Machine Translation with Large Language Models: Decoder Only vs. Encoder-Decoder

Team: Machine Translators
Mentor: Yash Bhaskar

Abhinav P M, University of Calicut
Sujay Kumar Reddy M, Vellore Institute of Technology, Vellore
C. Oswald, Assistant Professor, National Institute of Technology, Tiruchirappalli.

International Summer School on Natural Language Processing
Project Presentation
IIIT Hyderabad
05 July 2024

Problem Statement

To perform Machine Translation Task with respect to:

Types of LLM: 1. Decoder Only (1-1 and 1-Many) Architecture

2. Encoder-Decoder (1-1, Many-1, 1-Many and Many-Many) Architecture

Objectives:

To investigate on the following:

1. Behaviour of the model with respect to bi-lingual (1-1) and multilingual (Many-1, 1-Many and Many-Many) language translations

2. Performance of Encoder-Decoder based Transformer models for Neural Machine Translation (NMT) compared with smaller Decoder-only models, such as LLMs, when trained using the same data and similar parameters.

3. To quantify the role of context (no. of tokens) in translation with these two architectures.

Literature Review

Multilingual Machine Translation

- Multilingual machine translation with large language models: Empirical results and analysis - Zhu Wenhao et al. (arXiv, 2023) [1]
- Massively multilingual neural machine translation - Aharoni et al. (arXiv, 2019) [2]
- Massively multilingual neural machine translation in the wild: Findings and challenges - Naveen Arivazhagan et al. (arXiv, 2019) [3]

Learning Resources

- BERT - J. Devlin et al. (arXiv, 2018) [4]
- Attention - A Vaswani et al. (NeurIPS 2017) [5]

Our Proposed Approaches for Multilingual NMT

1. In-Context Learning (ICL) using Few-Shot Learning
2. Fine-Tuning of LLMs
3. Baseline Model Development from Scratch

1. In-Context Learning (ICL) using Few-Shot Learning

In-Context Learning – How it works?

- A way to use language models to learn tasks given only a few examples. [6]
- Prompt Engineering Tasks for Few-Shot Learning.
- Examples of MT pairs (<X>=<Y>) with a template T, X - Source Sentence, Y - Target Sentence. [1]
- **In-Context Exemplars**:
  <X>=<Y> - Strong recipe for best outputs from the model - Wu et al. (2023) [7]
- Prompt P = $T(X_1,Y_1) \oplus T(X_2,Y_2) \oplus \ldots \oplus T(X_n,Y_n)$ where $\oplus$ - concatenation,

  n - number of samples [1]

Wu, Zhenyu, et al. "Openicl: An open-source framework for in-context learning." *arXiv preprint arXiv:2303.02913* (2023).

Our Approach - ICL

How a prompt to the model is structured?

{
"1": "After submitting your application, you should receive a registration certificate within several business days. आवेदन जमा करबाक बाद, अहाँकेँ किछु व्यावसायिक दिनक भीतर एक टा पंजीकरण प्रमाण-पत्र भेटि सकैत अछि।"
“2": "Early pregnancy bleeding is usually from a maternal source, rather than a fetal one, प्रारंभिक गर्भावस्था में रक्तस्राव आमतौर पर भ्रूण के बजाय मातृ स्रोत से होता है।",
“3": "The bride's father symbolically offers to the bridegroom a cow as a present. एहन रोगीक लेल, जनिक संक्रमण आ कैंसरक नैदानिक संदेह न्यून रहैत अछि, एक्स-रे एकटा कम महग प्रारम्भिक विकल्पक अछि आ प्रयोगशालाक अध्ययनक सङ्ग एकर व्याख्या होइत अछि।",
}

- Evaluation:

Low-Resource Languages: Hindi, Telugu, Malayalam, Marathi and Tamil

LLM’s: On 3 sentence pairs using XGLM and MT5.

Metric: BLEU.

Datasets

BPCC Wiki MT Dataset [9]:

- 16k – 50k Translation Samples.
- English to 22 Indian Language Pairs.
- Context Length of each sentence pair is 40-200 characters long.

In-Context Learning – Experimentation Results (1/2)

Models Chosen:

1. Decoder Only - XGLM

Why? Generates moderate translation with 500 million parameters,
Builds bi-lingual mapping between non-English to English. [1]

2. Encoder-Decoder - mT5

Why? Capability for multilingual translation (mT5-base),
Contains 300 million parameters.

Reference	**Predicted (mT5)**	**Predicted (XGLM)**
▁After ▁sub mitting ▁your ▁application , ▁you ▁should ▁receive ▁a ▁registration ▁certificate ▁within ▁several ▁business ▁days .	▁अपना ▁आवेदन ▁जमा ▁करने ▁के ▁बाद, ▁आपको ▁कई ▁व्यावसायिक▁ दिनों ▁के ▁भीतर ▁पंजीकरण ▁प्रमाणपत्र ▁प्राप्त होना ▁चाहिए। ▁हिंदी में	▁आवेदन ▁जमा ▁कर बा क ▁बाद , ▁अ हाँ के ँ ▁कि छु ▁व्यावसायिक ▁दिन क ▁भीतर ▁एक ▁टा ▁पं जी करण ▁प्रमाण - पत्र ▁भेट ि ▁सक ै त ▁अ छि ।

In-Context Learning – Experimental Results (2/2)

Language-Pair	BLEU - Decoder Only	BLEU - Encoder-Decoder
eng-hin	**1.096078**	**1.04442**
hin-eng	0.443236	0.384031
eng-tel	0.307206	0.292951
tel-eng	0.379685	0.378953
eng-tam	0.307206	0.292951
tam-eng	0.349053	0.364626
eng-mal	0.307206	0.292951
mal-eng	0.330114	0.364626

Implementation Courtesy: https://github.com/NJUNLP/MMT-LLM

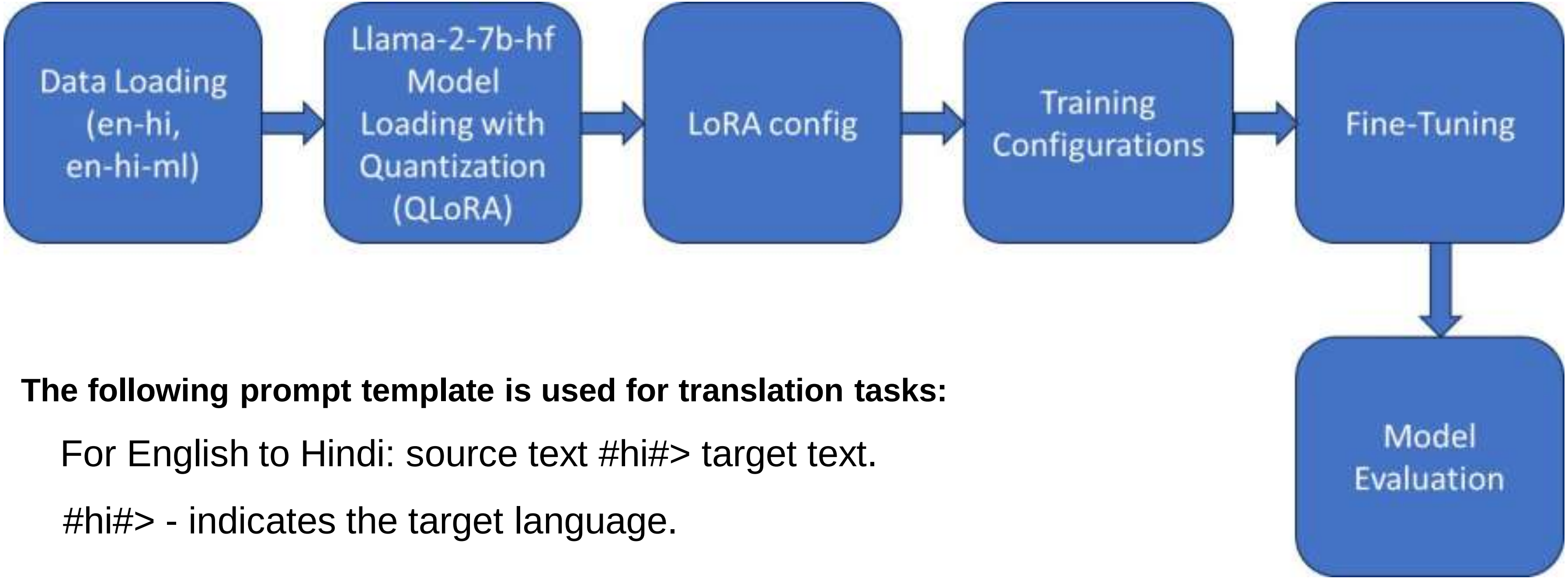
2. Fine-Tuning of LLMs

Work Flow of Encoder-Decoder Model: mT5 Fine-Tuning

1-1 Task : ‘en-hi’ from BPCC-Wiki dataset [9]
1-Many Task: ‘en-hi-bg’ from ALT dataset [8]

Figure 1. Work Flow of mT5 Fine-Tuning

Work Flow of Decoder-Only Model: Llama2 Fine-Tuning



The following prompt template is used for translation tasks:

For English to Hindi: source text #hi#> target text.

#hi#> - indicates the target language.

Figure 2. Work Flow of LlaMA2 Fine-Tuning

Experimental Results: mT5 Fine-Tuning

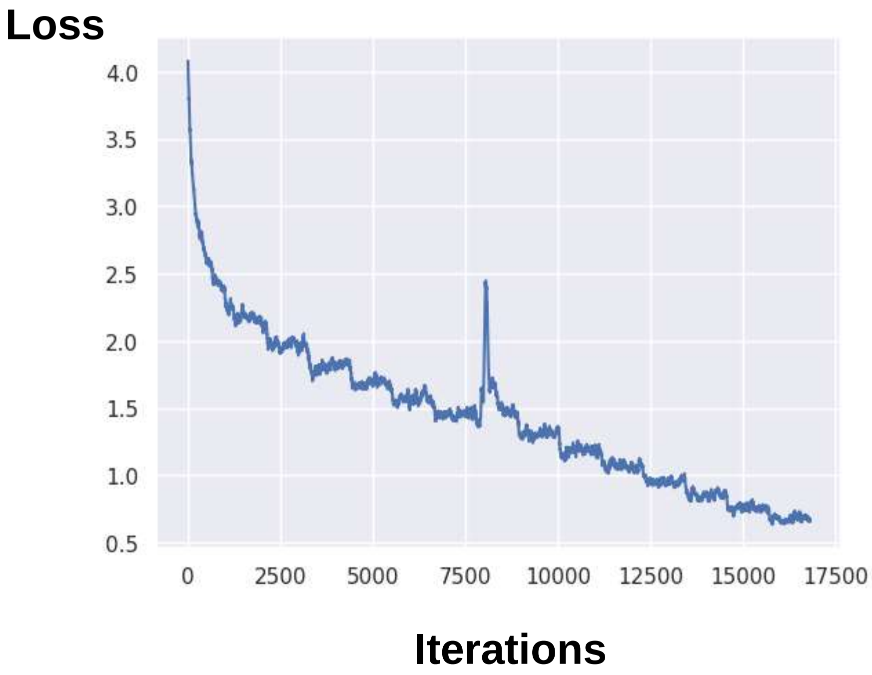


Figure 3. mT5 Multilingual (en-hi-bg)

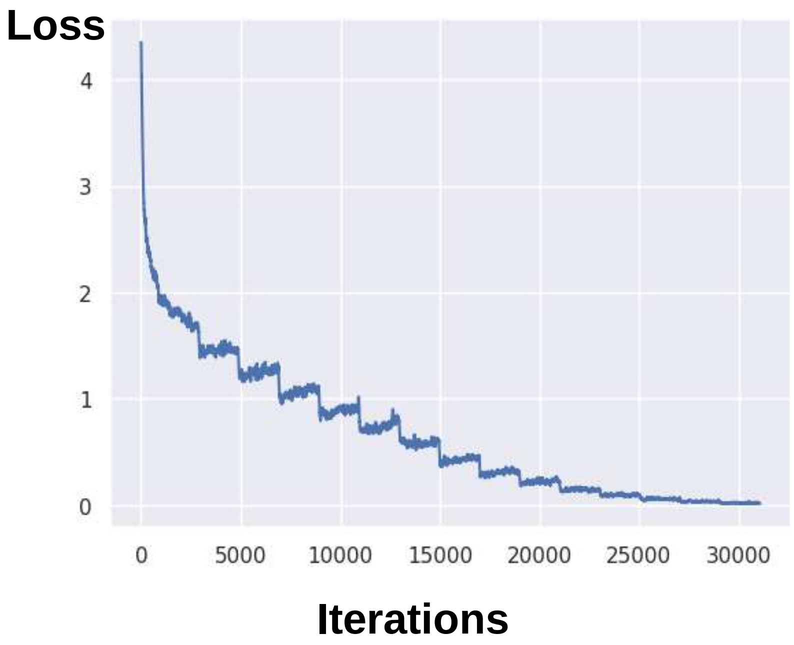


Figure 4. mT5 Bi-lingual (en-hi)

Experimental Results: Llama2 and mT5 Fine-Tuning

Model	BLEU	chrF	TER
Llama2-finetuned-one-many(en-hi)	0.0265	7.1217	94.0950
Llama2-finetuned-one-many(en-ml)	0.0409	6.8530	96.4312
Llama2-finetuned-one-one(En-Hi)	0.0955	9.2282	90.4864
mT5-bi-lingual(en-hi)	11.7107	31.0639	74.1626
mT5-bi-lingual(hi-en)	**14.1444**	**33.8278**	74.7157
mT5 many-many(en-hi)	3.4802	19.6184	84.7821
mT5 many-many(en-bg)	1.0885	16.2382	91.9398
mT5 many-many(hi-bg)	0.7545	15.6990	92.9326
mT5 many-many(hi-en)	5.2237	23.2258	84.6685
mT5 many-many(bg-en)	3.9469	21.5855	86.7228
mT5 many-many(bg-hi)	2.1458	16.9235	88.2083

Sample Results: mT5 Fine-Tuning (1/2)

Model	Type	Source Text	Translated Text	Target Text
mT5	**Bi-Lingual**	यह किताब मुझे बहुत पसंद है।	This book I really like.	I like this book very much.
		He loves his parents.	वह अपने माता-पिता को पसंद करती है।	वह अपने माता-पिता से प्यार करता है.
mT5	**Many-Many**	He loves his parents.	उसे अपने माता-पिता की प्यास है।	वह अपने माता-पिता से प्यार करता है.
		He has bought a new car.	তিনি একটি নতুন গাড়ি কিনে কিনেছেন।	তিনি একটি নতুন গাড়ি কিনেছেন।
		यह किताब मुझे बहुत पसंद है।	This book is very suitable for me.	I like this book very much.
		वह अपने माता-पिता से प्यार करता है।	তিনি তার বাবার কাছে খুবই পছন্দ করেন।	সে তার বাবা-মাকে ভালোবাসে।
		তুমি কেমন আছো?	तुम कौन हो?	आप कैसे हैं
		আমি আজকে বাসায় থাকব।	I will be at home today.	I will stay at home today.

Sample Results: mT5 Fine-Tuning (2/2)

Model	Type	Source Text	Translated Text	Target Text
Llama2	**1-Many**	Kerala, a state on India's tropical Malabar Coast	ഇന്ത്യയിലെ മലബാർ കോ	ഇന്ത്യയുടെ ഉഷ്ണമേഖലാ മലബാർ തീരത്തുള്ള ഒരു സംസ്ഥാനമാണ് കേരളം
			'भारत की तटीय मलबार कोष्ठ'	केरल, भारत के उष्णकटिबंधीय मालाबार तट पर स्थित एक राज्य
Llama2	**1-1**	Hello, what is your name?	नमस्ते, आपका नाम क्या है?	नमस्कार, आपका नाम क्या है?
		Today is my birthday	आज मेरी जन्मदिन है	आज मेरा जन्मदिन हे
		Click the "Search" button to begin the search	खोज करने के लिए "खोज" बटन क	खोज शुरू करने के लिए "खोजें" बटन पर क्लिक करें

Analysis: Llama2 and mT5 Fine-Tuning

- mT5 is trained with 32,216 data points for the English to Hindi (en-hi) pair.
  Reason: Increased data.

- For the decoder-only model (Llama2 finetuned), it performs poorly on 1-Many tasks.
- For 1-1 tasks, the Llama2 finetuned model performs better compared to the Llama2 1-Many model.
  Reason: Llama2 needs to be trained with more high-quality data.

3. Baseline Model Development from Scratch

Challenges so, far …

- Pre-trained models are trained on large data. For example: mT5

Now,

- To compare the Encoder-decoder and decoder-only models with similar training setting to evaluate the model's performance in the multi-task learning paradigms.
- To compare the context length of both the models by some quantitative metrics which provides some Interpretation of the models.

Proposed Methodology of our Baseline Model Development

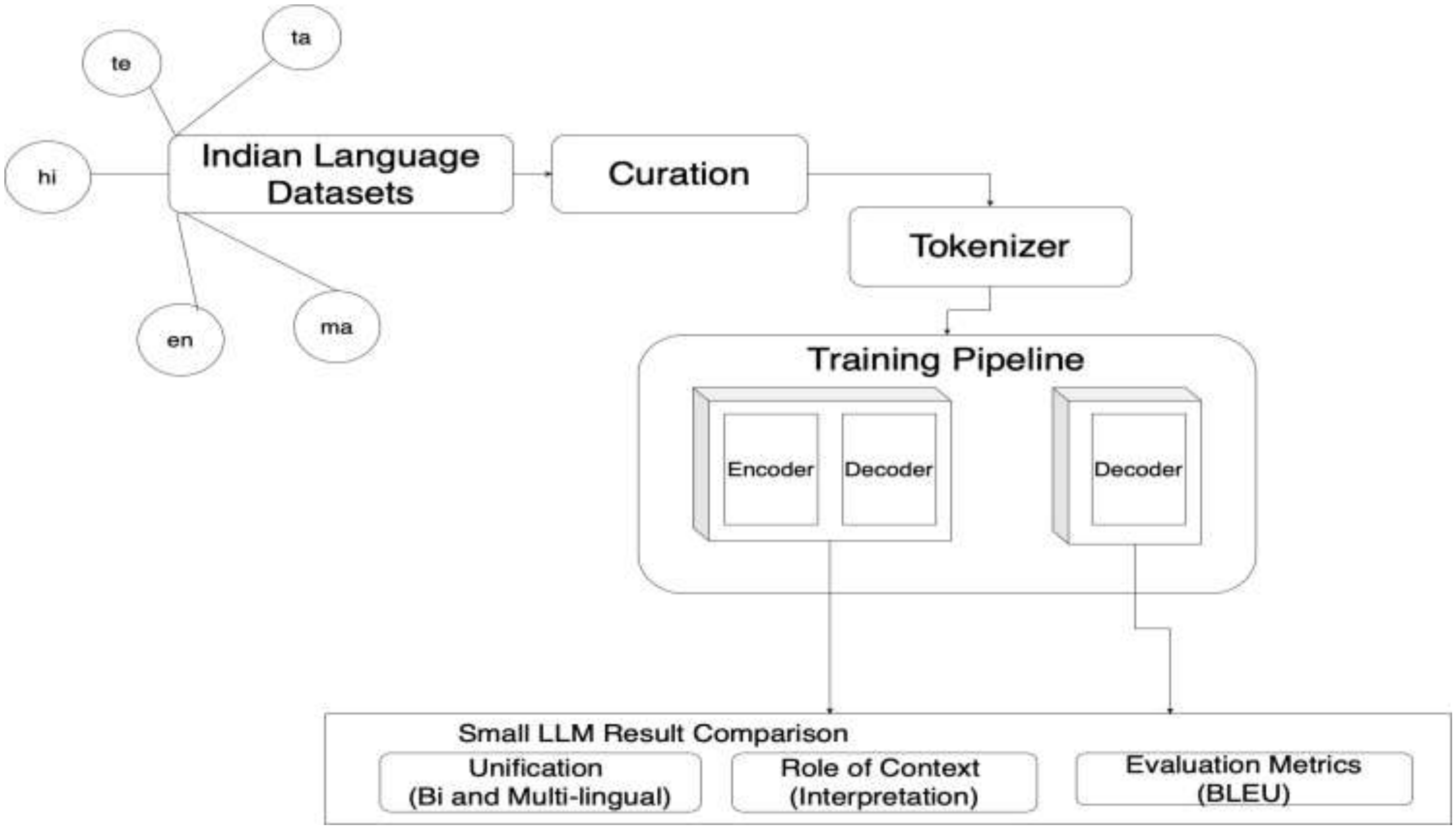


Figure 2. Proposed Methodology of our Baseline Model Development

Baseline Model Development

- To train a model from scratch - Pretrained model is more black boxed and less interpretable.
- Took stable baseline models and equated the parameters.
- Decoder-Only Model - XLNet as a base model (Wu et al., 2021) [10]
- Encoder-Decoder model - IndicBART as a base model (Dabre et al., 2021) [11]
- Tokenizer is shared across both the architectures.

Model Name	**Trainable Parameters**
XLNet Baseline	147,490,318
Indic-BART Baseline	145,339,392

Gitub Baseline Implementation : https://github.com/sujaykumarmag/iasnlp

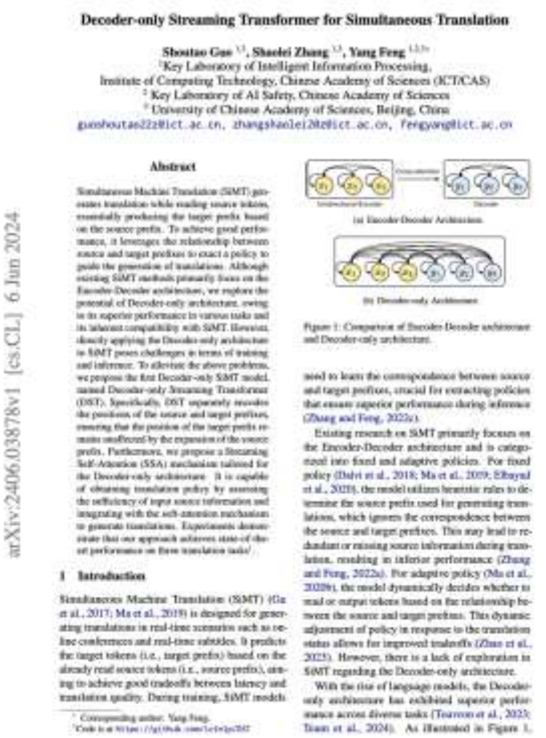
Take Aways and Future Prospects

- Encoder-Decoder model provides trustable results, while the Decoder-only models are trained differently as next word/char.

- The learning paradigms for both the Architectures are different:
  How do we converge for Multilingual Machine Translation?

- The Decoder-only model treats the starting tokens of the source text and the translated text separately.

- A recent new method - Streaming Self-Attention (SSA) helps the model decide when it has enough of the original text to start translating accurately.


arXiv:2406.03878v1 [cs.CL] 6 Jun 2024

Decoder-only Streaming Transformer for Simultaneous Translation

Shoutao Guo [1,3], Shaolei Zhang [1,3], Yang Feng [1,2,3*]
[1]Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] Key Laboratory of AI Safety, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
guoshoutao22z@ict.ac.cn, zhangshaolei20z@ict.ac.cn, fengyang@ict.ac.cn

Abstract

1 Introduction

Guo, Shoutao, Shaolei Zhang, and Yang Feng. "Decoder-only Streaming Transformer for Simultaneous Translation." *arXiv preprint arXiv:2406.03878* (2024).

References

1. Zhu, Wenhao, et al. ”Multilingual machine translation with large language models: Empirical results and analysis.” *arXiv preprint arXiv:2304.04675* (2023).
2. Aharoni, Roee, Melvin Johnson, and Orhan Firat. "Massively multilingual neural machine translation." *arXiv preprint arXiv:1903.00089* (2019).
3. Arivazhagan, Naveen, et al. "Massively multilingual neural machine translation in the wild: Findings and challenges." *arXiv preprint arXiv:1907.05019* (2019).
4. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
5. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
6. https://ai.stanford.edu/blog/understanding-incontext/
7. Wu, Zhenyu, et al. "Openicl: An open-source framework for in-context learning." arXiv preprint arXiv:2303.02913 (2023).
8. https://huggingface.co/datasets/mutiyama/alt
9. https://ai4bharat.iitm.ac.in/bpcc/
10. Wu, Nier, et al. "Low-Resource Neural Machine Translation Using XLNet Pre-training Model." Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30. Springer International Publishing, 2021.
11. Dabre, Raj, et al. "IndicBART: A pre-trained model for indic natural language generation." arXiv preprint arXiv:2109.02903 (2021).

Acknowledgements

- IIITH Administration for the excellent resource and amenities
- IASNLP Coordinators – Dr. Parameswari and Dr. Rahul Mishra
- Our Mentor: Mr. Yash Bhaskar
- Pre-school Speakers: Prashanth Kodali, Aparajitha, Priyanka Dasari, Aadya Ranjan, Sankalp Bahad

Thank you for your attention