CMU 11-711 Fall 2023 Advanced NLP Assignment 4
Low-resource Machine Translation for Dravidian Languages

Aboli Marathe[1], **Vinay Nair** [2], **Dheeraj Pai** [2]
[1] Machine Learning Department, [2] Language Technologies Institute
Carnegie Mellon University
(abolim, vinayn, dmohanda) @andrew.cmu.edu

Abstract

The adverse impact of differential representation of low-resource languages can be mitigated by solving existing challenges with novel NLP techniques. Here we present our work on two tasks: Neural Machine Translation and Named Entity Recognition for Indic languages, particularly low-resource in nature. We present **state-of-the-art performance** on Kannada-Sanskrit translation and Oriya Entity Recognition and discuss our findings from the tasks. Additionally, we apply the translation models for two unique applications: Sanskrit text translation and a general knowledge chatbot extension for speakers of Dravidian languages. The code and models are available at [1].

1 Introduction

The rich history of the world has been inscribed in ancient scrolls, narrated in dialects and passed down from generation to generation, changing and uniquely interpreted with each passing holder. As the distribution of speakers (Figure 1) varied with anthropological factors over time, certain languages became widespread and commonly spoken while others remained constrained to their regions and unknown to the larger public.

With the rapid advancements in technology, natural language processing was rapidly adopted for the former languages, a trend followed by an increase in resources, access and technological development. In this work, we aim to bridge the gap between different-resource languages by training models and solving different tasks that can be used for generating more datasets and improving access to language-specific texts.

Some notable challenges in this field include limitations in sourcing authentic texts and reliably constructing corpora for NLP tasks. Due to the variety in scripts (Figure 3) and diversity in scribes, often similar scripts could also be written in unique styles making them difficult to parse. For translation, further steps are required to construct parallel corpora and verify the translations. Thus finding architectural improvements to solve low-resource tasks is needed to mitigate the limitations of data-based constraints.

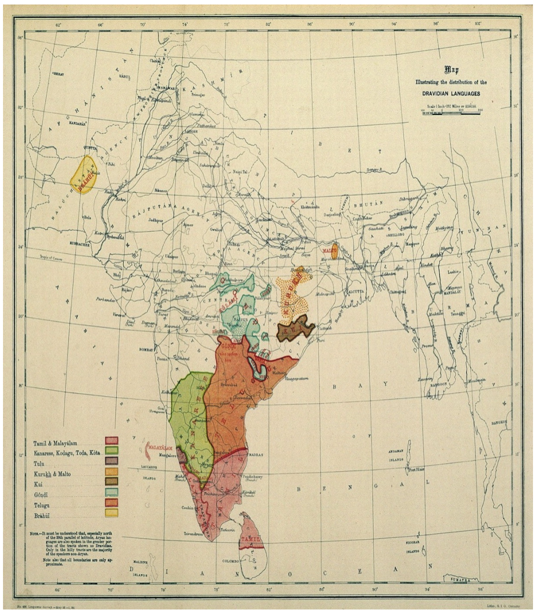

Map
Illustrating the distribution of the
DRAVIDIAN LANGUAGES
Tamil & Malayāḷam
Kanarese, Koḍagu, Toḍa, Kōta
Tuḷu
Kurukh & Malto
Kui
Gōṇḍī
Telugu
Brāhūī
NOTE.—It must be understood that, especially north of the 20th parallel of latitude, Aryan languages are also spoken in the greater portion of the tracts shown as Dravidian. Only in the hilly tracts are the majority of the speakers non-Aryan.
Note also that all boundaries are only approximate.
BOMBAY
MALAYALAM
LACCADIVE ISLANDS
MALDIVE ISLANDS
B E N G A L
I N D I A N O C E A N

Figure 1: Dravidian Language Geographical Distribution on Map [27]

Task Description: We select the 5 language pairs provided in the ACL 2022 Machine Translation in Dravidian languages Competition, which are: **Kannada, Sanskrit, Malayalam, Tulu, Tamil and Telugu** with source language as Kannada and target-X language pairs,from the Dravidian Languages listed above.

Parallelly, we also explore named entity recognition for Indic languages: **Assamese, Oriya, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and, Telugu**.

[1] https://github.com/Infernolia/11711_Assignment_IV

1.1 Dravidian Languages

The history and properties of Dravidian languages [27, 70] that influence translation modelling include:

1. Dravidian languages are approximately over 4000 years old.
2. Some of the early Dravidian inscriptions were found on cave walls in Southern India. Ancient inscriptions dating back to 2nd Century BCE were found as some of the earliest known Dravidian scripts.
3. Dravidian languages were the indigenous languages spoken widely in the Indian subcontinent.
4. Dravidian languages also influenced other Indic languages grammatically.
5. There are approximately 17 reported Dravidian languages including Kannada, Telugu , Tamil, Malayalam , Brahui , Tulu, Gondi and Coorg.
6. There are different groups of Dravidian languages including: Southern, South Central, Central and North.
7. Dravidian languages often loan words from Sanskrit, which can impact translation as we observe in the results.
8. These languages are agglutinative and follow subject-object-verb order in grammar.
9. They follow a five-vowel system.
10. There are singular and plural forms of numbers which are marked by suffixes.
11. There are transitive, intransitive and causative verbs in Dravidian grammar.
12. The gender of the words decides the gender (grammatical) across the Dravidian languages, each of which have their own gender rules and systems.
13. The grammar follows head-final and left-branching.
14. There are individual scripts for the selected languages including Southern Brahmic and Northern Brahmic for Sanskrit.
15. The vocabulary has simple, derived and compound words.
16. The pronunciations vary across Dravidian languages with dialect.

Figure 2: Sanskrit Shlokas Text Data Word Cloud

1.2 Contributions

We implement the following modules in this project:

1. **Neural Machine Translation** We present our trained model results on translation for 5 Dravidian language pairs from the low-resource taxonomy. We present state-of-the-art results on Kannada-Sanskrit translation task.
2. **Named Entity Recognition** We train across cross-lingual NER corpora to take the advantage of the structural similarity between languages and overcome low-resource data scenario limitations. We present state-of-the-art results on Oriya Entity Recognition task.
3. **Application in Domain Adaptation** We translate ancient texts in the form of Sanskrit Shlokas to spoken Dravidian languages.
4. **Application in Conversational AI** We leverage post-hoc machine translation for a multilingual chatbot specifically in Dravidian Languages to promote access to wider materials to these speakers.

2 Literature Survey

Human translators were for a long time, the only reliable medium of translation between languages, and were favoured for their interpretability, accuracy and generalizability. However, as the scale of computation grew and computer became ubiquitous, human translation started seeming inefficient and expensive to source. Sophisticated neural machine translation methods were able to automate several high-resource translation tasks with access to large training sets and sequential data processing models. One must note however, that humans are still employed in the verification and creation of translation corpora due to their fluency and versatile knowledge. Particularly for low-resource tasks which have very limited datasets, humans fluent in spoken languages are a valuable component of the dataset curation process.

Classical methods in machine translation proposed statistical modelling and techniques [28] that modelled language as distributions of words and studied the recurring patterns. With the creation of sophisticated seq-to-seq learning paradigms [67], the translation modelling quickly adopted deep-learning based methods including RNNs [58, 30], Conv2Seq [57], LSTMs [15] and BiLSTMs [24] which worked well with longer sequences that occur in language. Newer methods emerged with the ground-breaking transformer model [68] and the encoder-decoder method [49] adapted rapidly for neural machine translation tasks. Although recent architectures present novel learning paradigms, the transformer model inspired most of these models and is an essential component of their backbones.

We explored many architectures over the course of this project including the BART model [40] (denoising autoencoder-based model), and adapted MBART [42], UMT5 [14] (T5 objective) , [52]XNLI [16], NLLB [17], and M2M100 [22] (encoder-decoder) model which are popularly used due to their support for multi-lingual training, extensive pretraining and benchmark performance.

Benchmarks The above architectures are often evaluated using benchmarks like WMT [9], XNLI [16], Flores-101 [25], MLQA [41], XQuAD [6], and PAWS-X [71].

Evaluation One of the most popular metrics for evaluating translation is BLEU score [49] measuring translation similarity across ground truth and predictions which was used in this work as well.

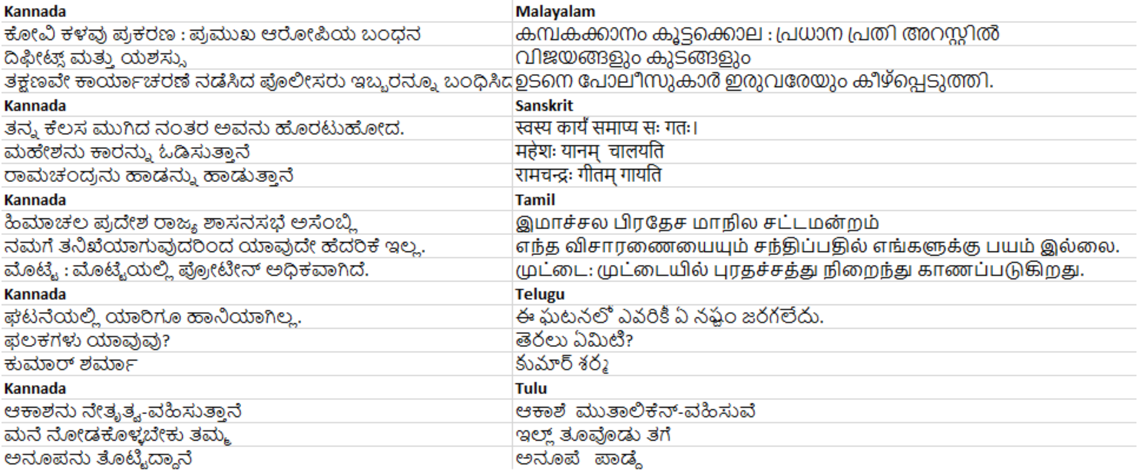
Low-Resource Machine Translation The broad taxonomy of low-resource tasks [43, 72, 50, 39, 51] can refer to corpora between 0.5- 0.1 M. Works in mitigating limitations of low resource [55] propose backtranslating corpora [60], performing data augmentation [45] , increasing pool size through unsupervised learning [20, 56], or semi-supervised learning paradigms [18], learning robustly through monolingual training [8], or incorporating transfer learning [5] and building zero-shot multi-NMT systems [23].

Named Entity Recognition Finding entities of named categories in low-resource datasets from different languages is a well-established language task[7, 21].Specifically for Dravidian languages: Kannada and Tamil, methods include LSTMs [46], zero shot (GPT-3.5-Turbo and GPT-4) LLM prompting [32] , mBERT [44] model-based approaches.

Indic Languages in NLP Increasing accessibility by solving Indic-language tasks is a field of growing interest [13, 35, 66, 10, 36]. Hindi being a widely spoken language has been widely incorporated into NLP research [62, 29, 61, 19] with applications in social good for farming [34] and legal language processing [33]. The development of large scale corpora and toolkits in this area include Indic NLP Library [38], Samanantar [53] , IndicTrans [54]. Among low-medium resource tasks, Dravidian languages, like Tamil [64, 65, 47, 59], and Malayalam [11, 12] are additionally being included in recent works.

Dravidian Languages Tamil [64, 63] is a Dravidian language spoken primarily by the Tamil population and commonly in Southern India (Tamil Nadu and Union Territory of Puducherry), Singapore and Sri Lanka. Many ancient texts and literature found in the form of cave wall inscriptions contain the earliest findings of Tamil inscriptions in Tamil Brahmi Script. Malayalam [12, 11] appears closely to Tamil historically, with the 13th century AD bringing more clear distinctions as branches in the languages.

Task The ACL Competition for Machine Translation in Dravidian languages 2022 [37] introduced 5 MT tasks: Kannada-Malayalam, Kannada-Sanskrit, Kannada-Tamil, Kannada-Telugu and Kannada - Tulu which are tested using BLEU score. We focus on the leading submissions for this competition [69, 26] and augment the experimentation to go beyond reproducing the competition results. For named-entity recognition we use the Naamapadam [44] benchmark.



Kannada	**Malayalam**
ಕೋವಿ ಕಳವು ಪ್ರಕರಣ : ಪ್ರಮುಖ ಆರೋಪಿಯ ಬಂಧನ	കമ്പകക്കാനം കൂട്ടക്കൊല : പ്രധാന പ്രതി അറസ്റ്റിൽ
ದಿಫೀಟ್ಸ್ ಮತ್ತು ಯಶಸ್ಸು	വിജയങ്ങളും കുടങ്ങളും
ತಕ್ಷಣವೇ ಕಾರ್ಯಾಚರಣೆ ನಡೆಸಿದ ಪೊಲೀಸರು ಇಬ್ಬರನ್ನೂ ಬಂಧಿಸಿದ	ഉടനെ പോലീസുകാർ ഇരുവരേയും കീഴ്പ്പെടുത്തി.
Kannada	**Sanskrit**
ತನ್ನ ಕೆಲಸ ಮುಗಿದ ನಂತರ ಅವನು ಹೊರಟುಹೋದ.	स्वस्य कार्यं समाप्य सः गतः।
ಮಹೇಶನು ಕಾರನ್ನು ಓಡಿಸುತ್ತಾನೆ	महेशः यानम् चालयति
ರಾಮಚಂದ್ರನು ಹಾಡನ್ನು ಹಾಡುತ್ತಾನೆ	रामचन्द्रः गीतम् गायति
Kannada	**Tamil**
ಹಿಮಾಚಲ ಪ್ರದೇಶ ರಾಜ್ಯ ಶಾಸನಸಭೆ ಅಸೆಂಬ್ಲಿ	இமாச்சல பிரதேச மாநில சட்டமன்றம்
ನಮಗೆ ತನಿಖೆಯಾಗುವುದರಿಂದ ಯಾವುದೇ ಹೆದರಿಕೆ ಇಲ್ಲ.	எந்த விசாரணையையும் சந்திப்பதில் எங்களுக்கு பயம் இல்லை.
ಮೊಟ್ಟೆ : ಮೊಟ್ಟೆಯಲ್ಲಿ ಪ್ರೋಟೀನ್ ಅಧಿಕವಾಗಿದೆ.	முட்டை: முட்டையில் புரதச்சத்து நிறைந்து காணப்படுகிறது.
Kannada	**Telugu**
ಘಟನೆಯಲ್ಲಿ ಯಾರಿಗೂ ಹಾನಿಯಾಗಿಲ್ಲ.	ఈ ఘటనలో ఎవరికీ ఏ నష్టం జరగలేదు.
ಫಲಕಗಳು ಯಾವುವು?	తెరలు ఏమిటి?
ಕುಮಾರ್ ಶರ್ಮಾ	కుమార్ శర్మ
Kannada	**Tulu**
ಆಕಾಶನು ನೇತೃತ್ವ-ವಹಿಸುತ್ತಾನೆ	ಆಕಾಶೆ ಮುತಾಲಿಕೆನ್-ವಹಿಸುವೆ
ಮನೆ ನೋಡಕೊಳ್ಳಬೇಕು ತಮ್ಮ	ಇಲ್ಲ್ ತೂವೊಡು ತಗೆ
ಅನೂಪನು ತೊಟ್ಟಿದ್ದಾನೆ	ಅನೂಪೆ ಪಾಡ್ಡೆ

Figure 3: Data Samples of Language-Translation Pairs used in this Task

3 Methodology

In this report, we quickly summarize our previous progress and related work from the Assignment 3 and build onto the results with our latest state-of-the-art results and comprehensive error analysis. For this error analysis, we use the Compare-MT [48]- A Tool for Holistic Comparison of Language Generation. Our evaluation methods were consistent with the implementation in Paper 1 [69]. The larger fine-tuned models have been made available at: https://huggingface.co/aboli-marathe.

3.1 Neural Machine Translation

We begin the experimentation by reproducing Conv2Seq baseline from [69]. This specific NMT architecture was specified in previous literature as built from scratch which we followed across the target tasks. We present this baseline in Table 2 with a error margin of (0.001) difference in our reproduction and the reported baseline.

kn-ml	kn-ta	kn-te	kn-tu	kn-sn
90,974	88,813	88,503	9,470	8,300

Table 1: Machine Translation Dataset Size

Data Processing Our implementation was largely benefited by the indicNLP library [38] which provides support for Indic language textual data processing tasks. Our data processing pipeline consisted of the components:

1. normalization
2. pre-tokenization
3. transliteration
4. BPE encoding
5. binarization
6. fairseq (dictionary-vocabulary build)

The competition datasets consist of 5 parallel corpora with train-dev-test dataset of sizes as in Table 1.

In the modelling phase, we experiment with 3 key architectures for the translation tasks: transformer-based model IT2 [4], NLLB [1] and a transformer model (trained from scratch- custom for each language pair). The transformer model has 18.5 M parameters, with 3 encoder-decoder layers, 8 attention heads, and dropout of 0.1. To improve performance and achieve state-of-the-art results we augment the low-resource datasets with samples from Bharat Parallel Corpus Collection (BPCC) [4], using the Indic-Trans translation model and convert english sentences to parallel Kannada Text to join the Sanskrit corpus. In the second half, we follow [26] and use Transformer from OpenNMT framework (fp16, lr = 2, 10k steps). This transformer has 6 encoder-decoder layers.

3.2 Named Entity Recognition

3.2.1 Dataset

For our dataset, we utilized the Naamapadam Dataset, which includes labeled Named Entity Recognition (NER) data for 11 Indian languages. Our primary focus was on the low-resource language, Oriya. The Table 3 displays the number of tokens available for each language in the Naamapadam dataset.

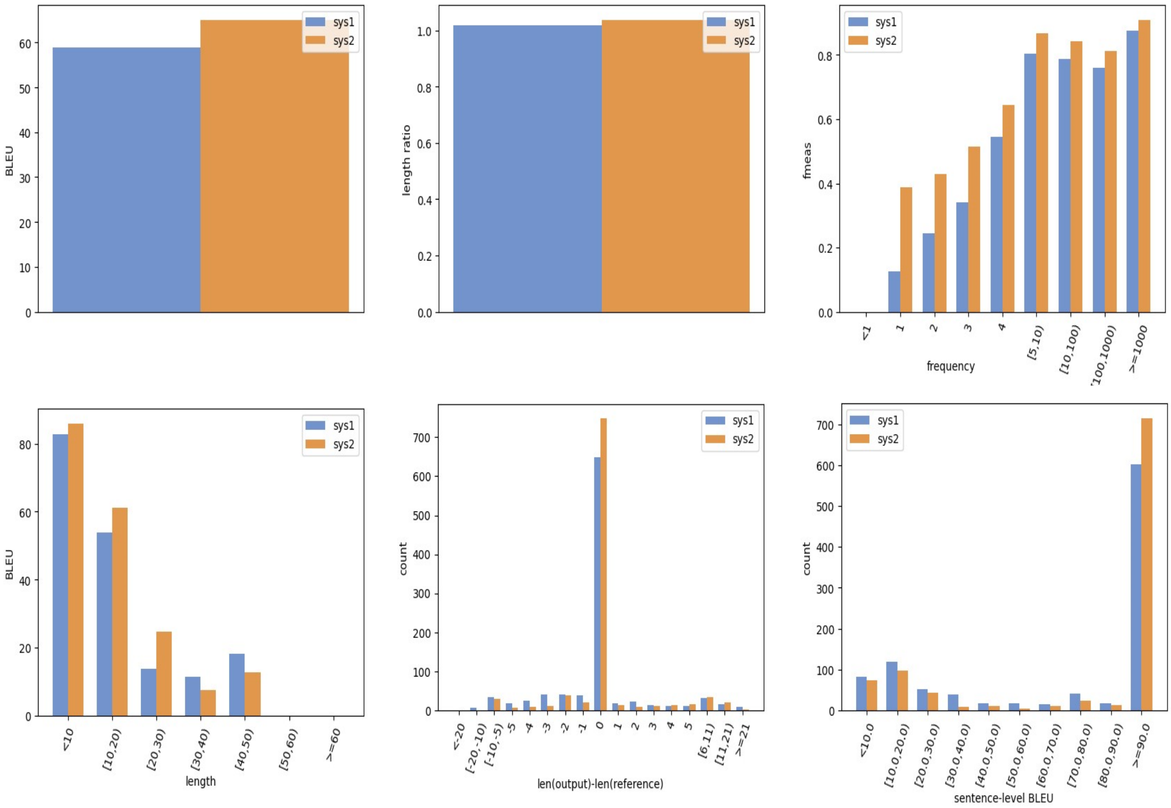
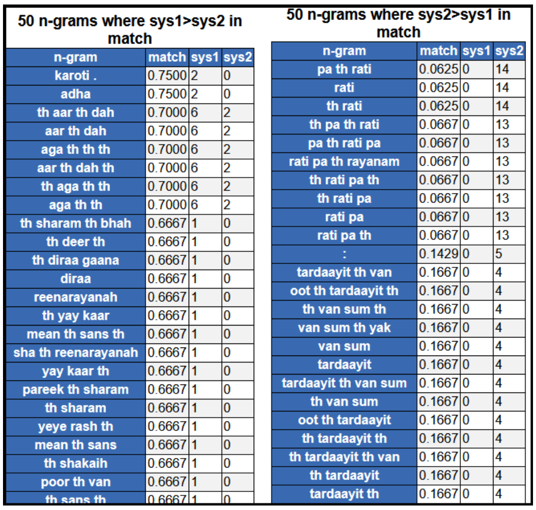


Figure 4: Error Analysis of Kannada-Sanskrit Translation Systems



50 n-grams where sys1>sys2 in match

n-gram	match	sys1	sys2
karoti .	0.7500	2	0
adha	0.7500	2	0
th aar th dah	0.7000	6	2
aar th dah	0.7000	6	2
aga th th th	0.7000	6	2
aar th dah th	0.7000	6	2
th aga th th	0.7000	6	2
aga th th	0.7000	6	2
th sharam th bhah	0.6667	1	0
th deer th	0.6667	1	0
th diraa gaana	0.6667	1	0
diraa	0.6667	1	0
reenarayanah	0.6667	1	0
th yay kaar	0.6667	1	0
mean th sans th	0.6667	1	0
sha th reenarayanah	0.6667	1	0
yay kaar th	0.6667	1	0
pareek th sharam	0.6667	1	0
th sharam	0.6667	1	0
yeye rash th	0.6667	1	0
mean th sans	0.6667	1	0
th shakaih	0.6667	1	0
poor th van	0.6667	1	0
th sans th	0.6667	1	0

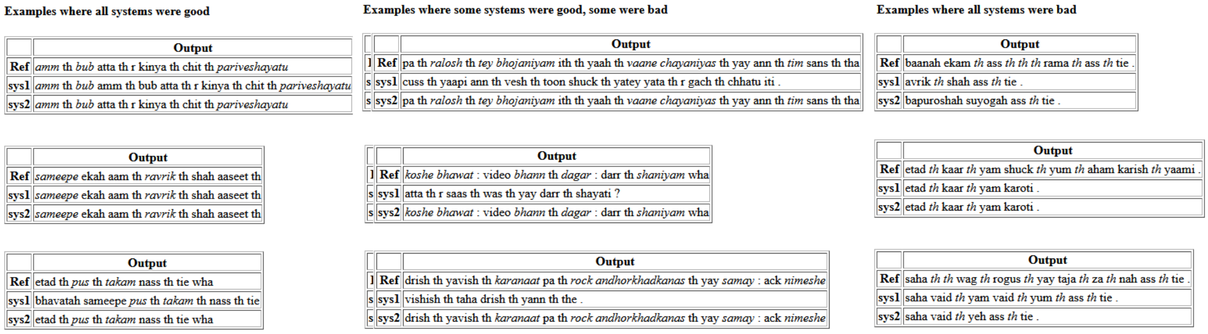
50 n-grams where sys2>sys1 in match

n-gram	match	sys1	sys2
pa th rati	0.0625	0	14
rati	0.0625	0	14
th rati	0.0625	0	14
th pa th rati	0.0667	0	13
pa th rati pa	0.0667	0	13
rati pa th rayanam	0.0667	0	13
th rati pa th	0.0667	0	13
th rati pa	0.0667	0	13
rati pa	0.0667	0	13
rati pa th	0.0667	0	13
:	0.1429	0	5
tardaayit th van	0.1667	0	4
oot th tardaayit th	0.1667	0	4
th van sum th	0.1667	0	4
van sum th yak	0.1667	0	4
van sum	0.1667	0	4
tardaayit	0.1667	0	4
tardaayit th van sum	0.1667	0	4
th van sum	0.1667	0	4
oot th tardaayit	0.1667	0	4
th tardaayit th	0.1667	0	4
th tardaayit th van	0.1667	0	4
th tardaayit	0.1667	0	4
tardaayit th	0.1667	0	4

Figure 5: N-gram Analysis for Kannada-Sanskrit Translation

3.2.2 Model

We fine-tune a model using the pre-trained IndicBert [31] for our Named Entity Recognition (NER) task. For named-entity recognition for low resource language like Oriya we explored an alternative approach for fine-tuning. Instead of the conventional method of fine-tuning a pre-trained model solely on the target low-resource language, we first fine-tuned the model on a multilingual dataset encompassing all the languages in Naamapadam dataset, followed by a secondary fine-tuning phase specifically on the low-resource language. This methodology showed a substantial improvement in NER performance for Oriya compared traditional fine-tuning techniques that focus on a single language. The highest performing results were



Examples where all systems were good

	Output
Ref	*amm* th *bub* atta th r kinya th chit th *pariveshayatu*
sys1	*amm* th *bub* amm th bub atta th r kinya th chit th *pariveshayatu*
sys2	*amm* th *bub* atta th r kinya th chit th *pariveshayatu*

	Output
Ref	*sameepe* ekah aam th *ravrik* th shah aaseet th
sys1	*sameepe* ekah aam th *ravrik* th shah aaseet th
sys2	*sameepe* ekah aam th *ravrik* th shah aaseet th

	Output
Ref	etad th *pus* th *takam* nass th tie wha
sys1	bhavatah sameepe *pus* th *takam* th nass th tie
sys2	etad th *pus* th *takam* nass th tie wha

Examples where some systems were good, some were bad

		Output
l	Ref	pa th *ralosh* th *tey bhojaniyam* ith th yaah th *vaane chayaniyas* th yay ann th *tim* sans th tha
s	sys1	cuss th yaapi ann th vesh th toon shuck th yatey yata th r gach th chhatu iti .
s	sys2	pa th *ralosh* th *tey bhojaniyam* ith th yaah th *vaane chayaniyas* th yay ann th *tim* sans th tha

		Output
l	Ref	*koshe bhawat* : video *bhann* th *dagar* : darr th *shaniyam* wha
s	sys1	atta th r saas th was th yay darr th shayati ?
s	sys2	*koshe bhawat* : video *bhann* th *dagar* : darr th *shaniyam* wha

		Output
l	Ref	drish th yavish th *karanaat* pa th *rock andhorkhadkanas* th yay *samay* : ack *nimeshe*
s	sys1	vishish th taha drish th yann th the .
s	sys2	drish th yavish th *karanaat* pa th *rock andhorkhadkanas* th yay *samay* : ack *nimeshe*

Examples where all systems were bad

	Output
Ref	baanah ekam *th* ass *th th th* rama *th* ass *th* tie .
sys1	avrik *th* shah ass *th* tie .
sys2	bapuroshah suyogah ass *th* tie .

	Output
Ref	etad *th* kaar *th* yam shuck *th* yum *th* aham karish *th* yaami .
sys1	etad *th* kaar *th* yam karoti .
sys2	etad *th* kaar *th* yam karoti .

	Output
Ref	saha *th th* wag *th* rogus *th* yay taja *th* za *th* nah ass *th* tie .
sys1	saha vaid *th* yam vaid *th* yum *th* ass *th* tie .
sys2	saha vaid *th* yeh ass *th* tie .

Figure 6: Best and Worst Case Translation Cases for Kannada-Sanskrit Translation

Task	kn-ml	kn-ta	kn-te	kn-tu	kn-sn
Paper Baseline - Conv2Seq	0.0233	0.0303	0.0701	0.3975	0.4401
Paper Best SOTA Results	0.2963	0.3536	0.3687	0.0054	0.0351
Competition Rank 1	0.2963	0.3536	0.3687	0.6149	0.7482
Competition Rank 2	0.1301	0.1791	0.1959	0.2788	0.6209
Competition Rank 3	0.0729	0.0798	0.1242	0.0071	0.0351
Our Baseline (Recreated)	0.0222	NR	NR	NR	NR
3-way Translation IT2	0.0854	0.0751	0.0770	NR	0.1431
NLLB	0.1327	0.1317	0.1507	NR	0.0854
Model (Best)	0.1327	0.1317	0.1507	NR	0.6235
Paper 2	0.0729	0.0798	0.1242	0.6149	0.7482
Model (Ours) Test Split	0.1024	0.1054	0.1176	0.3598	0.7192
Model (Ours) Public Test	0.0886	0.0857	0.0861	0.0546	0.1428
Custom Transformer					**0.7557**

Table 2: Benchmark and SOTA Machine Translation Results Results for Dravidian Languages (Error deviation 0.003)

achieved by fine-tuning the entire model. Additionally, we investigated the application of Low Rank Adapters for this task. The F1 scores for the model fine-tuned with Low Rank Adapters are presented in Table 7.

4 Results

4.1 Neural Machine Translation

As presented in Table 2, the results across the 3 architectures vary greatly in magnitude. The 3-way IT2 translation model was able to achieve results within (0.1/0.2 error on 4/5 tasks) of the competition reported best models (Paper 1) in the table. The second model (NLLB) (approx 0.1 on 2/5 tasks and 0.2 on 2/5 tasks (Paper 1)). Our previous best Sanskrit translation model was a fine-tuned MT5 model. The second and third-last rows of the

lang	train	valid	test
as	10266	52	51
bn	961679	4859	607
gu	472845	2389	1076
hi	985787	13460	867
kn	471763	2381	1019
ml	716652	3618	974
mr	455248	2300	1080
or	196793	993	994
pa	463534	2340	993
ta	497882	2795	758
te	507741	2700	847

Table 3: Number of labeled tokens in 11 Languages on Naamapadam [44] dataset.

Language	SOTA	Ours
as	60.19	41.03
or	25.91	33.21
pa	71.81	69.24
mr	81.13	79.93
gu	81.10	80.14
bn	80.74	76.45
hi	82.93	81.49
kn	81.07	78.54
ml	81.13	80.49
ta	74.11	68.24
te	82.20	80.14

Table 4: Results for Multilingual NER Implementation

dataset correspond to model performance from the training procedure of paper 2 [26].

Finally, we obtain **state-of-the-art** performance using the transformer for scratch trained on Kannada-Sanskrit task, attaining **0.7557** on the test set. This model attained best performance due to the special data augmentation step (described in Methodology) that we performed with en-SN data.

4.2 Named Entity Recognition

In comparison to the State-of-the-Art (SOTA) model, our model demonstrates nearly equivalent performance across 10 of the 11 evaluated languages. Notably, upon fine-tuning, our model surpasses the SOTA model in low-resource language performance. Our experiments in Low-Rank Adaptation techniques demonstrates that the model maintains learning efficiency, achieving close to similar performance with an increase of approximately 9.47% in the number of weights. On the other hand we observe that our model has a subpar performance in Assamese we investigate this in our error analysis section. We present results in Tables 4,5 and 6.

4.3 Error Analysis

We compare the Fine-tuned Translation models (Sys2) with previous trained weaker models (Sys1) to understand the common errors and improvement in our machine translation system. The first step is system metric analysis in which we observe that the System 2 BLEU score is over 6 points above System 1. Across the test samples, we perform Statistical Error Distributions (Figure 4), n-grams analysis (Figure 5), success and failure cases (Figure 6) all of which demonstrate the positive cases of System 2 over System 1. We see that System 2

lang	F1 Score
as	39.76 ± 0.89
bn	76.55 ± 0.12
gu	79.86 ± 0.30
hi	81.13 ± 0.26
kn	78.11 ± 0.30
ml	80.56 ± 0.13
mr	80.07 ± 0.19
or	31.13 ± 1.48
pa	69.40 ± 0.23
ta	67.17 ± 0.80
te	80.06 ± 0.36

Table 5: Multi-lingual NER Results (with Deviation) Across 11 Languages on Naamapadam [44] benchmark.

lang	F1 Score
as	41.03
bn	76.45
gu	80.14
hi	81.49
kn	78.54
ml	80.49
mr	79.93
or	33.21
pa	69.71
ta	68.24
te	80.14

Table 6: Best Multi-lingual NER Results for 11 Languages on Naamapadam [44] benchmark.

lang	F1(Before LoRA)	F1(After LoRA)
as	39.13	39.13
bn	76.48	76.93
gu	80.00	79.45
hi	81.03	80.61
kn	77.84	77.39
ml	80.75	79.54
mr	79.94	79.61
or	29.93	31.20
pa	69.17	68.81
ta	66.31	67.72
te	79.59	80.86

Table 7: F1 scores of our NER model with and without the LoRA adapter on 11 Languages on Naamapadam [44] benchmark.

Lang	Entity	Precision	Recall	F1
as	LOC	50.00	57.14	53.33
as	ORG	42.86	54.55	48.00
as	PER	33.33	16.67	22.22
as	overall	44.00	45.83	44.90
bn	LOC	78.41	71.30	74.68
bn	ORG	73.58	68.60	71.00
bn	PER	79.46	76.69	78.05
bn	overall	77.91	73.22	75.49
gu	LOC	83.47	78.05	80.67
gu	ORG	63.14	71.12	66.89
gu	PER	87.60	84.62	86.08
gu	overall	79.48	78.93	79.21
hi	LOC	78.50	78.50	78.50
hi	ORG	66.05	75.24	70.35
hi	PER	85.80	87.22	86.50
hi	overall	77.72	81.18	79.41
kn	LOC	80.57	70.85	75.40
kn	ORG	67.78	76.11	71.70
kn	PER	87.82	78.54	82.92
kn	overall	80.39	75.65	77.95
ml	LOC	86.45	76.76	81.32
ml	ORG	65.56	57.28	61.14
ml	PER	86.11	83.33	84.70
ml	overall	82.22	75.88	78.92
mr	LOC	84.33	79.16	81.66
mr	ORG	65.40	59.85	62.50
mr	PER	89.74	84.52	87.05
mr	overall	82.18	76.79	79.40
or	LOC	67.41	34.21	45.39
or	ORG	32.20	8.30	13.19
or	PER	32.24	18.33	23.37
or	overall	43.05	20.41	27.69
pa	LOC	74.59	72.95	73.76
pa	ORG	57.69	49.18	53.10
pa	PER	76.18	74.69	75.43
pa	overall	70.83	66.78	68.74
ta	LOC	71.51	67.10	69.23
ta	ORG	54.85	53.95	54.39
ta	PER	81.06	66.32	72.95
ta	overall	70.38	63.38	66.70
te	LOC	84.67	78.88	81.67
te	ORG	72.09	70.72	71.40
te	PER	86.41	78.33	82.17
te	overall	82.86	77.05	79.85

Table 8: Entity-wise NER Model Performance Metrics upon finetuning on Assamese

Lang	Entity	Precision	Recall	F1
as	LOC	50.00	42.86	46.15
as	ORG	33.33	45.45	38.46
as	PER	100.00	16.67	28.57
as	overall	40.91	37.50	39.13
bn	LOC	78.74	71.60	75.00
bn	ORG	73.44	68.12	70.68
bn	PER	81.19	79.96	80.57
bn	overall	78.84	74.72	76.73
gu	LOC	83.36	76.66	79.87
gu	ORG	63.62	68.02	65.74
gu	PER	87.05	88.46	87.75
gu	overall	79.71	79.16	79.44
hi	LOC	78.98	78.34	78.66
hi	ORG	69.79	75.24	72.41
hi	PER	86.23	90.38	88.26
hi	overall	79.38	82.43	80.87
kn	LOC	80.69	70.35	75.17
kn	ORG	62.61	73.72	67.71
kn	PER	87.76	82.76	85.19
kn	overall	79.01	76.95	77.97
ml	LOC	85.71	77.18	81.22
ml	ORG	65.33	57.93	61.41
ml	PER	86.81	88.52	87.66
ml	overall	82.38	78.60	80.45
mr	LOC	84.49	81.09	82.75
mr	ORG	67.67	61.60	64.49
mr	PER	88.56	85.80	87.16
mr	overall	82.38	78.40	80.34
or	LOC	72.52	35.71	47.86
or	ORG	42.37	10.92	17.36
or	PER	35.78	19.26	25.04
or	overall	48.10	21.92	30.12
pa	LOC	75.51	74.15	74.82
pa	ORG	60.46	49.41	54.38
pa	PER	73.45	75.94	74.67
pa	overall	70.97	67.72	69.30
ta	LOC	71.35	65.30	68.19
ta	ORG	55.07	50.00	52.41
ta	PER	78.56	71.90	75.08
ta	overall	70.14	64.06	66.96
te	LOC	84.28	79.92	82.04
te	ORG	71.37	69.20	70.27
te	PER	85.96	81.44	83.64
te	overall	82.48	78.52	80.45

Table 9: Entity-wise NER Model Performance Metrics upon finetuning on Oriya

consistently is better at n-grams and matches reference sentences in both length ratio and overall translation. Overall, we see much better Kannada-Sanskrit translation matches and fewer errors than the error analysis for other languages performed in Assignment 3.

For our NER Model we obtain an Entity-wise performance for each language Table the final metrics for the model fine-tuned on Oriya. We observe that for most languages our model shows weaker performance in identifying organizations ('ORG'), indicating this as a common area of difficulty.

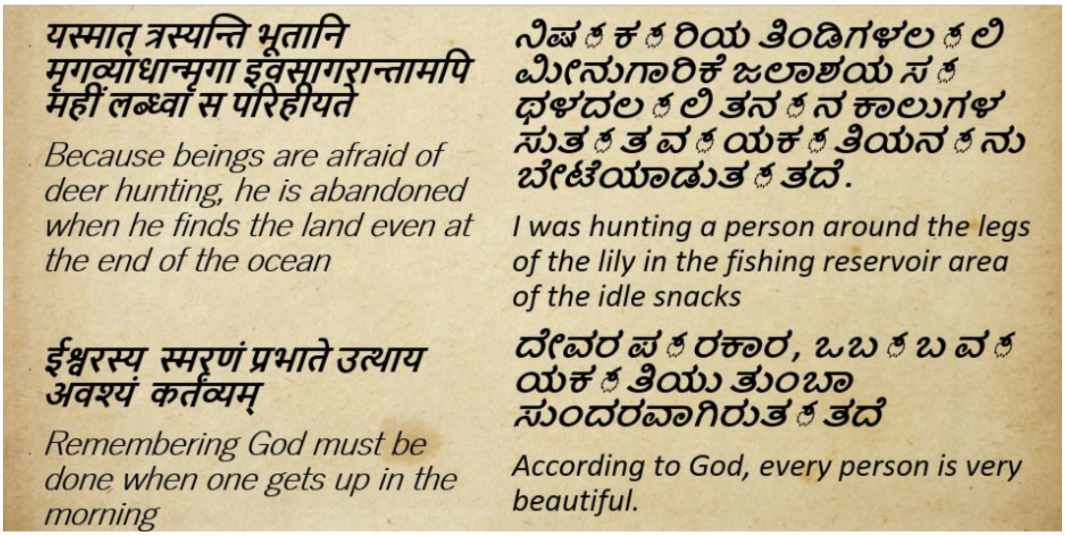
Low resource languages like Oriya (or) demonstrate notably lower performance across all categories, which may be attributed to various factors such as data quality or language complexity. We believe that multilingual training could be an effective strategy to mitigate these issues.

During our hyperparameter search, we noted that while performance for Assamese improved, there was a significant drop in performance for other languages. This can be observed in Table 8. This could potentially be due to the model overfitting on Assamese, which has substantially less data compared to Oriya. Consequently, we focused our hyperparameter search on Oriya rather than Assamese, considering that Assamese has only 52 tokens in the validation and test set. This decision was made to avoid the risk of overfitting on the test dataset, which could skew our experimental results. This can be observed in Table 9.

5 Applications

5.1 Sanskrit Shlokas Translation

Many ancient Indian scriptures and the vast knowledge stored in it were scripted in Sanskrit as it was the scholarly language and is held as sacred. Due to limited parallel corpora, translating Sanskrit texts (Figure 2) to Dravidian languages is a difficult task which we aim to solve by training on the competition data but in a reversed language order (Sanskrit-X) using transformer models. We consider a single-language Sanskrit Dataset [2] of Vidur Niti Shlokas, Chanakya Shlokas and Sanskrit-slogans. We train the from-scratch transformer on the same dataset as the main results, but test on Sanskrit Shlokas dataset. As visible in Figure 7, the models obtaining approximately 0.70 BLEU Score are good at translating key phrases but not the actual meaning of ancient shlokas, which are more complex in nature. As there are no parallel corpora for evaluation, we evaluate the correspondences, translation accuracy and semantics using human evaluation by translating the ancient texts and translations to English. We observe that learning deeper philosophical meanings of these Shlokas may require more sophisticated techniques than pure parallel



यस्मात् त्रस्यन्ति भूतानि
मृगव्याधान्मृगा इवसागरान्तामपि
महीं लब्ध्वा स परिहीयते

Because beings are afraid of deer hunting, he is abandoned when he finds the land even at the end of the ocean

ईश्वरस्य स्मरणं प्रभाते उत्थाय
अवश्यं कर्तव्यम्

Remembering God must be done when one gets up in the morning

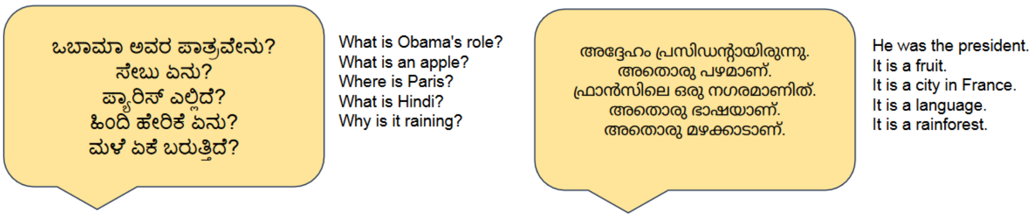
ನಿಷ್ ಕ್ ರಿಯ ತಿಂಡಿಗಳಲ್ ಲಿ
ಮೀನುಗಾರಿಕೆ ಜಲಾಶಯ ಸ್
ಥಳದಲ್ ಲಿ ತನ್ ನ ಕಾಲುಗಳ
ಸುತ್ ತ ವ್ ಯಕ್ ತಿಯನ್ ನು
ಬೇಟೆಯಾಡುತ್ ತದೆ.

I was hunting a person around the legs of the lily in the fishing reservoir area of the idle snacks

ದೇವರ ಪ್ ರಕಾರ, ಒಬ್ ಬ ವ್
ಯಕ್ ತಿಯು ತುಂಬಾ
ಸುಂದರವಾಗಿರುತ್ ತದೆ

According to God, every person is very beautiful.

Figure 7: Application 1: Translation of Sanskrit Shlokas to Dravidian Languages (English translations [Generated for understanding samples using Google translate] are given below for reference)



ಒಬಾಮಾ ಅವರ ಪಾತ್ರವೇನು?
ಸೇಬು ಏನು?
ಪ್ಯಾರಿಸ್ ಎಲ್ಲಿದೆ?
ಹಿಂದಿ ಹೇರಿಕೆ ಏನು?
ಮಳೆ ಏಕೆ ಬರುತ್ತಿದೆ?

What is Obama's role?
What is an apple?
Where is Paris?
What is Hindi?
Why is it raining?

അദ്ദേഹം പ്രസിഡന്റായിരുന്നു.
അതൊരു പഴമാണ്.
ഫ്രാൻസിലെ ഒരു നഗരമാണിത്.
അതൊരു ഭാഷയാണ്.
അതൊരു മഴക്കാടാണ്.

He was the president.
It is a fruit.
It is a city in France.
It is a language.
It is a rainforest.

Figure 8: Application 2: Chatbot extension for Dravidian Languages to promote accessibility of translation models.

translation which can be explored in future works.

5.2 Chatbot Extension

Although we have built several models in this project, they are not directly accessible to the speakers of Dravidian languages, which we try to solve in the form of conversational agents.. We try building this conversational chatbot by using the DialoGPT [3] and leveraging post-hoc Dravidian language translation using the en-Indic and Indic-en Indic-Trans [53] models. As visible in Figure 8, we see that this model works very well for general knowledge questions and can be used for improving accessibility of stronger translation models to the general public.

6 Conclusion

Across the many tasks, techniques and results presented in this work, the goal of this project was to explore low-resource tasks for Dravidian languages. Critically, we present state-of-the-art translation performance on Kannada-Sanskrit Translation on the presented benchmark and state-of-the- NER performance on the Oriya language benchmark.

Our findings across these tasks suggest that back-translation and cross-lingual training promises performance improvement as models can learn representations across larger corpora and transfer this knowledge to do well on smaller tests. We find that certain languages (Sanskrit) that have similar scripts to Hindi (script conversion was a step in data processing) show better results which may be due to underlying script similarity and thus corresponding absence of lossy processing. We find that translation models are good at picking up parallel word translations but may lose the semantic meaning in complicated sentences which can be explored in future work.

We present an application of a multi-lingual chatbot which can be incorporated with domain-specific information and fine-tuning across medical, agriculture and legal corpora for social good. *We hope to benefit the speakers of Dravidian languages through this work, and seek to improve accessibility to the previously under-represented languages in computing for future works.*

References

[1] facebook/nllb-200-distilled-600M · Hugging Face — huggingface.co. https://huggingface.co/facebook/nllb-200-distilled-600M. [Accessed 12-12-2023].

[2] iNLTK Sanskrit Shlokas Dataset — kaggle.com. https://www.kaggle.com/datasets/disisbig/sanskrit-shlokas-dataset. [Accessed 12-12-2023].

[3] microsoft/DialoGPT-large · Hugging Face — huggingface.co. https://huggingface.co/microsoft/DialoGPT-large. [Accessed 12-12-2023].

[4] AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv: 2305.16307*, 2023.

[5] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

[6] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

[7] Vinayak Athavale, Shreenivas Bharadwaj, Monik Pamecha, Ameya Prabhu, and Manish Shrivastava. Towards deep learning in hindi ner: An approach to tackle the labelled data scarcity. *arXiv preprint arXiv:1610.09756*, 2016.

[8] Christos Baziotis, Barry Haddow, and Alexandra Birch. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in NLP*, pages 7622–7634, 2020.

[9] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics, 2016.

[10] Asoka Chakravarthi and Bharathi Raja. *Leveraging orthographic information to improve machine translation of under-resourced languages*. PhD thesis, NUI Galway, 2020.

[11] Bharathi Raja Chakravarthi. 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, 2020.

[12] Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 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, 2021.

[13] Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. A survey of orthographic information in machine translation. *SN computer science*, 2(4):330, 2021.

[14] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023.

[15] Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*, 2018.

[16] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[17] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[18] Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, 2017.

[19] Nikita P Desai and Vipul K Dabhi. Taxonomic survey of hindi language nlp systems. *arXiv preprint arXiv:2102.00214*, 2021.

[20] Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, 2020.

[21] Asif Ekbal and Sivaji Bandyopadhyay. Named entity recognition in bengali and hindi using support vector machine. *Lingvisticæ Investigationes*, 34(1):35–67, 2011.

[22] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020.

[23] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020.

[24] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

[25] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

[26] Piyushi Goyal, Musica Supriya, Dinesh U, and Ashalatha Nayak. Translation techies @DravidianLangTech-ACL2022-machine translation in Dravidian languages. In Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Parameswari Krishnamurthy, Elizabeth Sherly, and Sinnathamby Mahesan, editors, *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 120–124, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[27] https://academic-accelerator.com/encyclopedia/dravidian languages. Dravidian languages, Accessed 2023.

[28] W John Hutchins. The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer, 2004.

[29] Leena Jain and Prateek Agrawal. Text independent root word identification in hindi language using natural language processing. *International Journal of Advanced Intelligence Paradigms*, 7(3-4):240–249, 2015.

[30] Michael I Jordan. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier, 1997.

[31] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pretrained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*, 2020.

[32] Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *arXiv preprint arXiv:2310.12321*, 2023.

[33] Arnav Kapoor, Mudit Dhawan, Anmol Goel, TH Arjun, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. Hldc: Hindi legal documents corpus. *arXiv preprint arXiv:2204.00806*, 2022.

[34] Soma Khan, Tulika Basu, Joyanta Basu, Madhab Pal, Rajib Roy, and Milton S Bepari. Data collection and development of bengali asr and tts for conversational ai-based automated advisories in the agriculture domain. In *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*, pages 1–6. IEEE, 2022.

[35] Vishnupriya Kolipakam, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science*, 5(3):171504, 2018.

[36] Atharva Kulkarni, Amey Hengle, and Rutuja Udyawar. An attention ensemble approach for efficient text classification of indian languages. *arXiv preprint arXiv:2102.10275*, 2021.

[37] Anand Kumar M, Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadarshini, Shashirekha Hosahalli Lakshmaiah, and John Philip McCrae. 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, April 2022.

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

[39] Surafel M Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. Adapting multilingual neural machine translation to unseen languages. *arXiv preprint arXiv:1910.13998*, 2019.

[40] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[41] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, 2019.

[42] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[43] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[44] Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy V, and Anoop Kunchukuttan. Naamapadam: A large-scale named entity annotated data for indic languages. *arXiv preprint arXiv:2212.10168*, 2022.

[45] Sreyashi Nag, Mihir Kale, Varun Lakshminarasimhan, and Swapnil Singhavi. Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation, 2020.

[46] Poojitha Nandigam, Abhinav Appidi, and Manish Shrivastava. Named entity recognition for code-mixed kannada-english social media data. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 43–49, 2022.

[47] Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 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, 2018.

[48] Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926, 2019.

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[50] Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, 2018.

[51] Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 529–535, 2018.

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[53] 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. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages, 2021.

[54] 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. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 02 2022.

[55] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

[56] Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. Explicit cross-lingual pre-training for unsupervised machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 770–779, 2019.

[57] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280, December 2020.

[58] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

[59] Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 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. IEEE, 2021.

[60] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, 2016.

[61] Richa Sharma, Sudha Morwal, Basant Agarwal, Ramesh Chandra, and Mohammad S Khan. A deep neural network-based model for named entity recognition for hindi language. *Neural Computing and Applications*, 32:16191–16203, 2020.

[62] Hewan Shrestha, Chandramohan Dhasarathan, Shanmugam Munisamy, and Amudhavel Jayavel. Natural language processing based sentimental analysis of hindi (sah) script an optimization approach. *International Journal of Speech Technology*, 23:757–766, 2020.

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

[64] CN Subalalitha. Information extraction framework for kurunthogai. *Sādhanā*, 44(7):156, 2019.

[65] CN Subalalitha and E Poovammal. Automatic bilingual dictionary construction for tirukural. *Applied Artificial Intelligence*, 32(6):558–567, 2018.

[66] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13, 2020.

[67] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[69] Aditya Vyawahare, Rahul Tangsali, Aditya Mandke, Onkar Litake, and Dipali Kadam. PICT@DravidianLangTech-ACL2022: Neural machine translation on Dravidian languages. In Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Parameswari Krishnamurthy, Elizabeth Sherly, and Sinnathamby Mahesan, editors, *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 177–183, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[70] Wikipedia contributors. Dravidian languages — Wikipedia, the free encyclopedia, 2023. [Online; accessed 12-December-2023].

[71] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*, 2019.

[72] Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, 2018.