# Multimodal Neural Machine Translation for Low-resource Language Pairs using Synthetic Data

**Koel Dutta Chowdhury**
**ADAPT Centre**
**School of Computing**
**Dublin City University**
**Dublin, Ireland**
koel.chowdhury@adaptcentre.ie

**Mohammed Hasanuzzaman**
**ADAPT Centre**
**School of Computing**
**Dublin City University**
**Dublin, Ireland**
mohammed.hasanuzzaman@adaptcentre.ie

**Qun Liu**
**ADAPT Centre**
**School of Computing**
**Dublin City University**
**Dublin, Ireland**
qun.liu@adaptcentre.ie

## Abstract

In this paper, we investigate the effectiveness of training a multimodal neural machine translation (MNMT) system with image features for a low-resource language pair, Hindi and English, using synthetic data. A three-way parallel corpus which contains bilingual texts and corresponding images is required to train a MNMT system with image features. However, such a corpus is not available for low resource language pairs. To address this, we developed both a synthetic training dataset and a manually curated development/test dataset for Hindi based on an existing English-image parallel corpus. We used these datasets to build our image description translation system by adopting state-of-the-art MNMT models. Our results show that it is possible to train a MNMT system for low-resource language pairs through the use of synthetic data and that such a system can benefit from image features.

## 1 Introduction

Recent years have witnessed a surge in application of multimodal neural models as a sequence to sequence learning problem (Sutskever et al., 2014; Kalchbrenner and Blunsom, 2013; Cho et al., 2014b) for solving different tasks such as machine translations (Huang et al., 2016), image and video description generation (Karpathy and Fei-Fei, 2015; Kiros et al., 2014; Donahue et al., 2015; Venugopalan et al., 2014), visual question answering (Antol et al., 2015), etc. However, neural machine translation (NMT), which is an inherently data-dependent procedure, continues to be a challenging problem in the context of low-resourced and out-of-domain settings (Koehn and Knowles, 2017). In other words, there is a concern that the model will perform poorly with languages having limited resources, especially in comparison with well-resourced major languages.

Although English(En) and Hindi(Hi) languages belong to the same family (Indo-European), they differ significantly in terms of word order, syntax and morphological structure (Bharati et al., 1995). While English maintains a Subject-Verb-Object (SVO) template, Hindi follows a Subject-Object-Verb (SOV) convention. Moreover, compared to English, Hindi has a more complex inflection system, where nouns, verbs and adjectives are inflected according to number, gender and case. These issues, combined with the data scarcity problem, makes Hi→En machine translation a challenging task.

Bilingual corpora, which are an important component for machine translation systems, suffer from the problem of data scarcity when one of the languages is resource-poor. To achieve better quality translation, a potential solution is to extend along the language dimension to construct bilingual corpora. In particular, for a distant language pair such as Hindi and English, building a bilingual corpus can prove to be a useful endeavor in multiple aspects.



*Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42

We are inspired by the recent successes of using visual inputs for translation tasks (see Section 2 for relevant studies). For translating image descriptions, given both the source image and it's description, it can be seen that both modalities can bring more useful information for generating the target language description. With the goal of preventing a low-resource language such as Hindi from being left behind in the advancement of multimodal machine translation, we take the first steps towards applying MNMT methods for Hi→En translation.

Our contributions in this study are as follows:

- To the best of our knowledge, we are the first to tackle the problem of multimodal translation from Hindi into English.
- We examine if visual features help to improve machine translation (MT) performance in low resource scenarios.
- We investigate whether the multimodal machine translation system for less-resourced language can benefit from synthetic data.
- We augment the Flickr30k dataset with synthetic Hindi descriptions, obtained from a MT system.
- We manually develop a validation and test corpus of the English counterpart in the Flickr30k dataset. We plan to release this dataset publicly for research purposes.

This paper is divided as follows: Section 2 provides the necessary background and establishes the relevance of the presented work, both in terms of low-resourced MT and MT in multimodal contexts. Section 3 describes the overall methodology. In Section 4 we outline the backgrounds of datasets used for training, validation and testing. Section 5 provides detailed descriptions of the multimodal models used in our experiments. Section 6 details the experimental set-ups. Results and analysis are presented in Section 7. Finally, in Section 8, we provide conclusions and indicate possible directions for future work.

## 2 Related Work

There has been some previous work on using visual context in tasks involving both neural machine translation (NMT) and image description generation (IDG) that explicitly uses an encoder-decoder framework as an instantiation of the sequence to sequence (seq2seq) learning problem (Cho et al., 2014a). Vinyals et al. (2015) proposed an IDG model that uses a vector, encoding the image as input based on the sequence-to-sequence framework. Specia et al. (2016) introduced a shared task to investigate the role of images in Multi-modal MT. Similarly, Huang et al. (2016) introduced a model to associate textual and visual features extracted with the VGG19 network for translation tasks (Simonyan and Zisserman, 2014). Elliott et al. (2015) generated multilingual image descriptions using image features transferred from separate non-attentive neural image description models. Calixto et al. (2017a) carried out experiments to incorporate spatial visual information into NMT using a separate visual attention mechanism. Although these approaches have demonstrated the plausibility of multilingual natural language processing with multiple modalities, they rely exclusively on the availability of a large three-way parallel corpus (bilingual captions corresponding to the image) as training data.

Having enough parallel corpora is a big challenge in NMT and it is very unlikely to have millions of parallel sentences for every language pair. Therefore, quite a few attempts have been made to build NMT systems for low-resource language pairs (Sennrich et al., 2016; Zhang and Zong, 2016) which focused on building NMT systems in a low-resource scenario. They incorporated huge monolingual corpus in the source or target side. Gulcehre et al. (2017) proposed two alternative methods to integrate monolingual data on target side, namely shallow fusion and deep fusion. In shallow fusion, the top $K$ hypotheses (produced by NMT) at each time step $t$ are re-scored using the weighted sum of the scores given by the NMT(trained on parallel data) and a recurrent neural network based language model (RNNLM). Whereas in deep fusion, hidden states obtained at each time step $t$ of RNNLM and NMT are concatenated

and output is generated from that concatenated state.

Sennrich et al. (2016) incorporated monolingual data on the target side to investigate two methods of filling the source side of the monolingual data. In the first method, they used a dummy source sentence for every target sentence, while in the second method synthetic source sentences were obtained via back-translation. Their results found that the second method is more effective. In a similar vein, Zhang and Zong (2016) explored the effect of incorporating large-scale source-side monolingual in NMT in many ways. In the first approach, inspired by Sennrich et al. (2016), they built a baseline system and then obtained parallel synthetic data by translating the monolingual data. This parallel data, along with the original data, is used again for training an attention-based encoder-decoder NMT system. Their second method involved the multi-task learning framework to generate the target translation and the reordered source-side sentences at the same time. They discovered that the use of source-side monolingual data in NMT is more effective than in SMT.

A few other popular approaches in this area involve using a method called transfer learning which focuses on sharing parameters, such as source side word-embeddings across related language pairs. Zoph et al. (2016) focus on training a model on high resource language pair and then using learned parameters to train the low resource language pair. However, it requires selecting closely related high and low resource language pairs. So this approach might not work if the language pairs are distant.

Most of the previous related work on this problem of low-resource NMT has tried to incorporate monolingual data in source or target side. The effect of adding monolingual data in NMT is similar to that of building language model (LM) on large-scale monolingual data in SMT. While in SMT it can make the output more fluent, adding monolingual data does not contribute much in improving adequacy for NMT.

## 3 Methodology Overview

We formulate the task of augmenting the Flickr30k dataset with Hindi descriptions as a multimodal NMT task. The task is defined as follows.

To produce a target side description of an image $i$ in Flickr30k dataset, a MT system may use unimodal information such as text in the form of description for image $i$ in the source language $En$, as well as multimodal information such as text plus visual features embedded in the image $i$ itself. Our overall approach consists of the following steps.

- Due to the unavailability of in-domain Hindi-English parallel corpus for our caption translation task, we use a general domain Hindi-English parallel corpus (referred as $Hi_c - En_c$ hereafter) which is compiled from a variety of existing sources. Details of the dataset are described in Section 4.

- Building a phrase based statistical machine translation(PBSMT) system using $Hi_c - En_c$ parallel corpus. To create a synthetic in-domain Hindi-English parallel corpus for the image descriptions translation task, we translate the English descriptions of Flickr30k dataset (referred to as $E_n$ (Manl.Trans.)) into Hindi, using a PBSMT system. We take motivation for using the PBSMT system over NMT from the work carried out by Kunchukuttan et al. (2017). For $Hi \rightarrow$En translation, their system achieves better results with PBSMT over NMT when trained on the same corpus.

- We divide the $E_n$ (Manl.Trans.) into training, validation and test set and call these as $En_t$ (Manl.Train.Trans.), $En_d$ (Manl.Dev.Trans.) and $En_r$ (Manl.Test.Trans.), respectively. We translate the $En_t$ (Manl.Train.Trans) into Hindi using the PBSMT system and call these as $Hi_t$ (Syn.Train.Trans).We manually translate $En_d$ (Manl.Dev.Trans) and $En_r$ (Manl.Test.Trans) into Hindi. We refer to these manually translated English descriptions as $Hi_d$ (Manl.Dev.Trans) and $Hi_r$ (Manl.Test.Trans).

- We use synthetic training data to build a text-only baseline NMT system. In particular we use $En_t$ (Manl.Train.Trans) and its automatically translated Hindi counter part $Hi_t$ (Syn.Train.Trans) to train the system. In addition to this, we use $Hi_d$ (Manl.Dev.Trans.) and $En_d$ (Manl.Dev.Trans.) to tune the system.

- Visual input may provide right-angled information that is free of the natural language ambiguities and can serve as extraneous information to textual features for machine translation in multimodal scenarios. This motivates us to extract deep visual semantic features from the entire image. We use a pre-trained-convolutional neural network(CNN) model to extract visual global features for all the images in Flickr30k dataset.

- We build MNMT system using the $En_t$ (Manl.Train.Trans.)-$Hi_t$ (Syn.Train.Trans.) parallel corpus and the extracted visual features. We use $Hi_d$ (Manl.Dev.Trans.) and $En_d$ (Manl.Dev.Trans.) to tune the system.

- Finally, we translate $Hi_r$ (Manl.Test.Trans.) into English and measure the performance with reference to $En_r$ (Manl.Test.Trans.)

## 4 Data

**$\mathbf{Hi_c - En_c}$**: In order to generate the synthetic data by means of back-translation, we use the general domain IITB English-Hindi Corpus to train a PBSMT system. The corpus is a compilation of parallel corpora collected from a various existing sources such as OPUS (Tiedemann, 2012), HindEn (Bojar et al., 2014b) and TED (Abdelali et al., 2014) as well as corpora developed at the Center for Indian Language Technology, IIT-B [1] over the years (Kunchukuttan et al., 2017).

**$\mathbf{Hi_t}$(Syn.Trans)** : We divide the English descriptions of Flickr30k dataset consisting of 158,915 sentences into training, development and test sets. The training dataset ($En_t(Manl.Trans)$) contains 156,915 sentences, the development dataset ($En_d(Manl.Trans)$) contains 1,000 sentences and the test set ($En_r(Manl.Trans)$) contains 1000 sentences. We use the PBSMT system trained on $Hi_c - En_c$ to translate the $En_t(Manl.Trans)$ set into Hindi by means of back-translation. Such a strategy represents the case where there is no parallel resources available but domain-specific monolingual data can be translated via an existing MT system and further provided as a training corpus to a new MT system. **$\mathbf{Hi_d}$(Manl.Trans)** and **$\mathbf{Hi_r}$(Manl.Trans)** : for manually curating the dataset, we were assisted by two bilingual speakers of Hindi and English. One of them translated the datasets $En_d(Manl.Trans)$ and $En_r(ManlTrans)$ into Hindi while the other speaker verified the same.

The examples of manually translated descriptions are shown in Table 4.

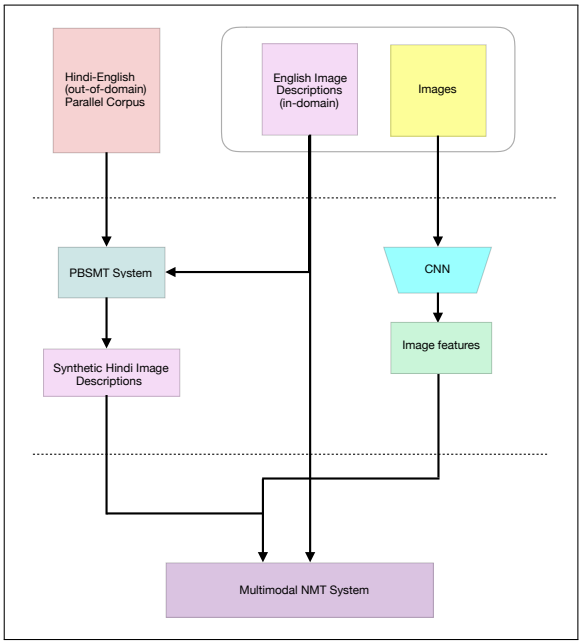


Figure 1: Architecture of the Hi-En MNMT System.

| Data | #Sentences | #Tokens | |
|---|---|---|---|
| | | En | Hi |
| Train | 1,492,827 | 20,667,259 | 22,171,543 |
| Dev | 3207 | 68459 | 74027 |

Table 1: Statistics of data sets used to train PBSMT system

[1] www.cfilt.iitb.ac.in

| Data–set | #sentences |
|---|---|
| Monolingual | 45,075,279 |

Table 2: Additional monolingual (Hi) text used for training the language model to create synthetic Hindi data

| Data–set | #sentences |
|---|---|
| Monolingual | 20,638,520 |

Table 3: Additional monolingual (En) text used for training the general domain PBSMT system

## 5 Multimodal NMT Architecture

In our experiments, we use models which can essentially be thought of as extensions of the attentive NMT framework of Bahdanau et al. (2015). However, following Calixto et al. (2017b) we have included an additional visual component for incorporating the visual features from images. For the encoder, we use a bi-directional recurrent neural network (RNN) with gated recurrent unit (GRU) (Cho et al., 2014a), while the concatenation of forward and backward hidden states, $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$ serves as the final annotation vector for a given source position $i$. In subsections 5.2 and 5.3 we describe the two multi-modal NMT models used in our experiments. For a detailed description of these models, we refer the reader to Calixto et al. (2017b).

### 5.1 Image feature extraction

For all the images, the global image feature vectors, which are the 4096D activations of the penultimate fully connected layer FC7, (henceforth referred to as $\mathbf{q}$), are extracted using a publicly available pre-trained model VGG19-CNN (Simonyan and Zisserman, 2014) which is trained for classifying images into one out of 1000 Imagenet classes (Russakovsky et al., 2015). In our experiment, we pass all images in our dataset through the pre-trained 19-layers VGG network (VGG19-CNN) to extract global image features and incorporate them - (i) to initialise the encoder hidden state and (ii) as additional input to initialise the decoder hidden state.

| **English Source Sentence** | **Hindi Translation** *(Manual)* |
|---|---|
| A man in an orange hat starring at something . | एक नारंगी टोपी में एक आदमी घूर रहा है । |
| People are fixing the roof of a house . | लोग एक घर की छत ठीक कर रहे हैं । |
| Group of Asian boys wait for meat to cook over barbecue . | एशियाई लड़कों का समूह बारबेक्यू पर खाना बनाने के लिए मांस का इंतजार करता है । |
| The person in the striped shirt is mountain climbing . | धारीदार शर्ट में व्यक्ति पहाड़ चढ़ाई कर रहा । |

Table 4: Examples of manual curated captions of the Flickr30k English descriptions in Hindi using PBSMT system. First column represents the original English captions. Second column represents the manually curated English captions in Hindi.

### 5.2 IMG$_\mathbf{E}$: Image for encoder initialization

Instead of initializing the hidden state of the encoder with the zero vector $\overrightarrow{0}$, as in the original attention-based NMT model of Bahdanau et al. (2015) we use two new single-layer feed-forward neural networks to compute the initial states of the forward and backward RNN, respectively.

We use Equation (1) to compute a vector $\mathbf{d}$ from the global image feature vector $\mathbf{q} \in \mathbb{R}^{4096}$:

$$\mathbf{d} = \mathbf{W}_I^2.(\mathbf{W}_I^1.\mathbf{q} + \mathbf{b}_I^1) + \mathbf{b}_I^2. \quad (1)$$

Here $\mathbf{W}$ and $\mathbf{b}$ denote the projection matrix and bias vector, respectively, such that $\mathbf{W}_I^1 \in \mathbb{R}^{4096\times4096}$ and $\mathbf{b}_I^1 \in \mathbb{R}^{4096}$ while $\mathbf{W}_I^2$ and $\mathbf{b}_I^2$ project the image features into the same dimensionality as the hidden states of the source language encoder.

The encoder hidden state is initialized by the feed-forward networks computed as follows:

$$\begin{aligned} \overleftarrow{h}_{\text{init}} &= \tanh(\mathbf{W}_f\mathbf{d} + \mathbf{b}_f), \\ \overrightarrow{h}_{\text{init}} &= \tanh(\mathbf{W}_b\mathbf{d} + \mathbf{b}_b), \end{aligned} \quad (2)$$

where $\mathbf{b}$ and $\mathbf{W}$ are respectively the bias vector and the multi-modal projection matrix for projecting the image features $\mathbf{d}$ into the encoder hidden state's dimensionality. The suffix '$f$' ('$b$') corresponds to forward (backward) states.

### 5.3 $\text{IMG}_\text{D}$: Image for decoder initialization

A new single-layer feed-forward neural network is used for incorporating an image into the decoder. Originally, the initial hidden state of the decoder is computed from the encoder's hidden states, often from concatenation of the last hidden states of the encoder forward RNN and backward RNN, respectively $\overrightarrow{h}_N$ and $\overleftarrow{h}_1$, or from the mean of the source-language annotation vectors $h_i$. However, here we compute the initial hidden state $\mathbf{s}_0$ of the decoder by including the image features as additional inputs as follows:

$$\mathbf{s}_0 = \tanh(\mathbf{W}_{di}[\overleftarrow{\mathbf{h}}_1; \overrightarrow{\mathbf{h}}_N]) + \mathbf{W}_m\mathbf{d} + \mathbf{b}_{di}, \quad (3)$$

where $\mathbf{W}_{di}$ and $\mathbf{b}_{di}$ are learned model parameters while the image feature $\mathbf{d}$ is projected into the decoder hidden state dimensionality by the multi-modal projection matrix $\mathbf{W}_m$.

As before, given the global image vector $\mathbf{q} \in \mathbb{R}^{4096}$, the vector $\mathbf{d}$ is calculated from Equation (1). However, in the present case, the image features are projected into the same dimensionality as the decoder hidden states by the parameters $\mathbf{W}_I^2$ and $\mathbf{b}_I^2$.

## 6 Experiment Set–Up

In this section, we briefly describe the experimental settings used to generate the synthetic Hindi data and further expand it into a multi-modal NMT framework.

The Hindi side of the $Hi_c - En_c$ is normalized using the Indic_NLP_Library[2] to ensure the canonical Unicode representation. We used the scripts from the above library to tokenize and normalize the Hindi sentences. For English, we used the scripts from the Moses tokenizer *tokenizer.perl*[3] to tokenize and lowercase the English representations for our experiments. We use settings similar to that of (Kunchukuttan et al., 2017) to develop $Hi_t$. They used the news stories from the WMT 2014 English-Hindi shared task (Bojar et al., 2014a) as the development(dev) and test corpora which we concatenate together to create our dev set. The training and dev corpora consist of 1,492,827 and 3,207 sentence segments respectively. We used the HindMono corpus (Bojar et al., 2014b) which contains roughly 45 million sentences to build our language model in Hindi. The corpus statistics are shown in Table.1 and Table.2. For training the $Hi_c - En_c$ corpus, we use the Moses SMT system (Koehn et al., 2007) . We use the SRILM toolkit (Stolcke, 2002) for building a language model and GIZA++ (Och and Ney, 2000) with the grow-diag-final-and heuristic for extracting phrases from $Hi_c - En_c$ .The trained system is tuned using Minimum Error Rate Training (Och, 2003). For other parameters of Moses, default values are used. If the sentences in English or Hindi are longer than 80 tokens, they are discarded. To measure the performance of the system, we also translate the $En_r$ testset into $Hi_r$ both manually and automatically.

We also perform Hindi→English (Hi→En) translation using a PBSMT system with the general domain $Hi_c - En_c$ corpus. We use the News Crawl articles 2016 from the WMT17 [4] as additional English monolingual corpora to train the 4-gram language model. This contain roughly 20 million sentence for English. (Table 3).

To build our Multi-modal NMT systems we use OpenNMT-py (the pytorch port of OpenNMT (Klein et al., 2017)) following the settings of Calixto et al. (2017b) which implements the encoder as a bi-directional RNN with GRU, one 1024D single-layer forward RNN and one 1024D single-layer backward RNN. Throughout the experiments, the models are parameterised using 620D source and target word embeddings, and both are trained jointly with the model. All non-recurrent matrices are initialised by sampling from a Gaussian distribution ($\mu = 0, \sigma = 0.01$), re-

[2] https://bitbucket.org/anoopk/indic_nlp_library

[3] https://github.com/moses-smt/mosesdecoder/blob/RELEASE-3.0/scripts/tokenizer/tokenizer.perl

[4] http://www.statmt.org/wmt17/translation-task.html

current matrices are random orthogonal and bias vectors are all initialised to 0. Dropout with a probability of 0.3 in source and target word embeddings, in the image features (in all MNMT models), in the encoder and decoder RNNs inputs and recurrent connections, and before the readout operation in the decoder RNN was applied. Following (Gal and Ghahramani, 2016), dropout to the encoder bidirectional RNN and decoder RNN using the same mask in all time steps are also applied. The models are trained for 25 epochs using Adam (Kingma and Ba, 2015) with learning rate 0.002 and mini-batches of size 40, where each training instance consists of one English sentence, one Hindi sentence and one image.

Finally, we evaluate translation quality quantitatively in terms of BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) and report statistical significance for the metrics using approximate randomisation computed with MultEval (Clark et al., 2011).

## 7 Results and Analysis

### 7.1 Quantitative Analysis

We develop the following five systems -

- **PBSMT**$_{out}$: a phrase based machine translation system trained on the general-domain $\mathbf{Hi_c} - \mathbf{En_c}$ corpus.
- **PBSMT**$_{In}$: a phrase based machine translation system trained on the in-domain $\mathbf{Hi_{it}} - \mathbf{En_{it}}$ corpus.
- **NMT**$_{text}$: a text-only NMT system trained on the in-domain $\mathbf{Hi_{it}} - \mathbf{En_{it}}$ corpus.
- **IMG**$_D$: the multimodal machine system that uses images as an additional input at the decoding stage.
- **IMG**$_E$: the multimodal machine system that uses images to initialise the encoder hidden state.

The comparative evaluation results of our systems are presented in Table 5.

Evaluation is performed against the English translations of the test set using standard MT evaluation metrics, with BLEU and METEOR (multeval implementation, but with METEOR 1.5).

| **Hi → En** | **BLEU** | **METEOR** |
|---|---|---|
| $\text{PBSMT}_{\text{out}}$ | 21.6 | 29.6 |
| $\text{PBSMT}_{\text{in}}$ | 22.7 | 30.2 |
| $\text{NMT}_{\text{text}}$ | 23.3 | 29.7 |
| $\text{IMG}_{\text{D}}$ | **24.2**(↑ 0.9) | **30.7**(↑ 1) |
| $\text{IMG}_{\text{E}}$ | 23.9 | 29.9 |

Table 5: Evaluation metrics scores Hi-En translation systems before and after applying the image features on manually curated dev data. Bold numbers indicate that improvements are statistically significant compared to **NMT**$_{text}$ with p = 0.05

Figure 2: Example from the Flickr30k dataset

We see from the results that the text-only NMT model outperforms phrase based SMT model in terms of BLEU score. Our results indicate that incorporating image features in multimodal models helps, as compared to our text-only SMT and NMT baselines. This is reflected in the fact that both the image models are shown to produce better results in terms of BLEU scores with respect to both the SMT and NMT text-only counterpart.

Although IMG$_E$ yields only little improvement over the text-only NMT counterpart, IMG$_D$ performs consistently better in terms of both metrics (BLEU by ↑ 0.9) and ( METEOR by ↑ 1) than the strong text-only NMT and SMT baseline.

### 7.2 Qualitative Analysis

In order to gain a qualitative insight into specific differences between the text-only and image NMT models, we highlight some instances as follows:

> **English reference:** two people wearing odd alien-like costumes , one blue and one

purple , are standing in a road .

**Manual Source:** तदो लोग अजीब विदेशी जैसी वेशभूषा पहनने, एक नीले और एक बैंगनी, एक सड़क में खड़े हैं।

**NMT:** two people dressed in exotic costumes wear a blue and one flag in a blue , are standing in a road .

**MNMT:** two people wearing funny foreign attire, one blue and one purple , are standing in a street .

In the first entry, although the NMT system without images incorrectly translated the color 'purple' (as can be seen from Figure. 2, where the costumes are clearly in two colors) the multi-modal model translated it correctly, yielding an improvement in the sentence-level BLEU **(↑ 21.47)** score. In terms of translations, we see that both the models extrapolate the reference and translate "alien-like costumes" into "exotic costumes" (text-only model) and as a "funny foreign attire" (multimodal model). We attribute this to the fact that the training set is small and contains different forms of biases and unwarranted inferences (van Miltenburg, 2016).

**English reference:** two young children are on sand.

**Manual Source:** दो छोटे बच्चे रेत पर हैं।

**NMT:** two little kids are on the sand.

**MNMT:** two small children are on sand.

For this particular example, the overall meaning of the source description has been correctly preserved into the target side description for the outputs generated by both models. However, if we closely look into each of the example, we note the difference in entity and its associated attribute. For example, the word-choice for the entity *children* in reference source changes to the term *kids* for text-only NMT but remains intact for MNMT model. Similar trend is observed for the attribute of the entity where the *young* in the reference source is replaced with *little* and *small* for the text and image models respectively. Although every target side entity-attribute pair is semantically close to the source side entity-attribute pair-they may vary in terms of their usage in conventional English language. Compared to the terminology obtained without the help of the image ( *little-kids*- 1417), the one obtained with the help of image ( *small children*-1595) tends to be more widely used in standard spoken English according to the 'Corpus of Contemporary American English' [5].

The above examples clearly asserts the positive impact of multimodal models in translation both in quantitative and qualitative sense.

## 8 Conclusion and Future Work

We presented the results of using synthetic Hindi descriptions of Flickr30k dataset generated via back-translation for multimodal machine translation and provided benchmark baseline results on this corpus.

Our study shows that despite being trained on the same in-domain En–Hi training data, there are inconsistencies in translation quality between the SMT and NMT system, at least in terms of evaluation metrics. These results are not necessarily surprising given that the grammatical syntax between the two languages is poorly represented in the synthetic Hindi training data. In addition to this, Hindi as a language presents many of the well-known issues that NMT currently struggles with (resource sparsity, rich morphology and complex inflection structure). An approach worth considering to address the divergence in word order of the En-Hi language pair is the pre-reordering approach such as the one taken by Ramanathan et al. (2008) to build stronger baseline systems. We will also investigate if incorporating local, spatial-preserving image features can provide more cues to an NMT model as an extension of this work.

In future, we will conduct a more structured study to extend this approach to different language pairs and data scenarios. In addition, we plan to include human evaluation rigorously in our studies to confirm that the MT systems are extended to enhance the translation quality and not simply be tuned to automatic evaluation metrics.

[5] https://corpus.byu.edu/coca/

## Acknowledgments

This research is supported by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology. The ADAPT Centre for Digital Content Technology is founded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. The authors would like to thank Longyue Wang and Meghan Dowling for providing many good suggestions of improvements, as well as our anonymous reviewers for their valuable comments and feedback.

## References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In *LREC*, volume 14, pages 1044–1054.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*, San Diego, California.

Bharati, A., Chaitanya, V., Sangal, R., and Ramakrishnamacharyulu, K. (1995). *Natural language processing: a Paninian perspective*. Prentice-Hall of India New Delhi.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., et al. (2014a). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Bojar, O., Diatka, V., Rychlỳ, P., Stranák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014b). Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC*, pages 3550–3555.

Calixto, I., Liu, Q., and Campbell, N. (2017a). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Conference of the Association for Computational Linguistics: Volume 1, Long Papers*, Vancouver, Canada (Paper Accepted).

Calixto, I., Liu, Q., and Campbell, N. (2017b). Incorporating global visual features into attention-based neural machine translation. *CoRR*, abs/1701.06521.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, Gothenburg, Sweden. The Association for Computer Linguistics.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Elliott, D., Frank, S., and Hasler, E. (2015). Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.

Gal, Y. and Ghahramani, Z. (2016). A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems, NIPS*, pages 1019–1027, Barcelona, Spain.

Gulcehre, C., Firat, O., Xu, K., Cho, K., and Bengio, Y. (2017). On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 639–645.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representation(ICLR)*.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2017). The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2000). Giza++: Training of statistical translation models.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Ramanathan, A., Hegde, J., Shah, R. M., Bhattacharyya, P., and Sasikumar, M. (2008). Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 543–553.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.

van Miltenburg, E. (2016). Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Zhang, J. and Zong, C. (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.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. pages 1568–1575.