TRANSLATION MODEL

by Group 2 (8:10)

INTRODUCTION

Model Details: Bilingual Translation System (Arabic-English / English-Arabic)

This project uses MarianMT, a state-of-the-art neural machine translation (NMT) framework based on the Transformer architecture. The models are pretrained on large multilingual datasets and fine-tuned on the Tatoeba parallel corpus for Arabic-English translation.

1. MODEL ARCHITECTURE

The MarianMT models follow a sequence-to-sequence (Seq2Seq) Transformer architecture, consisting of:

- Encoder: Processes the input sequence (source language).
- Decoder: Generates the translated sequence (target language).
- Attention Mechanism: Helps the model focus on relevant parts of the input when generating each word in the output.

Key Components

Encoder Layers: 6 layers of self-attention and feed-forward networks

Decoder Layers: 6 layers with self-attention, encoder-decoder attention, and feed-forward networks

Attention Heads: 8 parallel attention heads per layer

Hidden Dimension: 512 units

Feed-Forward Dimension: 2048 units (expands before projecting back to 512)

Positional Encoding: Learned embeddings to represent word order

Layer Normalization: Applied after each sub-layer for stable training

2. MODEL PARAMETERS

Pretrained Models Used

Helsinki-NLP/opus-mt-ar-en :
*Arabic (ar) to English (en)
*parameters: ~85M
*Size: ~298MB

Helsinki-NLP/opus-mt-en-ar:
*English (en) to Arabic (ar)
*parameters: ~85M
*Size: ~298MB

Key Hyperparameters

Batch Size:

value: 32 (GPU) / 16 (CPU) **What is it:** Number of samples processed per batch

Learning Rate:

value: 3e-5 **What is it:** Optimizer step size

Warmup Steps

value: 500 **What is it:** Gradually increases learning rate early in training

Weight Decay

value: 0.01 **What is it:** L2 regularization to prevent overfitting

Label Smoothing

value: 0.1 **What is it:** Helps generalization by softening target labels

Max Sequence Length

value: 128 tokens. **What is it:** Truncates longer sentences

Gradient Accumulation

value: 2 (GPU) / 8 (CPU) **What is it:** Accumulates gradients over multiple steps for a larger effective batch size

3. TOKENIZATION

- Uses SentencePiece subword tokenization.
- Vocabulary size: ~65,000 tokens (shared between source and target languages for MarianMT).
- Handles out-of-vocabulary (OOV) words by breaking them into subword units.

4. TRAINING PROCESS

Fine-Tuning on Tatoeba Dataset

- Data Preprocessing
  Normalizes Arabic text (removes diacritics, unifies characters).
  Cleans special characters, emojis, and extra spaces.
  Filters sentences (1–128 words).

- Training Loop
  Uses AdamW optimizer with weight decay.
  Cross-entropy loss with label smoothing.
  Mixed-precision training (FP16) if GPU is available.
  Early Stopping (if enabled):
  Monitors validation loss to avoid overfitting.

5. EVALUATION METRICS:

- BLEU (Bilingual Evaluation Understudy): Measures n-gram overlap.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation):
  - ROUGE-1 (unigram overlap)
  - ROUGE-2 (bigram overlap)
  - ROUGE-L (longest common subsequence)

6. LIMITATIONS

- **Sentence Length:** Works best on short-to-medium sentences (<25 words).
- **Dialects:** Trained on Modern Standard Arabic (MSA), struggles with regional dialects.
- **Rare Words:** May produce suboptimal translations for uncommon terms.
- **Idioms/Cultural References:** Literal translations may not capture intended meaning.
- **Domain Adaptation:** General-purpose model; may need fine-tuning for specialized domains (e.g., medical, legal).

THANK YOU