For this homework, I used a LSTM model that can be found with this link.

Configuration

```
batch_size = 64  # Batch size for training.
epochs = 3  # Number of epochs to train for.
latent_dim = 256  # Latent dimensionality of the encoding space.
num_samples = 3000  # Number of samples to train on.
# Path to the data txt file on disk.
data_path = "jsonDataset.txt"
```

Tokenization

I used the JSON format to parse the file and get the sentences one by one from the translation object until the sample number.

```
input_texts = []
target_texts = []
input_characters = set()
target_characters = set()
data_file = open(data_path, encoding="utf-8")
dataset = json.loads(data_file.read())
data_file.close()
for i in range(0, min(num_samples, len(dataset["dataset"]))):
  input_text = dataset["dataset"][i]["translation"]["ot"]
  target_text = dataset["dataset"][i]["translation"]["tr"]
  target_text = "\t" + target_text + "\n"
  input_texts.append(input_text)
  target_texts.append(target_text)
  for char in input_text:
      if char not in input_characters:
          input_characters.add(char)
  for char in target_text:
      if char not in target_characters:
          target_characters.add(char)
```

Training

I used 3 epochs and 3000 samples for the minimalization of training period.

```
Epoch 1/3
38/38 [==============================] - 300s 8s/step - loss: 0.8553 - accuracy: 0.8596 - val_loss: 0.8679 - val_accuracy: 0.8774
Epoch 2/3
38/38 [==============================] - 301s 8s/step - loss: 0.6166 - accuracy: 0.8602 - val_loss: 0.5184 - val_accuracy: 0.8824
Epoch 3/3
38/38 [==============================] - 295s 8s/step - loss: 0.4470 - accuracy: 0.8872 - val_loss: 0.4269 - val_accuracy: 0.8823
```

BLEU Score

I got BLEU score of 0. I think it's because of low sample number.

```
Input sentence: senesi mayısının 19 uncu günü samsuna çıktım .
Decoded sentence: eeee
Reference sentence:     senesi mayisinin 19 uncu günü samsuna çiktim .

BLEU score: 0
```

Burak Yıldırım
1901042609