Dravidian-T5: A Neural Machine Translation Approach for Dravidian Languages

Sujal T S, Satish G T, Sudeep Patil, Rohit S, and Dr.Uday Kulkarni

KLE Technological University, Hubballi, Karnataka, India
{01fe22bci051, 01fe22bci053, 01fe22bci055, 01fe23bci407,
uday_kulkarni}@kletech.ac.in
https://www.kletech.ac.in/

Abstract

Neural Machine Translation (NMT) for low-resource Dravidian languages is constrained by the lack of high-quality parallel corpora, hindering progress in developing effective translation systems. To address this, we propose a pivot-based data augmentation framework that leverages English as an intermediary to construct a multilingual parallel corpus. Starting with an existing Kannada-English parallel dataset, we generate parallel data by translating English sentences into multiple target languages, resulting in high-quality aligned sentence pairs. We employ a sequence-to-sequence NMT model and incorporate diverse decoding strategies during translation to ensure lexical diversity and semantic integrity. Additionally, we release a Text-To-Text Transfer Transformer (T5)-based model fine-tuned on the augmented corpus, facilitating further research and applications in multilingual Natural Language Processing (NLP). We also release DraPara, a 100k parallel corpus dataset for Dravidian languages, to support future advancements in low-resource NMT. Evaluation on an unseen subset of the data yields an average Bilingual Evaluation Understudy (BLEU) score of 0.69 and a sentence similarity score of 0.7764, demonstrating the effectiveness of our approach. This framework contributes significant resources and insights for NMT in low-resource Dravidian languages.

1 Introduction

Translation involves converting text from one language to another while preserving both the meaning and context. The goal of machine translation is to generate translations that are lexically diverse yet semantically accurate. Figure 1 illustrates how ideal translations should appear. In an optimal scenario, the ideal translation would aim for higher Bilingual Evaluation Understudy (BLEU) scores [16] while maintaining high semantic similarity[29] to the original text. Translation is pivotal in several applications, such as cross-lingual information

retrieval [20], multilingual dialogue systems [6], and neural machine translation [11]. However, ensuring accurate translations that maintain lexical variety and semantic integrity is a challenging task, especially when dealing with low-resource languages like Dravidian languages, which have complex grammar, morphology, and vocabulary.

Figure 1. *Example Translation*

To address the challenges in machine translation, recent approaches leverage sequence-to-sequence (seq2seq) Transformer models to generate high-quality translations [10]. These models are typically fine-tuned on large-scale parallel corpora to adapt effectively for translation tasks. However, they struggle when applied to low-resource settings due to the scarcity of parallel data, particularly in the case of Dravidian languages such as Kannada, Tamil, Telugu, and Malayalam. While multilingual models like Multilingual Text-to-Text Transfer Transformer (mT5) [1] and Multilingual Bidirectional and Auto-Regressive Transformer (mBART) [6] have made some progress, they often fall short in handling the linguistic richness and structural complexity of Dravidian languages. Moreover, most existing research focuses predominantly on high-resource languages like English, overlooking the data sparsity and specific linguistic challenges associated with low-resource languages [19],[22]. There remains a significant gap in effectively utilizing data augmentation strategies to improve translation quality for these underrepresented languages.

To overcome this, we propose a pivot-based data augmentation framework using Neural Machine Translation (NMT) to create a large, lexically diverse parallel corpus. Starting with a Kannada-English dataset, we translate the English sentences into Tamil, Telugu, and Malayalam using English as a pivot [3]. The resulting parallel corpus is filtered using semantic similarity and BLEU score to ensure high quality.

We further optimize performance by compressing the pretrained mT5 model [1] by retaining only the vocabulary relevant to Dravidian languages. Inspired by Kumar et al. [23], we also insert language tags to guide multilingual decoding. This results in a smaller, efficient Dravidian-T5 model tailored for low-resource NMT tasks.Evaluation on unseen datasets demonstrates that our model pro-

duces diverse and contextually appropriate translations, achieving an average BLEU score of 0.69 and a sentence similarity score of 0.7764. Our key contributions include A novel pivot-based data augmentation framework for NMT, A Dravidian-T5 model[1] based on T5 architecture, and A new 100k Dravidian multilingual parallel corpus[2] *DraPara*, has been released publicly.

The remainder of this paper is as follows. Section 2 describes related work and highlights existing research gaps. Section 3 elaborates on our methodology. Section 4 presents results. Section 5 discusses limitations and future work.

2 Related Work

NMT for low-resource languages has gained significant attention with the rise of multilingual pretrained models. Architectures like mBART [6], mT5 [1], and NLLB-200 [2] have enabled zero-shot [30] and few-shot [31] translation capabilities across diverse language families. However, the Dravidian language group—comprising Kannada, Tamil, Telugu, and Malayalam—remains critically underrepresented in model training corpora and evaluation, leading to suboptimal translation performance. This gap is further exacerbated by the lack of standardized datasets and linguistic resources tailored for these languages. To mitigate data scarcity, previous work has explored synthetic backtranslation [32] and data augmentation methods, though their effectiveness is often constrained by noise and quality concerns.

While the FLoRes benchmark [19] provides multilingual evaluation, its coverage of Dravidian languages remains limited. Even large-scale models like NLLB-200 offer insufficient evaluation for Dravidian languages due to inadequate language-specific training data. Chakravarthi et al.[22] contributed Dravidian datasets for code-mixed tasks but not for parallel corpus construction. This highlights a critical need for curated bilingual resources and task-specific adaptation strategies. Inspired by Kumar et al.[23], we adopt a lightweight adaptation of mT5 that involves embedding layer pruning and language tagging for efficient multilingual decoding. This improves both training efficiency and translation accuracy by reducing computational overhead and increasing language-specific representation.

Additionally, research by Tang et al.[21] and Aji et al.[26] emphasized the importance of balancing semantic similarity and lexical diversity. Our framework is the first to combine pivot-based corpus generation, rigorous filtering, and compressed model adaptation into a unified approach specifically designed for Dravidian low-resource translation. This end-to-end solution significantly narrows the gap in multilingual NMT for these underrepresented languages and sets a strong foundation for future research in inclusive language technologies.

[1] https://huggingface.co/Sujalts/Dravidian-T5

[2] https://huggingface.co/datasets/Sujalts/DraPara

2.1 Text-To-Text Transfer Transformer (T5)

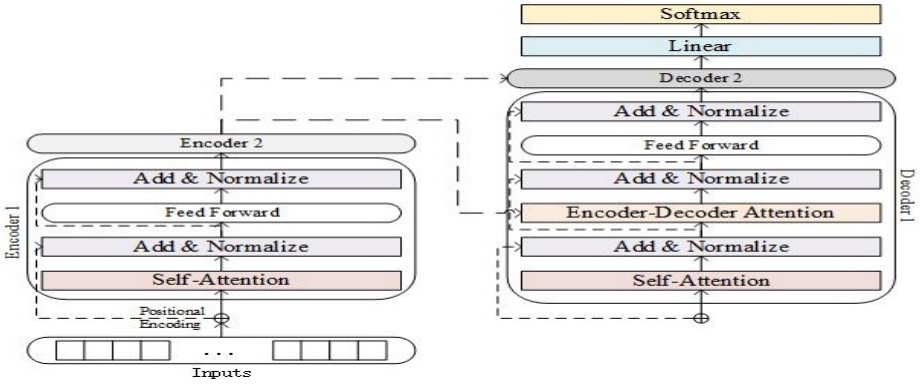

Figure 2. *Architecture of T5*[13]

The T5 (Text-to-Text Transfer Transformer) model [13] is a Transformer-based encoder-decoder architecture, as shown in Fig. 2. The encoder processes the input sentence to generate dense contextual representations using two main components: multi-head self-attention, which captures relationships between tokens within the input sequence, and feedforward layers that transform token embeddings to enhance context representation. The decoder generates the output sentence token-by-token, leveraging cross-attention to focus on the encoder's output for guidance and masked selfattention to ensure that each token is generated based only on previously generated tokens. Together, these components enable T5 to perform text-to-text transformations effectively across a variety of natural language processing tasks.

2.2 No Language Left Behind (NLLB-200)

The NLLB-200 (No Language Left Behind) model [2] is a Transformer-based multilingual translation system supporting over 200 languages, including low-resource ones. It uses attention mechanisms and a Mixture of Experts (MoE) to assign specialized layers to language groups, enhancing performance and reducing interference. A shared multilingual embedding space enables knowledge transfer from high- to low-resource languages. Curriculum learning and self-supervised techniques like backtranslation help address data scarcity. Sparse MoE computation ensures scalability, while dynamic expert activation and multilingual embeddings improve contextual accuracy during inference.

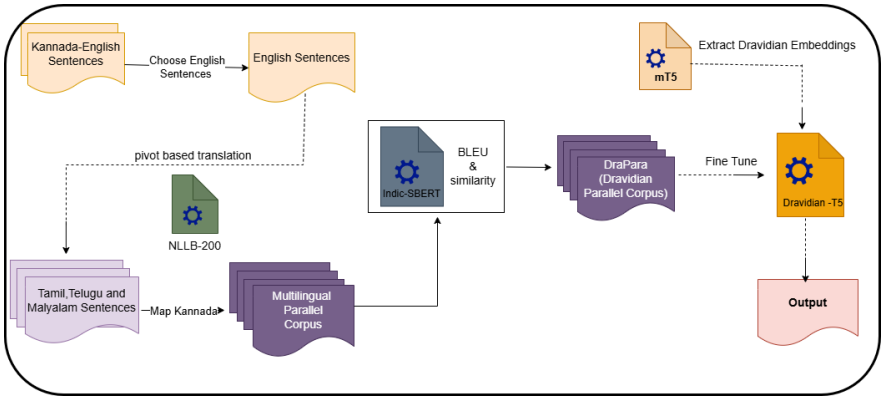

Figure 3. Comprehensive Architecture of the Proposed System Framework

3 Proposed Work

This section details the overall proposed methodology of our work. We first outline a framework to construct a parallel Dravidian language corpus by utilizing a Kannada-English parallel dataset and translating English sentences into Tamil, Telugu, and Malayalam as mentioned in subsection 3.1. Data is filtered based on lexical diversity and semantic similarity metrics to ensure high-quality sentence pairs across the languages in subsection 3.2. Further, we fine-tune a pretrained sequence-to-sequence transformer (mT5) on the augmented corpus to create a Dravidian language translation model in subsection 3.3.

3.1 Data Augmentation Framework

In this section, we present a comprehensive overview of the data augmentation framework employed to enhance the quality and diversity of multilingual paraphrase data. We initiate the process by randomly sampling 200,000 sentence pairs from the Kannada-English (Kn-En) Samanantar dataset [25], a widely used benchmark for low-resource Indian language translation tasks. To ensure relevance and richness of content, the selected sentences are filtered to retain only those with a minimum token length of four, effectively eliminating overly short and less informative examples.

Following the initial selection, we apply a preprocessing stage designed to normalize textual data and eliminate inconsistencies. This involves removing extraneous punctuation such as commas, periods, question marks, and exclamation marks, which may introduce noise during the translation and paraphrasing phases. We adopt a pivot-based translation approach to generate paraphrased data across multiple Dravidian languages. Specifically, the English counterparts of the Kannada-English pairs are used as a pivot to produce translations in Tamil, Telugu, and Malayalam, leveraging the multilingual capabilities of the NLLB-200 model as shown in figure 3 and entire process is summarized and

Algorithm 1: Pivot-Based Multilingual Paraphrase Generation using NLLB-200

Input *Samanantar Kn-En dataset* $\mathcal{D}_{Kn\text{-}En}$*, Target Languages* $L = \{Ta, Te, Ml\}$ **Output** *Multilingual paraphrase corpus* $\mathcal{D}_{augmented}$
Sample 200,000 sentence pairs from $\mathcal{D}_{\text{Kn-En}}$ at random;
Filter out pairs where Kannada sentence has < 4 tokens;
Normalize text: remove punctuation (commas, periods, question marks, exclamations);
Extract English sentences from selected pairs into set $\mathcal{D}_{\text{En}}$;
Initialize empty set $\mathcal{D}_{\text{augmented}}$;
for *each sentence* $s_{En} \in \mathcal{D}_{En}$ **do**
    for *each language* $l \in L$ **do**
        $s_l \leftarrow \text{NLLB}_{\text{En}\rightarrow l}(s_{\text{En}})$;
        Add s_l to $\mathcal{D}_{\text{augmented}}$;
    end
end
Map each translated s_l to its corresponding original Kannada sentence;
return $\mathcal{D}_{\text{augmented}}$ (Dravidian parallel corpus);

detailed in Algorithm 1. This approach enables controlled generation of parallel sentences across languages that may not have abundant direct parallel data.

3.2 Data filtering

We end up with 1,30,000 sentence pairs. we apply a multi-stage filtering process to ensure high-quality Translation data using Indic-SBERT[28]. Each sentence pair is evaluated using semantic similarity and BLEU score metrics to identify and remove outliers. Pairs with a semantic similarity score below 0.8 are discarded to maintain semantic integrity, while those with BLEU scores below 0.5 are removed to eliminate lexically dissimilar outputs as mentioned in equation 1. Additionally, duplicate sentence pairs are filtered out to enhance dataset uniqueness. This rigorous filtering pipeline results in a refined Dravidian language corpus comprising 1,07,057 high-quality and diverse sentence pairs in Kannada, Tamil, Telugu, and Malayalam.

$$\text{Filter}(s_{\text{src}}, s_{\text{tgt}}) = \begin{cases} & \text{if} \\ & \text{BLEU}(s_{\text{src}}, s_{\text{tgt}}) > 0.5 \\ & \text{and} \\ \text{keep}, & \text{Sim}_{\cos} > 0.8 \\ \\ \text{discard}, & \text{otherwise} \end{cases} \tag{1}$$

s_{src} and s_{tgt} are the source and target sentences respectively.

3.3 Fine tuning Dravidian-T5

We employed a multi-step fine-tuning procedure on an augmented dataset to adapt the pretrained mT5 model specifically for Dravidian languages. The dataset was first preprocessed by appending language-specific tokens such as <kannada>, <tamil>, <telugu>, and <malayalam> at the beginning of each sentence. This explicit language tagging helps the model disambiguate and better learn language-specific patterns during training.

Algorithm 2: Training Loop for Fine-Tuning

Input *Training steps T_s, Model M, Tokenizer T, Accumulation Steps A, Save Steps S, Max Length L* **Output** *Fine-tuned Model M*
for $i \leftarrow 1$ **to** T_s **do**
  $(x, y) \leftarrow$ `get_batch()` ; `// Get a batch of data`
  Tokenize x, y with T using L ; `// Tokenize with truncation and padding`
  Mask padding tokens in y with -100;
  Compute loss: `loss` $\leftarrow M(x, y)$;
  Backpropagate: `loss.backward()`;
  if i mod $A == 0$ **then**
    Perform optimizer and scheduler step ; `// Apply accumulated gradients`
    Zero gradients;
    Clear cache;
  end
  if i mod $S == 0$ ***and*** $i > 0$ **then**
    Save model and tokenizer;
  end
end
return M;

We trained the model using joint multilingual fine-tuning, where a single model is fine-tuned on all Dravidian languages simultaneously in a many-to-many fashion. This approach leverages cross-lingual transfer, allowing low-resource languages to benefit from higher-resource ones. Language tags play a critical role in guiding the model in effectively identifying input and output languages during training and inference. For fine-tuning, the Dravidian-T5 model was trained on the augmented data using the Adam optimizer with a learning rate of 3×10^{-5}. Training was carried out for 5 epochs with a batch size of 8, optimizing the model with cross-entropy loss[18]. The training process comprised 56,250 steps and required approximately 5 hours of runtime on the available GPU infrastructure. The final model checkpoint was selected based on the lowest validation loss, achieving improved performance across the Dravidian languages in the machine translation task.

Before that, as the original mT5 model was pretrained on 101 languages, we reduced its size by extracting only the embeddings corresponding to the Dravidian languages as shown in figure 3. This reduction was accomplished by updating the tokenizer and trimming the embedding layer to retain only the relevant vocabulary and embeddings using Algorithm 3, resulting in a Dravidian-T5 model that is more than 20% smaller than the original mT5. This size reduction facilitates faster training and inference, while focusing the model's capacity on the target language family.

Algorithm 3: Converting mT5 to Dravidian-T5

Input: mT5 model M with vocab size V_{orig}, corpus C, target vocab size V_{target}
Output: Dravidian-T5 with reduced vocabulary

`// 1. Load mT5`
$tokenizer, model \leftarrow$ load_pretrained("google/mt5-small");
`// 2. Token Frequency Analysis`
$count \leftarrow$ Counter();
foreach $x \in C$ **do**
  $count$.update($tokenizer.encode(x)$)
end
$kept\_ids \leftarrow$ Top-V_{target} tokens from $count$;
`// 3. Update Model Weights`
Let d be the hidden size, $V' = |kept\_ids|$;
Initialize $E' \in \mathbb{R}^{V' \times d}, \quad W' \in \mathbb{R}^{V' \times d}$;
foreach $(i, j) \in enumerate(kept\_ids)$ **do**
  if $j < V_{orig}$ **then**
    $E'[i] \leftarrow E[j]$;
    $W'[i] \leftarrow W[j]$;
  end
end
Set $E \leftarrow E'$, $W \leftarrow W'$, update $V \leftarrow V'$;
`// 4. Update Tokenizer`
Let $\mathcal{P} = [p_j \mid j \in kept\_ids]$ be new token pieces;
Replace tokenizer's vocabulary with $\mathcal{P}$;
`// 5. Save`
Save updated model, tokenizer, and SentencePiece;

4 Results and Evaluation

In this section, we analyze the results of our proposed data augmentation framework. The augmented data is evaluated using sentence-BLEU [16] and sentence similarity[29] metrics, with the latter calculated using Indic-SBERT. To assess Dravidian-T5's performance, we used it to generate translations for 1,000 unseen

sentences from each Dravidian language (Kannada, Tamil, Telugu, and Malayalam). These generated translations were evaluated using sentence-BLEU and sentence similarity scores, and details are shown in Tables 1 and 2.

Table 1. BLEU Scores Matrix

	Kannada	Tamil	Telugu	Malayalam
Kannada	10.00	6.52	6.75	6.39
Tamil	6.94	10.00	5.98	5.98
Telugu	6.10	6.08	10.00	6.25
Malayalam	6.07	5.96	5.69	10.00

Table 2. Similarity Scores Matrix

	Kannada	Tamil	Telugu	Malayalam
Kannada	1.0000	0.7203	0.7484	0.7168
Tamil	0.7013	1.0000	0.7146	0.6963
Telugu	0.7413	0.7105	1.0000	0.7143
Malayalam	0.6910	0.7137	0.6958	1.0000

Our model demonstrates results comparable to those of previous works. It generates translations that not only retain the original meaning but also exhibit creativity in their phrasing. This validates the reliability of our data augmentation framework. Table 3 presents these findings, showing that our translations are both accurate and innovative.

Table 3. Example Translations using Dravidian-T5

Language	**Sentence**	**English Meaning**
Kannada	ನಾನು ಊಟ ಮಾಡುತ್ತಿದ್ದೇನೆ, ಅಪ್ಪ ಅಡುಗೆ ಕೆಲಸ ಮಾಡುತ್ತಿದ್ದಾನೆ.	I am eating, my father is cooking.
Tamil	நான் சமையல் செய்கிறேன், அப்பா சமையல் அறையில் வேலை செய்கிறார்.	I am cooking, father is working in the kitchen.
Telugu	నేను వంట చేస్తున్నాను, నాన్న వంటగదిలో పని చేస్తున్నారు.	I am cooking, father is working in the kitchen.
Malayalam	ഞാൻ ഭക്ഷണം തയ്യാറാക്കി, അച്ഛൻ അടുപ്പിൽ ജോലി ചെയ്തു.	I cooked food, father worked near the stove.

Our framework also addresses the challenge of translation generation for low-resource languages by incorporating decoding strategies like greedy search and beam search. These strategies are well-suited to the nature of Dravidian languages, which often face challenges due to limited data availability.

Additionally, the generated data has significant potential for future work in the NLP domain, as it can be used to improve other tasks. The model developed using this augmented data shows promising results in translation generation, further validating the utility of our framework.

Table 4 illustrates the computational cost comparison between mT5 (small model) and Dravidian-T5, a model built upon the mT5 architecture. As shown in the table, our model size is reduced by over 20% by retaining only the Dravidian tokens from mT5's embeddings (Algorithm 3). This reduction makes our model more lightweight and easier to deploy. Additionally, the FLOPs (Floating Point Operations) per forward pass are significantly lower compared to mT5, enhancing the inference efficiency of our model.

Table 4. Comparison of model size, and FLOPs

Model	**Model Size**	**FLOPs**
Dravidian-T5	889.04 MB	70.05 GFLOPS
mT5-small	1145.08 MB	76.85 GFLOPs

However, our approach relies on the use of existing NMT models, luckily newer models are being introduced time by time which can interpret more and more languages like NLLB-200.

5 Conclusion and Future Work

We present Dravidian-T5, a translation model tailored for Dravidian languages (Kannada, Tamil, Telugu, and Malayalam), supported by a large parallel corpus. Our pivot-based data augmentation framework enables the generation of diverse, high-quality paraphrases that preserve meaning and enhance performance in downstream tasks like machine translation, summarization, and sentiment analysis.

Future work will extend the pivot-based approach to more low-resource languages, explore multilingual fine-tuning for zero- and few-shot translation, integrate advanced decoding methods, perform human quality evaluations, and optimize the model for real-time translation

References

1. L. Xue, N. Constant, A. Roberts, M. Kale, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," *ArXiv*, Oct. 2020. https://arxiv.org/abs/2010.11934

2. NLLB Team, S. Ainslie, H. Amrhein, C. Agarap, D. Ball, M. Barrón-Cedeño, et al., "No Language Left Behind: Scaling Human-Centered Machine Translation," *ArXiv*, 2022. https://arxiv.org/abs/2207.04672
3. M. Utiyama and H. Isahara, "A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation," *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007. https://aclanthology.org/D07-1085
4. Pluščec, D., Šnajder, J., "Data Augmentation for Neural NLP," *ArXiv*, 2023. https://arxiv.org/abs/2302.11412
5. Abudouwaili, A., et al., "Strategies to Improve Low-Resource Agglutinative Languages Morphological Inflection," In Proceedings of CoNLL 2023. *ACL Anthology*. https://aclanthology.org/2023.conll-1.34
6. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L., "Multilingual Denoising Pre-training for Neural Machine Translation," *ArXiv*, 2020. https://arxiv.org/abs/2001.08210
7. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *ArXiv*, 2019. https://arxiv.org/abs/1910.13461
8. Kunchukuttan, A., Kakwani, D., Golla, S., N., G., Bhattacharyya, A., Khapra, M. M., Kumar, P., "AI4Bharat-IndicNLP Corpus: Monolingual Corpora and Word Embeddings for Indic Languages," *ArXiv*, 2020. https://arxiv.org/abs/2005.00085
9. Khanuja, S., Bansal, D., Mehtani, S., et al., "MuRIL: Multilingual Representations for Indian Languages," *ArXiv*, 2021. https://arxiv.org/abs/2103.10730
10. M. I. Azad, R. Rajabi, and A. Estebsari, "Sequence-to-Sequence Model with Transformer-based Attention Mechanism and Temporal Pooling for Non-Intrusive Load Monitoring," *arXiv preprint*, 2023. https://arxiv.org/abs/2306.05012
11. Bahdanau, D., Cho, K., Bengio, Y., "Neural Machine Translation by Jointly Learning to Align and Translate," *ArXiv*, 2014. https://arxiv.org/abs/1409.0473
12. Sutskever, I., Vinyals, O., Le, Q. V., "Sequence to Sequence Learning with Neural Networks," *ArXiv*, 2014. https://arxiv.org/abs/1409.3215
13. Raffel, C., Shazeer, N., Roberts, A., et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *ArXiv*, 2019. https://arxiv.org/abs/1910.10683
14. Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention Is All You Need," *ArXiv*, 2017. https://arxiv.org/abs/1706.03762
15. Yenduri, G., M., R., et al., "Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," *ArXiv*, 2023. https://arxiv.org/abs/2305.10435
16. Papineni, K., Roukos, S., Ward, T., Zhu, W., "BLEU: a Method for Automatic Evaluation of Machine Translation," *ACL*, 2002. https://aclanthology.org/P02-1040
17. A. R. Lahitani, et al., "Cosine similarity to determine similarity measure: Study case in online essay assessment," 2016 4th International Conference on Cyber and IT Service Management, 2016. 10.1109/CITSM.2016.7577578
18. Mao, A., Mohri, M., Zhong, Y., "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," *ArXiv*, 2023. https://arxiv.org/abs/2304.07288
19. F. Guzmán, C. Cheung, M. Cettolo, V. Stoyanov, and M. Federico, "The FLoRes Evaluation Datasets for Low-Resource Machine Translation," *EMNLP*, 2019. https://aclanthology.org/D19-1632/

20. E. M. Ponti, S. E. Kim, A. T. Vaswani, T. Schneider, and N. C. P. Smith, "XCOPA: A Cross-lingual Causal Commonsense Reasoning Dataset," *EMNLP*, 2020. https://aclanthology.org/2020.emnlp-main.185/
21. B. Thompson and M. Post, "Paraphrase Generation as Zero-Shot Multilingual Translation: Disentangling Semantic Similarity from Lexical and Syntactic Diversity," *arXiv preprint*, 2020. https://arxiv.org/abs/2008.04935
22. B. R. Chakravarthi, S. Mishra, S. Murthy, and S. J. S., "Overview of Dravidian-CodeMix Shared Task at EACL 2021," *EACL Workshops*, 2021. https://aclanthology.org/volumes/2021.dravidianlangtech-1/
23. G. Blackwood, M. Ballesteros, and T. Ward,"Multilingual Neural Machine Translation with Task-Specific Attention," *arXiv preprint*, 2018. https://arxiv.org/abs/1806.03280
24. Y. Tang, et al., "Evaluating NMT Architectures," *ACL*, 2020. [Link not available]
25. M. Razdak, S. R. Suresh, S. S. Rao, A. Kunchukuttan, P. Bhattacharyya, and G. N. Jha, "Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages," *LREC*, 2020. https://aclanthology.org/2020.lrec-1.751
26. L. Burchell, A. Birch, and K. Heafield, "Exploring Diversity in Back Translation for Low-Resource Machine Translation," *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, 2022. https://aclanthology.org/2022.deeplo-1.8
27. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," *NIPS*, 2014. https://arxiv.org/abs/1409.3215
28. M. Raj, A. Kunchukuttan, V. Kamath, P. Bhattacharyya, and A. Talukdar, "IndicSBERT: A Transformer-based Framework for Indic Languages Sentence Embeddings," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. https://aclanthology.org/2021.emnlp-main.165
29. P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya,"A Comparison of Semantic Similarity Methods for Maximum Human Interpretability," *arXiv preprint*, 2019. https://arxiv.org/abs/1910.09129
30. Y. Guo, Y. Liao, X. Jiang, Q. Zhang, Y. Zhang, and Q. Liu, "Zero-Shot Paraphrase Generation with Multilingual Language Models," *arXiv preprint*, 2019. https://arxiv.org/abs/1911.03597
31. A. Abaskohi, S. Rothe, and Y. Yaghoobzadeh, "LM-CPPF: Paraphrasing-Guided Data Augmentation for Contrastive Prompt-Based Few-Shot Fine-Tuning," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023. https://aclanthology.org/2023.acl-short.59
32. J. Xu, Y. Ruan, W. Bi, G. Huang, S. Shi, L. Chen, and L. Liu, "On Synthetic Data for Back Translation," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. https://aclanthology.org/2022.naacl-main.32