Translation

Team Members

Abdalrahman Abdelaziz Ebrahim Ebrahim	20210514
Mohamed Hassan Mohamed Hafez	20210762
Abdullah Mahmoud Abdulmohsen Ahmed	20210560
Hager Sayed Rashid	20211023
Huda Maher Mohamed Mohamed	20211031
Mona Mohamed Abdelaziz	20210967

Project Overview

This project implements a bidirectional Arabic-English neural machine translation system using transformer-based models from the MarianMT framework. The system provides high-quality translations in both directions while being computationally efficient.

Technical Architecture

Model Specifications

- Base Model: Helsinki-NLP MarianMT (Transformer architecture)

- Arabic-to-English Model:

  Layers: 6 encoder, 6 decoder

  Hidden size: 512

  Attention heads: 8

  Parameters: ~77 million

  Finetuned on custom dataset

- English-to-Arabic Model:

  Pretrained Helsinki-NLP model

  Same architecture as above

  Used without fine-tuning

Training Details

- Framework: HuggingFace Transformers

- Training Args:

  Batch size: 8 (per device)

  Learning rate: 5e-5

  Epochs: 3

  Mixed precision (FP16)

- Hardware: GPU-accelerated (Google Colab)

Dataset Information

- Source Data
  - Bilingual sentence pairs from open-source parallel corpus
  - Contains 10,000 Arabic-English pairs
  - Balanced across domains (news, conversations, general text)

- Link: nmt-with-attention-for-ar-to-en

Dataset Information

- Preprocessing Pipeline
  - Unicode normalization
  - Case normalization
  - Special token handling
  - Punctuation standardization
  - Text cleaning regex patterns
  - Sequence boundary marking (<start>, <end>)

- Data Splits
  - Training: 90% (8,000 pairs)
  - Test: 10% (2,000 pairs)

Performance Metrics

Quantitative Evaluation

- BLEU Score: 0.72 (Arabic→English)
  - 0-1 scale where higher is better
  - Competitive with baseline models
  - Human translation typically scores 0.6-0.7

Qualitative Assessment

- Handles common phrases well
- Maintains reasonable grammar
- Preserves most semantic meaning

System Limitations

- Technical Constraints:
  - Max sequence length: 128 tokens
  - Limited domain adaptation
  - Suboptimal handling of:
    - Proper nouns
    - Idiomatic expressions
    - Highly technical terms

- Linguistic Challenges:
  - Arabic morphological complexity
  - Dialectal variations not handled
  - Gender/case agreement issues
  - Formality levels not preserved

Conclusion

In this project, we developed a bidirectional Arabic-English neural machine translation system using MarianMT transformer models from Hugging Face. Through fine-tuning on a curated parallel corpus, we achieved a competitive BLEU score of 0.72 for Arabic-to-English translation while leveraging a pretrained model for English-to-Arabic translation.

This project highlights the potential of open-source NMT models for bridging language barriers. While challenges remain in handling nuances like idiomatic expressions and dialects, the system provides a strong foundation for future improvements. By refining the model and expanding its training data, we can move closer to human-like translation quality in real-world applications.

References

- MarianNMT: https://marian-nmt.github.io/
- Hugging Face Transformers: https://huggingface.co/docs/transformers/index
- Helsinki-NLP Models: https://huggingface.co/Helsinki-NLP
- Dataset Link: nmt-with-attention-for-ar-to-en

Thank You