Zero-shot translation among Indian languages

Rudali Huidrom
Graduate School of IPS
Waseda University
Kitakyushu, Japan
rudali.huidrom@ruri.waseda.jp

Yves Lepage
Graduate School of IPS
Waseda University
Kitakyushu, Japan
yves.lepage@waseda.jp

Abstract

Standard neural machine translation (NMT) allows a model to perform translation between a pair of languages. Multilingual neural machine translation (NMT), on the other hand, allows a model to perform translation between several language pairs, even between language pairs for which no sentences pair has been seen during training (zero-shot translation). This paper presents experiments with zero-shot translation on low resource Indian languages with a very small amount of data for each language pair. We first report results on balanced data over all considered language pairs. We then expand our experiments for additional three rounds by increasing the training data with 2,000 sentence pairs in each round for some of the language pairs. We obtain an increase in translation accuracy with its balanced data settings score multiplied by 7 for Manipuri to Hindi during Round-III of zero-shot translation.

1 Introduction

End-to-end neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014) can be applied to low resource languages with the risk that small amounts of training data result in low translation accuracy (Koehn and Knowles, 2017). Improvement in translation of low resource languages has been reported with the use of multilingual models (Ha et al., 2016; Johnson et al., 2017), back-translation (Sennrich et al., 2016a) and unsupervised learning (Lample et al., 2018).

Initially, MT systems were designed for one single language pair (Johnson et al., 2017). However, NMT systems can be trained simultaneously on many language pairs. This enables translation from and into any of the languages used during training. Dong et al. (2015) first modified an attention-based encoder-decoder model so as to perform multilingual translation from one language to many languages while Luong et al. (2015) used multitask learning for multilingual training.

Firat et al. (2016) introduced the notion of multilingual NMT, by sharing the attention mechanism across several languages. Gu et al. (2018) introduced *universal machine translation,* where a universal representation space is used for all languages. Johnson et al. (2017) introduced *zero-shot translation*: training on multiple source and target languages enables to translate arbitrarily between any of the languages used during training, even between languages for which no sentence pair was ever seen during training. The authors characterised zero-shot translation as "a working example of transfer learning within neural translation models".

Our work consists in testing the use of zero-shot translation for a very low resource language, an Indian language called Manipuri, locally known as Meiteilon, in the context of training with other Indian languages between which no parallel data may exist. Again, Manipuri is a low resource language. It is spoken by about two million people predominantly in the state of Manipur, India. It is an endangered language (Moseley and Nicolas, 2010) from the Sino-Tibetan language family and it shows highly agglutinating word structure. With its language status as endangered, it is one of the two endangered languages of the 8th Schedule of the Indian Constitution. Machine translation for this language is at its infant stage due to the very limited amount of resources available.

We make use of the pmindia dataset[1] (Haddow and Kirefu, 2020). This data set provides monolingual and parallel corpora with English for thirteen Indian languages. We take the following language pairs into consideration: Assamese–English,

[1] http://data.statmt.org/pmindia/

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 47–54

Bengali–English, Hindi–English and Manipuri–English[2].

Our main objective is to measure how much accuracy can be achieved in translation from Manipuri into the three other Indian languages (Assamese, Bengali and Hindi), without using any data from these language pairs, thanks to zero-shot translation. Additionally, we use the JW300 dataset[3] (Agić and Vulić, 2019; Tiedemann, 2012) for Assamese–English language pairs for two rounds of the experiment due to the limited number of data present in the pmindia data set for this language pair. In our experiments, we use only the above-mentioned resources.

Our goal is to improve the translation quality of our zero-shot translation system among the low resourced languages. We propose to control the translation quality by introducing the notion of balanced data in the respective language pairs as a parameter.

The reason why we concentrate on Manipuri is because it is an extremely low resource language: only 7,000 sentence pairs in Manipuri–English are available in the pmindia data set. Developing MT systems with such a small amount of data is a true challenge. Our experiments consist in increasing the training data by groups of 2,000 sentence pairs (Indian language–English), in three rounds. We measure the translation accuracy between Manipuri and other Indian languages in zero-shot translation.

The structure of the paper is as follows. Section 2 describes previous work. Section 3 gives details about the data set used. Section 4 presents the methodology. Section 5 describes the experiments, their results and provides an analysis. Section 6 concludes and proposes future directions.

2 Related work

NMT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014) for a single language pair has been explored extensively over the years. It has been extended to multilingual models (Dong et al., 2015; Luong et al., 2015; Ha et al., 2016; Firat et al., 2016; Johnson et al., 2017) on available multilingual data. One of the approach is that of zero-shot translation (Johnson et al., 2017; Arivazhagan et al., 2019) between language pairs for which no parallel data has been seen during training. Another interesting work addressed by (Johnson et al., 2017; Ha et al., 2016) is the introduction of artificial tokens. It helps in minimizing the architectural changes in the decoder.

Zero-shot machine translation has been explored for low resource languages. Zoph and Knight (2016) proposed an approach for multi-source translation. Their model consists in multiple encoders with a different attention mechanism for each source language. However, this model requires a multi-way parallel corpus for every language pairs, which is hard to obtain, especially for languages with low resource.

NMT is capable of cross-lingual learning (Kim et al., 2019; Zoph and Knight, 2016). This is the motivation for zero-shot translation. Firat et al. (2017) introduced the notion of zero-resource translation. They used a pre-trained multi-way multilingual model and performed fine-tuning with the pseudo parallel data generated by the model. Madaan and Sadat (2020) introduced an approach for improving multilingual NMT for Indian languages. They showed that their model is able to improve the translation for low resource language pairs by leveraging high resource language pairs, thanks to transfer learning.

Our work is closely related to (Johnson et al., 2017): we analyse the performance of multilingual models on our data and perform zero-shot translation as well. The originality in our work is that we aim to improve the translation quality of our model and since we deal with low resource languages, we propose to control the translation quality such that we train, validate and test our model on balanced data sets across all the language pairs.

3 Dataset

We use the pmindia dataset (Haddow and Kirefu, 2020).[4] This data set contains the official documents from the Prime Minister Office of the Government of India. It contains monolingual and parallel corpora. There are 13 Indian languages and English in it.

We use data for four language pairs from the parallel corpus found in the data set: from Assamese, Bengali, Hindi and Manipuri into English. The Indian languages used belong to different language

[2]The codes from ISO 639-2 for these languages are as follows: Assamese (asm), Bengali (ben), Hindi (hin), Manipuri (mni) and English (eng)

[3]http://opus.nlpl.eu/JW300.php

[4]https://www.pmindia.gov.in/en/pm-india-language-banner/

Language pair	sentence pairs	words / sent.	word types
Assamese	9,732	17	26,649
English		20	22,900
Bengali	29,584	15	55,150
English		17	38,781
Hindi	56,831	20	52,441
English		19	59,061
Manipuri	7,419	15	22,289
English		19	18,502

Table 1: Statistics on the data set used.

families, yet they have high lexical similarities because of regional influences. Assamese (ISO 639-2 asm), Bengali (ben) and Hindi (hin) belong to the Indo-Aryan language family, Manipuri (mni) belongs to the Sino-Tibetan language family and English (eng) belongs to the Indo-European language family. Assamese, Bengali and Manipuri share the same writing system, the Eastern Nagari script. Bengali has high language influence on Assamese, and some influence on Manipuri as well. Hindi, on the other hand, influences all the other languages, with a large number of words borrowed from it.

Statistics about the data used are presented in Table 1. The largest number of sentence pairs is for Hindi–English (almost 60,000). Only half is available for Bengali–English and less than 10,000 for Assamese–English and Manipuri–English, the latter one having only 7,419 sentence pairs. The number of words per sentence in all languages ranges from 15 to 20. The number of word types in each language reflects the number of sentences and the structure of the language: it is natural that the more the sentence pairs, the higher the number of word types; it should however be observed that, although the number of sentence pairs in Bengali–English is half of that in Hindi–English, the number of word types in Bengali is higher than in Hindi.

Additionally, we use the JW300 dataset (Agić and Vulić, 2019; Tiedemann, 2012) for Assamese–English language pairs for two rounds due to the limited number of sentence pairs present in the pmindia data set.

4 Methodology

We propose to first measure the effect of using zero-shot translation on balanced data sets for all language pairs. We then expand our experiments and increase the training data with the aim of inspecting how efficient or not transfer learning can be. The increase will be performed by groups of 2,000 sentence pairs in training for language pairs excluding Manipuri. There will be three rounds of increase of data.

To deal with numerous source and target languages during multilingual or zero-shot translation training, we classically introduce an artificial token at the beginning of all the source language sentences. The artificial token contains the information about the source language and the target language for the sentence pair at hand.

Our first series of experiments consists in measuring translation quality in a balanced data set setting. Our model is trained on an equal amount of training data for all the languages. We start with 5,000 sentences for training across all language pairs. The amount of test and validation data is 1,000 sentences each for all the languages in our model. Each language uses a balanced data set for training, validation and test in the Balanced round. The data is randomly selected. We do not perform pivoting through the English language because of the limited number of sentences in common.

Our second series of experiments measures translation quality when increasing the data for other languages than Manipuri. We showed in Table 1 that the Manipuri–English language pair has the least number of sentence pairs (7,419) among all language pairs. Because of this, we first create a balanced data set of 7,000 sentences in total for all the language pairs. We then increment the training data by 2,000 sentence pairs in all language pairs, except Manipuri–English, in three rounds.

In total, we report translation quality for three types of models.

- single models: these are models trained on a single language pair, one for each language pair.
- multilingual models: these are models trained on our data on different types of multilingual data, i.e., one-to-many, many-to-one and many-to-many (Johnson et al., 2017).
- zero-shot models: these are the models for testing zero-shot translation on the language pairs for which no parallel data was seen during training.

An example of a zero-shot model is described below. Suppose that we have trained a model on

the following language pairs: Manipuri–English and Assamese–English, in both directions, hence in 4 language directions. The zero-shot model will translate between Manipuri and Assamese in both directions, although no Assamese–Manipuri or Manipuri–Assamese sentence pair has been seen during training.

One of our experiments focuses on zero-shot translation with balanced data set. For that, we test the three possible different combination of language pairs with our data. For one of the three Indian languages, Assamese, Bengali or Hindi, call it X, we build a system to perform zero-shot translation from Manipuri into X by using sentence pairs from Manipuri–English in both directions and X–English, in both directions too.

We change the conditions of the above experiments by increasing the training data with groups of 2,000 sentence pairs in the three Indian language to English language pairs for three rounds, for all of our models.

5 Experiments and results

5.1 Experimental setup

We briefly outline the experimental setup used in all of our experiments in this section.

We firstly introduce the three types of models used in our experiments. The single models are trained on a single language pair, the multilingual models are trained on different types of multilingual data and the zero-shot models are trained on language pairs which exclude the language pair to be tested. In a first series of experiments, we measure the translation accuracy of all our models on balanced data sets for all of our language pairs. For that, we randomly select 5,000 sentences for training and 1,000 sentences each for validation and testing from our data set for all the language pairs. Later, we expand our experiments by increasing the training data by 2,000 for all the language pairs excluding Manipuri-English language pair in three rounds.

5.1.1 Single models

We train the single models on a single language pair, one for each language pair. There are a total of eight language pairs in our experiments: from each of Assamese (asm), Bengali (ben), Hindi (hin) and Manipuri (mni), into English and vice-versa. We then measure the effects of balanced data set on our single models. Each language pair has 5,000 sentences for training and 1,000 sentences each for validation and testing.

5.1.2 Multilingual models

We train our multilingual models (Johnson et al., 2017) on different types of configurations. They are listed below.

- **One-to-Many**: A One-to-Many multilingual model is trained on language pairs that has only one type of source language and different types of target languages. In simple terms, it is a model which translates one source language into many target languages. Because of our data set, our source language is English and the target languages are the Indian languages.
- **Many-to-One**: In a Many-to-One multilingual model, only language pairs that have several source languages and only one target language are used for training. This is the other direction than One-to-Many. Again, because of our data set, the source languages are the Indian languages and the target language is English.
- **Many-to-Many**: Many-to-Many multilingual model is trained on language pairs that have several source languages and several target languages. We train our model on language pairs that have source and target languages as Assamese (asm), Bengali (ben), Hindi (hin), Manipuri (mni) and English (eng) respectively.

5.1.3 Zero-shot models

Lastly, we train zero-shot models for testing zero-shot translation as described in Section 4 on language pairs without parallel data. Our model translates between Manipuri into the other Indian languages, i.e., Assamese, Bengali and Hindi.

5.1.4 Pre-processing and tools

Before preprocessing the data, we use Joint Byte-Pair Encoding (Sennrich et al., 2016b) to address the problem of rare words by using sub-word segmentation. We apply Byte-Pair Encoding (BPE) and perform sub-word segmentation on all of our selected data set with 10,000 merge operations so as to obtain a vocabulary representation of all our language pairs.

For all of our experiments, we use the OpenNMT-py toolkit (Klein et al., 2017). We preprocess the training and validation data set for all

the language pairs after applying BPE. We train our model on a 2-layered RNN model with a bidirectional RNN as encoder and a simple RNN as decoder. We measure the translation accuracy of all our experiments using BLEU with a confidence at 95 % (Koehn, 2004).

5.2 Model configuration

There are many RNN architectures available for NMT. We choose the default model provided by OpenNMT-py toolkit (Klein et al., 2017). It is a seq2seq architecture with attention mechanism (Luong et al., 2015). In our models, both the encoders and decoders are long short-term memory cells (Hochreiter and Schmidhuber, 1997). The hyper-parameters are mostly the default ones provided by the toolkit. The exact values for the hyper-parameters are listed in Table 2.

It is known that the Transformer architecture (Vaswani et al., 2017) usually leads to better translation accuracy in comparison to the RNN architectures. For instance, Lakew et al. (2018) report that their Transformer architecture outperforms the recurrent ones in all their systems. In our settings and with our datasets, this is not the case. For example, in the many-to-one multilingual experiments, the translation accuracy (measured using BLEU with confidence at 95%) for the Transformer architecture lies in the range of 0.9 ± 0.2 to 8.6 ± 1.0, whereas the results with recurrent architecture range from 2.2 ± 0.5 to 9.6 ± 1.1 (see Table 3). This justifies why we use the RNN architecture.

5.3 Training settings

In all of our experiments, the hyper-parameters are uniform throughout all the models. The model is trained on a 2 layered RNN model having layer size of 64 for embedding and 500 for inner layers. The LSTM has encoder type as bidirectional RNN and decoder as a simple RNN. Since our data set is very small, we use a drop-out (Srivastava et al., 2014) rate of 0.3 (Gal and Ghahramani, 2016). We also use the general typed global attention mechanism onto the models. The models are trained with 10,000 training steps with checkpoints at every 5,000 steps.

For optimization of the model during training, we use the Adam (Kingma and Ba, 2015) optimizer with a learning rate of 0.001. The number of steps before dropping the learning rate is set to 50,000 and the decay frequency which is the number of steps at which the learning rate starts to drop at each training step is taken as 10,000.

RNN model	
Embed Dim	500
RNN Type	LSTM
Num Layers	2
Hidden Dim	500
Input Feeding	True
Attention	Global
Attention type	General
Dropout	0.3
Encoder Type	brnn
Decoder Type	rnn
Optimization	
Batch size	64
Batch type	Sentences
Optimizer	adam
Init learning rate	0.001
Learning rate schedule	
# steps before decay	50,000
Decay frequency	10,000
Learning rate decay	lcurr * 0.5

Table 2: Parameters used for RNN model. They are mostly from openNMT-py toolkit suggestions.

5.4 Results and analysis

In this particular experimental setting for balanced data set of single and multilingual models, we observe that the multilingual models perform comparatively better than the single model in the case of Indian language–English language pairs, excluding Manipuri to English. See results in Table 3.

As for zero-shot translation, the BLEU scores reported are very low. In Table 4, the results of the experiments on balanced data are shown under the label Balanced, while the results obtained when increasing the training data by 2,000 sentences for three rounds, are shown under the labels of Round-I, Round-II and Round-III.

We observe that the translation accuracies in Round-I to Round-III are slightly higher in comparison to the results in Balanced. We also observe that zero-shot translation between Manipuri–Bengali performs comparatively better than the rest, with translation accuracy more than twice that of Manipuri–Assamese and 1.4 times that of Manipuri-Hindi, with statistical significance. Additionally, the scores of Manipuri–Bengali increases by twice in Round-III, with statistical significance,

Language Pair	single	multilingual		
		one-to-many	many-to-one	many-to-many
asm–eng	5.3 ± 0.3	—	5.9 ± 0.9	6.0 ± 0.9
ben–eng	0.9 ± 0.2	—	2.2 ± 0.5	1.8 ± 0.4
hin–eng	4.1 ± 0.5	—	4.6 ± 0.6	4.6 ± 0.6
mni–eng	9.7 ± 1.1	—	9.6 ± 1.1	8.1 ± 1.0
eng–asm	2.0 ± 0.4	2.4 ± 0.4	—	1.7 ± 0.3
eng–ben	1.4 ± 0.4	3.7 ± 0.5	—	3.7 ± 0.5
eng–hin	3.6 ± 0.5	3.3 ± 0.5	—	3.3 ± 0.5
eng–mni	50.5 ± 0.7	10.3 ± 0.9	—	10.6 ± 0.9

Table 3: Experiment results from balanced data set setting for single models and multilingual models (one-to-many, many-to-one, many-to-many). The translation accuracy is measured with BLEU with a confidence at 95 %.
Note: "—" represents that there is no experiment conducted for this language pairs on the model.

Language Pair	Zero-shot			
	Balanced	Round-I	Round-II	Round-III
mni–asm	0.2 ± 0.1	2.6 ± 0.4	4.6 ± 0.6	6.1 ± 0.9
mni–ben	6.1 ± 0.9	8.3 ± 1.0	10.9 ± 0.9	13.2 ± 1.3
mni–hin	1.3 ± 0.4	5.5 ± 0.3	7.6 ± 0.9	9.7 ± 1.1

Table 4: Experiment results of the zero-shot translation on balanced data set setting and, Round-I and Round-II and Round-III where the data for training is increased by 2,000 in Assamese to English (asm–eng), Bengali to English (ben–eng) and Hindi to English (hin–eng) language pairs for three rounds. Translation accuracy is measured using BLEU (in the range of 0-100) with confidence at 95 %.

in comparison to Balanced. Lastly, scores in Manipuri–Assamese increase by 2 points with statistical significance, progressively from Balanced to Round-III. For Manipuri–Hindi, the scores of Round-III is Balanced multiplied by 7.

The observed sentence behaviours of the translated sentences in Balanced in terms of average number of words per sentence for each language pair is half of the length of its reference sentences. In order to understand this characteristic, we looked into the training sentences added in each round. For all the language pairs, the average number of words in a sentence for training sentences before addition is lesser than the length of the test sentences. For example, in Assamese to English, the average number of words in a training sentence is 16 words initially and that of our added sentences for each round is 18 (20 for the test sentences). Thus, the average length of sentences added in each round becomes closer to the length in the test set.

Our NMT system did not perform well in Balanced: it was not good at learning short sentences. This could explain why the length of the translated sentences is only half of that of reference sentences with repeated words. For example, there are only 584 unique words out of 14,893 words (4 % unique words) in the translated sentences of Manipuri–Assamese in Balanced.

As we progress with the rounds, the translation accuracy increases and the behaviour of the translated sentences changes as well. The average length of sentences becomes closer to the length of the references as the rounds increase; the accuracy increases too. The average sentence length of Manipuri to Bengali is equal to the length of the reference sentences in Round-III (17 words). This language pair gives the highest BLEU scores among all other pairs. Manipuri–Bengali outperforms the rest because of Bengali having lexical influence over Manipuri. It is followed by Manipuri–Hindi, Manipuri–Assamese exhibiting the least score. This may be explained by the fact that they do not influence each other directly: Bengali influences Assamese and Manipuri but Assamese does not influence Manipuri.

6 Conclusion

This work provided an investigation in the use of zero-shot translation between some Indian languages, in the context of low resource.

Firstly, we studied the influence of the balance in data sets across the considered language pairs. We observed that a multilingual model performs comparatively better than a baseline single model, in terms of BLEU scores. In addition, we observed that, in zero-shot translation, a balanced configuration does not perform well. As observed in other works, the use of NMT on a very small amount of data for training, validation and testing, results in low translation accuracy, because NMT has a steep learning curve with respect to amount of data (Koehn and Knowles, 2017).

The translation accuracy when incrementing the data size is comparatively better than in the balanced data set settings. We observed a very small increase in BLEU scores in the balanced settings from Round-I to Round-III, although there is no statistically significant difference. In zero-shot translation, Manipuri–Bengali recorded the highest BLEU score among all language pairs, while Manipuri–Assamese recorded the least score.

In the future, we would like to inspect the possibility of increasing the size of our data by using back-translation. We expect that synthetic data will help our models in improving the translation accuracy (Sennrich et al., 2016a). We would also like to inspect the use of the unsupervised learning approach with adversarial training to learn a mapping from source to target languages without any parallel data or anchor points (Lample et al., 2018) on our models.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, M. Johnson, M. Krikun, M. Chen, Yuan Cao, G. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Y. Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kyunghyun Cho, B van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Comput. Speech Lang.*, 45(C):236–252.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of Machine Learning Research*, volume 48, pages 1050–1059, New York, New York, USA. PMLR.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Thanh Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *In Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle.

Barry Haddow and Faheem Kirefu. 2020. Pmindia - a collection of parallel corpora of languages of india. *ArXiv*, abs/2001.09907.

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

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.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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.

Pulkit Madaan and Fatiha Sadat. 2020. Multilingual neural machine translation involving Indian languages. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 29–32, Marseille, France. European Language Resources Association (ELRA).

Christopher Moseley and Alexandre Nicolas. 2010. *Atlas of the world's languages in danger*, 3 edition. UNESCO, France.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.