Towards Efficient Bilingual Machine Translation for a Morphologically Diverse Language Pair

Danish Ebadulla , Rahul Raman , Hridhay Kiran Shetty , Ashish Shenoy , Natarajan S.
PES University , Bangalore , India
{danishebadulla, rahulraman, hridhayshetty, ashishshenoy}@pesu.pes.edu
natarajan@pes.edu

Abstract

Indic-English translation has always been a challenging problem. Varying language morphology and scripts stop us from leveraging shared embeddings and vocabularies. This problem is compounded for Dravidian languages that are considered low resource and are agglutinative in nature. Multilingual models are currently the state of the art, but come at the cost of large parameter counts and access to high-end compute resources as a necessity to build and train them. We attempt to build bilingual models that are compact and produce high quality translations by incorporating several enhancements into the translation pipeline. We train models for the Kannada language, translating to and from English. We test the effect of popular subword tokenizers, transliteration and vocabulary optimization to find the best possible data pipeline for translation. We also study the effect of transliteration on subword generation and optimal transport based vocabulary learning on model performance and efficiency. Our models achieve competitive BLEU scores while being compact and easy to train and deploy.

1 Introduction

Transformers have revolutionized many domains across Deep Learning and Natural Language Processing in particular, helping achieve state of the art results on several understanding and generation tasks. Neural machine translation has seen advances in leaps and bounds with the advent of encoder-decoder based models (Sutskever et al., 2014; Bahdanau et al., 2015; Wu et al., 2016). Progress for Indic languages has been limited despite these advancements due to lack of high quality corpora and the high morphological complexity of many Indic languages.

Multilingual models are currently the ideal choice for neural machine translation for low resource and Indic languages. Multilingual NMT models let you leverage shared encoder-decoder embeddings, joint vocabularies and a richer embedding space over all languages. This makes access to high compute resources a necessity for attempting to beat the state of the art models.

Most of the recent research on Indic language translation uses the base variant of the Transformer (Vaswani et al., 2017), without any modification regardless of the task or language choice, with research primarily focused on improving the quality and form of input data, leveraging transliteration, pre-training and subword tokenization to enhance model performance.

In this paper, we attempt to find the best transformer parameter configuration for bilingual translation between Kannada and English and conduct a detailed study comparing the effect 2 subword tokenization libraries i.e. SentencePiece (Kudo and Richardson, 2018) and subword-nmt (Sennrich et al., 2015), transliteration and VOLT (Xu et al., 2021) on model efficiency and performance. We present (i) The optimal parameter settings for the transformer model to perform bilingual translation for Kannada to English and English to Kannada. (ii) The effect of transliteration on subword generation. (iii) The effect of VOLT on vocabulary size and model performance and its correlation with transliteration.

2 Related Work

Over the course of time, research in machine translation has evolved from linguistic rules to data driven methods using neural networks. Especially in Indic translation where linguistic rules and complexity varied for each language, research relied on dictionaries (Sindhu and Sagar, 2017) and phrase pair (Kumar and Chitra, 2014; HONNASHETTY et al., 2017) based methods for machine translation. With the advent of sequence models, works utilised LSTMs (Goyal and Sharma, 2019) or GRUs (Ramesh and Sankaranarayanan,

2018) for translating from Indic languages to English.

Even after the advent of attention (Bahdanau et al., 2015) and transformers (Vaswani et al., 2017), Indic language translation and particularly Dravidian languages lagged behind the rest of the world due to a lack of good quality parallel corpora. Many works (Goyal and Sharma, 2019; Madaan and Sadat, 2020; Dhar et al., 2020) leveraged monolingual corpora with backtranslation or looked at iterative training (Philip et al., 2020) to synthetically increase the amount of parallel corpora. Multilingual models came to the forefront during this period (Sen et al., 2018; Madaan and Sadat, 2020; Philip et al., 2019) as they produced higher accuracy by taking advantages of rich vector spaces, joint vocabularies and language relatedness, slightly overcoming the problems caused by lack of data. Up until the release of the Samanantar (Ramesh et al., 2021) corpus, an aggregation of all the open source parallel corpora available gave you approximately 400K Kannada-English parallel sentences.

The lack of a standardised validation and test corpus to compare models against was also a persistent problem, with most research for the Kannada language using small custom test sets or splitting their training corpus (Chimalamarri et al., 2020) to evaluate their models. This changed with the release of parallel corpora and the test and validation sets by the Workshop for Asian Translation in 2021 for their shared task[1]. The latest research on Indic language translation has used this corpus to validate and test their models, giving researchers a much needed common benchmark to compare their model performance against.

2.1 Subword Tokenizers

Subword tokenizers are widely used in research on Indic languages, as almost all Indic languages are morphologically rich. There are many popular subword tokenizers that are used in Indic language research but the 2 most popular choices across existing literature are SentencePiece (Kudo and Richardson, 2018) and subword-nmt (Sennrich et al., 2015).

2.2 VOLT

Vocabulary Learning via Optimal Transport (VOLT) (Xu et al., 2021) is a technique to determine the optimal vocabulary and size for a corpus without any trial training. VOLT was found to achieve upto 70% vocabulary size reduction while still managing to achieve higher BLEU scores on various translation tasks.

2.3 Transliteration

Transliteration of Indic languages and its effectiveness in Multilingual NMT was first demonstrated by (Ramesh et al., 2021). Transliteration is made possible due to the fact that Unicode points of various Indic scripts are at corresponding offsets from the base codepoint for that script. We use the IndicNLP library [2] to transliterate Kannada to the Devanagari script.

3 Methodology

In this section we describe our data and training pipeline in detail. We train 2 types of models, Kn-En which translates from Kannada to English and En-Kn, which translates from English to Kannada. We describe our model architecture, the steps taken to prepare the data for training and the training process itself throughout this section.

3.1 Dataset

All our models are trained using the data from the Samanantar[3] corpus. Samanantar provides 4M parallel sentences for Kannada-English. The validation and test data is sourced from the WAT-2021 shared MultiIndicMT task in order to use a standardized validation and testing metric, allowing us to compare our scores with ongoing and future research. All corpora were deduplicated and we ensured that there is no overlap between the training corpus and the validation or test corpus

3.2 Preprocessing

All Kannada sentences are tokenized using the IndicNLP trivial tokenizer while English sentences use the Moses tokenizer. Sentences are converted to lowercase to optimize the vocabulary. Transliteration is applied on the Kannada sentences to convert them to the Devanagari script. We experiment with both subword-nmt and SentencePiece for subword tokenization to determine which achieves the best performance. Both tokenizers use BPE with a fixed vocabulary size of 32000. We also use VOLT to learn a smaller vocabulary using optimal transport. We set the vocabulary size threshold to 10000

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/

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

[3] https://indicnlp.ai4bharat.org/samanantar/

Parameters	Kn-En	En-Kn
Encoder layers	4	4
Decoder layers	4	4
Embedding size	768	512
Feed-forward dim	1536	1024
Attention heads	16	16
Dropout	0.1	0.1

Table 1: Model configurations for both translation directions

and let VOLT recommend the optimal vocabulary and size. VOLT outputs both the recommended size and the vocabulary, allowing us to retrain the subword-nmt model with the recommended vocabulary as an explicit input to the model. Since SentencePiece does not support taking a pre-generated vocabulary as an input while training the SentencePiece model, VOLT recommends giving its recommended input vocabulary size to the model, letting it generate a new vocabulary using the new vocabulary size threshold.

3.3 Model Training

All models are trained using Facebook's fairseq[4] with a custom transformer configuration given in Table 1. To avoid a high percentage of unknown words, we do not specify a source or target dictionary while preparing the data with fairseq, this results in a final vocabulary size approximately 3-5% larger than the hard limits set during subword generation. We train our models to 300K steps. 300K steps is empirically chosen as we observed that the BLEU score varies by less than 0.5 after 200-250K steps. Our models have 4 encoder-decoder layers with 16 attention heads. The exact parameter configuration of each model is given in Table 1. The optimal parameters were found through iterative fine tuning of the parameters of the base Transformer. All variants of the models were tested on these configurations. All models were trained on a Tesla P100 GPU. The models were trained using the Adam (Kingma and Ba, 2015) optimizer with label smoothing set to 0.1 and inverse square root learning rate schedule. Our models use learning rate of 5e-4, with 4000 warmup steps and an initial learning rate of 1e-7 as proposed in (Vaswani et al., 2017). The En-Kn models trained faster than the Kn-En models as the input embedding sizes were smaller for the former.

[4]https://github.com/pytorch/fairseq

3.4 Evaluation

Model performance is evaluated using BLEU scores. We average the last 5 checkpoints of all our models and use this averaged checkpoint to run evaluation metrics. To ensure consistency we use tokenized sentences after removing BPE and run sacrebleu[5] with beam search by setting the beam size to 5 to generate our BLEU scores.

4 Results and Discussion

Our BLEU scores are documented in Table 2 for the Kannada-English direction and Table 3 for the English-Kannada direction. In both directions, scores were close with less than 1.2 BLEU score difference between the best and the worst model, but SentencePiece with VOLT and without transliteration was found to be the best model configuration achieving a BLEU score of 27.75 in the Kannada-English direction and a BLEU score of 14.38 in the English-Kannada direction.

4.1 Model architecture

Our models have 4 encoder-decoder layers unlike the Base-Transformer architecture which has 6 layers. We found that our models converge quicker when we reduce our encoder-decoder layers. The attention heads are set to 16 similar to the model in (Ramesh et al., 2021). Models in the English-Kannada direction were found to perform better when using a smaller embedding size of 512 when compared to the Kannada-English models which used an embedding size of 768.

4.2 SentencePiece

SentencePiece models in both directions were superior when used with VOLT or with Transliteration. VOLT produced a decrease in vocabulary size of 72-75% for Kannada and 78% for English.

4.3 Subword-nmt

Subword-nmt was inferior to SentencePiece when using VOLT or Transliteration but a model that is trained without using VOLT or transliteration performs better with subword-nmt.

4.4 Transliteration

While results with transliteration were inconsistent when considering all our models, transliterated data consistently produced approximately 3% smaller vocabularies when thresholded with VOLT,

[5]https://github.com/mjpost/sacrebleu

Subword Tokenizer	Transliteration	VOLT	Parameters	SRC Vocab	TGT vocab	BLEU
SentencePiece	Yes	Yes	66M	8000	7000	27.64
		No	122.4M	32000	32000	27.31
	No	**Yes**	**67M**	**9000**	**7000**	**27.75**
		No	122.6M	32000	32000	27.53
Subword-nmt	Yes	Yes	58.6M	7000	3000	26.84
		No	122.3M	32000	32000	27.20
	No	Yes	59.7M	8000	3000	27.07
		No	122.5M	32000	32000	27.40

Table 2: BLEU scores for the Kn-En models

Subword Tokenizer	Transliteration	VOLT	Parameters	SRC Vocab	TGT vocab	BLEU
SentencePiece	Yes	Yes	29.9M	7000	8000	13.91
		No	54.8M	32000	32000	13.81
	No	**Yes**	**30.0M**	**7000**	**9000**	**14.38**
		No	55.1M	32000	32000	13.25
Subword-nmt	Yes	Yes	27.0M	3000	7000	13.66
		No	54.8M	32000	32000	13.67
	No	Yes	27.7M	3000	8000	13.82
		No	54.8M	32000	32000	13.96

Table 3: BLEU scores for the En-Kn models

despite applying the same parameters to generate the vocabulary for both transliterated and non-transliterated versions of our training corpus.

4.5 VOLT

Models trained with VOLT managed to keep up with base models despite using only a fraction of the vocabulary that the base models were trained on. Total parameters of models with VOLT were on average half of the base models. We obtained opposite results with VOLT when comparing our subword tokenizers, with SentencePiece showing an improvement in BLEU while subword-nmt scores dropped when VOLT was applied.

5 Conclusions

In this paper, we compare 2 popular subword tokenization techniques and their effectiveness on the Kannada language. We study the effect of transliteration on these subword tokenization techniques and the effectiveness of optimal transport based vocabulary learning techniques on the Kannada language. We present a compact and effective transformer architecture and data pipeline to train models to translate between English and Kannada. Our models achieve a BLEU score of 27.75 for Kannada-English and 14.38 for English-Kannada. We hope our work will help cement the effectiveness of VOLT, and serve as a proof of concept to quickly and effectively train bilingual models for other Indic languages without needing access to high compute resources. In the future, we plan on extending our data pipeline by using monolingual corpora to enhance our models' performance, testing fine tuning vs. upsampling when synthetic data sources are involved, and extending our study to other Indic languages.

References

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*.

Santwana Chimalamarri, Dinkar Sitaram, Rithik Mali, Alex Johnson, and K A Adeab. 2020. Improving transformer based neural machine translation with source-side morpho-linguistic features. In *2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 1–5.

Prajit Dhar, Arianna Bisazza, and Gertjan van Noord. 2020. Linguistically motivated subwords for english-tamil translation: University of groningen's submission to WMT-2020. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 126–133. Association for Computational Linguistics.

Vikrant Goyal and Dipti Misra Sharma. 2019. LTRC-MT simple & effective Hindi-English neural machine translation systems at WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 137–140, Hong Kong, China. Association for Computational Linguistics.

SHARANBASAPPA HONNASHETTY, Mallamma Reddy, and Prajwal Hanumanthappa. 2017. Phrase structure based english to kannada sentence translation. *International Journal of Computer and Communication Technology*, pages 96–100.

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*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

K. M. S. Kumar and C. Chitra. 2014. A comprehensive study of statistical machine translation for english to kannada language.

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).

Jerin Philip, Vinay P. Namboodiri, and C. V. Jawahar. 2019. A baseline neural machine translation system for indian languages. *CoRR*, abs/1907.12437.

Jerin Philip, Shashank Siripragada, Vinay P. Namboodiri, and C. V. Jawahar. 2020. Revisiting low resource status of indian languages in machine translation. *CoRR*, abs/2008.04860.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, 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. *CoRR*, abs/2104.05596.

Sree Harsha Ramesh and Krishna Prasad Sankaranarayanan. 2018. Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 112–119, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2018. IITP-MT at WAT2018: Transformer-based multilingual indic-English neural machine translation system. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

D V Sindhu and B M Sagar. 2017. Dictionary based machine translation from kannada to telugu. *IOP Conference Series: Materials Science and Engineering*, 225:012182.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, Online. Association for Computational Linguistics.