Analysis of Machine Learning Techniques for Translating Educational Documents from English to Spanish

Aaron Hassan Robinson
U3224074
Engineering (honours) – Software and Network
University of Canberra
Canberra, ACT
u3224074@uni.canberra.edu.au

***Abstract*—Machine translation remains pivotal for bridging linguistic barriers, particularly in underserved communities reliant on English-Spanish document conversion for educational and cultural integration. This study evaluates hyperparameter optimisation in a transformer based seq2seq model for English-to-Spanish translation, employing a grid search across vocab sizes (16,000-25,000), embedding dimensions (256-512) and attention heads (4-8) on a compact corpus of 118,964 sentence pairs. Evaluation harness consisted of BLEU and chrF2 scores. The optimal configuration found had a vocab size of 25,000, embedding dimension of 512, and 4 attention heads, and got a BLEU score of 19.30 and a chrF2 score of 37.41, demonstrating the effect of expanding the lexical coverage in language translation tasks, and mitigating out of vocabulary challenges inherent to Spanish's morphological richness. Conversely, reduced vocab sizes decreased performance (BLEU: 9.86-10.55) underscoring the need for more robust tokenisation. Larger embeddings enhanced BLEU when paired with higher vocab, whilst increasing heads to 8 produced inconsistent results, occasionally boosting chrF2 while decrementing BLEU, hypothesised to be from overfitting.**

Keywords—AI, Machine learning, Spanish, English, seq2seq, transformer, encoder, decoder

I. Introduction

Language translation has been a staple part of connecting different cultures globally, allowing the transfer of knowledge and empowering individuals to take advantage of resources they otherwise would not have access to. However, in the modern age, there are still communities that suffer from isolation from the rest of society and do not have access to resources to translate important documents, which can lead to a halt in progress within their communities. With the latest AI boom, deep learning techniques have emerged that have greatly improved the way we translate documents, allowing for deeper capturing of semantic and cultural nuances. Development of these models means that they can be deployed in remote areas and assist locals with further incorporating educational documents into their curriculum. This paper aims to evaluate techniques for translating English documents into Spanish, by preserving the original meaning.

II. Dataset

A. Description

The dataset was sourced from [1] hosted on Kaggle. Initial analysis on the website shows:

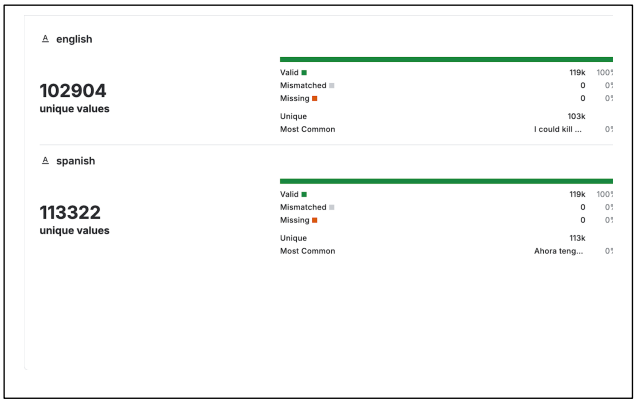


Fig. 1. Screenshot from Kaggle website showing initial analysis of dataset

From this we can see we have 102,904 unique English values, and 113,322 unique Spanish values. Normally for classification tasks, an imbalanced dataset is undesirable for training; however, for language translation, this is often not the case.

B. Dataset justification

This corpus serves as an effective baseline for training the transformer. The balanced data set has no mismatched columns, and has enough rows that exceed our computational needs.

C. Dataset exploration

Dataset analysis:

Name	***Value***
Total dataset size	118,964 sentence pairs
Training set size	95,171 (80.0%)
Validation set size	23,793 (20.0%)
Validation split	20.0%

Sentence length analysis – English:

TABLE I. ENGLISH SENTENCE LENGTH ANALYSIS

Name	***Value***
Count	118964.000000
Mean	6.310363
STD	2.611586

Name	*Value*
min	1. 000000
25%	4.000000
50%	6.000000
75%	8.000000
Max	47.000000
Name	english_length, dtype: float64

Sentence length analysis – Spanish:

TABLE II. SPANISH SENTENCE LENGTH ANALYSIS

Name	*Value*
Count	118964.000000
Mean	6.083874
STD	2.764453
min	1. 000000
25%	4.000000
50%	6.000000
75%	7.000000
Max	49.000000
Name	spanish_length, dtype: float64

Table 1 and 2 tell us, after processing, we have 118,964 pairs. The average English sentence length is 6.31 words, while Spanish is 6.08 (rounded 2 dp). This indicates that English sentences are slightly longer, and that most sentences are quite short. The standard deviation for English is 2.61, while Spanish is 2.76. This shows that Spanish sentences have slightly more variability in length.

The percentile quartile analysis tells us that, for English and Spanish, 25% of sentences have less than or equal to four words. 50% have less than or equal to six words. For English, 75% of sentences have eight or less words whereas Spanish has seven or less.

Analysing the interquartile range of our data (IQR), we can see that for English, it comes to 8 - 4 = 4 words, and for Spanish, 7 - 4 = 3 words. This tells us that 50% of the data spans only 3-4 words, therefore, is very compact. The minimum sentence length for English and Spanish is one word, whereas the longest sentence length is 47 words (English) and 49 words (Spanish).

After analysing the data, we know the minimum and maximum are outliers, as most data is much shorter. This means is that the original sequence_length=20 we gave in Assignment 2 is generous.

Vocab analysis:

TABLE III. ENGLISH VOCAB ANALYSIS

Unique words	23,848
Total words	750,706
Vocabulary richness	0.0318

TABLE IV. SPANISH VOCAB ANALYSIS

Unique words	41,720
Total words	723,762
Vocabulary richness	0.0576

TABLE V. MODEL INITIAL VOCAB ANALYSIS

Model vocabulary size (config)	16,000
Coverage English	67.1%
Coverage Spanish	38.4%

Next, vocabulary analysis shows a significant imbalance between source and target languages. The English corpus contains 23,484 unique words whereas the Spanish corpus contains 41,720 unique words. This imbalance could be due to Spanish, as a language, having a much richer morphological structure; Spanish has more verb conjugations, and gender/number agreement words than English [2].

The vocabulary richness metric (unique words over total words) helps quantify this disparity: Spanish, at 0.0576, demonstrates roughly 1.8x greater lexical diversity than English at 0.0318.

With our current hyperparameter of vocabulary set to 16,000, we achieve 67.1% coverage for English but only 38.4% for Spanish. This presents a critical limitation of our current model, where 61.6% of the Spanish vocabulary is out-of-vocabulary (OOV), which may degrade translation quality for complex or infrequent terms.

Sequence length analysis:

TABLE VI. SENTENCES WITHIN 20 WORDS

English	118,857 (99.91%)
Spanish	118,810 (99.87%)

TABLE VII. SENTENCES EXCEEDING 20 WORDS

English	107 (0.09%)
Spanish	154 (0.13%)

Top ten words:

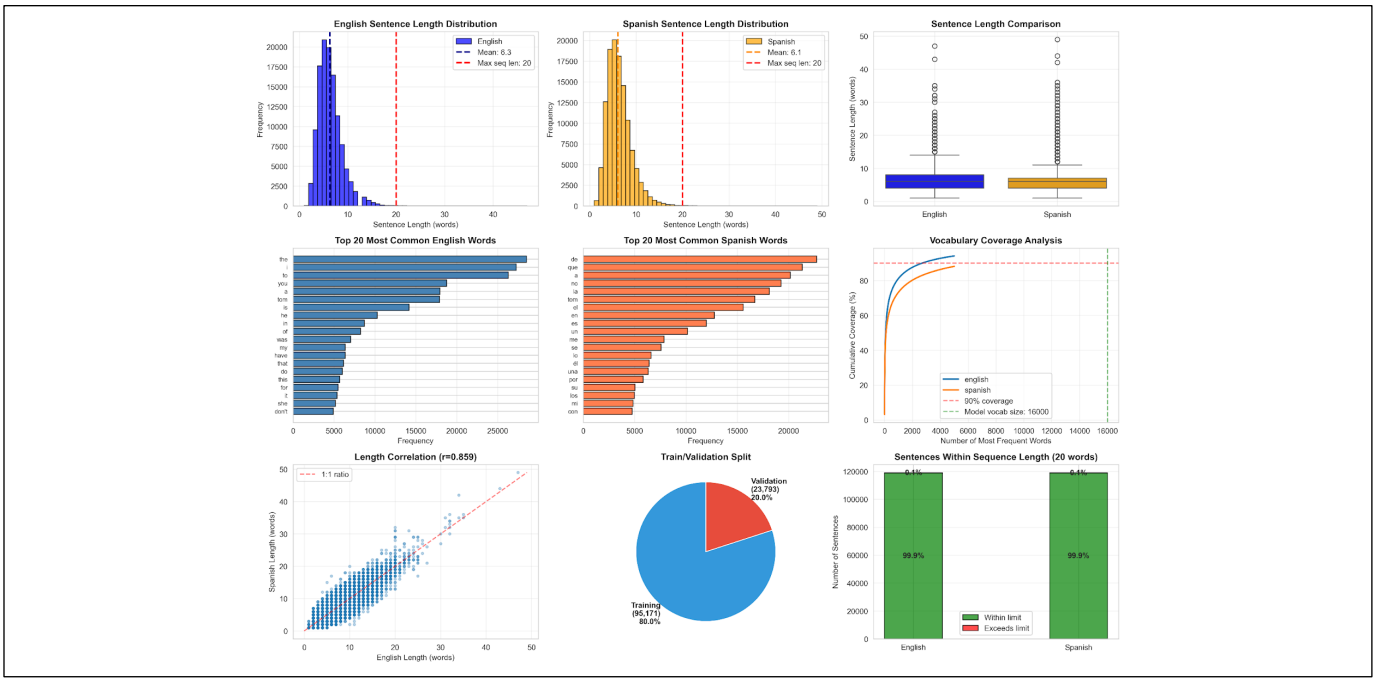
English: ['the', 'i', 'to', 'you', 'a', 'tom', 'is', 'he', 'in', 'of']

Spanish: ['de', 'que', 'a', 'no', 'la', 'tom', 'el', 'en', 'es', 'un']

Finally, analysis of sequence length demonstrates that the chosen hyperparameter maximum length of 20 accommodates 99.91% of English sentences and 99.87% of Spanish

sentences, with only 107 English sentences and 154 Spanish sentences exceeding the limit.

Frequency analysis of the most common words reveals typical linguistic patterns, with function words (i.e., articles, prepositions, and pronouns) dominating both vocabularies. It is worth noting that the noun 'Tom' appears in both top 10 most frequent words. This may potentially be a limitation of the chosen dataset, as it suggests that the corpora contain artificially constructed example sentences rather than naturally occurring translations. Whilst such a dataset may be useful for establishing a baseline translation capability, it may not fully capture complex or diverse real world translation scenarios. This nature should be considered when interpreting model performance.



III. Computational Approach

For this translation task, the transformer architecture was selected and is heavily based on the implementation in [3]. This comprises a seq2seq model that relies entirely on attention mechanisms rather than recurrent connections. The transformer has become the standard for neural machine translation tasks due to its superior performance and parallelisable training compared to traditional RNN-based approaches. The implementation consists of three main components:

A. Encoder

The encoder processes the input English sentence into a contextual representation. It utilises multi-head self-attention to capture relationships between all words. Furthermore, it contains a feed-forward network. Finally, layer normalisation and residual connections are used for training stability.

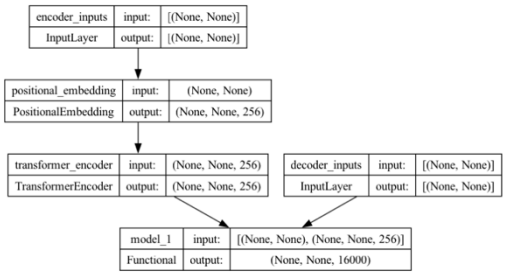
B. Decoder

The decoder generates Spanish translations autoregressively, meaning one token at a time. It utilises masked self-attention to prevent information leakage from future tokens. Furthermore, it has a cross-attention mechanism to attend to encoder outputs. Finally, the architectural patterns are identical to the encoder in terms of attention + feed forward network (FFN) + normalisation.

C. Positional encoding

This is used to inject sequence order information as transformers have no notion of position. This uses learned positional embeddings rather than sinusoidal encoding and is applied to both the encoder and decoder inputs.

D. Overall architecture



IV. Implementation

Tensorflows `TextVectorization` layer was used for converting text to integer sequences.

A. Preprocessing decisions

Firstly, `[start]` and `[stop]` tokens were added to Spanish sequences to signal. In order to do this, we defined a custom standardisation for Spanish that excludes the square brackets. Additional special character handling was also applied to remove the inverted question mark '¿', common in Spanish corpus. A frequency-based vocabulary was utilised: words are ranked by frequency; OOV words map to a padding token.

1) Sequence formatting

The encoder receives the full English input [w1, w2, …, wn]. The decoder receives the Spanish sequence shifted right [start, w1, w2, …, wn-1].

Decoder target is the Spanish sequence shifted left: [w1, w2, …, wn, stop].

This teacher-forcing strategy enables the decoder to learn from gold-standard previous tokens during training.

2) Data optimisation

Several data optimisation techniques were utilised:
Shuffling: prevents memorsation of sequence order.
Batching: groups examples for parallel processing.
Prefetching: overlaps data loading with model training.
Caching: stores preprocessing data in memory.

B. Model configuration

These are the baseline hyperparameters for our base model, balancing model capacity with computational constraints:

TABLE VIII. MODEL HYPERPARAMETERS

Hyper-parameter	Value	Justification
Vocabulary sse	16,000	Covers 67% English, 38% Spanish vocab
Sequence length	20	Accommodates 99.9% of dataset
Embedding dimension	256	Standard size for moderate-scale tasks
Latent dimension (FFN)	256	Matches embedding dimension
Num of attention heads	4	Allows diverse attention patterns
Batch size	512	Maximises GPU utilisation
Dropout rate	0.5	Prevents overfitting

C. Training strategy

The optimiser chosen was Adaptive learning rate (Adam)
The loss function chosen was sparse categorical cross-entropy (appropriate for word prediction).
Metrics include accuracy at the token level and edit distance at the sequence level.

Regularisation:
Dropout applied after positional embeddings and in decoder (rate=0.5).
Early stopping enabled with a patience level =10.
Model checkpointing enabled.

Hardware utilisation:
Training performed on Apple M3 Max with metal GPU acceleration.
Tensorflow 2.15.0 with metal backend for optimised performance on Apple Silicon

D. Algorithm selection rational:

The reason why the transformer was selected over other architectures.

Transformers enable parallel processing of all sequence positions simultaneously through self-attention mechanisms, which dramatically reduces training time compared to the sequential processing inherent in LSTMs and GRUs which suffer from computational bottlenecks

Second, the direct connections established by using attention mechanisms between all input-output positions eliminates the vanishing gradient problem that affects RNNs when capturing long-range dependencies.

Transformers represent the current state-of-the-art in neural machine translation.

Adam optimiser was selected for its adaptive learning rates and robust performance across various deep learning tasks.

V. Evaluation harness

To properly assess the translation model, an evaluation harness incorporating traditional string-matching metrics and modern quality assessment approaches was developed. Whilst BLEU remains the most widely used metric for machine translation evaluation, research has shown that traditional metrics show poor performance in capturing semantic similarity between MT outputs and human reference translations, and therefore, we employed complementary metrics to provide a holistic view of the model performance.

A. BLEU: Bilingual Evaluation Understudy

Automatic metric that measures n-gram overlap between machine translations and human references. It computes a

modified precision score for n-grams, and applies a brevity penalty to discourage overly short translations. BLEU has been the industry standard for MT.

$$BLEU\,Score = BP * exp(\sum_{i=1}^{N}(w_i * ln(p_i))$$

Where:

- BP = brevity penalty
- Pn = modified n-gram precision
- Wn = uniform weights

Limitations:

- Ignores semantic similarity
- Biased toward precision over recall
- Insensitive to word order beyond n-gram overlap

B. chrF2: Character n-gram F-score

Measures character n-gram overlap correlation. Has been shown to correlate better than BLEU with human judges. This would have advantages for our task as Spanish has a rich morphology that might not be captured by the BLEU score.

VI. HYPER PARAMETER SELECTION

To address the vocabulary coverage limitation, we experimented with increasing vocabulary sizes (16k and 25k) to evaluate the trade-off between model capacity and Spanish lexical coverage.

Addressing sequence length:
A sequence length of 20 words was selected based on empirical analysis showing 99.6% dataset coverage: balancing computational efficiency with minimal truncation-induced info loss.

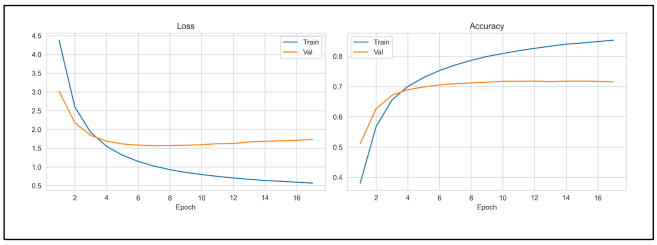
Final param grid settled:

Vocab_size	16000, 25000
Sequence_length	20
Batch_size	512
Embed_dim	256, 512
Num_heads	4, 8

Total configurations: 8. Estimated time: 20.0 hours (assuming 2.5h per config on my device (Apple m3)

VII. RESULTS

Initial training curves:



Models were named in a format that reflected the hyperparameters used to tune them: "*model_vs{VOCAB_SIZE}_s1{SEQ_LEN}_bs{BATCH_SIZE}_ed{EMBED_DIM}_nh{NUM_HEADS}*"

A. Top 5 models

1. model_vs25000_sl20_bs512_ed512_nh4
   BLEU: 19.30 | chrF2: 37.41
   Config: vocab=25000, embed=512, heads=4
2. model_vs16000_sl20_bs512_ed512_nh8
   BLEU: 10.55 | chrF2: 36.06
   Config: vocab=16000, embed=512, heads=8
3. model_vs25000_sl20_bs512_ed256_nh4
   BLEU: 10.25 | chrF2: 42.65
   Config: vocab=25000, embed=256, heads=4
4. model_vs16000_sl20_bs512_ed512_nh4
   BLEU: 9.86 | chrF2: 39.56
   Config: vocab=16000, embed=512, heads=4
5. model_vs25000_sl20_bs512_ed256_nh8
   BLEU: 9.86 | chrF2: 42.31
   Config: vocab=25000, embed=256, heads=8

B. Sample translations on unseen data (5 random examples, utilising best model):

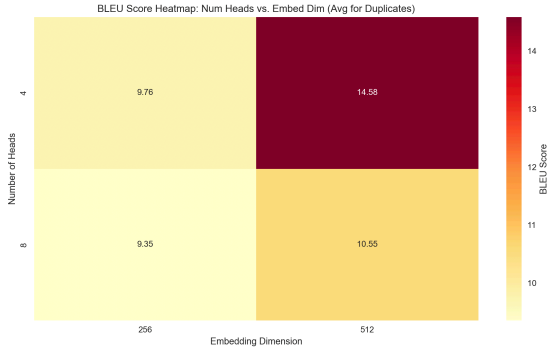
English	I know it in my heart.
Reference	Lo sé con mi corazón.
Translated	sé que me gusta

English	Even during work, I secretly indulge my Internet addiction.
Reference	Incluso durante el trabajo, secretamente satisfago mi adicción por Internet.
Translated	incluso hora trabajo me permito secretamente complacer mi adicción a internet

English	Tell them to come here.
Reference	Deciles que vengan aquí.
Translated	diles que vengan aquí

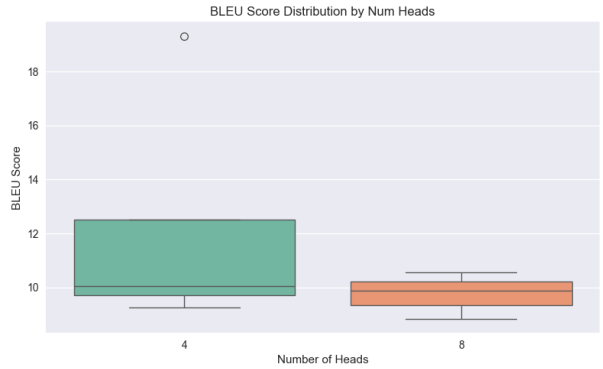
English	She's six years older than me.
Reference	Ella es seis años mayor que yo.
Translated	ella tiene seis años más que yo

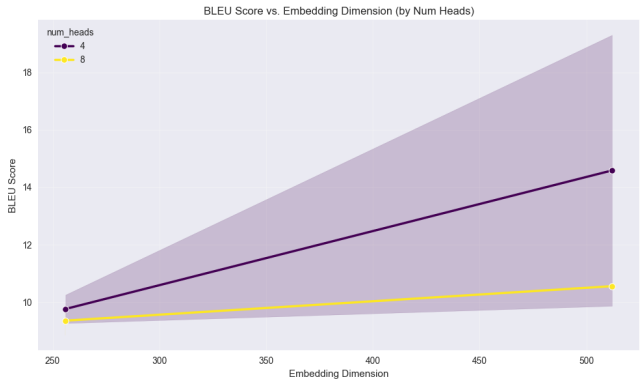
English	Let it rest, Tom.
Reference	Déjalo descansar, Tom.
Translated	deja que tom se vaya a dejar descansar



Above: BLEU score heatmap: number of heads vs embed dim

Below: BLEU score distribution by number of heads





Above: BLEU vs Embedding dimension (2 series, number of heads)

C. Discussion

The grid search results reveal patterns in the transformer model's performance for English to Spanish translation, with the top configuration featuring a vocabulary size of 25,000, an embedding dimension of 512, and four attention heads. This model achieved the highest BLEU score of 19.30 and a chrF2 score of 37.41, demonstrating the benefits of expansive lexical coverage and deeper representational capacity in capturing semantic nuances. In contrast, models with a reduced vocabulary size of 16,000 exhibited notably lower BLEU scores (for example 10.55 for the 512-dimensional embedding with 8 heads and 9.86 for 4 heads), highlighting the critical role of mitigating out of vocabulary issues in Spanish, which has inherently greater morphological diversity.

It was noted that while larger embedding dimensions (512 vs 256) consistently boosted BLEU in tandem with higher vocabulary sizes, increasing the number of attention heads from 4 to 8 yielded mixed outcomes, sometimes enhancing chrF2 (for example, 42.65 for 256-dim with 4 heads) but diminishing overall fluency as measured by BLEU. This is hypothesised to be from overfitting on the compact dataset.

These findings suggest that balanced capacity scaling is key to preserving translation correctness.

D. Model persistence and reusability

The final transformer model was saved using TensorFlow's native serialisation tools to ensure seamless reusability. Specifically, the model architecture and weights were saved via the line: *tf.keras.models.save_model(transformer, 'best_transformer_model', save_format='tf').*

This approach was preferred over tools such as pickle due to its native compatibility with the TensorFlow ecosystem.

In order to predict future unseen data the saved model can be reloaded with the code:
tf.keras.models.load_model('best_transformer_model')
And then deployed in inference mode. English sentences are tokenised using the pre-adapted English TextVectorization layer, passed through the encoder-decoder pipeline, and then decoded to generate Spanish translations, enabling real-time

applications like document processing in low-resource settings.

E. Ethical and privacy considerations

Several ethical and privacy issues emerged in this task, particularly around linguistic bias, cultural misrepresentation and data handling in a global context. The training corpus, derived from public parallel sentence pairs, risks echoing biases from overrepresented standard dialects, which may potentially marginalise indigenous or Latin American variants. Privacy risks are minimal as the dataset contains no personally identifiable information; however, if we were to extend the training to a larger dataset, this should be considered.

FUTURE WORK

Due to time constraints the full grid search could not be realised, meaning the more complex 25000 vocab size models could not be tested. If there was greater computational power and time, a more extensive grid search would be conducted. It is hypothesised that further hyperparameter tuning, exploring deeper layers, or optimising learning rates could further elevate performance, enabling more robust deployment in resource constrained settings for culturally sensitive document translation.

REFERENCES

[1] Lonnie, "English-Spanish Translation Dataset," Kaggle.com, 2025. https://www.kaggle.com/datasets/lonnieqin/englishspanish-translation-dataset

[2] D. Martinez et al., "The effects of explicit morphological analysis instruction in early elementary Spanish speakers," Journal of Experimental Child Psychology, vol. 246, pp. 106004–106004, Jul. 2024, doi: https://doi.org/10.1016/j.jecp.2024.106004.

[3] lonnieqin, "English-Spanish Translation: Transformer," Kaggle.com, Nov. 03, 2024. https://www.kaggle.com/code/lonnieqin/english-spanish-translation-transformer/notebook