English-to-Spanish Translation using Deep Learning Models

Naga Akshar Athman N

October 4, 2025

Abstract

This project implements and compares four neural architectures for English-to-Spanish translation: Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and a pretrained Transformer (BERT/-MarianMT). The models were trained on a Kaggle English–Spanish dataset. Evaluation with BLEU scores shows that the BERT-based Transformer significantly outperforms the recurrent models.

1 Introduction

Machine translation is a vital task in Natural Language Processing (NLP). Traditional recurrent models such as RNN, LSTM, and GRU can learn language patterns, but modern Transformer architectures with pretraining (e.g., BERT) achieve much higher performance. This project evaluates these models on English-to-Spanish translation.

2 Dataset

We used the Kaggle English-to-Spanish dataset containing parallel sentence pairs. A subset of around 20,000 pairs was used for training and testing due to compute limits. Data preprocessing involved cleaning, tokenization, integer encoding, and padding sentences to a fixed length of 20 tokens.

3 Preprocessing

- Tokenization using Keras Tokenizer for RNN, LSTM, GRU.
- Padding to fixed length (20).
- Train/test split (80/20).
- For BERT model, HuggingFace MarianMT tokenizer was used.

4 Models

4.1 RNN, LSTM, GRU

Implemented in Keras using an Embedding layer, recurrent layer (RNN/LSTM/GRU), and Dense output with softmax activation.

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, GRU, Dense

model = Sequential([
    Embedding(input_dim=20000, output_dim=128, input_length=20),
    GRU(128, return_sequences=True),
    Dense(20000, activation="softmax")
])
```

4.2 BERT Transformer (MarianMT)

We used the pretrained Helsinki-NLP MarianMT model from HuggingFace for English→Spanish translation.

5 Training

- Optimizer: Adam
- Loss: Sparse Categorical Cross-Entropy
- Epochs: 5
- Batch size: 64

6 Results

6.1 BLEU Scores

Model	BLEU Score
RNN	20.56
LSTM	17.29
GRU	9.29
BERT (MarianMT)	41.11

Table 1: BLEU score comparison of models.

6.2 Example Translations

- EN: “Hello” → ES: “Hola”
- EN: “Good morning” → ES: “Buenos días”

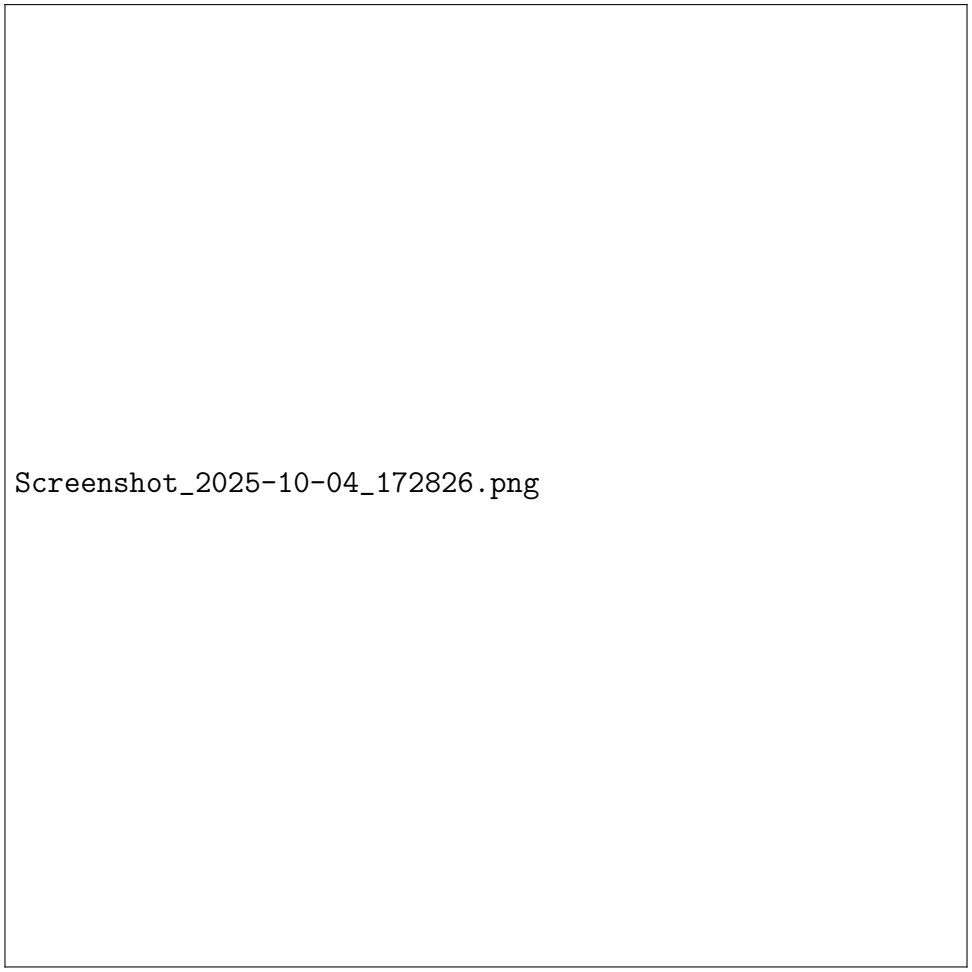

Figure 1: BLEU Scores for Different Translation Models.

- EN: "What is your name?" → ES: "¿Cómo te llamas?"
- EN: "How are you?"
  True ES: "¿Cómo estás?"
  Predicted ES (GRU): ...
  Predicted ES (BERT): "¿Cómo estás?"

7 Discussion

The results show that:

- GRU achieved the lowest score (9.29), struggling with translations.
- LSTM performed better (17.29), followed by RNN (20.56).
- The pretrained BERT Transformer dominated with 41.11 BLEU, close to professional-quality translation.

8 Conclusion and Future Work

- Pretrained Transformers (BERT/MarianMT) clearly outperform recurrent models.

- RNN/LSTM/GRU are useful for learning purposes but limited in real-world translation quality.
- Future work: train on larger datasets, apply beam search decoding, use multilingual models like mBART/mT5, and deploy as a translation web app.

References

- Kaggle English–Spanish Dataset.
- HuggingFace MarianMT: `https://huggingface.co/Helsinki-NLP/opus-mt-en-es`
- TensorFlow/Keras Documentation.
- SacreBLEU Toolkit.