Multilingual Multimodal Machine Translation for Dravidian Languages utilizing Phonetic Transcription

Bharathi Raja Chakravarthi, Bernardo Stearns, Mihael Arcan, Manel Zarrouk, and John P. McCrae
Insight Centre for Data Analytics
National University of Ireland Galway
Galway, Ireland
name.surname@insight-centre.org

Ruba Priyadharshini
Saraswathi Narayanan College,
Madurai, India
Arun Jayapal
Smart Insights from Conversations,
Hyderabad, India
S. Sridevy
Tamil Nadu Agricultural University,
Coimbatore, India

Abstract

Multimodal machine translation is the task of translating from a source text into the target language using information from other modalities. Existing multimodal datasets have been restricted to only highly resourced languages. In addition to that, these datasets were collected by manual translation of English descriptions from the Flickr30K dataset. In this work, we introduce MMDravi, a Multilingual Multimodal dataset for under-resourced Dravidian languages. It comprises of 30,000 sentences which were created utilizing several machine translation outputs. Using data from MMDravi and a phonetic transcription of the corpus, we build an Multilingual Multimodal Neural Machine Translation system (MMNMT) for closely related Dravidian languages to take advantage of multilingual corpus and other modalities. We evaluate our translations generated by the proposed approach with human-annotated evaluation dataset in terms of BLEU, METEOR, and TER metrics. Relying on multilingual corpora, phonetic transcription, and image features, our approach improves the translation quality for the under-resourced languages.

1 Introduction

The development of a Multilingual Multimodal Neural Machine Translation (MMNMT) system requires multilingual parallel corpora and images which are aligned with the parallel sentences for training. The largest existing dataset containing captions, images, and translations for English, German, French and Czech is the WMT shared task Multi30K dataset which is derived from the Flickr30k dataset (Plummer et al., 2015; Plummer et al., 2017). Typically this data is manually created with the help of bilingual annotators (Elliott et al., 2016), however, for many languages, such resources are not available. In those cases, machine translation can be a useful tool for the quick expansion to new languages by producing candidate translation (Dutta Chowdhury et al., 2018). In order to reduce the amount of time, we pose translation as a post-editing task. We automatically translated the English sentences from the WMT corpus using a pre-trained general domain Statistical Machine Translation (SMT) and Neural Machine Translation (NMT).

Multilingual NMT models (Firat et al., 2016) have been shown to increase the translation quality for under-resourced languages. Closely related Dravidian languages such as Tamil (ISO-639-1: ta), Kannada (ISO-639-1: kn), and Malayalam (ISO-639-1: ml) exhibit a large overlap in their vocabulary and strong syntactic and lexical similarities. Dravidian languages are a family of languages spoken primarily in the southern part of India and spread over South Asia and are considered as under-resourced languages. However, the scripts used to write these languages are different and they differ in their morphology. Recently Chakravarthi et al. (2019) have shown that phonetic transcription of a corpus into Latin script improves the multilingual NMT performance for under-resourced Dravidian languages.

In this paper, we propose applying Multilingual Multimodal NMT for translating between closely

related Dravidian languages and English. We created multimodal data using SMT and NMT methods, trained on a general domain corpus for under-resourced languages. Combining multilingual and multimodal data along with a phonetic transcription of the corpus improves translation performance for closely related Dravidian languages is shown in the results.

2 Related Work

To capture rich information from multimodal content available on the Web, especially images with descriptions in English was explored in the content of NMT (Specia et al., 2016a) in the WMT shared task. The WMT shared task also provided resources for other popular languages German, Czech and French (Elliott et al., 2017). Most of those data were expensive, for example, English-German corpus was created by Elliott et al. (2016), and cost €23,000 for data collection (€0.06 per word). Such resources are not available for under-resourced languages. Recent work by (Dutta Chowdhury et al., 2018), carried out experiments by utilizing synthetic data for Hindi-English language pair. In contrast, we created MMDravi as a translation post-editing task by utilizing translations of the English sentences associated with images using SMT and NMT trained on general domain data.

The shared task on Multimodal NMT (MNMT) was introduced by Specia et al. (2016b) to generate image descriptions for a target language, given an image and/or a description in the source language. In previous works on MNMT, the researchers utilized visual context by involving both NMT and Image Description Generation (IDG) features that explicitly uses an encoder-decoder (Cho et al., 2014). However, the encoder-decoder architecture encodes the source sentence into a fixed-length vector. To overcome this drawback (Bahdanau et al., 2015) introduced attention mechanism to focus on parts of the source sentence. The work by Calixto and Liu (2017), carried out different experiments to incorporate visual features into NMT by projecting an image feature vector as words into the source sentence, using the image to initialize the encoder hidden state, and using image features to initialize the decoder hidden state. In Calixto et al. (2017), the author incorporated features through a separate encoder and doubly-attentive attention of the decoder to depend on the image feature. This allowed them to predict the next word and showed that the image feature improved the translation quality. Although all these approaches have demonstrated the possibility of MNMT, they rely on manually collected corpora but under-resourced languages do not have such resources. Our work follows the doubly-attentive model (Calixto et al., 2017) with MMDravi data for the multilingual model by phonetic transcription.

In Ha et al. (2016) and Johnson et al. (2017), the authors have demonstrated that multilingual NMT improves translation quality. For this, they created multilingual NMT without changing the architecture by introducing special tokens at the beginning of the source sentence indicating the source language and target language as shown in Figure 1. We follow this by introducing special tokens in the source sentence to indicate the target language. Phonetic transcription to Latin script and the International Phonetic Alphabet (IPA) was studied by (Chakravarthi et al., 2019) and showed that Latin script outperforms IPA for the Multilingual NMT of Dravidian languages. We propose to combine multilingual, phonetic transcription and multimodal content to improve the translation quality of under-resourced Dravidian languages. Our contribution is to use the closely related languages from the Dravidian language family to exploit the similar syntax and semantic structures by phonetic transcription of the corpora into Latin script along with image feature to improve the translation quality.

3 Background

3.1 Dravidian Languages

Dravidian languages have individual writing scripts and have been assigned a unique block in the Unicode computing industry standard. The similarity of these languages is that they are all written from left to right, consist of sequences of simple or complex characters and follow an alpha-syllabic writing system in which the individual symbols are syllables (Bhanuprasad and Svenson, 2008). The languages also have different sets of vowels and consonants. Vowels and consonants are atomic but when they are combined with each other they form consonant ligatures. Dravidian languages such as Tamil do not represent differences between aspirated and unaspirated stops, while other Dravidian languages such as Kannada

and Malayalam have a large number of loan words from Indo-Aryan languages and support a large number of compound characters resulting from the combination of two consonants symbols (Kumar et al., 2015).

3.2 Phonetic Transcription

Phonetic transcription is the use of phonetic symbols such as IPA or non-native script. As the Dravidian languages under study are written in different scripts, they must be converted to some common representation before training the MM-NMT to take advantage of closely related language resources. Phonetic transcription to Latin script and the International Phonetic Alphabet (IPA) was studied by (Chakravarthi et al., 2019) and showed that Latin script outperforms IPA for the Multilingual NMT Dravidian languages. The improvements in translation performance were shown in terms of the BLEU (Papineni et al., 2002) metric. We used the Indic-trans library by Bhat et al. (2015) for phonetic transcription of corpora into the Latin script, which brings all the languages into a single representation by a phoneme matching algorithm. The same library was used to back-transliterate from Latin script to the corresponding Dravidian language to evaluate the translation performance.

3.3 Neural Machine Translation

Neural Machine Translation is a sequence-to-sequence approach (Sutskever et al., 2014) using an encoder-decoder architecture with an attention mechanism (Bahdanau et al., 2015). Given a source sentence X=x_1, x_2, x_3,...x_n and target sentence Y=y_1, y_2, y_3,...y_n the bidirectional encoder transforms the source sentence into annotation vectors C=h_1, h_2, h_3,...h_n. At each time step t, the source context vector c_t is computed based on the annotation vector and the decoder's previous hidden state s_{t-1}. The decoder generates one target word at a time by computing the probability of $P(y_t = k|y_{<t}, c_t)$ given a hidden state s_t as follows

$$P(y_t = k|y_{<t}, c_t) \propto \exp(L_0 tanh(L_s s_t + L_w E_y[y_{t-1}] + L_c c_t)) \quad (1)$$

The L_0,L_s,L_W and L_c are transformation matrices.

The attention model calculates c_t as the weighted sum of the source side context vectors:

$$c_t = \sum_{i=1}^{N} \alpha_{t,i}^{src} h_i \quad (2)$$

$$\alpha_{t,i}^{src} = \frac{\exp(e_{t,i}^{src})}{\sum_{j=1}^{N} \exp(e_{t,j}^{src})} \quad (3)$$

$\alpha_{t,i}^{src}$ is the normalized alignment matrix between each source annotation vector h_i and word y_t to be emitted at a time step t. Expected alignment $e_{t,i}^{src}$ between each source annotation vector h_i and the target word y_t is computed using the following formula:

$$e_{t,i}^{src} = (V_a^{src})^T tanh(U_a^{src} s_t^{'} + W_a^{src} h_i) \quad (4)$$

V_a^{src}, U_a^{src} and W_a^{src} are model parameters.

3.4 Multimodal Neural Machine Translation

The Multimodal NMT (MNMT) (Calixto et al., 2017) model is an extension of the encoder-decoder framework, by incorporating visual information. To incorporate the visual features extracted from the pre-trained model the authors have integrated another attention mechanism to the decoder. The doubly-attentive decoder Recurrent Neural Network is conditioned on the previous hidden state, previously emitted word, source sentence and the image via attention mechanism (Calixto et al., 2017). In the original attention-based NMT model described in Section 3.3, a single encoder for the source sentence, a single decoder for the target sentence and the attention mechanism are conditioned on the source sentence. MNMT integrates two separate attention mechanism over the source language and visual features associated with the source and target sentence. The decoder generates a target word by computing a new probability $P(y_t = k|y_{<t}, C, A)$ given a hidden state s_t, the previously emitted word $y_{<t}$, and the two context vectors c_t from encoder of source sentence and i_t from image features.

$$P(y_t = k|y_{<t}, C, A) \propto \exp(L_0 tanh(L_s s_t + L_w E_y[y_{t-1}] + L_{cs} c_t + L_{ci} i_t)) \quad (5)$$

L_0, L_s, L_w, E_y, L_{cs},and L_{ci} are projection matrices. The mechanism in MNMT is similar to NMT

Source	__opt_src__en __opt_src__kn a group of people standing in front of an igloo .
Target (ISO-639-1: kn)	ಇಗ್ಲೂ ಮುಂದೆ ನಿಂತಿರುವ ಜನರ ಗುಂಪು.
Source	__opt_src__en __opt_src__ta a group of people standing in front of an igloo .
Target (ISO-639-1: ta)	ஒரு குடில் முன் நின்று மக்கள் குழு.
Source	__opt_src__en __opt_src__ml a group of people standing in front of an igloo .
Target (ISO-639-1: ml)	ഇഗ്ലൂ മുന്നിൽ നിൽക്കുന്ന ഒരു കൂട്ടം ആളുകൾ.

Figure 1: Example of sentences with special tokens to indicate the source and target languages.

with an attention model, except for the source sentence and previous hidden state, it also takes the context vector a from the image features using a double attention layer to calculate the current hidden state. The doubly-attentive model calculates the time-dependent vector i_t as follows:

$$i_t = \beta_t \sum_{l=1}^{L} \alpha_{t,l}^{img} a_l \quad (6)$$

Where,

$$\beta_t = \sigma(W_\beta s_{t-1} + b_\beta) \quad (7)$$

The expected alignment vector of image is given by

$$\alpha_{t,l}^{img} = \frac{\exp\left(e_{t,l}^{img}\right)}{\sum_{j=1}^{L} \exp\left(e_{t,j}^{img}\right)} \quad (8)$$

$$e_{t,l}^{img} = (V_a^{img})^T tanh(U_a^{img} s_t^{'} + W_a^{img} a_l) \quad (9)$$

V_a^{img}, U_a^{img} and W_a^{img} are model parameters.

	Corpus Statistics		
Lang pair	***sent***	***s-tokens***	***t-tokens***
En-Ta	0.8M	6.4M	13.3M
En-Kn	0.5M	2.6M	4.5M
En-Ml	1.4M	16.7M	23.5M

Table 1: Statistics of the parallel corpora used to train the general domain translation systems. sent: Number of sentences, s-tokens: Number of source tokens, and t-tokens: Number of target tokens.

4 Experimental Setup

4.1 Data

The images required for our work were collected from Flicker by Plummer et al. (2015). The Multi30K dataset contains parallel corpora for English and German. There were two types of multilingual annotations released by Multi30K dataset (Elliott et al., 2016). The first one is an English description for each image and its German translation. The second is a corpus of five independently collected English and German description pairs for each image. Synthetic data or back-transliterated data have been widely used to improve the performance of NMT and MNMT. To produce a target side description of an image, we create a general domain SMT and NMT for English-Tamil, English-Kannada, and English-Malayalam pairs. We collected the general domain parallel corpora for the Dravidian languages from the OPUS website (Tiedemann and Nygaard, 2004) and (Chakravarthi et al., 2018). The corpus statistics are shown in Table 1. The corpus is tokenized and standardized to lowercase. The general domain SMT was created with Moses (Koehn et al., 2007) while the NMT system was trained with OpenNMT (Klein et al., 2017). After tokenization, we fed the parallel corpora to Moses and OpenNMT. Preprocessed files are then used to train the models. We used the default OpenNMT parameters for training, i.e. 2 layers LSTM with 500 hidden units for both, the encoder and decoder.

	BLEU Score	
Lang pair	***SMT***	***NMT***
En-Ta	30.29	35.52
En-Kn	28.81	26.86
En-Ml	36.73	38.56

Table 2: Results of general domain SMT and NMT translation systems on general domain evaluation set

The SMT and NMT system results on general

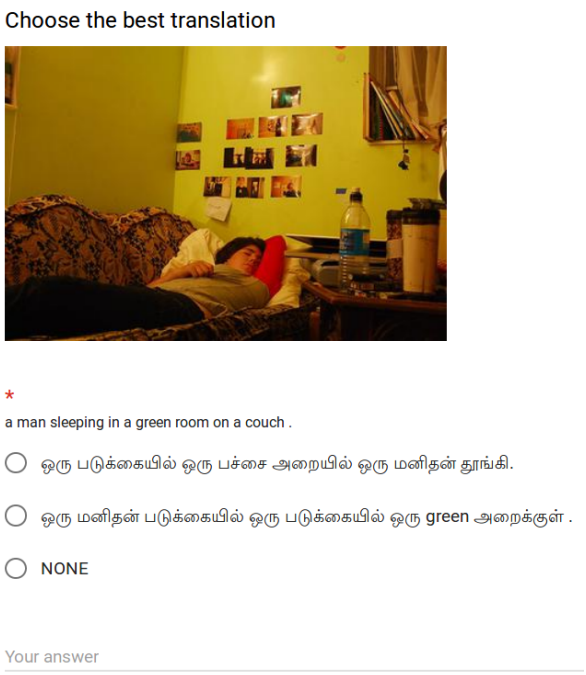

Figure 2: Example of sentence and image with candidate translation to choose.

domain evaluation set are shown in Table 2. The development and test set of the multimodal corpus was collected with the help of volunteer annotators. To reduce the annotation time, we posed the translation task of the development and test set as a post-editing task. We provided the candidate translation of the English sentence from SMT, NMT, and an option to choose the best translation or provide an original translation. Eighteen annotators participated in this annotation process, with different backgrounds, they all are native speakers of the language that they annotated. The data for the Malayalam language was collected from three different native speakers. Ten Tamil native speakers participated in creating data for the Tamil language and five Kannada native speakers annotated for the Kannada language. Since voluntary annotators are scarce and annotate little data, each sentence was annotated by only one annotator. We then selected the system that performed better based on the choice of annotators. We designed an annotation tool to meet the objective of method. We decided to use Google Forms to collect the data from the voluntary annotator's. An example is shown in Figure 2. We chose NMT and used the general domain NMT to post-edit the translation for the training set of MMDravi.

For our tasks, all descriptions in English were converted to lowercase and tokenized, while we

Table 3: Results are expressed in BLEU score: Baseline is Multimodal NMT, MMNMT is trained on native script, and MMNMT-T is trained utilizing phonetic transcription.

Lang pair	BLEUscore		
	Baseline	***MMNMT***	***MMNMT-T***
En-Ta	50.2	51.0	**52.3**
En-Ml	35.6	36.0	**36.5**
En-Kn	44.5	45.1	**45.9**
Ta-En	45.2	47.4	**48.9**
Ml-En	34.3	36.2	**37.6**
Kn-En	50.0	50.2	**50.8**

did not have to bother about the case correction for Dravidian languages (as they do not have cases). We tokenized the Dravidian language using the OpenNMT tokenizer with *segment alphabet* options for Tamil, Kannada, and Malayalam. For the sub-word level representation, we chose the 10,000 most frequent units to train the BPE (Sennrich et al., 2016) model. We used this model for the sub-word level segmentation for the training, development, and evaluation set. We trained the MMNMT model to translate from English into Dravidian languages as well as from Dravidian languages into English. Visual features were extracted from publicly available pre-trained CNN's. Specifically, we extract spatial image features using the VGG-19 network (Simonyan and Zisserman, 2014). In our experiment, we pass all the images in our dataset through the pre-trained VGG-19 layered network to extract global information and use them in a separate visual attention mechanism as described in Calixto et al. (2017).

4.2 Multilingual Multimodal Neural Machine Translation

Since we translate between closely related languages and English, we set up the translation setting in two scenarios, 1) One-to-Many and 2) Many-to-One.

4.2.1 One-to-Many Approach

In this setting, we create a model to translate from English into Tamil, Malayalam, and Kannada. The source language sentence was replicated three times for the three languages with a token indicating target language. Figure 1 shows the example of sentences.

src	a black dog runs on green grass with a toy in his mouth .
ref	ஒரு கருப்பு நாய் வாயில் ஒரு பொம்மையுடன் பச்சை புல்லில் ஓடுகிறது.
MMNMT	ஒரு கருப்பு நாய் வாயில் ஒரு பொம்மையுடன் பச்சை புல் மீது இய ங்கும் .
MMNMT -T	ஒரு கருப்பு நாய் வாயில் ஒரு பொம்மையுடன் பச்சை புல் மீது ஓடுகிறத.

Figure 3: Example showing improvement of translation quality and readability of the translation over baseline model. Errors are shown in red color.

src	a woman and two men , that are dressed professionally, are having a discussion.
ref	ஒரு பெண் மற்றும் இரண்டு ஆண்கள், தொழில்முறை உடையணிந்து, ஒரு விவாதத்தில் உள்ளார்கள்.
MMNMT	ஒரு பெண் , மற்றும் இரு ஆண்கள் professionally ஆயத்தம்பண்ணி , ஒரு விவாதத்திலை வேண்டினோம் .
MMNMT -T	ஒரு பெண் மற்றும் இரண்டு ஆண்கள், professional உடையணிந்து, ஒரு விவாதத்தில் உள்ளார.

Figure 4: Example showing translation with accurate transfer of important information. Errors are shown in red color.

4.2.2 Many-to-One Approach

In the many-to-one MMNMT system, we create a model to translate from Tamil, Malayalam, and Kannada (Dravidian languages) to English. We replicated the English sentence three times for three languages on the target side of the corpus. We then train the MNMT system with a visual feature for individual language level with the MMDravi data. We compared the results with the MMNMT for one-to-many and many-to-one models.

4.3 Results

We applied the baseline bilingual Multimodal NMT systems with respect to the MMDravi data created from the Multi30k dataset. Then we trained our MMNMT and MMNMT-T (phonetic transcription of corpus) for English into Dravidian languages and vice versa. Results are presented in BLEU (Papineni et al., 2002) (BiLingual Evaluation Understudy), which measures the n-gram precision with respect to the evaluation set.

Table 3 provides the BLUE scores for the MMNMT model. We observed that the translation performance of MMNMT is higher compared to the Bilingual Multimodal NMT model in BLEU. Translation from Dravidian to English has the highest improvement in terms of BLEU Score. Our experiments show that the MMNMT system compared with the bilingual system has an improvement in several language directions, which are likely gained from phonetic transcription, image features, and transfer of parameters from different languages.

The results show that for MMNMT with phonetically transcribed corpora, helps more in Dravidian to English than English to Dravidian. An explanation for this is that in the dataset, each source sentence has three targets, which encourages the language model to improve the translation results. In Table 3, we compare the BLEU scores with a baseline approach and our method. In order to evaluate the effectiveness of our proposed model, we have explored MMNMT trained on original scripts and MMNMT trained on a single script. Our empirical results show that the best result is achieved when we phonetically transcribed the corpus and brought it to a single script for both English to Dravidian and Dravidian to English translation tasks.

Figure 3 shows the examples of where the MMNMT model improves the translation quality and readability of the translation over the baseline model. The results given by the human evaluation confirm the results observed in evaluation BLEU metric. The second example for English-Tamil translation of MMNMT system outperforming the baseline is shown in Figure 4. The first example showns an almost perfect translation obtained with the MMNMT system for English to Tamil. In the second example, translation obtained with the MMNMT system is acceptable with the accurate transfer of important information (Coughlin, 2003). This suggests the synthetic data with our MMNMT model can be used in an under-resourced language setting to improve the translation quality.

5 Conclusion

We introduced a new dataset, named MMDravi and proposed a MMNMT method for closely related Dravidian languages to overcome the resource issues. Compared to the baseline approach, the results show that our approach can improve translation quality, especially for Dravidian languages. Our evaluation, using phonetic transcription, multilingual and multimodal NMT, has shown that the proposed MMNMT-T outperforms the existing approach of multimodal, multilingual in low-resource neural machine translation across all the language pairs considered. We plan to release multilingual translations as an addition to Flickr30k set, and explore the effect of the quality of this synthetic data in our future work.

Acknowledgments

This work is supported by a research grant from Science Foundation Ireland, co-funded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289 and the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure and grant agreement No 825182, Prêt-à-LLOD.

References

Bahdanau, Dzmitry, 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*.

Bhanuprasad, Kamadev and Mats Svenson. 2008. Errgrams – a way to improving ASR for highly inflected Dravidian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Bhat, Irshad Ahmad, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. IIIT-H System Submission for FIRE2014 Shared Task on Transliterated Search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

Calixto, Iacer 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. Association for Computational Linguistics.

Calixto, Iacer, 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 (Volume 1: Long Papers)*, pages 1913–1924. Association for Computational Linguistics.

Chakravarthi, Bharathi Raja, Mihael Arcan, and John P. McCrae. 2018. Improving Wordnets for Under-Resourced Languages Using Machine Translation. In *Proceedings of the 9th Global WordNet Conference*. The Global WordNet Conference 2018 Committee.

Chakravarthi, Bharathi Raja, Mihael Arcan, and John P. McCrae. 2019. Comparison of Different Orthographies for Machine Translation of Under-resourced Dravidian Languages. In *Proceedings of the 2nd Conference on Language, Data and Knowledge*.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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, October. Association for Computational Linguistics.

Coughlin, Deborah. 2003. Correlating automated and human assessments of machine translation quality. In *In Proceedings of MT Summit IX*, pages 63–70.

Dutta Chowdhury, Koel, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42, Melbourne, July. Association for Computational Linguistics.

Elliott, Desmond, 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. Association for Computational Linguistics.

Elliott, Desmond, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233. Association for Computational Linguistics.

Firat, Orhan, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875. Association for Computational Linguistics.

Ha, Thanh-Le, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Johnson, Melvin, 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, December.

Klein, Guillaume, 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. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Kumar, Arun, Lluís Padró, and Antoni Oliver. 2015. Joint Bayesian morphology learning for Dravidian languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 17–23, Hissar, Bulgaria, September. Association for Computational Linguistics.

Papineni, Kishore, 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*.

Plummer, Bryan A., 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 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 2641–2649, Washington, DC, USA. IEEE Computer Society.

Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vision*, 123(1):74–93, May.

Sennrich, Rico, 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. Association for Computational Linguistics.

Simonyan, Karen and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Specia, Lucia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016a. 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*, pages 543–553. Association for Computational Linguistics.

Specia, Lucia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016b. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553, Berlin, Germany, August. Association for Computational Linguistics.

Sutskever, Ilya, 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.

Tiedemann, Jörg and Lars Nygaard. 2004. The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).