Translation Techies @DravidianLangTech-ACL2022-Machine Translation in Dravidian Languages

Piyushi Goyal **Musica Supriya** **Dinesh Acharya U** **Ashalatha Nayak**
Department of Computer Science & Engineering
Manipal Institute of Technology
Manipal Academy of Higher Education, Manipal, Karnataka 576104 , India
piyoo.goyal@gmail.com
{musica.supriya, dinesh.acharya, asha.nayak}@manipal.edu

Abstract

This paper discusses the details of submission made by team Translation Techies to the Shared Task on Machine Translation in Dravidian languages- ACL 2022. In connection to the task, five language pairs were provided to test the accuracy of submitted model. A baseline transformer model with Neural Machine Translation(NMT) technique is used which has been taken directly from the OpenNMT framework. On this baseline model, tokenization is applied using the IndicNLP library. Finally, the evaluation is performed using the BLEU scoring mechanism.

1 Introduction

A multilingual country such as India has a diversified population. Several languages are spoken at various parts of the country (Chakravarthi et al., 2019, 2018). Human spoken languages of India are divided into various groups. Indo-Aryan and Dravidian languages are the two primary families. For almost 2600 years, there has been a recorded Tamil literature (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019, 2020a,b, 2021). The earliest period of Tamil literature, known as Sangam literature, is said to have lasted from from 600 BC to AD 300. Among Dravidian languages, it possesses the oldest existing literature. The earliest epigraphic documents discovered on rock edicts and 'hero stones' date from the sixth century BC. In Tamil Nadu, the Archaeological Survey of India discovered over 60,000 of the 100,000 odd inscriptions discovered in India (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). However, the English language dominates the content available on the Internet (Chakravarthi, 2020; Chakravarthi and Muralidaran, 2021). It is a difficult task to have a human translator who can translate texts in across all language pairs. This forms the basic purpose of this shared task. We need constructive and precise computer algorithms that need minimal human intervention to bridge this massive language divide. Machine translation can be used to complete this task effectively.

With various conversational AIs and voice assistants taking the world by storm, translation of native and low-resource languages has become imperative. The Dravidian languages are morphologically rich in nature and are hence, difficult to deal with (Chakravarthi et al., 2020). The scripts are different when compared to the Western scripts and require more attention (Sampath et al., 2022; Ravikiran et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). This task is an attempt to utilise the existing tools for translation of low-resource Dravidian languages. The goal here is to develop a smooth algorithm which will help in knowledge dissemination, and end-to-end speech translation. We have used Neural Machine Translation in our approach towards this problem. The rest of the paper is structured as follows: Section 2 Literature Survey, Section 3 Methodology, Section 4 Results obtained from the shared task and Section 5 Conclusion and Future Scope.

2 Literature Survey

As a result of improved processing capabilities and training data, intensive research on MT began in the early 1950s (Hutchins, 2004), and it has progressed significantly since the 1990s. To accomplish more and more accurate machine translation, a variety of methodologies have been proposed (Hutchins, 2004). Statistical Machine Translation(SMT), a subtype of Corpus-based translation, was the most extensively applied of them because it produced better results previous to the NMT systems.

The statistics-based method to machine translation does not employ traditional language data. It functions on the basis of the probability principle. In this situation, the word in the source language corresponds to the comparable word in the target

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 120 - 124

language. However, a large corpus of trustworthy translations in both the source and target languages is required. This strategy is comparable to that of IBM's research group in the early 1990s, which had some success with speech recognition and machine translation.

NMT approaches are data driven and demands language resources such as a parallel corpora for translation. When it comes to large-scale translation projects like English to German and English to French, (Wu et al., 2016), it outperformed typical MT models. In recent years, several architectures for neural network-based machine translation have been proposed, including a simple encoder-decoder based model, an RNN based model, and an LSTM model that learns problems with long-range temporal dependencies, as well as an Attention mechanism-based model, which is machine translation's most powerful neural model.

The evolution of Machine translation approaches on Indian Languages was surveyed in detail giving an overview from rule based methods used, Statistical machine translation methods implemented on the major Indian languages(J., 2013).

A sequence-to-sequence model based machine translation system for the Hindi language was proposed (Shah et al., 2018) which encouraged the use of NMT architecture on Indian languages.

The neural based approaches in Machine Translation have gained more scope as the accuracy improves based on the quality of the parallel corpora and it may be beneficial to develop an extension of the encoder–decoder paradigm that learns to align and translate together (Bahdanau et al., 2016).

Transformer computes input and output representations using self-attention rather than sequence aligned RNNs or convolution (Vaswani et al., 2017).

This shared task addresses this issue and we have implemented the Transformer model using OpenNMT platform (Klein et al., 2017). The essential principles of n-gram precision are used by BLEU (Papineni et al., 2002) to calculate similarity between the reference and created phrases. Since it employs the average score of all discoveries in the test dataset rather than presenting results for each sentence. Hence, we have used the BLEU metric for the model in this paper.

The base model chosen is Transformer architecture on OpenNMT framework and we have further enhanced this model and applied to given five language pairs. The results are tabulated based on BLEU metric.

3 Methodology

This task explores the transformer approach in OpenNMT framework. With less resources in hand, the OpenNMT framework offers best models to experiment upon. The baseline model was a Transformer architecture directly borrowed from the OpenNMT framework and used on the Dravidian Language pairs. [OpenNMT-py toolkit with commands] The model was used for five language pairs with different sizes of training, validation and testing data as shown in Table 1.

Table 1: The five language pairs and the sizes of their training, validation and testing files (Kumar M et al., 2022).

Source	Target	Dataset size (in lines)		
		Train	Valid	Test
Kannada	Malayalam	90974	2000	2000
Kannada	Sanskrit	9470	1000	1000
Kannada	Tamil	88813	2000	2000
Kannada	Telugu	88503	2000	2000
Kannada	Tulu	8300	1000	1000

The baseline model used the parallel corpora without pre-processing and it was observed that most of the words were tagged as unknown in the output prediction file on the test set. So, the configuration file was altered. In the configuration file, the learning rate is set as 2, training steps as 10,000, valid steps as 500 and checkpoints to save the model was created at every 500 steps. This file was used without any further modification across all given language pairs. It contains the paths to the training source and target files, and the validation files of the same.

On both the encoder and decoder, this configuration will run the default 2-layer LSTM model containing 500 hidden units. The supplied parameters worldsize = 1 and gpu ranks[0], which operates on a single GPU.

The vocab is built using the 'onmt_build_vocab' command present in the OpenNMT-py package installed in the first step. In this, '-n_sample' represents the amount of lines extracted from each corpus, used to create vocabulary.

Without any tokenization or transforms, this is the simplest configuration conceivable. Using this, many unknown tokens and less translated words

were obtained. We used the same hyperparameters for all the five language pairs.

In order to get better results, the input datasets were tokenized before training, using the IndicNLP library (Kunchukuttan, 2020). This helped to get way better results for all the language pairs as more translated words, and lesser unknown tokens were produced.

4 Results

The sample text for five language pairs based on training data is shown in Figure 1.

Figure 1: Sample data of all five language pairs from the training set

```
Kannada   : ಭಿನ್ನವಾದದ್ದು ಮಾಡುವ ಬಯಕೆ ಇತ್ತು.
Malayalam : വ്യത്യസ്തമായി എന്തെങ്കിലും ചെയ്യണമെന്ന് ആഗ്രഹിച്ചിരുന്നു.

Kannada   : ಅನುಜನು ವಾಹನವನ್ನು ಚಲಾಯಿಸುತ್ತಿದ್ದಾನೆ
Sanskrit  : अनुजः यानम् चालयति

Kannada   : 40 ಕೋಟಿ ಮಂಜೂರಾಗಿದೆ.
Tamil     : 40 கோடி !

Kannada   : ಆಗುವುದು ಎಲ್ಲಾ ಸಿನಿಮಾ ಇಂಡಸ್ಟ್ರಿ ಗಳಲ್ಲೂ ಸಹಜ.
Telugu    : సినిమా ఇండస్ట్రీలోనైతే ఆ విషయం నూటికి నూరుపాళ్లు నిజం.

Kannada   : ದೇವರು ಕೊಡುವಾಗ ಎಲ್ಲವನ್ನೂ ಕೊಡುತ್ತಾರೆ
Tulu      : ದೇವೆರ್ ಕೊರ್ನಗ ಮಾತಲ ಕೊರ್ಪೆರ್
```

After running the model on all the five language pairs of Kannada-Malayalam, Kannada-Sanskrit, Kannada-Tamil, Kannada-Telugu , and Kannada-Tulu; the following BLEU scores in Table 2 were obtained:

Table 2: The BLEU scores calculated for the prediction files of the respective language pairs.

Source	Target	BLEU Score
Kannada	Malayalam	0.0729
Kannada	Sanskrit	0.7482
Kannada	Tamil	0.0798
Kannada	Telugu	0.1242
Kannada	Tulu	0.6149

Despite using the same model and parameters for all the pairs, different BLEU scores were obtained. It can be observed that the model gave best results for Sanskrit and Tulu despite the fact that the dataset was smaller for these two. This is because the test set has similar kind of sentences when compared to train set. There is an overlap of sentences and words used in the source and target sets. As for Telugu, it performed fairly well. The datasets used were small and limited. Hence, our results do not give much insights into the performance of the model. However, the scores can be further improved by enhancing the quality of the dataset and enhancing the model. Better transforms and pre-processing techniques need to be applied on the datasets before training to achieve the same. Some techniques can be byte-pair encoding (Sennrich et al., 2015) and data augmentation (Wei and Zou, 2019) to get more translated words.

5 Conclusion and Future Scope

This paper describes the details of submission made by team Translation Techies to the Shared Task on Machine Translation in Dravidian languages-ACL 2022. The Transformer architecture present in OpenNMT framework along with modifications is implemented in this shared task. The current model can be further improved by providing larger datasets and pre-processing them in detail. We can use data augmentation and byte-pair encoding techniques as well. Subword tokenization is also a good technique to alleviate the problem with such low-resource language pairs (Dhar et al., 2021). The efficient translation of the Dravidian languages is necessary as the need for smart systems are rising rapidly.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Work-*

shop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Navaneethan Rajasekaran, Mihael Arcan, Kevin McGuinness, Noel E. O'Connor, and John P. McCrae. 2020. Bilingual lexicon induction across orthographically-distinct under-resourced Dravidian languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 57–69, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Prajit Dhar, Arianna Bisazza, and Gertjan van Noord. 2021. Optimal word segmentation for neural machine translation into Dravidian languages. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 181–190, Online. Association for Computational Linguistics.

W. John Hutchins. 2004. The georgetown-ibm experiment demonstrated in january 1954. In *Machine Translation: From Real Users to Research*, pages 102–114, Berlin, Heidelberg. Springer Berlin Heidelberg.

Antony P. J. 2013. Machine translation approaches and survey for Indian languages. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 1, March 2013*.

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.

Anand Kumar M, Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadarshini, Shashirekha Hosahalli Lakshmaiah, and John Philip McCrae. 2022. "findings of the shared task on Machine Translation in Dravidian languages". In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

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

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishor Kumar Ponnusamy, and Santhiya Pandiyan. 2022. Findings of the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. 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.

Parth Shah, Vishvajit Bakrola, and Supriya Pati. 2018. *Neural Machine Translation System for Indic Languages Using Deep Neural Architecture*, pages 788–795.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

C. N. Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based part of speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196.

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.