Machine Translation for a very Low-Resource Language - Layer Freezing approach on Transfer Learning

Amartya Roy Chowdhury and **S. R. Mahadeva Prasanna**
Indian Institute of Technology, Dharwad

Deepak K.T
Indian Institute of Information Technology, Dharwad

Samudra Vijaya K
Koneru Lakshmaiah Education Foundation, Vijayawada

Abstract

This paper presents the implementation of Machine Translation (MT) between Lambani, a low-resource Indian tribal language, and English, a high-resource universal language. Lambani is spoken by nomadic tribes of the Indian state of Karnataka and there are similarities between Lambani and various other Indian languages. To implement the English-Lambani MT system, we followed the transfer learning approach with English-Kannada as the parent MT model. The implementation and performance of the English-Lambani MT system are discussed in this paper. Since Lambani has been influenced by various other languages, we explored the possibility of getting better MT performance by using parent models associated with related Indian languages. Specifically, we experimented with English-Gujarati and English-Marathi as additional parent models. We compare the performance of three different English-Lambani MT systems derived from three parent language models, and the observations are presented in the paper. Additionally, we will also explore the effect of freezing the encoder layer and decoder layer and the change in performance from both of them.

1 Introduction

Machine Translation started way back in the 1950s as a way to bridge the communication gap. The techniques are broadly classified in three types (a) Rule-Based Machine Translation(RBMT) (Charoenpornsawat et al., 2002) (b) Statistical Based Machine Translation(SMT) (Zens et al., 2002) and (c) Neural-based approaches (NMT). Warren Weaver created the first computer-generated Machine Translation (Hutchins, 1997) during the 1980s by using Statistical methods using 'Shannon's Information Theory' (Stone). In the last couple of years, neural-based Machine Translation has achieved state-of-the-art performance where large amounts of parallel data are available. With the introduction of the encoder-decoder-based architecture (Eriguchi et al., 2016; Vaswani et al., 2018), there was a surge of interest and a lot of research has been conducted. However, it was quickly realized that these initial systems require a huge amount of data to get a performance close to that of a SMT system. (Koehn, 2009). Transfer learning has proved successful for low resource settings (Yi et al., 2018; Tits et al., 2019; Maimaiti et al., 2019; Imankulova et al.,

Figure 1: Distribution of Lambani language in the state of Karnataka

Proceedings of the 29th International Conference on Computational Linguistic, pages 48–55
October 12–17, 2022.

2019) and achieves higher translation performance. In this paper, we will specifically be focusing on NMT although transfer learning has been used for SMT in the past. Specifically for low-resourced languages SMT seem to give better performance in case of domain mismatch (Kumar et al., 2018)

In this paper, we focus on Lambani language (Chandramouli and General, 2011) which is generally spoken by the banjaras (Varady, 1979; Childers et al., 2003) and study how the language draws its influence from various other languages. We show how morphological similarity can improve the performance of a language. We focus on three different languages and how are they related to Lambani. However, it is a major challenge to collect a large amount of data for languages which are not spoken by a lot of people. Despite the recent emphasis on low resource languages, we are not aware of any research that has done any work in the Lambani language.

The paper is organized as follows. A summary of the background work is given in section 2. The proposed approach is explained in section 3. The details of the dataset used are given in section 4. The effect of layer freezing is presented in Section 5.

2 Background

In this section we give an overview of the transfer learning approach in the context of Machine Translation.

2.1 Transfer Learning

Transfer learning was first conceptualized in 2016 (Do and Ng, 2005; Zoph et al., 2016) and was mainly used for the text classification task. Transfer learning is the transfer of knowledge from one model to another. We apply the same concept in our work for MT between various languages.

2.2 Transfer learning in Machine Translation

(Zoph et al., 2016) used transfer learning for MT between four languages, viz. Uzbek, Hausa, Turkish, and Urdu. In the paper, the parent model was trained on a high-resource data set and the model parameters were transferred to the low-resource setting. By using this method Zoph et al. were able to improve the BLEU (Papineni et al., 2002) score by an additional 5 to 6, on average. In the case of Urdu, we see the largest change in BLEU score from 5.2 to 13.5 was seen in case of English to Urdu MT. An increase in BLEU score of 16 was observed in case of a Spanish to English MT when transfer learning approach was used with English to French as the parent model Based on the above results we can be sure that using transfer learning we get a performance improvement. Also the study showed that performance depends on the proximity of the languages.

(Kocmi and Bojar, 2018) in 2018 explored a very similar scenario where they have trained multiple parent models having no relations between them. By this method, the child model was performing significantly better as compared to baseline models. In the paper, the improvement was also noticed for unrelated languages that are languages that don't show any similarities like Czech and Estonia. There was an improvement of +3.38 BLEU for the EN-ET pair when EN-CS was taken as the parent model. This is in direct contradiction with what Zoph et. al (Zoph et al., 2016) reported that more related the models are better will be the translation. The paper also explored completely unrelated languages like Arabic and Russian, Although there were some improvements, the gains are very small (+0.49 to +0.78). Therefore, compared to the baseline models it is preferable to do the transfer learning from the related parent model to target model.

(Maimaiti et al., 2021) tried to improve the performance of transfer learning models by incorporating lexicon information as well as lexical embedding of low-resource child languages. In this work, the parent model was trained using a hybrid approach where the lexical information was shared between the parent and child model before fine-tuning. Using this method, there was an improvement in BLEU score of +0.25 on the Azerbijan-Chinese child pair and an improvement of +0.38 on the Farsi to Chinese language pair. But the method of incorporating the lexical information doesn't

give better performance with morphologically poor language.

3 Proposed Approach

In this proposed work we will first train a parent pair containing a large number of sentences for a given number of iterations and then switch to the child language pair without changing any of the hyperparameters. Then subsequently the performance is improved by freezing some of the layers where weights of some layers are frozen.

Transfer learning in Machine translations was first proposed by (Zoph et al., 2016). We will be applying the same principle in this work . The Lambani language has no script of its own and it is generally written in Kannada script. Whereas, Marathi and Gujarati generally follow Devanagari script. To avoid script mismatch we will be transliterating both Marathi and Gujarati to Kannada script. To the best of our knowledge this work demonstrates the effectiveness of transfer learning for very low resource Indian tribal language. The novel part of this paper is that we will not be sharing any vocabulary instead we will use distinct vocabularies for the parent and child models. A shared vocabulary will not work in our case as some of the parent models don't share lexical features with the child model. We will also incorporate encoder and decoder layer freezing and how they impact the performance of our child model.

During our training, we train our NMT model on high resource data and this is called our parent model. Then we will be using the parent model to train the child model on low-resource data using the transfer learning approach.

3.1 Model Architecture

We will use the Transformer Sequence-to-Sequence model (Kalchbrenner and Blunsom, 2013) as proposed by (Vaswani et al., 2017) Initially we will train three different parent models namely English-Kannada, English-Marathi, and English-Gujarati. As there was no existing data available on Lambani, so

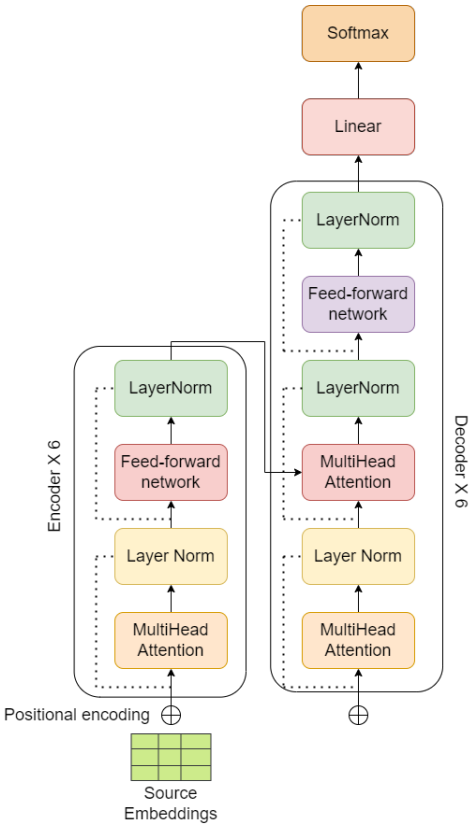

Figure 2: Transformer (Vaswani et al., 2017) architecture

it was mostly a manual process. The parent model will be used to fine-tune the child model. We will also use the parent model to try and understand the role of language relatedness in transfer learning.

For both the models we will be using Transformer model (Vaswani et al., 2017) containing six encoder and six decoder layers and eight attention heads. The tokenization method is SentencePiece (Kudo and Richardson, 2018) which produces a vocabulary of 32,000 for every parent model and 4000 for the child model. The parent languages pair are chosen based on similarity. As explained above we have two models. All our languages are summarized in the table below. Without modifying the architecture of the MT models, the architecture of the parent models is identical to the child model. As for hyper-parameters we have a beam search width of five. The batch size is set to 25. The parent models are trained for 500000 steps on Samanantar dataset (Ramesh et al., 2021). The average checkpoint with the lowest validation loss is then selected. For all our experiments we will be using OpenNMT-tf

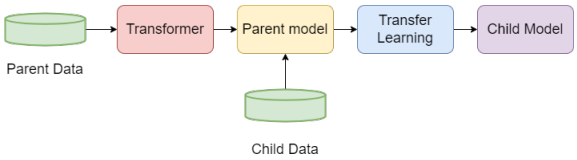

Figure 3: Block diagram of the entire process

Language		Sentence	Vocab	
Source	Target		Source	Target
English	Kannada	4M	32K	32K
English	Marathi	3.32M	32K	32K
English	Gujrati	3M	32K	32K
English	Lambani	6K	4K	6K

Table 1: Details of the dataset used for our experiment. Here, vocab means vocabulary of the Source and target language

by Google (Klein et al., 2017).

4 Dataset

In our experiment, we are be mainly working on a low-resource dataset. We consider Lambani as a very low-resource language reason being that no text-based resources are available to the best of our knowledge. While Kannada, Marathi, and Gujrati are considered to be medium resources languages. The size of the dataset is given in table 1. Preparing the Lambani dataset was mostly a manual process. (a) Firstly raw sentences were extracted from NCERT books (Upreti et al., 2014) and Wikipedia articles (Wikipedia contributors, 2022). (b) Then the sentences were pre-processed and longer sentences were removed. Most of the sentences in our dataset are within 6-10 words. We have also removed any semantically or syntactically incorrect sentences (c) Then the sentences were translated by a Lambani native speaker and was quality checked by other Lambani native speakers. Almost similar level of prepossessing was followed for the parent Pairs, sentences with less than three words and with more than 100 words are removed from the parent dataset, along with that any 'URLs' and unknown characters are also removed.

Language	Role	Train	Test	Valid
Kannada	Parent	3.5M	0.5M	0.5M
Marathi	Parent	3M	0.15M	0.15M
Gujrati	Parent	2.7M	0.15M	0.15M
Lambani	Child	5.4K	0.3K	0.3K

Table 2: Details of the number of sentences in the Train, Validation and Test

Language Pair		Transfer		Baseline
Parent-Pair	Child-Pair	Valid	Test	Parent-only
EN-KN	EN-LA	9.90	13.28	17.2
EN-MR	EN-LA	8.44	10.25	14.4
EN-GU	EN-LA	9.88	12.24	15.5

Table 3: Our transfer learning method applied to various parent models. Note that we are getting the best BLEU(Papineni et al., 2002) score when kannada is treated as the Parent model.

Language Pair		Transfer			
		Encoder		Decoder	
Parent-Pair	Child-Pair	Valid	Test	Valid	Test
EN-KN	EN-LA	12.42	14.25	7.64	11.78
EN-MR	EN-LA	11.44	14.43	7.56	10.23
EN-GU	EN-LA	9.93	14.83	7.37	9.93

Table 4: BLEU (Papineni et al., 2002) score obtained by freezing the first five layers of encoder and the decoder. If we compare it with our previous transfer result from Table 2. we can see that we are getting better performance while encoder is frozen.

4.1 Experiments

For our experiments, we are using Kannada, Marathi, and Gujrati models as our parent models. All three of these parent models has almost similar dataset size. While our child pair contains only 6000 sentence pairs. As mentioned above the parent model was trained for 500K steps while the child model is trained for 50K steps.We are representing the models with a pair of source and target codes. For example, the English-to-Kannada is denoted by EN-Kn and transfer learning models will be represented as EN-XX-LA (where XX represents the target code). The size of the vocabularies used for all these models are also given in Table 1.

For both the parent and child model we have used English as the common language (that

	English	Lambani
Test	7.0%	6.7%
Validation	6.3%	5.2%

Table 5: Details of vocabulary overlap of the Test and Validation set with the training set

means EN-XX). Table 3. summarizes the various results from both the high-resource and low-resource languages. From the table we can see that we get the best performance when EN-KN is used as the parent model. with a BLEU (Papineni et al., 2002) score of almost 9.9 on the validation set and 13.28 on the test set. This score is expected as the data used in this experiment was collected from Lambani speakers located in Karnataka. So, their language would be influenced by Kannada language. Further, the score is not restricted to related language when EN-GU pair we reach a score of 9.88 a -0.02 over the best performing pair. Now we interpret that these two BLEU scores are almost comparable. For the EN-MR we are seeing the worst performance which is almost -1.44 degradation over the best model indicating that EN-MR is the least related language as compared to Lambani.

Freezing analysis in Automatic Speech Recognition (ASR) shows an improvement in performance (Eberhard and Zesch, 2021). Motivated by the study we have applied it in the current Machine Translation study. Details of freezing and experimental setup is explained in section 5. Figs 4, 5 and 6 show the BLEU score curves on the validation set for all the three parent models. In all of the three plots we can see that we are getting better performance when we are freezing the first five layers of the encoder (represented in 'orange' color). Whereas freezing the layers in the decoder may not help in improving the performance as can be noticed from the plot (represented in 'green' color) over the baseline performance (represented by 'blue' color)

The baselines are models trained entirely on parent data. Table 3. also summarizes the results on the Test which are quite higher compared to the validation set, we think this may be due to higher vocabulary overlap between the Training and Test sets as given in table 5.

5 Freezing

5.1 Freezing encoder layers

We are interested to measure the overall change in performance upon freezing the encoder layers. We perform continued training while freezing the layers of the encoder(i.e. keeping the layers fixed to the values while training the child model while adapting the rest of the components). The results are shown in Table 4. For all of the language pairs, we are seeing a performance improvement. For the EN-MR-LA model, we are seeing the largest improvement in performance followed by EN-KN-LA (+3 BLEU and +2.52 BLEU respectively) on the validation set. This increase in performance may be because the initial few layers of a model are generally well trained. This shows that by freezing the encoder during training, the model can find a local minimum that is better than the one when the models are transfer learned.

Sr.No.	Sentence	
1	Source	I do nothing on Sundays.
	Ground Truth	ಮ ರವಿವಾರೆರ್ ಕಾಂಯಿ ಕರುನಿ.
	Transliterated Sentence	ma ravivaarer kaanyi karuni.
	EN-KN-LA	ಕಾಂಯಿ ರವಿವಾರೇರ ಕಾಂಯಿ ಕರೆನಿ.
	EN-MR-LA	ರವಿವಾರೇರ ಖಾಲಿ ರೆಚಿ ಕ?
	EN-GU-LA	ಮ ರವಿವಾರೇರ ಕಾಂಯಿ ಕರು?
2	Source	I get up early in the morning
	Ground Truth	ಮ ಪರ್ಬಾತಿ ಜಲ್ದಿ ವುಟುಚು.
	Transliterated Sentence	ma parbaati jaldi vutuchu.
	EN-KN-LA	ಮ ಪರ್ಬಾತಿ ಜಲ್ದಿ ವುಟುಚು
	EN-MR-LA	ಮ ಪರ್ಬಾತಿ ಜಲ್ದಿ ವೂಟೇನಿ
	EN-GU-LA	ಮ ಪರ್ಬಾತಿ ಜಲ್ದಿ ವೂಟೇನಿ

Table 6: Some example sentences from all the three parent models along with transliterated sentences

5.2 Freezing the decoder layers

If we freeze the entire decoder layer it is noticed that the results are inferior. From Table 4. we can see that for all the models we are getting degradation in performance when the decoder is frozen. We can see the largest drop in performance occur in the case of EN-KN (-3.8) followed by EN-GU (-2.56) on the validation set. One interesting thing to note here and also can be seen from the curves Fig 4, 5, and 6 is that the BLEU score on the validation set for all the models are very close to one another (an experiment we keep for future study). This reduction in the performance may be because the layers in the decoder need more training as compared to the Encoder as the final layers of the model are more task-specific.

6 Future work

Although there may have been a couple of research on transfer language of related and unrelated languages there is very little research

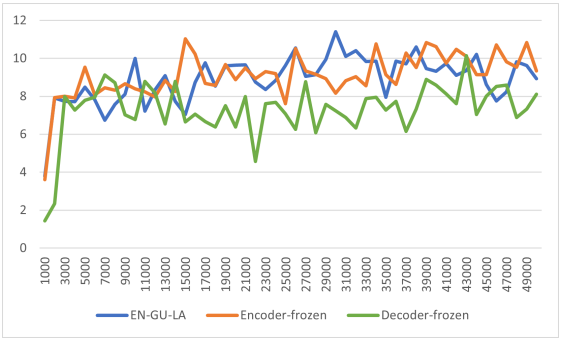

Figure 4: BLEU score curves on the val set for EN-GU as parent

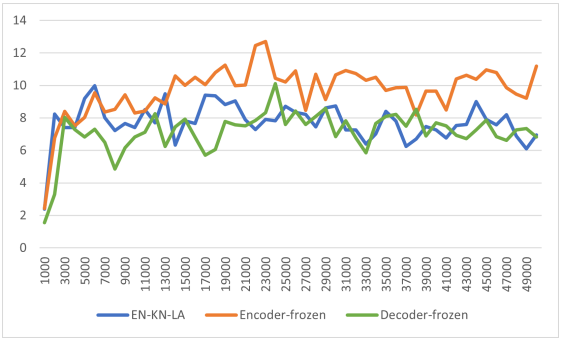

Figure 5: BLEU score curves on the val set for EN-KN as parent

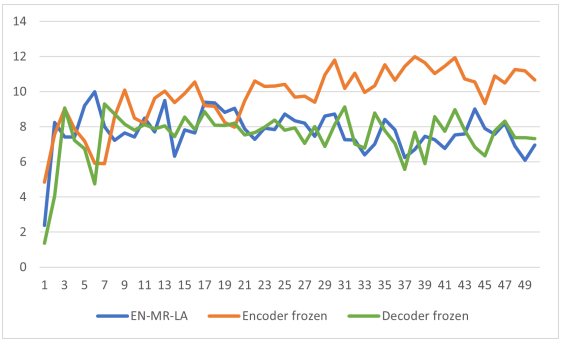

Figure 6: BLEU score curves on the val set for EN-MR as parent

as to why transfer learning is giving better results for related languages from a linguistic perspective. As in our case Lambanis a nomadic tribe before they settled in the modern state of Karnataka. As a result, the language is morphologically rich and may share some linguistic similarities with other language. According to (Edunov et al., 2018) adding noise to the training data has improved Neural Machine Translation. The same idea can be applied to our model. We can randomly drop words from the training data and replace them with filler words in order for the model to learn better. Noisy sentence help in learning as it makes it harder to predict translation.

7 Conclusion

Our experiment is limited to a transfer learning method between closely related languages. From our experiments, we are seeing much better performance when similar languages are taken for transfer learning while for unrelated languages we are not seeing a drastic change in BLEU (Papineni et al., 2002) score which may be because of our dataset size of all the parent models is almost similar. We have further improved our model performance by incorporating encoder freezing and reached a performance improvement of +3 over the EN-MR-LA model. From our experiments we also notice that freezing the decoder is reducing the performance. This may be because the decoder needs more data than an encoder.

8 Acknowledgement

The authors like to thank "Anatganak", high-performance computation (HPC) facility, IIT Dharwad, for enabling us to perform our experiments. And Ministry of Electronics and Information Technology (MeitY), Govt. of India, for supporting us through the "Speech to Speech translation for tribal languages" project. We would also like to thank Tonmoy Rajkhowa for his valuable help in setting up the paper

References

C Chandramouli and Registrar General. 2011. Census of india. *Rural urban distribution of population, provisional population total. New Delhi: Office of the Registrar General and Census Commissioner, India.*

Paisarn Charoenpornsawat, Virach Sornlertlamvanich, and Thatsanee Charoenporn. 2002. Improving translation quality of rule-based machine translation. In *COLING-02: machine translation in Asia.*

CH Childers et al. 2003. *Banjaras.* Oxford University Press.

Chuong B Do and Andrew Y Ng. 2005. Transfer learning for text classification. *Advances in neural information processing systems*, 18.

Onno Eberhard and Torsten Zesch. 2021. Effects of layer freezing when transferring deepspeech to new languages.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Character-based decoding in tree-to-sequence attention-based neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 175–183, Osaka, Japan. The COLING 2016 Organizing Committee.

John Hutchins. 1997. From first conception to first demonstration: the nascent years of machine translation, 1947-1954. a chronology. *Machine Translation*, 12(3):195–252.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1907.03060*.

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.

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.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

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

Nirmal Kumar, K. Mrinalini, and P. Vijayalakshmi. 2018. Improving the performance of low-resource smt using neural-inspired sentence generator. In *2018 International Conference on Computer, Communication, and Signal Processing (ICCCSP)*, pages 1–4.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–26.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2021. Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation. *Tsinghua Science and Technology*, 27(1):150–163.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

James V Stone. Information theory: A tutorial introduction.

Noé Tits, Kevin El Haddad, and Thierry Dutoit. 2019. Exploring transfer learning for low resource emotional tts. In *Proceedings of SAI Intelligent Systems Conference*, pages 52–60. Springer.

K. Upreti, G. Khanna, and SK Singh. 2014. *NCERT Solutions - Science for Class X*. Arihant Publication India Limited.

Robert Gabriel Varady. 1979. North indian banjaras: Their evolution as transporters. *South Asia: Journal of South Asian Studies*, 2(1-2):1–18.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Wikipedia contributors. 2022. India — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=India&oldid=1100411060. [Online; accessed 30-July-2022].

Jiangyan Yi, Jianhua Tao, Zhengqi Wen, and Ye Bai. 2018. Language-adversarial transfer learning for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):621–630.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *KI 2002: Advances in Artificial Intelligence*, pages 18–32, Berlin, Heidelberg. Springer Berlin Heidelberg.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.