Report

1. Introduction

Machine translation using a sequence-to-sequence (seq2seq) encoder-decoder model has revolutionized the field of language translation. By leveraging the power of deep learning and recurrent neural networks, this approach has enabled the automatic translation of text from one language to another. The seq2seq model consists of an encoder network that processes the source language input and captures its semantic representation, followed by a decoder network that generates the corresponding translated output. This paradigm shift in machine translation has paved the way for more accurate and contextually relevant translations, empowering individuals and organizations to bridge linguistic barriers and foster global communication.

2. Dataset

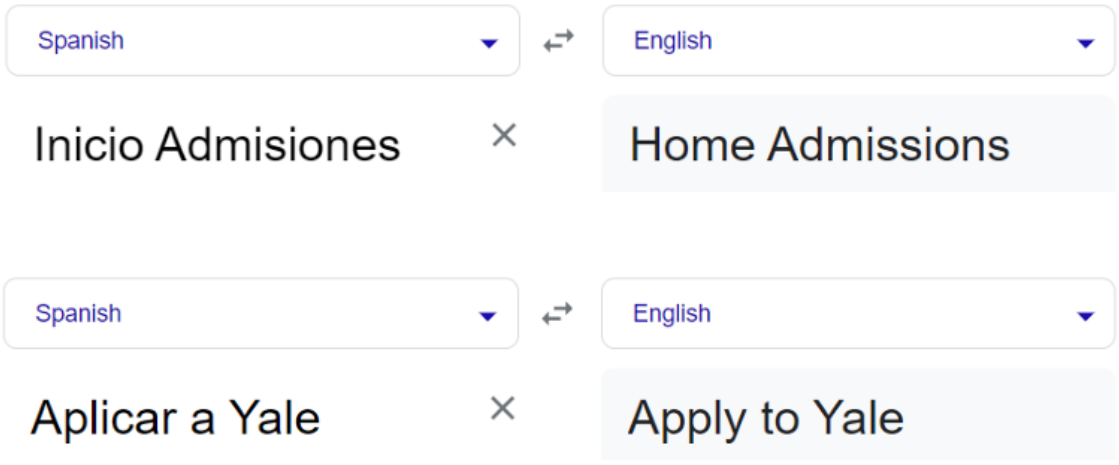
2.1. Data Collection

For the data collection, we have taken the sentences from the Yale University admissions page and have collected a lot of 136 samples. The data that we collected is as follows:

	A	B
1	**English**	**Spanish**
2	With 22% of its student body coming from abroad, Yale University offers a diverse and exciting global environment in which to study	Con el 22 % de su alumnado procedente del extranjero, la Universidad de Yale ofrece un entorno global diverso y emocion
3	Yaleâ€™s history of including international students is a long one, starting back in the 1800s.	La historia de Yale de incluir estudiantes internacionales es larga, ya que comenzÃ³ en el siglo XIX.
4	Today, Yale welcomes the largest international community in its history, with a current enrollment of 2,841 international students fr	
5	Applying to Yale	Aplicar a Yale
6	To be considered for acceptance to Yale, interested applicants must apply directly to the school, college, or program where the degr	Para ser considerado para la aceptaciÃ³n en Yale, los solicitantes interesados â€‹â€‹deben presentar su solicitud directam
7	yale College for undergraduate degrees; the Graduate School of Arts and Sciences for doctoral programs and some masterâ€™s deg	Yale College para tÃ-tulos universitarios; la Escuela de Graduados en Artes y Ciencias para programas de doctorado y algu
8	Each program has its own procedures for international applicants and for applying for financial assistance.	Cada programa tiene sus propios procedimientos para solicitantes internacionales y para solicitar asistencia financiera.
9	Graduate School of Arts & Sciences	Escuela de Graduados en Artes y Ciencias
10	Yale Summer Session	SesiÃ³n de verano de Yale
11	Information on programs of study, academic requirements, and financial aid are specific to each school.	La informaciÃ³n sobre programas de estudio, requisitos acadÃ©micos y ayuda financiera es especÃ-fica para cada escuela
12	Visit for information about doctoral programs	Visite para obtener informaciÃ³n sobre los programas de doctorado
13	Visit for information about Yale College and advice for applicants.	Visite para obtener informaciÃ³n sobre Yale College y consejos para los solicitantes.
14	Fellowships and Financial Assistance	Becas y asistencia financiera
15	Programs Not Granting Degrees	Programas que no otorgan tÃ-tulos
16	Yale offers significant financial assistance to international students to cover tuition costs as it does with students from the U.S.	
17	Fox International Fellowship	Fox International Fellowship
18	The Fox International Fellowship is a graduate student exchange program between Yale University and 21 world-renowned academi	Fox International Fellowship es un programa de intercambio de estudiantes graduados entre la Universidad de Yale y 21 sc
19	The Fox Fellowship is working to sustain and expand that global network.	Fox Fellowship estÃ¡ trabajando para sostener y expandir esa red global.
20	Open Yale Courses	Cursos abiertos de Yale
21	The aim of the project is to expand access to educational materials for all who wish to learn.	
22	Yale College credit are offered online through Yale Summer Online including OYC professors John Rogers and Craig Wright.	Los crÃ©ditos de Yale College se ofrecen en lÃ-nea a travÃ©s de Yale Summer Online, incluidos los profesores de OYC Joh
23	The aim of the project is to expand access to educational materials for all who wish to learn.	El objetivo del proyecto es ampliar el acceso a materiales educativos para todos los que deseen aprender.
24	Registration is not required	No es necesario registrarse
25	Jackson School of Global Affairs	Escuela Jackson de Asuntos Globales
26	The Jackson School of Global Affairs promotes education and scholarship on global affairs at Yale.	La Escuela Jackson de Asuntos Globales promueve la educaciÃ³n y la erudiciÃ³n sobre asuntos globales en Yale.

2.2. Annotation

We have used Google Translator as the annotation tool to translate the Spanish text to English.



2.3. Pre-Processing

We pre-processed the English and Spanish text data contained in a pandas DataFrame by cleaning the data using the functions defined and then splitting the cleaned data into lists of individual words. This pre-processing step is often necessary before using the data for natural language processing or machine learning tasks.

Then, the function addTokens(), takes in a list of tokens x and adds special start and end tokens to the beginning and end of the list, respectively. By adding these tokens to the beginning and end of a list of tokens, the model can be trained to recognize the start and end of a sentence or sequence during training and inference.

We then defined the vocab class, to create a vocabulary from a given dataset of sentences. It takes in two arguments - data and token. What happens here is that it computes the maximum length of the sentences in the dataset, iterates through each sentence in the dataset and adds words to the vocabulary if they are not already present, converts each sentence to a tensor where each word is replaced by its corresponding integer index, pads the tensor to the maximum length of the sentences in the dataset, appends the tensor to the x list. Once the

vocab class is instantiated, the x attribute is used as input for further processing, such as training a neural network language model or a machine translation model.

3. Model

Encoder: Here, we defined an LSTM encoder module in PyTorch. The module takes in an input sentence represented as a sequence of words and produces the hidden and cell states of the LSTM. The __init__ function initializes the embedding layer and the LSTM layer. The embedding layer converts the input tokens to a dense vector representation.

Decoder: The decoder module is used to decode the hidden state of the encoder module and generate the output sequence word by word. We use LSTM in the decoder module because it can handle variable-length sequences and can retain important information from previous time steps.

In the forward method of the decoder module, we take the input word, which is a single word and not the whole sentence like in the encoder module. We reshape the input word to have a shape of [batch_size, 1], which means we consider the input as a sequence of length 1. Then we pass it through an embedding layer and LSTM layer to generate the output sequence. The output of the LSTM layer is passed through a fully connected layer to obtain a probability distribution over the Spanish vocabulary.

SEQ2SEQ: The seq2seq model combines the encoder and decoder modules to perform machine translation from English to Spanish. The forward function takes as input a batch of English sentences and a batch of corresponding Spanish sentences, both of which are represented as sequences of word indices.

Next, we created a DataLoader object loader which will be used to load data in batches during the training process. Then comes, the model architecture and the optimization and loss functions to train the model. First, an encoder is initialized with the English vocabulary size, embedding size, hidden size, and the number of layers. Similarly, a decoder is initialized with the Spanish vocabulary size, embedding size, hidden size, and the number of layers. Then, a seq2seq model is created by passing the encoder and decoder to it. An Adam optimizer is initialized with the model parameters and a learning rate. A CrossEntropyLoss

criterion is initialized with the argument ignore_index=0, which indicates that the padding index should be ignored during the loss calculation.

Training the seq2seq model, the outer loop runs for several epochs, which is a hyperparameter that determines how many times the entire training set is processed. The inner loop uses the DataLoader object to load batches of input-output pairs from the training dataset. For each batch, the input x and target y sequences are converted to tensors and moved to the specified device (CPU or GPU) using the to() method. Then, the model object is used to generate predictions for the target sequence given the input sequence, using a teaching force of 1 (meaning that the model is always given the ground truth next word during training).

The output sequence is reshaped and used to compute the loss using the criterion object (which is a CrossEntropyLoss function with the ignore_index parameter set to 0, indicating that the padding token should be ignored). The loss is backpropagated through the model, and the optimizer is used to update the model parameters based on the gradients computed during backpropagation. Finally, the training loss for the epoch is printed and added to a list train_loss.

3.2. Experiments and Results

We have passed 3 learning rates i.e., 0.0001, 0.0005, and 0.001, and trained the model. We came to the conclusion 0.001 is the best learning rate

For the evaluation, we used the BLEU score and got an average of 0.82 approximately. Hence, the model is performing better and can be improved with more data.