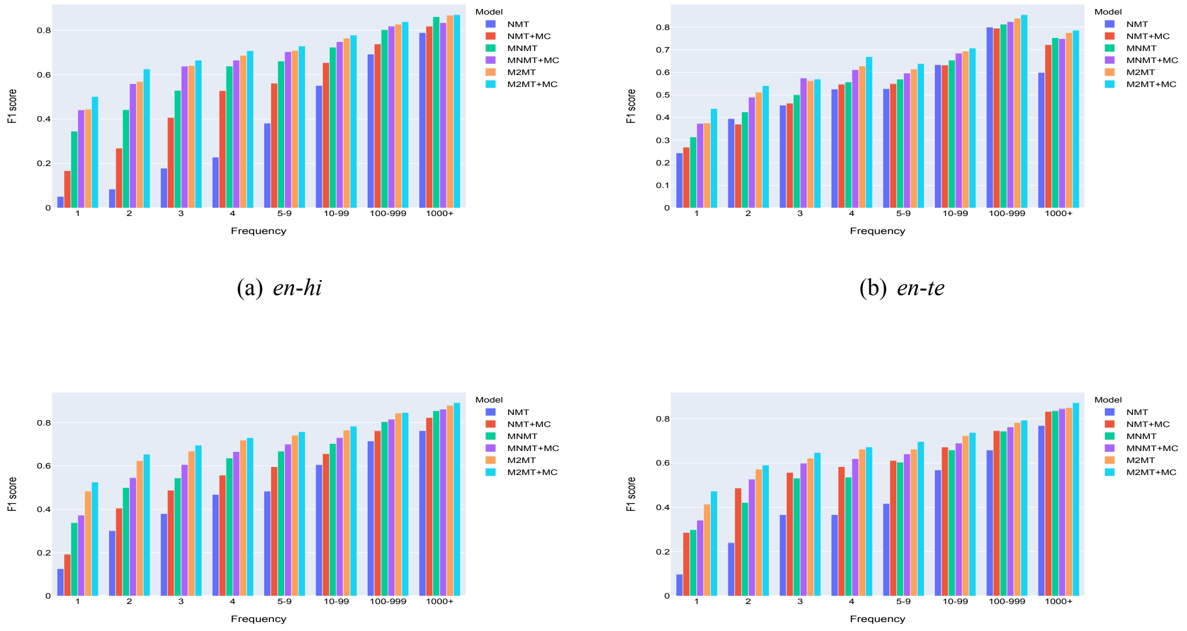
On the other hand, the Hindi to English translation presented in Figure 4 highlights the effectiveness of the multiple captions (MC) by maximizing the coverage of the reference words in the output, thus making the translation more adequate. Additionally, it is observed that except for the **MNMT+MC** all other systems wrongly generates the pronoun "his" instead of "its", which further highlights the gender biasedness of the systems. However, the translated output of **M2MT+MC** covers most of the reference words. It is syntactically correct (apart from the wrong pronoun "his"), thus making the translation more adequate and fluent than the outputs of the other baselines.

5.4 Error Analysis

We conduct the error analysis of the systems considering the translation accuracy of the words based on their frequency in the training corpus, which is illustrated in Figure 1. The translation accuracy provides an insight into the word coverage and how well the system handles the translation of rare words. The proposed multimodal system **M2MT+MC** improves the translation accuracy for the entire vocabulary and the rare words to the training corpus in particular. The **NMT**

Input Image	**Model Outputs**
 en-hi Translation	**English Source:** Three little children in a grassy yard running towards the camera. **Reference:** एक घास के मैदान में तीन छोटे बच्चे कैमरे की ओर दौड़ रहे हैं। TT: ek ghaas ke maidaan mein teen chhote bachche kaimare kee or daud rahe hain. Gloss: one grass's field in three little children camera towards running. ET: Three little children are running towards the camera in a grassy field. **NMT:** घास में तीन छोटे बच्चे कैमरे की ओर पीठ करते हैं। TT: ghaas mein teen chhote bachche kaimare kee or peeth karate hain. Gloss: grass in three little children camera towards back do. ET: Three little children in the grass with their backs to the camera. **NMT+MC:** घास के मैदान में तीन छोटे बच्चे कैमरे के साथ दौड़ते हैं। TT: ghaas ke maidaan mein teen chhote bachche kaimare ke saath daudate hain. Gloss: grass's field in three little children camera with run. ET: Three little children run with the camera in the grassy field. **MNMT:** घास के मैदान में तीन छोटे बच्चे कैमरे की ओर दौड़ रहे हैं। TT: ghaas ke maidaan mein teen chhote bachche kaimare kee or daud rahe hain. Gloss: grass's field in three little children camera towards running. ET: Three little children are running towards the camera in the grass field. **MNMT+MC:** घास के मैदान में तीन छोटे बच्चे कैमरे की ओर दौड़ते हैं। TT: ghaas ke maidaan mein teen chhote bachche kaimare kee or daudate hain. Gloss: grass's field in three little children camera towards run. ET: Three little children run towards the camera in the grassy field. **M2MT:** घास के यार्ड में तीन छोटे बच्चे कैमरे की ओर दौड़ रहे हैं। TT: ghaas ke yaard mein teen chhote bachche kaimare kee or daud rahe hain. Gloss: grass's yard in three little children camera camera towards running. ET: Three little children are running towards the camera in the grassy yard. **M2MT+MC:** घास के एक मैदान में तीन छोटे बच्चे कैमरे की ओर दौड़ रहे हैं। TT: ghaas ke ek maidaan mein teen chhote bachche kaimare kee or daud rahe hain. Gloss: grass's one field in three little children camera camera towards running. ET: Three little children are running towards the camera in a grassy field.
 hi-en Translation	**Hindi Source**: भूरे रंग का कुत्ता अपने दांतों के साथ दूसरे कुत्ते के गाल पर पकड़ रखता है। TT: bhoore rang ka kutta apane daanton ke saath doosare kutte ke gaal par pakad rakhata hai. Gloss: brown colour's dog its teeth with other dogs cheeks on holds. ET: The brown dog holds the other dog's cheek with its teeth. **Reference**: The brown dog has a hold of the other dogs cheek with its teeth. **NMT:** The brown dog is holding the other dog on the other dog with his teeth. **NMT+MC:** The brown dog has his teeth on the cheek of the other dog. **MNMT:** The brown dog holds on the cheek of his teeth with his teeth. **MNMT+MC:** The brown dog holds on to another dog with its teeth. **M2MT:** The brown dog grabs the other dog with his teeth on the cheek. **M2MT+MC:** The brown dog holds on to the cheek of another dog with his teeth.

Table 4: Sample input and output for the *en-hi* translation.



(a) *en-hi*

(b) *en-te*

(c) *hi-en*

(d) *te-en*

Figure 1: F1 accuracy of target word translation based on the word frequency in the training corpus.

without any embellishments lags behind the accuracy of **NMT+MC** for the rare word translation. Similarly, the performance of all other model variants degrades when the multiple captions are not included and further strengthens the role of **MC** in the improvement of the translation quality as a whole. Additionally, the multilingual system (**MNMT**) has higher accuracy for rare word translation than both the bilingual variants (**NMT** and **NMT+MC**) for all translation directions. However for the *te-en* translation, **NMT+MC** gives a competitive performance with **MNMT** for the rarest word and sometimes better for relatively frequent words.

6 Conclusion

This work presents the multimodal machine translation embellished with multiple captions in a multilingual setup. We find that the multiple captions benefit both the text-only and the multimodal models by introducing lexical diversity, making the system more robust to handle the rare word translation and thus increasing the translation quality. Additionally, the visual features which provide the context information further alleviate the translation quality. However, this work is experimented on synthetic data, hence the trained systems possesses the accumulated errors present in the training data. In our future work, we intend to address this issue with some counter measures.

Acknowledgments

This work is supported by Scheme for Promotion of Academic and Research Collaboration (SPARC) Project Code: P995 of No: SPARC/2018-2019/119/SL(IN) under Ministry of Education (erstwhile MHRD), Govt. of India. The authors would like to thank the anonymous reviewers for providing critical suggestions. The authors also thank the volunteers for their help in human evaluation tasks.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–

1548, Hong Kong, China. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost Van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? *arXiv preprint arXiv:1605.09186*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017a. 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)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017b. Incorporating global visual features into attention-based neural machine translation. *arXiv preprint arXiv:1701.06521*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Bernardo Stearns, Arun Jayapal, Sridevy S, Mihael Arcan, Manel Zarrouk, and John P McCrae. 2019. Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland. European Association for Machine Translation.

Raj Dabre, Chenhui Chu, Fabien Cromieres, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. Large-scale dictionary construction via pivot-based statistical machine translation with significance pruning and neural network features. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 289–297, Shanghai, China.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *ArXiv*, abs/2010.11125.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 639–649, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2020. How to ask good questions? try to leverage paraphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6130–6140, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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.

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.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, 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. *arXiv preprint arXiv:1907.08948*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Alok Singh, Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021a. Generation and evaluation of hindi image captions of visual genome. In *Proceedings of the International Conference on Computing and Communication Systems*, pages 65–73, Singapore. Springer Singapore.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021b. An encoder-decoder based framework for hindi image caption generation. *Multimedia Tools and Applications*.

Salam Michael Singh and Thoudam Doren Singh. 2020. Unsupervised neural machine translation for English and Manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78, Suzhou, China. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Christoph Tillmann and Hermann Ney. 2003. Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, 29(1):97–133.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2019. Paraphrases as foreign languages in multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 113–122, Florence, Italy. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.