Google - Unlock Global Communication with Gemma

DATA-255 Sec 22 - Deep Learning Technologies

Team 2
Ananya Varma Mudunuri
Sai Sahithi Bethapudi
Sandeep Reddy Potula
Shreya Chilumukuru
Sowmya Neela
Vinay Bhati

Introduction

- The evolution of large language models (LLMs) has revolutionized natural language processing, particularly in machine translation.
- High-resource languages like English and Spanish benefit significantly, but low-resource languages like Hindi remain underrepresented.

Challenges in Hindi Translation: Hindi, a low-resource language, poses translation challenges due to its unique structure (SOV), idiomatic expressions, and complex word forms.

This project aims to bridge the gap in Hindi-English translations by fine-tuning advanced LLMs for better fluency, accuracy, and cultural authenticity.

Related work

- **Gemma 2 Models**: Scalable architecture (2B–27B parameters) designed for multilingual tasks and low-resource languages.
- **Fine-Tuning Techniques**: RLHF, LoRA applied to improve translation accuracy and fluency.
- **Multilingual Adaptations**: Pivot-based translation methods enhance performance for low-resource languages like Hindi.
- **Inclusion Focus**: Studies emphasize handling syntactic differences, idiomatic expressions, and cultural nuances in translation.
- **Notable Achievements**: Improved results in benchmarks like MMLU and GSM8K, highlighting adaptability and inclusivity.

Proposed solution

Models

- **Gemma 2-2B**: A lightweight model (~2 billion parameters) optimized for efficient and straightforward translations with advanced mechanisms like **RoPE** and **GQA**.
- **Gemma 2-9B**: A robust model (~9 billion parameters) designed for handling complex and culturally nuanced translations with deeper contextual understanding.

Fine-Tuning Techniques

- **LoRA**: Enables parameter-efficient updates (<1%) and dynamic task switching while optimizing resource usage for large models.
- **Supervised Fine-Tuning (SFT)**: Aligns the model with task-specific objectives using labeled datasets, ensuring fluent and contextually accurate translations.

This combination ensures efficient and high-quality translations for Hindi-English and English-Hindi language pairs.

Data sources

IIT Bombay Dataset:

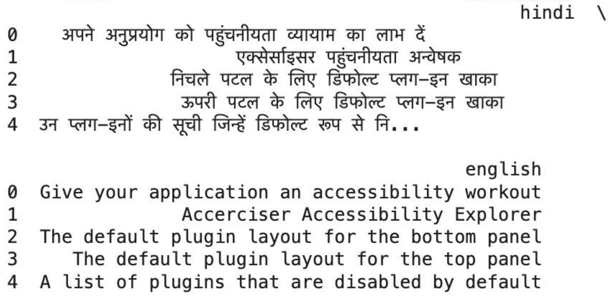
- Comprehensive parallel corpus with diverse Hindi-English sentence pairs.
- Covers technical, formal, and colloquial contexts for robust model training.

OPUS-100 Dataset:

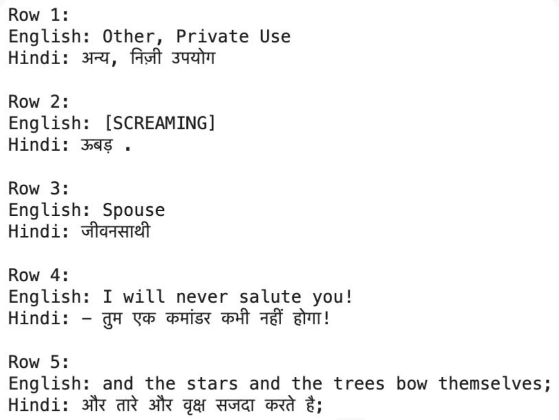
- Multilingual corpus with varied domains like literature, media, and technical content.
- Ensures diversity in sentence structures and vocabulary.

Integrated Dataset:

- Combined IIT Bombay and OPUS-100 datasets for a larger, more diverse corpus.
- Preprocessed for quality and consistency, ideal for machine translation tasks.



```
                                                              hindi  \
0    अपने अनुप्रयोग को पहुंचनीयता व्यायाम का लाभ दें
1                       एक्सेसर्साइसर पहुंचनीयता अन्वेषक
2                  निचले पटल के लिए डिफोल्ट प्लग-इन खाका
3                   ऊपरी पटल के लिए डिफोल्ट प्लग-इन खाका
4  उन प्लग-इनों की सूची जिन्हें डिफोल्ट रूप से नि...

                                        english
0  Give your application an accessibility workout
1                Accerciser Accessibility Explorer
2  The default plugin layout for the bottom panel
3     The default plugin layout for the top panel
4  A list of plugins that are disabled by default
```



```
Row 1:
English: Other, Private Use
Hindi: अन्य, निज़ी उपयोग

Row 2:
English: [SCREAMING]
Hindi: ऊबड़ .

Row 3:
English: Spouse
Hindi: जीवनसाथी

Row 4:
English: I will never salute you!
Hindi: - तुम एक कमांडर कभी नहीं होगा!

Row 5:
English: and the stars and the trees bow themselves;
Hindi: और तारे और वृक्ष सजदा करते है;
```

Cleansing

- Removed duplicates.
- Fixed mixed content (e.g., combined Hindi-English rows).
- Addressed missing values by removing.
- Normalized text by removing unwanted characters and ensuring consistent formatting.

Transformation

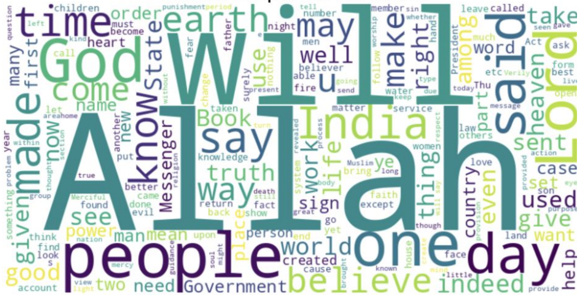
- Combined datasets to create a larger, diverse corpus.
- Standardized formats for compatibility with machine translation models.

Feature engineering

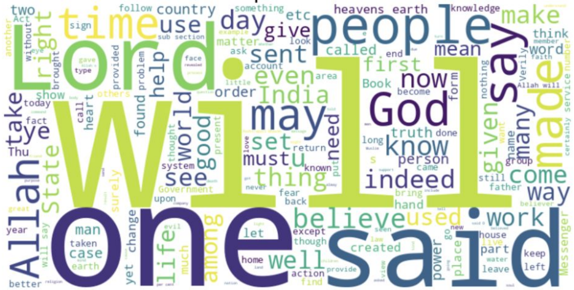
- **Sentence Length Analysis**:
  - Calculated word counts for Hindi and English sentences.
  - Filtered extreme values (95th percentile) to ensure balanced training.
- **Word Frequency Analysis**:
  - Removed stop words to focus on meaningful vocabulary.
  - Identified key terms using word clouds and frequency bar plots.
- **Alignment Validation**:
  - Correlated Hindi and English sentence lengths via scatter plots and heatmaps.
  - Ensured structural alignment for effective model learning.
- **Data Standardization**:
  - Normalized text to remove inconsistencies and improve dataset uniformity.

Visualization

Before Stopword Removal



After Stopword Removal

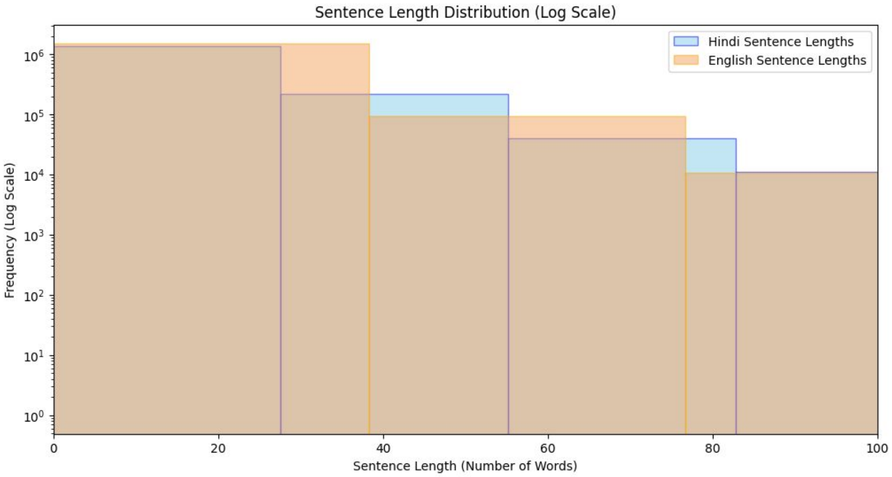


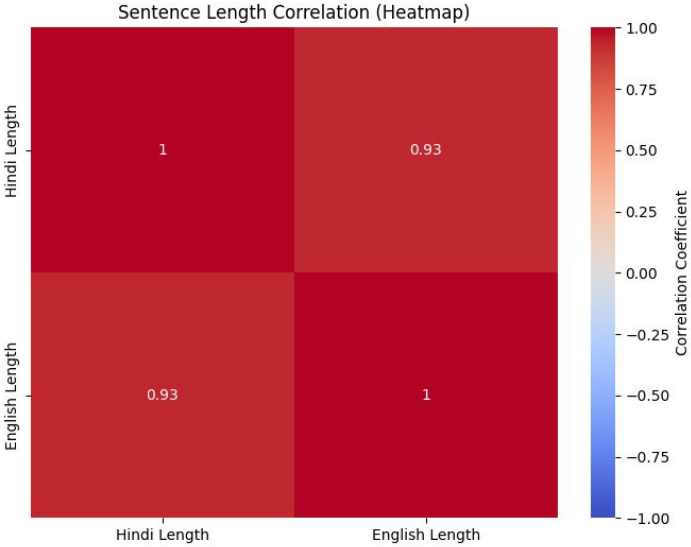
Before Stopword Removal (Hindi)

क
र
त
ह
स
य
म
ल
ज
न
द
और
अपन
करन
यह
पर
इस
उस

After Stopword Removal (Hindi)

क
र
ह
न
स
त
य
म
व
ज
करन
अपन
वह
उस
करत


Sentence Length Distribution (Log Scale)
Hindi Sentence Lengths
English Sentence Lengths
Frequency (Log Scale)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
20
40
60
80
100
Sentence Length (Number of Words)


Sentence Length Correlation (Heatmap)
Hindi Length
English Length
1
0.93
0.93
1
Hindi Length
English Length
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Correlation Coefficient

Model Fine Tuning

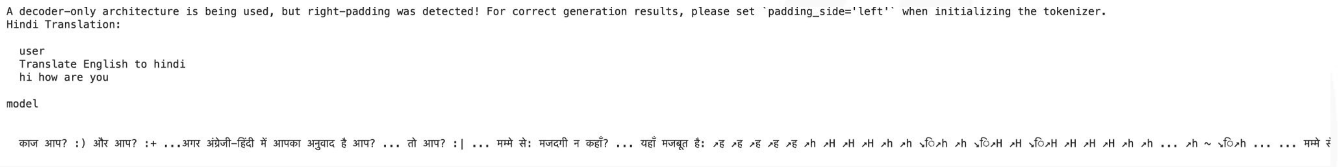
Gemma 2 Model: A multilingual transformer with advanced features like Grouped Query Attention (GQA) and Rotary Position Embeddings (RoPE), optimized for tasks from standard NLP to complex multilingual translations.

Low-Rank Adaptation (LoRA):

- Updates only a small subset of parameters (low-rank matrices) while keeping the base model frozen.
- Efficient for large models; reduces memory usage and computational requirements.
- Ideal for fine-tuning models like Gemma 2-2B and 9B on resource-constrained hardware.

Supervised Fine-Tuning (SFT):

- Aligns the model’s output with the specific task using labeled datasets.
- Ensures the model learns task-specific objectives, such as Hindi-English translations.
- Enhances contextual understanding and semantic accuracy.

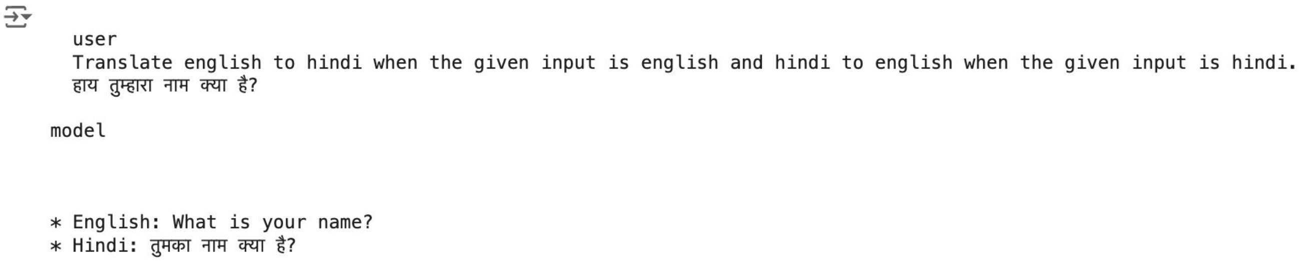


```
A decoder-only architecture is being used, but right-padding was detected! For correct generation results, please set `padding_side='left'` when initializing the tokenizer.
Hindi Translation:

  user
  Translate English to hindi
  hi how are you

model

  काज आप? :) और आप? :+ ...अगर अंग्रेजी–हिंदी में आपका अनुवाद है आप? ... तो आप? :| ... मम्मे से: मजदगी न कहाँ? ... यहाँ मजबूत है: ↗ह ↗ह ↗ह ↗ह ↗ह ↗h ↗H ↗H ↗H ↗h ↗h ↘ि↗h ↗h ↘ि↗H ↗H ↘ि↗H ↗H ↗H ↗H ↗h ↗h ... ↗h ~ ↘ि↗h ... ... मम्मे सं
```

Justify Models

GEMMA 2 - 2B:

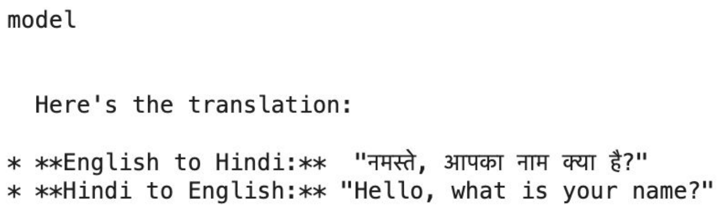
- Smartly captures subtle meanings using advanced deep learning techniques.
- Handles tricky words like "light," translating it as "रोशनी" (brightness) or "हल्का" (not heavy) based on context.
- Learns from bilingual dictionaries and parallel texts for accurate translations.



```
  user
  Translate english to hindi when the given input is english and hindi to english when the given input is hindi.
  हाय तुम्हारा नाम क्या है?

model



* English: What is your name?
* Hindi: तुमका नाम क्या है?
```

GEMMA 2 - 9B:

- Brings translations to life by using culturally appropriate idioms and phrases.
- Understands formal (आप) vs. informal (तुम) tone and adjusts translations accordingly.
- Tailors translations to regional Hindi preferences for a natural feel.



```
model


  Here's the translation:

* **English to Hindi:**  "नमस्ते, आपका नाम क्या है?"
* **Hindi to English:** "Hello, what is your name?"
```

GEMMA 2 - 27B:

- Breaks barriers by supporting less common regional Hindi dialects.
- Learns from diverse multilingual data to handle complex linguistic situations.
- Delivers fast, real-time translations, perfect for live captions or instant messaging.

Model evaluation

Metric: BLEU Score

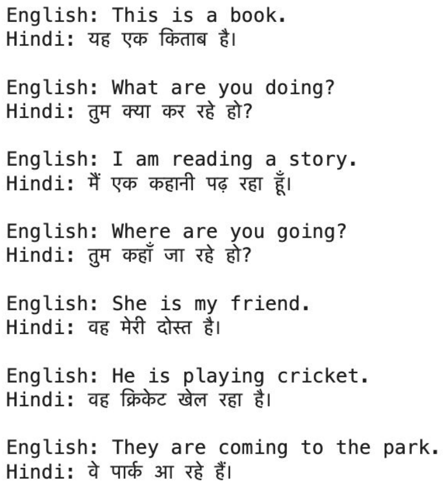
- Measures how similar the model's translations are to human-created ones.
- Ensures translations are complete and make sense in context with a brevity penalty.

Key Features:

- **n-gram Precision**: Checks how well word patterns match the reference translation.
- **Brevity Penalty**: Prevents translations from being too short.
- **Translation Length**: Looks at the total words in the model’s translation.
- **Reference Length**: Compares the model's output length to the human translation.

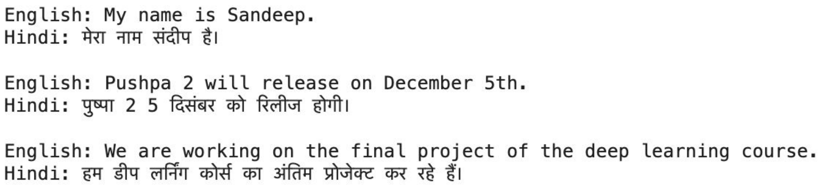
Result analysis and Visualization

Results with 2B , without formal usage

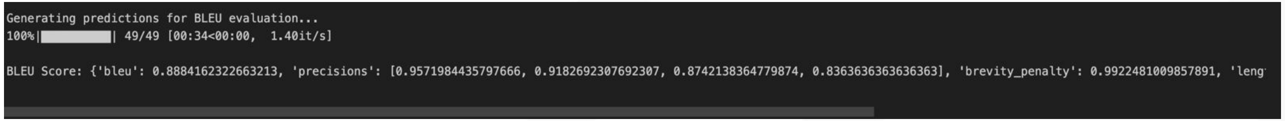


```
English: This is a book.
Hindi: यह एक किताब है।

English: What are you doing?
Hindi: तुम क्या कर रहे हो?

English: I am reading a story.
Hindi: मैं एक कहानी पढ़ रहा हूँ।

English: Where are you going?
Hindi: तुम कहाँ जा रहे हो?

English: She is my friend.
Hindi: वह मेरी दोस्त है।

English: He is playing cricket.
Hindi: वह क्रिकेट खेल रहा है।

English: They are coming to the park.
Hindi: वे पार्क आ रहे हैं।
```

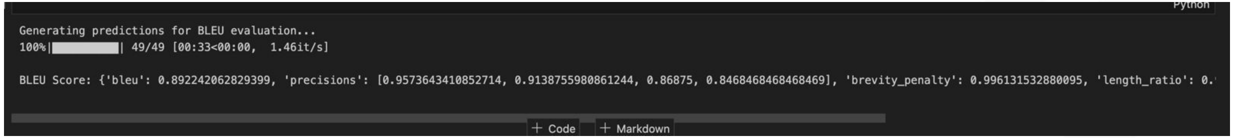
Results with 9B , with formal usage



```
English: My name is Sandeep.
Hindi: मेरा नाम संदीप है।

English: Pushpa 2 will release on December 5th.
Hindi: पुष्पा 2 5 दिसंबर को रिलीज होगी।

English: We are working on the final project of the deep learning course.
Hindi: हम डीप लर्निंग कोर्स का अंतिम प्रोजेक्ट कर रहे हैं।
```



```
Generating predictions for BLEU evaluation...
100%|██████████| 49/49 [00:34<00:00,  1.40it/s]

BLEU Score: {'bleu': 0.8884162322663213, 'precisions': [0.9571984435797666, 0.9182692307692307, 0.8742138364779874, 0.8363636363636363], 'brevity_penalty': 0.9922481009857891, 'leng
```



```
Generating predictions for BLEU evaluation...
100%|██████████| 49/49 [00:33<00:00,  1.46it/s]

BLEU Score: {'bleu': 0.892242062829399, 'precisions': [0.9573643410852714, 0.9138755980861244, 0.86875, 0.8468468468468469], 'brevity_penalty': 0.996131532880095, 'length_ratio': 0.
```

Innovation and Existing Market Usage

Innovation:

- Focuses on Hindi, a low-resource language.
- Combines IIT Bombay and OPUS-100 datasets for diversity.

Market Usage:

- Applications in education, media, business, and public services.
- Empowers Hindi-speaking communities with inclusive AI.

Conclusion

Efficient Fine-Tuning:

- LoRA made fine-tuning lightweight and efficient with minimal computational effort.
- SFT ensured translations stayed accurate, fluent, and meaningful.

Performance Highlights:

- **2B Model**: BLEU score of 0.89; strong performance but struggled with context in some cases.
- **9B Model**: BLEU score of 0.88; handled context better but required more epochs.

Future Potential:

- **27B Model**: Designed to excel with complex phrases and subtle translations.
- Expanding training epochs and refining techniques could further enhance performance.

LoRA and SFT together showcased how efficient and high-quality translations can be achieved, setting the stage for even more advanced solutions for low-resource languages.

Thank You