English to French
NLP
Final project

Dataset : english-french.csv

Mia Koppen

WHAT IS NLP?

- **NLP stands for Natural Language Processing. It is an area of artificial intelligence/machine learning and computer science. It focuses on communication between a computer and its user. The goal of NLP is to get the computer to understand a human language through training, and then apply it through testing.**
  - **Examples are: Google Translate, virtual assistants, speech recognition software, Siri, Alexa, etc.**

- This is difficult to do because human language is complex. It can be hard for the computer to understand as there are many interpretations of speech.

Reference: https://builtin.com/data-science/introduction-nlp

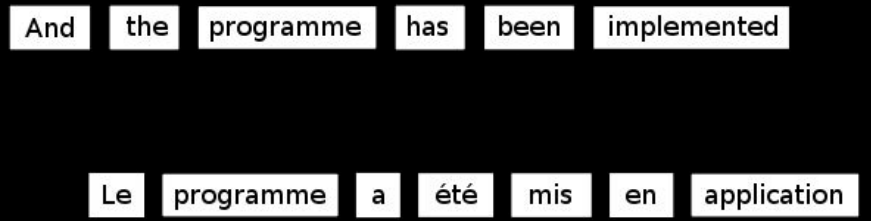
WHY IS NLP IMPORTANT?

- ★ **NLP is used for many different purposes. Most common are language translation and virtual assistants. The United Nations uses NLP for assemblies and conferences.**
  - **The United Nations Parallel Corpus, was,** "...created as part of the United Nations commitment to multilingualism and as a reaction to the growing importance of statistical machine translation (SMT) within the Department for General Assembly and Conference Management (DGACM)" (UN, 2024).
  - The purpose of the program was to use it to aid in research and language translation.
    - The corpus is posted on the United Nations website for all to use, as long as you acknowledge the UN as the source.

Reference: https://conferences.unite.un.org/uncorpus

HOW DOES NLP WORK?

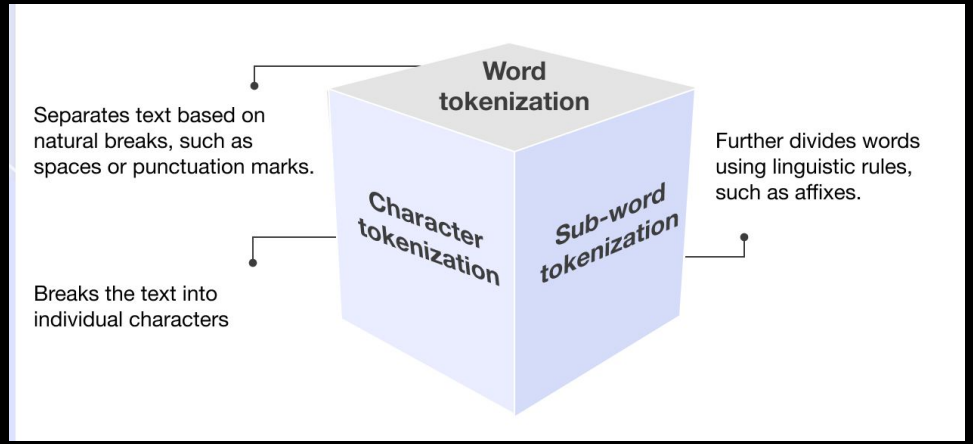
- ★ NLP takes text and transforms it into numbers/patterns that the computer will be able to understand. The text comes from a dataset. The computer is trained to associate certain words in one language with words in another language, and give that as output.
  - The more data you feed the machine, the more it learns, and the more accurate the machine's predictions will become.



References: https://monkeylearn.com/natural-language-processing/

WHAT IS TOKENIZATION?

★ **Tokenization is essential for NLP. If you want to create an NLP, you need to be able to break apart strings of words into tokens. Tokenization is needed because it takes data and presents it in a way that a computer can understand, and without losing context.**



https://www.simform.com/wp-content/uploads/2023/04/types-of-tokenization.png

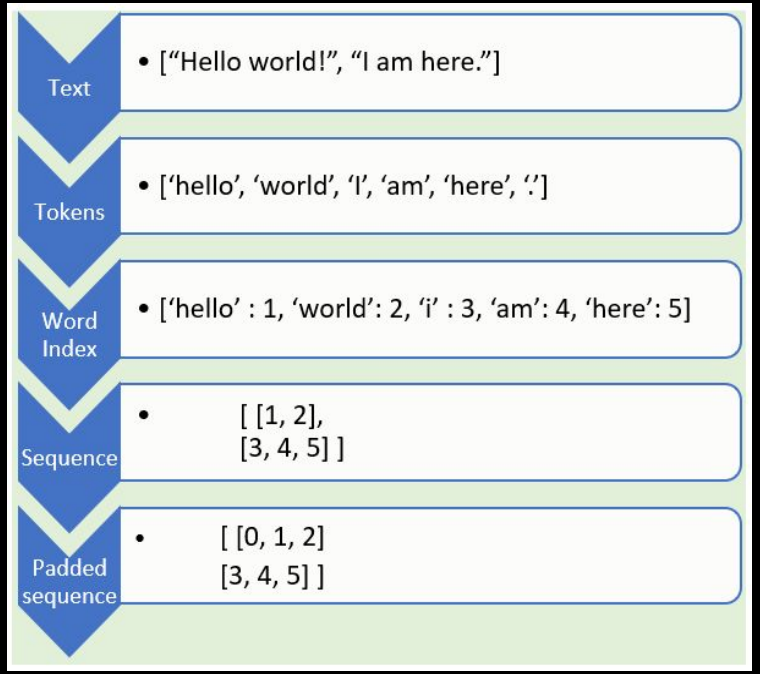
WHAT IS PADDING?

Padding is the process of making all sequences in a batch of text data the same length. Recurrent neural networks and transformers, require input sequences of uniform length.

1. Padding Tokens: Add a special padding token to shorter sequences until they reach the maximum length.

2. Sequence Length: Determine the maximum length of sequences in your dataset

https://miro.medium.com/v2/resize:fit:1218/1*zsIXWoN0_CE9PXzmY3tIjQ.png

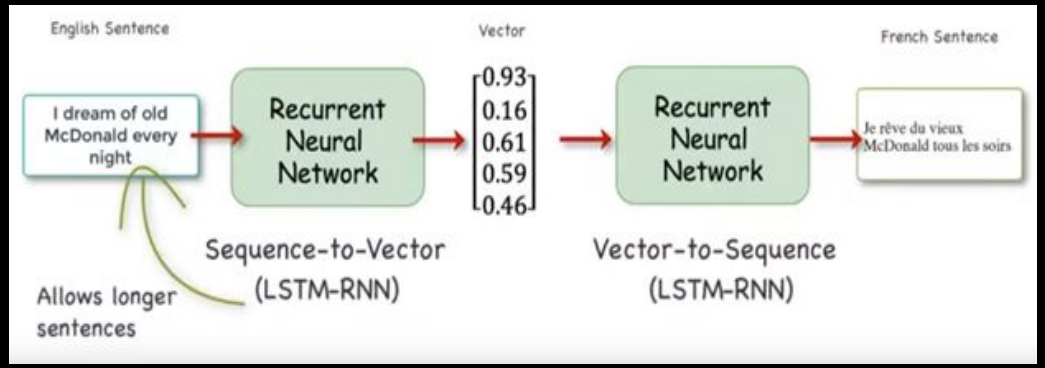


ENCODING/DECODING?

Encoding

Converting text into a numerical format that machine learning models can process.

1. **Text to Tokens:** Convert sentences to a list of tokens.
2. **Tokens to Integers:** Map each token to an integer.
3. **Integer Sequences to Vectors.**



Decoding

Converting numbers back into human language.

1. **Integers to Tokens.**
2. **Tokens to Text: Combine tokens to form the original sentence.**

The **Encoder-Decoder** architecture is a common framework for sequence-to-sequence tasks like translation. It consists of two parts:

Encoder

- ★ **Process Input Sequence:** Reads and encodes the entire input sequence into a fixed-size context vector.
- ★ **Hidden State:** The final hidden state of the encoder summarizes the input sequence.
- ★ **Decoder**
- ★ **Generate Output Sequence:** Uses the context vector to generate the target sequence one token at a time.
- ★ **Attention Mechanism:** Enhances the context vector by focusing on relevant parts of the input sequence at each decoding step.

WHAT EXACTLY IS MY PLAN?

- I ...
  - Will read the english-french csv file using pandas.
  - Use a sample size of 4,000 phrases and sentences (full size is 5,000)
  - Make all the letters in the dataset lowercase to make it easier to translate. Then tokenized all of the data. I then converted the sentences to sequences.
  - Pad the sequences.
  - Create a sequential model and trained it, then compile it and show its summary.
  - After, I will test the model using user input, and print it to the screen.

MY MODEL

I chose to use a Sequential model for my translation task.

About Sequential():

> Sequential models in NLP are designed to process data where the order of the elements is crucial, such as sentences in language. These models handle sequences of words or tokens in a way that captures the dependencies and structures inherent in natural language

OTHER OPTIONS FOR MODELING

Model	Description	Pros	Cons	Use Cases
Sequential (RNN, LSTM, GRU)	Models designed for sequential data processing	- Captures temporal dependencies - Good for time-series data	- Training can be slow - Prone to vanishing gradient problem	- Language modeling - Sequence prediction
Linear Regression	A basic regression model for predicting a continuous outcome	- Simple to implement - Easy to interpret results	- Assumes linearity - Not suitable for complex relationships	- Sentiment analysis (basic) - Text regression
Decision Trees	Tree-like model of decisions	- Easy to interpret - Handles non-linear relationships	- Prone to overfitting - Not well-suited for large datasets	- Text classification (basic)
Random Forest	Ensemble of decision trees	- Reduces overfitting - Robust to outliers	- Can be slow for large datasets - Less interpretable	- Text classification - Entity recognition
Support Vector Machines (SVM)	Model for classification and regression	- Effective in high-dimensional spaces - Robust to overfitting	- Computationally expensive - Requires careful tuning	- Text classification - Sentiment analysis
Naive Bayes	Probabilistic classifier based on Bayes' theorem	- Simple and fast - Works well with small datasets	- Assumes feature independence - Limited accuracy for complex tasks	- Spam filtering - Document classification
Logistic Regression	Regression model for binary classification	- Simple to implement - Good for binary outcomes	- Assumes linearity - Not suitable for complex relationships	- Binary text classification - Sentiment analysis (binary)
K-Nearest Neighbors (KNN)	Instance-based learning method	- Simple to implement - No training phase required	- Computationally expensive - Sensitive to irrelevant features	- Text classification - Document retrieval
Neural Networks (DNN)	Deep learning models with multiple layers	- Can model complex relationships - Highly flexible	- Requires large datasets - Computationally intensive	- Text generation - Machine translation
Transformers (BERT, GPT, T5)	Advanced neural network architecture with attention mechanism	- Handles context well - State-of-the-art performance	- Requires extensive computational resources - Complex to implement	- Language modeling - Text generation - Question answering

	A	B	C	I
1	English wo	French words/sentences		
2	Hi.	Salut!		
3	Run!	Coursâ€¯!		
4	Run!	Courezâ€¯!		
5	Who?	Qui ?		
6	Wow!	Ã‡a alorsâ€¯!		
7	Fire!	Au feu !		
8	Help!	Ã€ l'aideâ€¯!		
9	Jump.	Saute.		
10	Stop!	Ã‡a suffitâ€¯!		
11	Stop!	Stopâ€¯!		
12	Stop!	ArrÃªte-toi !		
13	Wait!	Attends !		
14	Wait!	Attendez !		
15	Go on.	Poursuis.		
16	Go on.	Continuez.		
17	Go on.	Poursuivez.		
18	Hello!	Bonjour !		

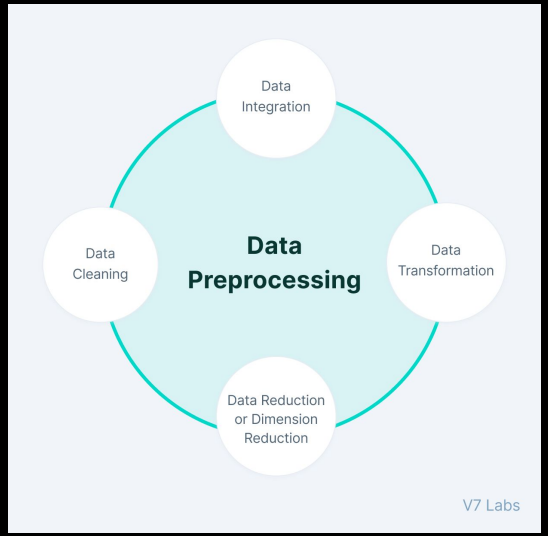
THE DATASET

- ★ This csv file contains a dataset, which contains two columns, English words and sentences, and French words and sentences.
- ★ I can feed my computer this data to train it. Then it will be able to make predictions based on the data given.

https://www.kaggle.com/datasets/devicharith/language-translation-englishfrench

WHAT IS PRE-PROCESSING?

- ★ Preprocessing is the process of identifying and fixing errors in a set of data. It can be. You are preparing the data for analysis.
  - Reading the data.
  - Removing duplicates or filling null values.
  - Tokenization or encoding.
  - Splitting the data into train and test groups.
  - etc...



Reference: https://cdn.prod.website-files.com/5d7b77b063a9066d83e1209c/6137488ee132f03deba81f9f_data-preprocessing.png

MY PRE-PROCESS IN DEPTH

- ★ The first step I took to pre-process my data after creating a pandas dataframe was, to take both columns and set every character to lowercase to make it easier for the model to translate.
- ★ Next I created instances of Tokenizer...
- ★ Then I converted the text (data) into sequences of numbers, so that the model could make predictions based on the data, and each sequence of numbers coded for a different word.

```
English sequences sample: [[21, 179, 388, 51, 977], [39, 46, 1432, 78, 245, 231],
French sequences sample: [[76, 147, 320, 46, 60, 1096], [52, 80, 77, 1764, 78, 22
```

PROBLEMS WITH THE DATASET

- some languages, like French have many variations of how a verb or noun looks depending on the person using it. Sometimes, the same phrase or word will be present in the dataset multiple times.

```
Go now.,"Va, maintenant."
Go now.,Allez-y maintenant.
Go now.,Vas-y maintenant.
```

- Because I tried to create an NLP model, and used a data set containing words and sentences instead of values, it was harder to figure out how to train the data in a way that would work.
    - In the future, I might consider choosing a different project for a final project, as this was probably not the best for a near beginner.

VISUALIZATION OF THE MODEL

```
✓ model.summary() ···

Model: "sequential_2"
```

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 32, 128)	376,576
lstm_4 (LSTM)	(None, 128)	131,584
repeat_vector_2 (RepeatVector)	(None, 31, 128)	0
lstm_5 (LSTM)	(None, 31, 128)	131,584
time_distributed_2 (TimeDistributed)	(None, 31, 4428)	571,212

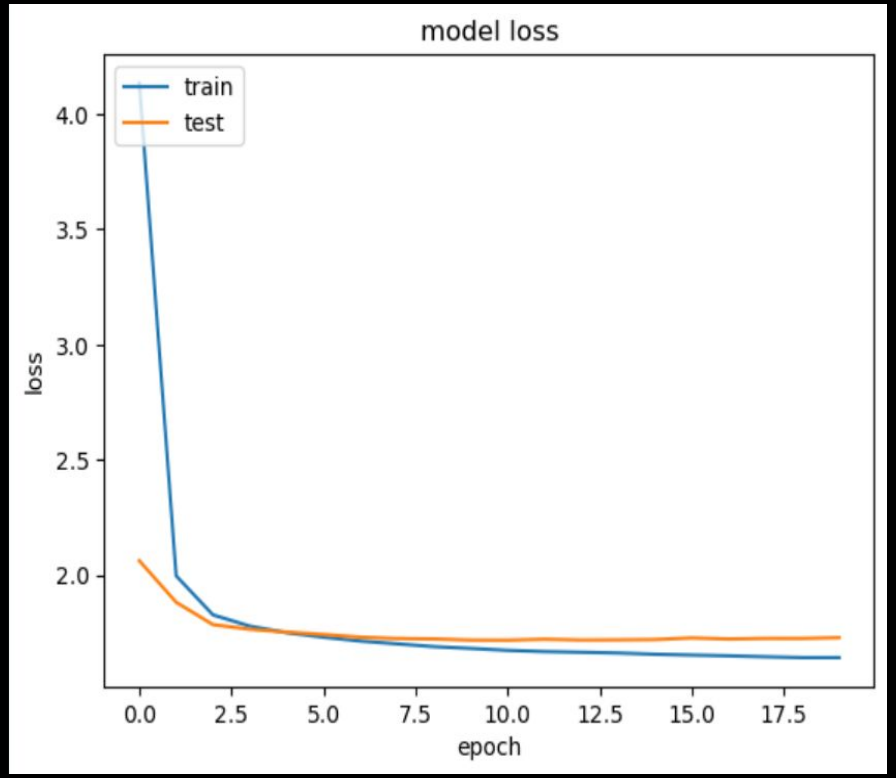
```
Total params: 1,210,956 (4.62 MB)

Trainable params: 1,210,956 (4.62 MB)

Non-trainable params: 0 (0.00 B)
```

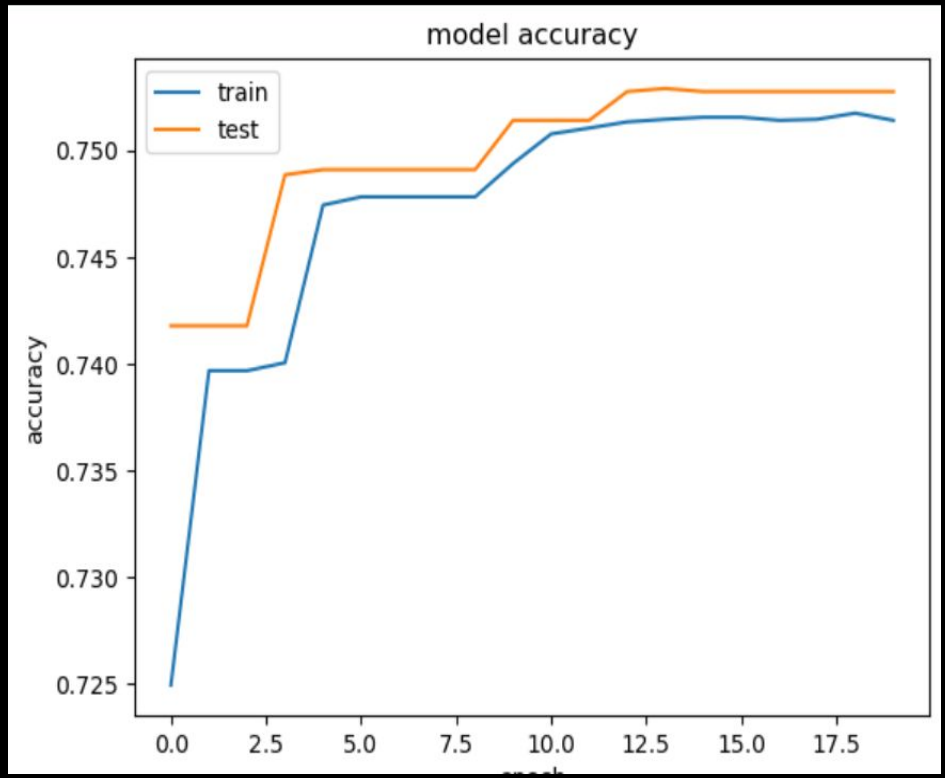
★ This is a summary of the model after compiling it.....

- It shows all of the layers and their names, the shape for each layer, and the weight of each parameter.
- Here, I called it to make sure the parameters were correct.

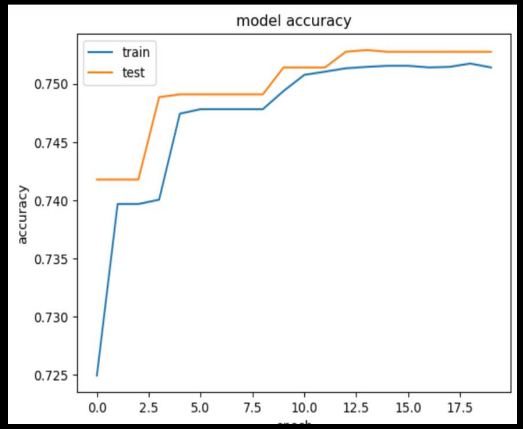
Reference:
https://www.tensorflow.org/js/guide/models_and_layers#:~:text=Model%20summary,-Call%20model.&text=summary()%20to%20print%20a,weight%20parameters%20of%20each%20layer.

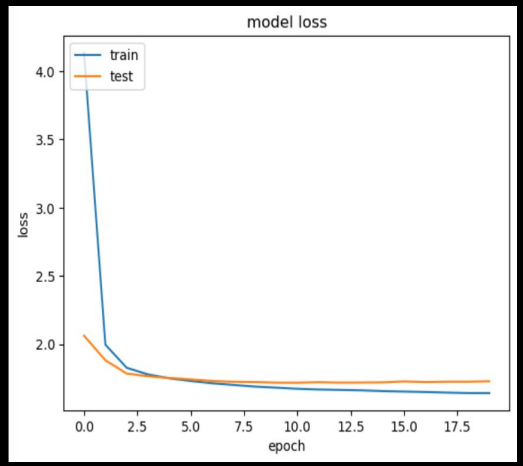
VISUALIZATION OF THE MODEL





WHAT DOES THE VIS. SHOW?





- ★ The plot image on the top shows the percentage of predictions/assumptions that my model got right.
  - The higher the percentage, the better your model is.
- ★ The plot image on the bottom shows the model loss. It measures how off your model behaves when going through optimization.
  - The lower the loss, the better your model is.
  - Sometimes when trying to reduce the loss values, you overfit your model, meaning you feed it so much that it becomes sort of ineffective.

AFTER THE MODEL...

★ After i'm done creating the model, I will then create functions to

1. Pre-process the data.

```
def preprocess_input(sentence, english_tokenizer, max_eng_len):
```

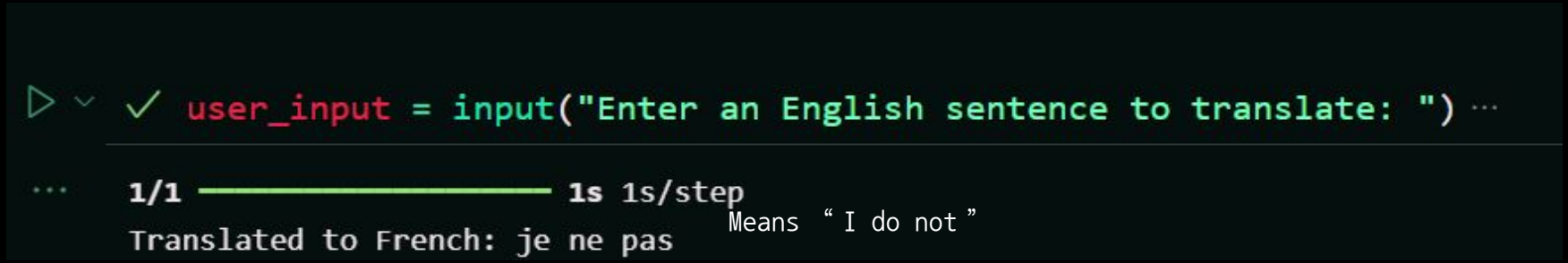
2. Decode the sequences during translation.

```
def decode_sequence(prediction, tokenizer):
```

3. Translate the user input from the keyboard.

```
def translate_user_input(user_input, model, english_tokenizer, french_tokenizer, max_eng_len, max_fr_len):
```

THE RESULT



This is what happens when you run the program...

The translator works, but...

1. the information/the translation is not correct
2. Sometimes, you get repeating words for output.

MORE EXAMPLES

```
Why did the chicken cross the road
Enter an English sentence to translate: (Press 'Enter' to confirm or 'Escape' to cancel)
1/1 ━━━━━━━━━━━━━━━━━━━━ 2s 2s/step
Translated to French: je ne pas
```

English: I do not

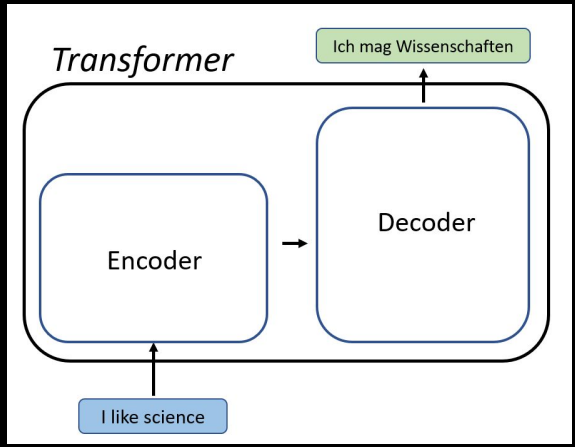
```
i will go to the store
Enter an English sentence to translate: (Press 'Enter' to confirm or 'Escape' to cancel)
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 1s/step
Translated to French: je je
```

English: I I

WHAT WOULD I DO NEXT TIME

I would create a different model, not using Sequential(), and I would chose to create a Transformer model instead. Transformer models can process language, generate text, and be trained to understand context.

For Example: Chat GPT



https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Ftransformers-an-overview-of-the-most-novel-ai-architecture-cdd7961eef84&psig=AOvVaw0xpwYYrpkrVFTbbuXro_CJ&ust=1718825158482000&source=images&cd=vfe&opi=89978449&ved=0CBEQjRxqFwoTCIiR9vmL5oYDFQAAAAAdAAAAABA5