

1. Introduction

In this project, I used a pretrained Transformer model from the HuggingFace transformers library to perform machine translation on text from Romeo and Juliet by William Shakespeare. The goal was to translate selected English passages from the play into Spanish, evaluate the model's performance using an appropriate metric, and explore how different hyperparameters affect translation quality. Machine translation is a core NLP task, and Transformer based models, especially encoder decoder architectures, are currently state of the art for this type of problem. This project demonstrates how modern pretrained models can be applied directly to real text and evaluated without building a model from scratch.

2. Dataset Description

The dataset for this project consists of text from the public-domain book Romeo and Juliet (Project Gutenberg Book #349).

The plain UTF-8 text file was accessed directly through the following public URL:

<https://www.gutenberg.org/files/349/349-0.txt>

The full book was downloaded programmatically in the Colab notebook using `requests.get()`, following assignment rules that forbid local file uploads. The text was cleaned by removing the Gutenberg header/footer and then split into paragraphs. For evaluation, a curated test set of 10 short passages from the play was selected. I manually created Spanish reference translations for each passage so the model's output could be compared using the BLEU score.

3. Methodology

3.1 Why Transformers Work for Machine Translation

Transformer models rely on a mechanism called self attention, which allows each token in a sentence to directly consider every other token. For translation, this is ideal because

the model must understand long range dependencies and reorder phrases into the target language.

Most translation models use an encoder decoder architecture:

- The encoder reads and embeds the English input.
- The decoder generates the Spanish output one token at a time, attending to both previous output tokens and the encoder's hidden states.

This architecture gives Transformers strong fluency and accuracy when generating target language sequences.

3.2 Model Used

The model used in this project was:

Helsinki-NLP/opus-mt-en-es

This is a pretrained encoder-decoder Transformer specifically trained for English to Spanish translation. It is lightweight, fast on CPU/GPU, and fully supported through HuggingFace's pipeline API.

3.3 Implementation

The core translation pipeline:

```
translator = pipeline("translation_en_to_es", model="Helsinki-NLP/opus-mt-en-es")
```

Input: English text (passages from the play)

Output: Spanish text

Evaluation metric: BLEU score using sacrebleu

Hyperparameters tested:

- Default settings
- num_beams = 4 (beam search)
- max_length = 128 (shorter outputs)

4. Experiments and Results

4.1 Evaluation Setup

To evaluate translation quality, I created a test set of 10 short passages taken directly from Romeo and Juliet. Each passage contained:

- The original English text
- A manually written Spanish reference translation
- The model's predicted translation

BLEU was used as the evaluation metric through the sacrebleu library. BLEU compares the model's output to the reference based on overlapping n-grams, with higher values indicating closer matches.

Three translation configurations were tested:

- Default pipeline settings
- Beam search (num_beams = 4)
- Shorter output limit (max_length = 128)

All experiments used the same model: Helsinki-NLP/opus-mt-en-es

4.2 Quantitative Results

Experiment	BLEU Score
Default	32.14
Beam Search (4 beams)	32.14
Shorter max_length (128)	32.14

All three settings produced the same BLEU score, even though the translations differed qualitatively. This result is expected due to:

1. Small evaluation set

BLEU is designed for large datasets.

With only 10 samples, BLEU becomes coarse and cannot capture subtle stylistic differences.

2. High lexical overlap

Shakespearean lines often translate into similar Spanish structures.

Even when synonyms differ (e.g., hogares vs familias), many n-grams still match the reference.

4.3 Qualitative Results

Example 1

EN: Two households, both alike in dignity, in fair Verona where we lay our scene.

REF: Dos familias, ambas iguales en dignidad, en la hermosa Verona donde situamos nuestra escena.

PRED (Beam): Dos hogares, ambos iguales en dignidad, en la bella Verona donde ponemos nuestra escena.

Differences:

- hogares vs familias
- ponemos nuestra escena vs situamos nuestra escena
Still accurate, but stylistically different.

Example 2

EN: From ancient grudge break to new mutiny, where civil blood makes civil hands unclean.

REF: De un antiguo rencor nace una nueva revuelta, donde la sangre de ciudadanos mancha manos ciudadanas

.

PRED (Beam): Desde el antiguo rencor hasta el motín nuevo, donde la sangre civil contamina las manos civiles.

Differences:

- More literal structure
- contamina instead of mancha

Example 3

EN: A pair of star-crossed lovers take their life.

REF: Una pareja de amantes malditos por las estrellas se quita la vida.

PRED (Beam): Una pareja de amantes cruzados por las estrellas se quitan la vida.

Differences:

- cruzados por las estrellas vs malditos por las estrellas
- Meaning remains correct

4.4 Observations

- Beam search produced the most complete and natural sounding translations.
- Default settings were slightly less literal but still accurate.
- Lower max_length sometimes simplified or shortened phrases.
- Despite Shakespeare's archaic language and inverted syntax, the model preserved meaning well.

5. Discussion

Even though the three experiments produced identical BLEU scores (~32.14), the outputs clearly differed when inspected manually. This reveals a limitation of using BLEU on small datasets and poetic language. Beam search produced the best translations, even though BLEU did not reflect the improvement. Default settings were still strong, while short max_length occasionally cut off important details.

Overall, the model handled Shakespearean English well. This demonstrates the power of pretrained encoder-decoder Transformers and shows that quantitative metrics must be combined with qualitative human evaluation to fully understand model performance.

6. Conclusion

In this project, I applied a pretrained Transformer model to translate English passages from Romeo and Juliet into Spanish. The HuggingFace transformers library made it easy to use a modern encoder decoder architecture without training a model from scratch. I evaluated translation quality using BLEU and explored the effects of different hyperparameters.

Although the BLEU scores for all configurations were identical, the qualitative differences showed that beam search generated the most complete and fluent translations. Shorter max lengths produced more compact outputs, sometimes at the cost of meaning. The model performed well despite the complexity of Shakespeare's language.

7. References

Project Gutenberg (Dataset Source)

Shakespeare, William. Romeo and Juliet. Project Gutenberg, Book #349.

<https://www.gutenberg.org/files/349/349-0.txt>

Transformers Library

Hugging Face. Transformers Documentation.

<https://huggingface.co/docs/transformers>

Translation Model

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT: Building open translation services for the world.

Model: <https://huggingface.co/Helsinki-NLP/opus-mt-en-es>