English to Arabic machine translation model

**Omar Adly Mahmoud 20201446550**

**Mohamed Ahmed Mohamed Eldokm 20201495711**

**Omar Hossam Abdulrazak 20201374189**

**Mazen Mohamed Elsayed 20201312620**

Our main goal is to make a **machine translation model** which translates from English to **Arabic** and English to **French** by using

Seq2seq model bidirectional **GRU** (Gated Recurrent Unit) architecture with TimeDistributed layers

**1) installing the dependencies:**

First installing all the libraries that will be used to help the model

We will be using TensorFlow and nltk for the model and the word translation system itself

(TensorFlow is an open-source machine learning library developed by the Google Brain team. It is

designed to facilitate the development and training of machine learning models, especially deep

neural networks. TensorFlow provides a comprehensive set of tools and abstractions for building and

deploying various machine learning applications)

(NLTK, or Natural Language Toolkit, is a powerful Python library for working with human language

data. It provides easy-to-use interfaces to over 50 corpora and lexical resources, such as WordNet,

along with a suite of text processing libraries for classification, tokenization, stemming, tagging,

parsing, and semantic reasoning)

**2) Preprocess Data:**

Downloaded the dataset for the English and Arabic words install it in our code and then shuffle it to make the

training phase as random as we can so the model doesn’t overfit

• Text Understanding:

• Word-Level Analysis:

• Feature Extraction:

• Text Cleaning

We added an extra function for padding that will help in making all the data be with the same length so it will

be easily combined in batches

For getting much features from the data we will have to get data with high values this means we will have to

keep data with meaning and delete any garbage from the data like non alphabetic characters and the stop

words because the machine mainly doesn’t have to understand them for its training

So with that being said we created a function to do that and here is the regular expression for it

re.sub(r'[^a-zA-Z\s]'

re.sub(r'[^a-zÀ-ÿ\s]' remove special characters, punctuation, and numbers in french

Then we adjusted the length of the sentence to be fixed for Memory Efficiency and to avoid data loss during training

**Eng to Arab :**

Max English sentence length: 10

Max Arabic sentence length: 10

English vocabulary size: 26062

Arabic vocabulary size: 57846

**Eng to French :**

Max English sentence length: 10

Max French sentence length: 10

English vocabulary size: 13946

French vocabulary size: 30769

The to\_text function is a utility function for converting model predictions (logits) back into humanreadable text using a provided tokenizer

**3) Initializing the sequential model:**

The model takes the input shape and the output shape\_sequence\_length , the English vocab size and

the Arabic vocab size

Putting the learning rate with 0.003

For the sequential layers we added

A Gated Recurrent Unit (GRU) -is a type of recurrent neural network (RNN) architecture that was

introduced to address some of the limitations of traditional vanilla RNNs. GRUs belong to the family

of neural network architectures designed for sequence modeling and processing, where the input and

output have an ordered structure, such as in time series or natural language-

Adds an embedding layer with english\_vocab\_size as the input dimension, embedding dimension of

256, and input length specified by input shape

**Bidirectional**

Adds a bidirectional GRU layer with 256 units. Bidirectional layers process the input sequence in both

forward and backward directions.

TimeDistributed

Adds a dense layer with 1024 units and ReLU activation. The TimeDistributed wrapper applies the

dense layer to each time step of the input sequence independently.

**Dropout**

Adds a dropout layer with a dropout rate of 0.5. Dropout helps prevent overfitting by randomly

setting a fraction of input units to zero during training

TimeDistributed(Dense (arabic\_vocab\_size, activation='SoftMax'))

Adds another time-distributed dense layer with arabic\_vocab\_size units and SoftMax activation. This

is often used for sequence-to-sequence tasks, where each time step outputs a probability distribution

over the Arabic vocabulary

This model defines a sequence-to-sequence model for machine translation. **The bidirectional GRU**

layer allows the model to capture contextual information from both past and future tokens in the

input sequence. The time-distributed dense layers enable the model to produce a sequence of

outputs at each time step. The model is trained using the sparse categorical crossentropy loss and the

Adam optimizer

Layer (type) Output Shape Param #

=================================================================

embedding (Embedding) (None, 10, 256) 6672128

bidirectional (Bidirection (None, 10, 512) 789504

al)

time\_distributed (TimeDist (None, 10, 1024) 525312

ributed)

dropout (Dropout) (None, 10, 1024) 0

time\_distributed\_1 (TimeDi (None, 10, 57847) 59293175

stributed)

=================================================================

Total params: 67280119 (256.65 MB)

Trainable params: 67280119 (256.65 MB)

Non-trainable params: 0 (0.00 Byte)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For the batch size its 64 and 50 epochs

**4) Testing the model**

The accuracy was 40%

When testing the model, it gave us good predictions

Prediction:

1/1 [==============================] - 0s 102ms/step

0 0 0 0 0 0 0 0 كاتبا قابلنا

Correct Translation:

قابلنا كاتبا .

Original text:

We met a writer.

Enter an English sentence: i am from egypt

1/1 [==============================] - 0s 30ms/step

Translated Arabic sentence: 0 0 0 0 0 0 0 مصر من أنا

Our real issue is that Arabic is one of the hardest languages that’s why the accuracy isn't the best .

Prediction:

1/1 [==============================] - 0s 72ms/step

ils semblent toutes deux mefiantes pad pad pad pad pad

Correct Translation:

ils semblent tous deux suspects

Original text:

both of them seem suspicious

**F1Score**

Eng to Arab:

A screenshot of a computer

Description automatically generated

Eng to French :

Eng to Arab:

A graph of a curve

Description automatically generated with medium confidenceA graph of a line

Description automatically generated with medium confidence

Eng to French : A graph showing the performance of a performance

Description automatically generatedA line graph showing the loss of a dog

Description automatically generated

**5) references**

https://www.tensorflow.org/guide/keras/sequential\_model

https://www.nltk.org/

https://www.kaggle.com/datasets/samirmoustafa/arabic-to-englishtranslation-sentences

https://www.geeksforgeeks.org/seq2seq-model-in-machinelearning/

https://sh-tsang.medium.com/review-seq2seq-sequence-tosequence-learning-with-neural-networks-bcb84071a670