Selected topic in ai

Project 2

Team members:

Marwan eslam

Omar Mohamed

Farid Mohamed

Karim osama

Bodour Mohamed

Nada kamal

# Under supervision of

Eng/ Mariam Mohamed

**Table of Contents**

1. **Introduction**
2. **Dataset and Preprocessing**
3. **Model Selection Rationale**
4. **Implementation Details**
5. **Results and Analysis**
6. **Conclusion**

**1. Introduction**

**1. Problem Statement**

**Goal**

The objective of this project is to develop a **Machine Translation (NMT)** system capable of translating English text into Vietnamese using the **MarianMT** architecture provided by Hugging Face.

**Use Case**

The system can be applied to a variety of real-world scenarios such as:

* **Multilingual Communication Platforms**: Facilitating real-time communication between English and Vietnamese speakers.
* **Educational Tools**: Assisting Vietnamese learners in understanding English educational material, or vice versa.
* **Content Localization**: Automatically translating web content, software interfaces, or digital media to suit Vietnamese-speaking audiences.

**2. Dataset and Preprocessing**

**Dataset: Opus-100**

* **Source**: The Opus-100 multilingual corpus curated by Tiedemann (2020), available via Hugging Face Datasets.
* **Language Pair**: English (en) ↔ Vietnamese (vi)
* **Training Samples**: 1,000,000 sentence pairs
* **Validation Set**: 2,000 sentence pairs
* **Test Set**: 2,000 sentence pairs

**Preprocessing Pipeline**

1. **Language Extraction**:
   * Extract translation["en"] as the source input.
   * Extract translation["vi"] as the target output.
2. **Sequence Length Management**:
   * Sentences are padded or truncated to a max\_length of **128 tokens**.
3. **Tokenization**:
   * Utilized the tokenizer from **Helsinki-NLP/opus-mt-en-vi**.
   * Applied consistent tokenization to both input and target sequences.
4. **Model Input Formatting**:
   * Converted tokenized text to model-compatible inputs and labels.
   * Used Hugging Face’s DataCollatorForSeq2Seq for dynamic padding and label shifting.

**3. Model Selection Rationale**

**Model Used: Helsinki-NLP/opus-mt-en-vi**

* A pretrained **MarianMT** model specifically fine-tuned on English↔Vietnamese pairs.
* Developed by the University of Helsinki as part of the OPUS-MT project.

**Why MarianMT?**

* **Lightweight** and optimized for **low-resource** machine translation.
* **Pretrained** on large multilingual corpora.
* **Open-source** and compatible with Hugging Face’s ecosystem.
* **Easily fine-tunable** for domain-specific applications or low-resource enhancements.

**4. Implementation Details**

**Training Framework**

* Used **Hugging Face Transformers** library with the Seq2SeqTrainer class.

**Hyperparameters**

| **Parameter** | **Value** |
| --- | --- |
| Batch Size | 16 |
| Learning Rate | 2e-5 |
| Epochs | 7 |
| Max Input Length | 128 tokens |
| Weight Decay | 0.01 |
| Evaluation Strategy | epoch |

**Evaluation Metrics**

* **BLEU**: Measures n-gram precision between generated and reference translations.
* **ROUGE-L**: Focuses on the longest common subsequence.
* **METEOR**: Incorporates synonym and stem matching for semantic evaluation.

**Demo Interface**

* Built with **Gradio** for real-time user interaction.

A screenshot of a computer

Description automatically generated

* Features:
  + Text input for English.
  + Instant Vietnamese translation.
  + Display of input/output in user-friendly layout.

**5. Results and Analysis**

**Training Progress**

Loss Curve: Training loss steadily decreased across all epochs.

A graph of training curve

Description automatically generated

Final Training Loss: ~.089, indicating good convergence.

**Evaluation Scores**

**A graph with a bar and a number of numbers

Description automatically generated with medium confidence**

Metric Score

BLEU 15.49

ROUGE-L 0.5474

METEOR 0.4992

**Interpretation**

* **BLEU Score**: A moderate score, suggesting that while translations are syntactically accurate, there's room for improvement in fluency and stylistic similarity.
* **ROUGE-L and METEOR**: Higher scores here indicate the model is performing well in preserving meaning and semantic content.
* **Possible Improvements**:
  + Increasing the number of training epochs.
  + Incorporating domain-specific or back-translated data.
  + Exploring larger or multilingual pretrained models for knowledge transfer.

**7. Conclusion**

**Key Insights**

In this project, we successfully built and fine-tuned a neural machine translation (NMT) system using the **MarianMT** architecture with the **Helsinki-NLP/opus-mt-en-vi** model to translate English text into Vietnamese. By leveraging the **Opus-100** dataset and Hugging Face's **Seq2SeqTrainer**, we developed a system capable of delivering reasonable translation quality with limited resources.

Key highlights include:

* Achieved a **BLEU score of 15.49**, indicating moderate translation quality.
* Strong **ROUGE-L (0.5474)** and **METEOR (0.4992)** scores suggest good semantic understanding and content preservation.
* The model demonstrated stable training behavior with a final training loss of approximately 0**.89**.
* A user-friendly **Gradio interface** was deployed for real-time translation, making the system accessible and interactive.