Ar-En Translation App

NLP Project

Project Overview

NMT: Ar-En

This project focuses on building an Arabic-to-English neural machine translation (NMT) system using a fine-tuned MarianMT model.

The system processes Arabic text, handles linguistic nuances such as diacritics and orthographic variations, and translates it into English.

The project uses the Hugging Face Transformers library and a custom dataset for training and evaluation.

Project Information

Dataset

- **Source**: Custom Arabic-English parallel corpus (ara_eng.txt) from GitHub
- **Size**: Not explicitly specified, but split into 80% training and 20% testing
- **Format**: Tab-separated text file with English and Arabic sentence pairs
- **Data Splitting**:
  - Train-test split: 80-20 ratio
  - Random state: 42 for reproducibility
- **Conversion**: Converted to Hugging Face Dataset format for training

- **Preprocessing**:
  - Arabic text cleaned by:
    - Removing diacritics (Unicode range \u064B-\u065F)
    - Normalizing characters (e.g., ى to ة ,ي to إ ,ه to ا)
    - Removing non-Arabic characters (except spaces and punctuation: ؟!.،)
    - Collapsing multiple spaces
  - English text stripped of leading/trailing whitespace

Model Parameters

- **Model**: MarianMT (Helsinki-NLP/opus-mt-ar-en)
- **Architecture**: Transformer-based sequence-to-sequence model
- **Tokenizer**: MarianTokenizer, customized for Arabic-English translation
- **Input/Output Max Length**: 256 tokens (adjusted from 128 for better context capture)
- **Training Hyperparameters**:
  - Per-device train batch size: 6
  - Per-device eval batch size: 6
  - Logging steps: 50
  - Predict with generate: Enabled
- **Generation Parameters**:
  - Num beams: 10 (evaluation), 5 (inference)
  - Length penalty: 1.5 (evaluation)
  - No repeat n-gram size: 3 (inference)
  - Early stopping: Enabled

Model Limitations

- **Context Length**: Limited to 256 tokens per input/output, which may truncate longer sentences or lose context.
- **Dataset Size**: The dataset size is not specified, and a small dataset may limit generalization.
- **Arabic Dialects**: The model may struggle with dialectal Arabic (e.g., Egyptian, Levantine) if the dataset primarily contains Modern Standard Arabic (MSA).
- **Orthographic Variations**: While preprocessing normalizes some variations, inconsistent spellings or rare characters may affect performance.
- **Long Text Handling**: Long texts are split into chunks (300 characters), which may disrupt coherence across sentence boundaries.

Methodology

Data Preprocessing

- **Text Cleaning**: Arabic text is cleaned using a custom clean_text function to normalize characters, remove diacritics, and retain only Arabic letters and select punctuation.
- **Data Loading**: Loaded into a Pandas DataFrame, with null values dropped and empty strings filtered out.
- **Splitting**: Train-test split performed using sklearn.model_selection.train_test_split.

Tokenization

- **Tokenizer**: MarianTokenizer from Helsinki-NLP/opus-mt-ar-en
- **Settings**:
  a. Max length: 256 tokens
  b. Padding: max_length
  c. Truncation: Enabled
  d. Padding side: Right
- **Process**: Arabic inputs and English targets tokenized separately, with English tokenized as the target using as_target_tokenizer.

Fine-Tuning

- **Objective**: Adapt the pre-trained MarianMT model (Helsinki-NLP/opus-mt-ar-en) to the custom Arabic-English dataset for improved translation performance.
- **Framework**: Hugging Face Seq2SeqTrainer
- **Dataset**: Tokenized training dataset (80% of ara_eng.txt) used for fine-tuning, with the test dataset (20%) for evaluation.
- **Training Configuration**:
  a. Batch size: 6 (per device for both training and evaluation)
  b. Logging: Every 50 steps, saved to ./logs
  c. Save strategy: No intermediate checkpoints (save_strategy="no")
  d. Output directory: ./results
  e. Reporting: Disabled (report_to="none")
- **Process**:
  a. The pre-trained model's weights are updated using the custom dataset over 10 epochs.
  b. The trainer.train() function executes the fine-tuning, optimizing the model for the specific Arabic-English sentence pairs.
- **Model Saving**: The fine-tuned model and tokenizer are saved to ./results for inference and evaluation.

- **Limitations**:
  a. Small dataset size (unspecified) may limit fine-tuning effectiveness.
  b. Risk of overfitting due to 10 epochs without early stopping or validation-based checkpointing.
  c. Fixed hyperparameters (e.g., batch size, epochs) may not be optimal.
- **Future Improvements**:
  a. Add a validation set for early stopping or checkpointing.
  b. Experiment with learning rates, batch sizes, or regularization (e.g., dropout).
  c. Incorporate a larger or more diverse dataset.
  d. Enable checkpointing to save intermediate models.

Training

- **Scope**: The fine-tuning process constitutes the core training phase, as described above.
- **Additional Notes**: The training leverages the tokenized train and test datasets, with evaluation performed post-training.

Evaluation

- **Metrics**:
  - **BLEU**: Measures n-gram overlap between predicted and reference translations.
  - **ChrF**: Character n-gram F-score, robust for morphologically rich languages like Arabic.
- **Process**:
  - Generated translations for test set using beam search (10 beams, length penalty 1.5).
  - Computed BLEU and ChrF scores using the evaluate library.
  - Printed sample translations for qualitative analysis.
- **Output**: Included Arabic input, predicted English translation, true English translation, and evaluation scores.

Inference

- **Long Text Handling**:
  - Text split into chunks of 300 characters using sentence boundaries (split_text_by_sentences).
  - Chunks processed in batches (batch size: 6).
  - Translated using beam search (5 beams, no repeat n-gram size 3).
- **Output Processing**:
  - Translated chunks joined with spaces.
  - Multiple spaces collapsed to ensure clean output.

Arabic → English Translator

Arabic → English Translator

Arabic

إلى تحسين طرق شهد الطب تطورًا مذهلاً في العقود القليلة الماضية بفضل التقدم التكنولوجي والاكتشافات العلمية. أدت هذه التطورات باء الآن استخدامالتشخيص والعلاج، وزيادة معدلات الشفاء من العديد من الأمراض التي كانت تعتبر مستعصية في الماضي. أصبح بإمكان الأط ضى. كما ساعدتتقنيات مثل الذكاء الاصطناعي والروبوتات في العمليات الجراحية، مما زاد من دقة الإجراءات الطبية وسرعة تعافي المر تكلفة الرعايةتطبيقات الصحة الرقمية في تمكين الأفراد من مراقبة حالتهم الصحية بشكل مستمر. ومع ذلك، لا تزال هناك تحديات تتعلق ب زاتالصحية وإمكانية الوصول إليها للجميع، مما يتطلب حلولًا مبتكرة وسياسات عادلة لضمان استفادة الجميع من هذه الإنجا

Source chars: 611

Translate Text

Translate File

English

medicine has seen an spectacular change in the past few decades, thanks to advances in technology and science. These advances have improved methods of diagnosis and treatment, increasing the levels of disease recovery that have been difficult in the years. Doctors are now able to use such techniques as artificial smarts and robots in the process, increasing the speed of medical action and the recovery of patients. Digital health applications have also helped to allow people to continue their health surveillance. There are, however, still some challenges regarding the cost of and universal access to health care, which require innovative responses and just policies to guarantee that these results are available for everyone.

Copy Text

Translated chars: 731