Improving Neural Machine Translation for Sanskrit-English

Ravneet Punia, Aditya Sharma, Sarthak Pruthi, Minni Jain
Delhi Technological University
Delhi, India
{ravneetpunia_bt2k16,adityasharma_bt2k17,sarthakpruthi_bt2k18it110,minnijain}@dtu.ac.in

Abstract

Sanskrit is one of the oldest languages of the Asian Subcontinent that fell out of common usage around 600 B.C. In this paper, we attempt to translate Sanskrit to English using Neural Machine Translation approaches based on Reinforcement Learning and Transfer learning that were never tried and tested on Sanskrit. Along with the paper, we also release monolingual Sanskrit and parallel aligned Sanskrit-English corpora for the research community. Our methodologies outperform the previous approaches applied to Sanskrit by various researchers and will further help the linguistic community to accelerate the costly and time-consuming manual translation process.

1 Introduction

Sanskrit is one of the oldest, extensively studied, and researched languages in the world.[1] It is the oldest Indo-Aryan Language prominently used in Indo-European studies and now used for interlingual translation to English and many other Indian languages, however, the fact that it is dead in today's time cannot be denied. English has emerged as the most popular language on the world level, and the advent of globalization has led to the need for cross-language translations. The developing regions still used the regional languages, and thus the translation of the English language into local languages can make information accessible. Machine Translation is one of the most onerous tasks in natural language processing. Sanskrit is unique as it does not work using a noun-phrase model.[2] It's strict grammar rules, and syllables match being a direct parent of Modern Hindi language. The challenges faced during machine translation of Sanskrit to other languages are translation divergence or the ambiguity phrase due to multiple-meaning, the lack of parallel language data. Lots of historical and cultural data such as Bhagavad Gita, Ramayana, Mahabharata, and Hindu Literature Vedas were originally written in the Sanskrit language, and most of them are untranslated to other languages. Despite its important part in Indian culture and history, not much work has been done for translation to or from the Sanskrit language.

Although the past few years, many efforts have been made to translate Sanskrit to other languages using various machine translation approaches. Mishra and Mishra (2008) and Gupta et al. (2013) implemented example-based and rule-based approaches for Sanskrit-English machine translation. Later Mishra and Mishra (2010) improved the Rule-Based Machine Translation approach by integrating with the Artificial Neural Network (ANN) model. Recently Koul and Manvi (2019) proposed an encoder-decoder based Neural Machine Transition approach for Sanskrit to English.

In recent years Neural Machine Translation techniques like Sequence to Sequence Learning, Encoder-Decoder attention-based architectures (Bahdanau et al., 2014), and Transformers have achieved State Of The Art (SOTA) results for supervised machine translation tasks. However, for low resource methods like Back translation (Edunov et al., 2018), Cross-Language Modeling, Phrase-Based Machine Translation (Lample et al., 2018), and Dual Learning Mechanism based upon reinforcement learning(He et al., 2016) takes the benefit of monolingual data to improve the quality of translations over supervised approaches. Unfortunately, none of the above methods has been used for Sanskrit's machine translation task due to the lack of linguistic resources.

Through this paper

[1] https://en.wikipedia.org/wiki/Sanskrit

[2] https://www.genpact.com/

Proceedings of the 17th International Conference on Natural Language Processing, pages 234–238
Patna, India, December 18 - 21, 2020.

- We test multiple machine translation approach based supervised methodology, Transfer Learning, and reinforcement learning approach that leverages monolingual data for Neural Machine Translation (NMT).

- We also release the collected parallel English - Sanskrit data as well as monolingual data for Sanskrit.

2 Related Work

Work by Mishra and Mishra (2009) mainly focuses on building tokenization, POS Tagger, and a Named Entity Recognition (NER) system for the Sanskrit language using statistical machine translation approach. Mane et al. (2010) introduced a dictionary-based approach for implementing machine translation on Sanskrit by parsing and replacing source word with the target using a bilingual dictionary.

Bahadur et al. (2012) developed Machine translation which primarily focused formulation of Synchronous Context-Free Grammar (SCFG) and a subset of Context-Free Grammar (CFG). The developed model firstly tokenize input data and then match the exact word or phrase from the dictionary. The developed model also gathers information about parts of speech (POS) of input sentences. The work by Rathod (2014) implemented a Rule-Based and Example-based approach for Machine translation using a bilingual dictionary and speech synthesizer that also converts speech to text. The designed model was capable of grammar and spell check too. An open-source web portal [3] collects data from domains like primary and secondary school Sanskrit literature books, also established by Govt. of India in 2015. It also implements statistical Machine Translation algorithms and even tries to solve Word Sense Disambiguation (WSD) problem.

Apart from Koul and Manvi (2019) encoder-decoder model, no such work has been done on Sanskrit's Neural Machine Translation in the best of author's knowledge.

3 Dataset

For this paper, we extracted parallelly allied English-Sanskrit data as well as monolingual data for each language. The parallel English-Sanskrit data, we obtained 2,100 sentences from OPUS[4], Sanskrit translation of Bible, Shlokas from Ramayana and more sentences from Gita. As data is extracted from multiple sources, sentences with the same source but multiple translations and sentences with the same translation, various sources are removed. Finally, a parallel dataset with 9000 parallel lines is extracted, further divided into the standard train, test, and validation set with a ratio of 80:10:10, respectively.

For the monolingual data, we collected the data from the Romanized version of Mahabharata, consisting of 130,000 lines (approx) and for English, we extracted Europarl dataset (Koehn, 2005)

4 Proposed Methodology

Previous Neural Machine Translation approaches for Sanskrit mainly focus on Rule-Based Approach and Encoder-Decoder Mechanism using LSTM units. The classical Rule-Based approach is time-consuming, requires much manual work by the linguist, and does not have good learning capabilities. In contrast, LSTMs based models tend to overfit faster, suffer from issues related to polysemy, and multiple word senses (Calvo et al., 2019; Huang et al., 2011).

To handle all these issues, we first established a baseline translation model using a multi-head self-attention mechanism using encoder-decoder architecture, as suggested by Vaswani et al. (2017). Further, we improved the baseline translator using a reinforcement learning approach by establishing language models and agents that leverage monolingual data. We further experimented with the Transfer Learning approach for Machine Translation to get the benefit of lexically similar Hindi language that is rich resource language.

4.1 Transformer Translator

Initially, the raw sentences were tokenized using SentancePiece tokenizer (Kudo and Richardson, 2018), which is an unsupervised and language-independent tokenization method. Further, the parallel and monolingual tokenized data was used to train word-vectors of length 128, using the word2vec (Mikolov et al., 2013) technique. As the transformer architecture doesn't maintain any word order, so along with the trained word-vectors, a positional encoding signal is mixed and given to the encoder as input. Introducing the positional

[3] http://sanskrit.jnu.ac.in/shmt/index.jsp

[4] http://opus.nlpl.eu/

encoding helps maintain the embedding information and gives the vital position information to the encoder. In the architecture, both encoder and decoder are formed by stacking four identical layers in the same manner as described by Vaswani et al. (2017). The encoder takes the representation of Sanskrit token through word embedding and positional encoding, which is then fed to a multi-head attention unit where feed-forward units with residual connections are employed between every other sublayer. This signal normalizes and is given to the decoder as input along with the output embeddings, positional encoding, and masked multi-head attention. The decoder works similar to the encoder and generates output word by word and finally makes a sequence.

4.2 Reinforcement Translator

This methodology is inspired by He et al. (2016), where we used our Transformer model as the translation model, and building the language model from the Recurrent Neural Network (RNN) using the monolingual data only. We define dual NMT as a combination of Sanskrit to English considered to be the primary task and English to Sanskrit being dual. For both primary and dual tasks, we set individual agents to perform two agent communication games where they correct each other through a reinforcement learning process. The reward system is a combination of Language model (r_1) reward and communication reward (r_2), which can be expressed using the equation:

$$r = \alpha(r_1) + (1 - \alpha)r_2. \quad (1)$$

Where α is a hyper-parameter which is set to 0.1. Further Transformer models are improved using a policy gradient method (Sutton et al., 2000) for maximum reward, which is a common methodology in reinforcement learning. The process iterated for 600 rounds and stoped when the translation model converges. Other parameters such as beam search size, learning rate, the individual reward for each agent r_1 and r_2 were taken same as defined by He et al. (2016)

4.3 Transfer Learning

The main idea of transfer learning is to transfer the knowledge learned by a model trained on a high resource language set, i.e., parent model, to train another model with a similar application, i.e., child model. For our experimentation, we firstly prepared a Hindi-English NMT model using Transformers, as the parent model on 1.56 Million parallel data provided by Kunchukuttan et al. (2017) and training Sanskrit - English NMT model as a child model. The Hindi data was firstly tokenized using Indic Tokenizer (Kunchukuttan, 2020), English using Moses tokenizer (Koehn et al., 2007), and Sanskrit using sentencepiece (Kudo and Richardson, 2018). Further Hindi-English NMT model was trained using the same training procedure as of Transformer model discussed in section 4.1

5 Result & Discussion

The baseline model in section 4.1 was implemented using the OpenNMT Framework (Klein et al., 2017). For the transfer learning implementation, we used the NEMATUS toolkit (Sennrich et al., 2017). The baseline and child model in the transfer learning approach were trained, tuned, and tested on the same data split set discussed in section 3. For the quantitative evaluation, we used the BLEU score (Papineni et al., 2002) for English translation generated by the model against the test set. The results obtained are shown in Table 1.

Architecture	BLEU	Rating
1. Transformer Translator	4.6	2.4
2. Reinforcement Translator	5.8	2.9
3. Transfer Learning	18.4	3.9

Table 1: Evaluation of different models with English translation using BLEU scores

For the qualitative analysis, five Sanskrit language experts were randomly given 50 sentences each from all three models for the rating based on the following rating schema:

- Good[5]: Sentence is interpretable by the language expert, having no incorrectly translated words.
- Helpful[3]: Sentence is interpretable by the language expert with some context knowledge, has some errors and wrong word order.
- Partially Helpful[1]: contains incorrectly translated content words, few UNK Tokens, but still interpretable by language experts.
- Wrong[0]: Sentence having many UNK Tokens or untranslated words and considered as not translated by a language expert.

All average ratings are shown in the last column of Table 1. Hyperparameters searched and best selected for the baseline model during the training are mentioned in the Table 2.

Hyperparameters	Experimented	Best
Epochs	200	200
Batch-Size	512,1024	1024
Number of Layer	4	4
Learning Rate	Dynamic	
Dropout	0.1	0.1
Dimensional Vectors	128,256	128

Table 2: Hyperparameter searching for the best results

Few Observations from results:

- Transfer Learning approach performs best among all three models. The lexical similarity between Hindi and Sanskrit helped in achieving a better result.
- Transformer translator performed worst, most likely due to small and sparse dataset from various domains and a large number of parameters of the model. However, Reinforcement learning made a slight improvement of 1.2 BLEU points.
- The dataset used Koul and Manvi (2019) is different and not available to the public domain for testing, so it won't be appropriate to compare results of Koul and Manvi (2019) with our experiments.

6 Conclusion

In this paper, we explored approaches that have never before been used for the translation of the Sanskrit language to English. Firstly we established a baseline with the Transformer architecture. Further, we improved the Transformer model with Dual Learning methodology and gained small improvement on BLEU Score. The best BLEU Score we observed was with the Transfer Learning method. Although we will not like to make an explicit comment that Transformers architecture is the first time explored in our research, a few unofficial repositories have worked and published the results. In the future, we would try to add more parallel data to improve the trained models' quality. We believe that our research would open the doors for many researchers, linguists, and students to work and explore Sanskrit.

Dataset, training subroutine, and trained model is available at: https://github.com/RavneetDTU/Improving-Neural-Machine-Translation-for-Sanskrit-English

References

Promila Bahadur, AK Jain, and DS Chauhan. 2012. Etrans-a complete framework for english to sanskrit machine translation. In *International Journal of Advanced Computer Science and Applications (IJACSA) from International Conference and workshop on Emerging Trends in Technology*. Citeseer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hiram Calvo, Arturo P Rocha-Ramirez, Marco A Moreno-Armendáriz, and Carlos A Duchanoy. 2019. Toward universal word sense disambiguation using deep neural networks. *IEEE Access*, 7:60264–60275.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

V. K. Gupta, N. Tapaswi, and S. Jain. 2013. Knowledge representation of grammatical constructs of sanskrit language using rule based sanskrit language to english language machine translation. In *2013 International Conference on Advances in Technology and Engineering (ICATE)*, pages 1–5.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in neural information processing systems*, pages 820–828.

Fei Huang, Alexander Yates, Arun Ahuja, and Doug Downey. 2011. Language models as representations for weakly supervised nlp tasks. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 125–134.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL*

on interactive poster and demonstration sessions, pages 177–180. Association for Computational Linguistics.

Nimrita Koul and Sunilkumar S Manvi. 2019. A proposed model for neural machine translation of sanskrit into english. *International Journal of Information Technology*, pages 1–7.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

DT Mane, PR Devale, and SD SURYAWANS. 2010. A design towards english to sanskrit machine translatio and sy thesizer syste 1 si grule base approach. *Int J Multidisp Res Adv Eng*, 1(1).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Vimal Mishra and RB Mishra. 2008. Study of example based english to sanskrit machine translation. *Polibits*, (37):43–54.

Vimal Mishra and RB Mishra. 2009. Divergence patterns between english and sanskrit machine translation. *INFOCOMP Journal of Computer Science*, 8(3):62–71.

Vimal Mishra and RB Mishra. 2010. Approach of english to sanskrit machine translation based on case-based reasoning, artificial neural networks and translation rules. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 2(4):328–348.

Kishore Papineni, 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*, pages 311–318.

Sarita G Rathod. 2014. Machine translation of natural language using different approaches. *International journal of computer applications*, 102(15).

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.