Arabic-to-English Machine Translation using Pretrained Models

NLP

Problem Statement & Objective

01 Arabic-to-English translation plays a key role in sharing Arabic content with the global community.

02 High-quality translation is essential for making Arabic news and media accessible worldwide.

03 Arabic (Fusha) presents unique linguistic challenges in machine translation.

04 **Project Goal**: Fine-tune a pretrained translation model using the Global Voices dataset to produce accurate Arabic-to-English translations.

OPUS Global Voices Dataset

Description

Dataset Size

63,071 Arabic-to-English sentence pairs

Content Type

News program-style sentences, structured with Arabic and English translations in parallel

Challanges

empty or irrelevant words

Some sentences contained no text or were corrupted

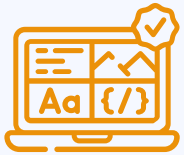
links or nonsensical values

Certain sentences included links or nonsensical values that needed to be cleaned

Preprocessing Steps

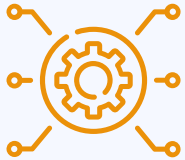
Cleaning & Filtering

- Removed empty or too-short sentence
- Filtered out noisy data: long numbers, unwanted symbols (e.g., @, {}, <, etc.)
- Final cleaned pairs: ~56078



Dataset Splitting

- Train: 80%
- Validation: 10%
- Test: 10%



Tokenization

- Applied pretrained tokenizer MarianMT
- Max token length 256
- Created PyTorch dataset class for efficient batching and training

Model Architecture

Model Used: Helsinki-NLP/opus-mt-ar-en

- A pretrained MarianMT model specialized for Arabic-to-English translation

Why this model?

- Specifically trained for many language pairs including ar-en
- Lightweight and fast, suitable for low-resource settings
- Open-source and easy to integrate via Hugging Face Transformers

Fine-Tuning Approach:

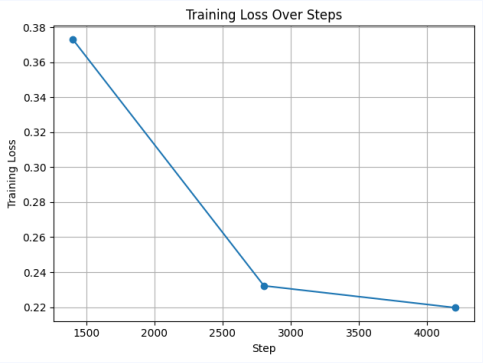
- Continued training on the cleaned Global Voices dataset
- Adjusted for domain-specific vocabulary and structure (news-style content)

Training Details

Seq2Seq Trainer

- **Google Colab with GPU**
  - **Training Configuration:**
    - ✓ Batch Size: 32
    - ✓ Learning Rate: 3e-5
    - ✓ Epochs: 3
    - ✓ Checkpoints saved per epoch
- **Time: ~1 hour**
  - **Observed Results:**
    - ✓ Training Loss: 0.275
    - ✓ Validation Loss: 0.233

Evaluation And Analysis



Training Evaluation

The training loss steadily decreased over time, indicating that the model was learning effectively during training but later he learn so slowly.

```
{'eval_loss': 0.2334008663892746,
 'eval_runtime': 39.855,
 'eval_samples_per_second': 140.71,
 'eval_steps_per_second': 4.416,
 'epoch': 3.0}
```

Vlidation evaluation

The model achieved a low eval loss of **0.233** in under **40 seconds**, showing good performance and fast inference.

```
Sentence 9:
🔸 Arabic:    صورة لعامل نيبالي مهاجر قتل في "حادث" في قطر.
🔹 Reference: A Nepali migrant worker killed in Qatar 'accident' being photographed by his brother.
✅ Predicted: A photograph of a Nepali migrant worker killed in an "incident" in Qatar.

🌎 BLEU score on first 10 sentences: 40.42408975026862

🌎 METEOR score on first 10 sentences: 0.7007658985445097
```

Test Evaluation

The model achieved a **BLEU score of 40.42** and a **METEOR score of 0.70** on the first 10 test sentences, indicating high-quality translations that closely match the references in both structure and meaning.

GUI and Example

Thanks !

Do you have any questions?

Team Members

- Bassam Emad Hamdy
- Buthaina Esam Mohamed
- Terevena Reda Amin
- Aya Mohamed Ali
- Eman Ehab Ebrahiem
- Gamil Mohamed Gamil
- Bassem yasser Ragab