Natural Language Processing
CSE 454

Hw#4

Ömer Faruk Akduman
1801042094

Content

Homework Concept

In this homework, I will develop a simple translator from Ottoman Turkish to modern day Turkish. Provided training set at the class Teams page. As an example if the input is “itilaf devletleri , mütareke ahkamına riayete lüzum görmüyorlar” the translated output will be “itilaf devletleri , ateşkes hükümlerine uymayı gerekli görmüyorlar ”.

Data Set

The name of the dataset file I used is jsonDataset.txt. This file contains information like this

{"dataset" : [{"translation":{"ot":"senesi mayısının 19 uncu günü samsuna çıktım .","tr":"senesi mayisinin 19 uncu günü samsuna çiktim ."}},…

“ot” means in Ottoman language.

“tr” means in Turkish language.

Data Preprocessing

Cleaning Data and remove unwanted characters from the sentences.

To make more standardize I remove Turkish characters and make string clean

```
#cleaning data and removes umwanted characters
def str_cleaning(a_str):
  a_str=a_str.lower()
  a_str = a_str.replace("ı","i").replace("ü","u").replace(".","").replace("ş","s").replace("ç","c").
  result = ''.join(c for c in a_str if c.isalpha() or c == " ")
  return re.sub(' +', ' ', result).rstrip()
```

At the end of the Data Preprocessing some samples

	Turkish	Ottoman
38451	eller agza gitti	eller agza gitti
8813	tut ki uzattim	tut ki uzattim
4513	belge	vesika
18006	oraclakilere soze bakin clemis	huzzara demis gorun kelami
31111	onun dostu evangeliyayi sirkecide bir gazinoda...	onun dostu evangeliyayi sirkecide bir gazinoda...
37647	prenses senetlerin arif beye goturulup goturul...	prenses senetlerin arif beye goturulup goturul...
25265	ruhun bazi anlasilmaz ifade edilmez sikâyetler...	ruhun bazi anlasilmaz ifade edilmez sikâyetler...
21236	ayagindan ipi gevsetmeyi akil etmez o da	ayagindan ipi gevsetmeyi akletmez o da
11833	elbette boyle bir iliskiye girismek kesin olar...	bittabi boyle bir munasebete girismek katiyyen...
16706	salih pasa bu durumu biliyor ve buna bilerek n...	salih pasa bu vaziyeti bilerek ve bu hale bile...

Encode and Tokenize

I tokenize sentences and use one hot encode to encode.

```
[ ]
    def create_tokenizer(lines):
        # fit a tokenizer
        tokenizer = Tokenizer()
        tokenizer.fit_on_texts(lines)
        return tokenizer

    def max_len(lines):
        # max sentence length
        return max(len(line.split()) for line in lines)

    def encode_sequences(tokenizer, length, lines):
        # encode and pad sequences
        X = tokenizer.texts_to_sequences(lines) # integer encode sequences
        X = pad_sequences(X, maxlen=length, padding='post') # pad sequences with 0 values
        return X

    def encode_output(sequences, vocab_size):
        # one hot encode target sequence
        ylist = list()
        for sequence in sequences:
            encoded = to_categorical(sequence, num_classes=vocab_size)
            ylist.append(encoded)
        y = np.array(ylist)
        y = y.reshape(sequences.shape[0], sequences.shape[1], vocab_size)
        return y
```

Splitting Data

I divided the dataset into 2 parts, train and test.

```
    # Prepare training data
    trainX = encode_sequences(src_tokenizer, src_length, train[:, idx_src])
    trainY = encode_sequences(tar_tokenizer, tar_length, train[:, idx_tar])
    trainY = encode_output(trainY, tar_vocab_size)

    # Prepare test data
    testX = encode_sequences(src_tokenizer, src_length, test[:, idx_src])
    testY = encode_sequences(tar_tokenizer, tar_length, test[:, idx_tar])
    testY = encode_output(testY, tar_vocab_size)
```

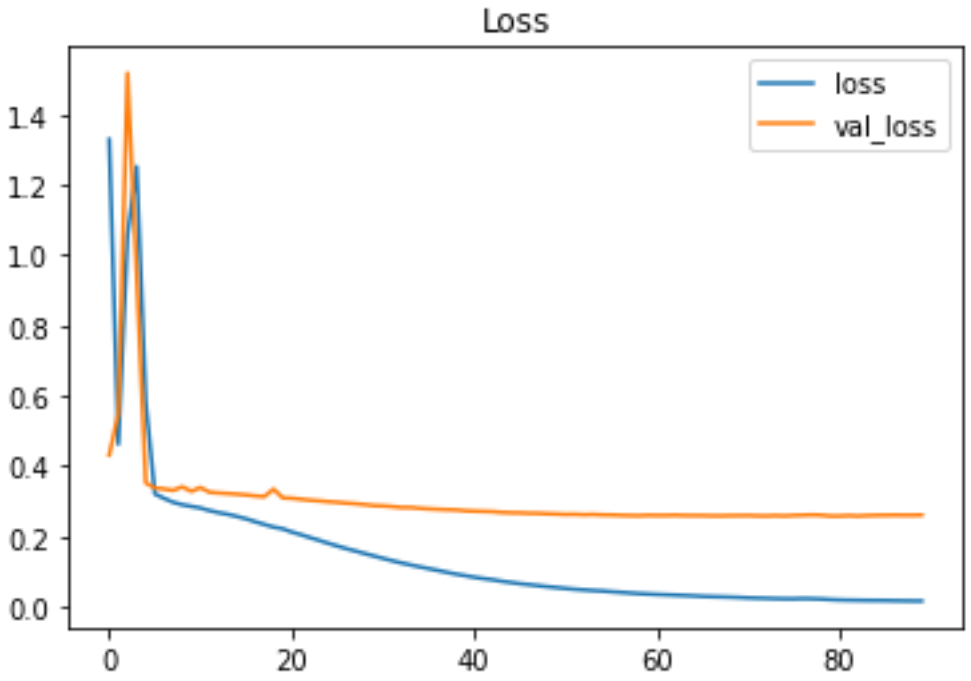
Model

I used LSTM model also used some features like Early stopping, due to *model training time "approximately it takes 8 hours for one model in colab"* I only train 3different model and put here best one with these features.

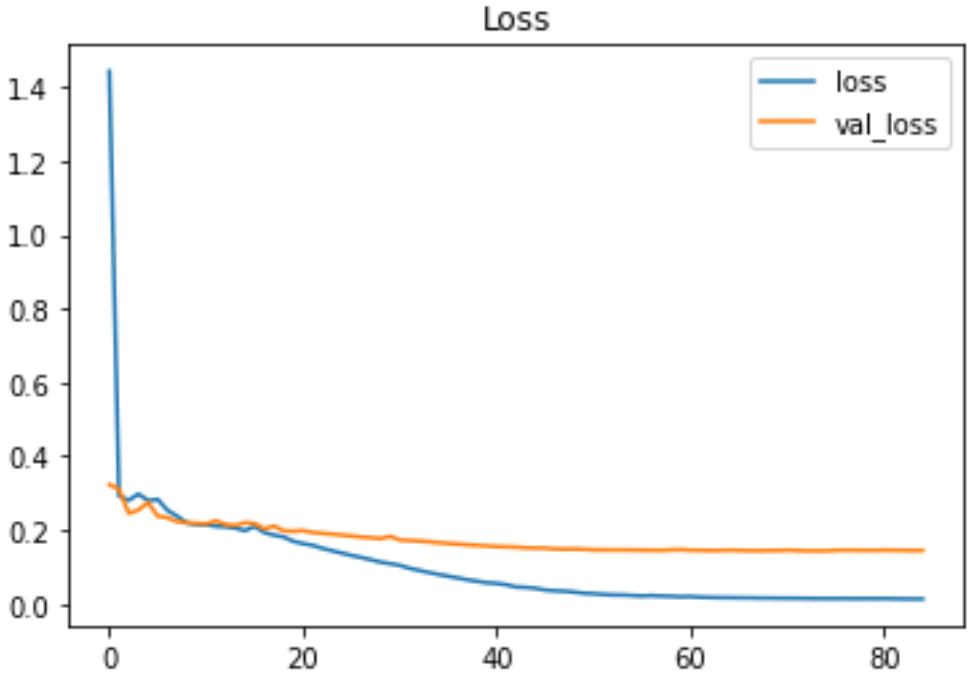
```
[ ]  def create_model(src_vocab, tar_vocab, src_timesteps, tar_timesteps, n_units):
         # Create the model
         model = Sequential()
         model.add(Embedding(src_vocab, n_units, input_length=src_timesteps, mask_zero=True))
         model.add(LSTM(n_units))
         model.add(RepeatVector(tar_timesteps))
         model.add(LSTM(n_units, return_sequences=True))
         model.add(TimeDistributed(Dense(tar_vocab, activation='softmax')))
         return model

     # Create model
     model = create_model(src_vocab_size, tar_vocab_size, src_length, tar_length, 256)
     model.compile(optimizer='adam', loss='categorical_crossentropy')

     history = model.fit(trainX,
               trainY,
               epochs=200,
               batch_size=64,
               validation_split=0.1,
               verbose=1,
               callbacks=[
                           EarlyStopping(
                           monitor='val_loss',
                           patience=10,
                           restore_best_weights=True
                       )
                 ])
```

Results

With at most 40 characters sentences



With at most 30 characters sentences



Result on the Training Set

Result on the Training Set

OTTOMAN (SOURCE)	TURKISH (TARGET)	AUTOMATIC TRANSLATION IN TURKISH
birkac bez parcasi cikardilar	birkac bez parcasi cikardilar	birkac bez parcasi cikardilar
at basindan belayi	at basindan belayi	at basindan belayi
nerede	nerede	nerede
birini sevmez mi	birini sevmez mi	birini sevmez mi
gece evime gittim	gece evime gittim	gece evime gittim
fayda cikmadi	fayda cikmadi	fayda cikmadi
kalbi kanaya kanaya yandi	kalbi kanaya kanaya yandi	kalbi kanaya kanaya yandi
sana para lazim	sana para gerek	sana para gerek
yazihanede	yazihanede	yazihanede
bir daha caldi	bir daha caldi	bir daha caldi
taniyamadi ama suphelendi	taniyamadi ama suphelendi	taniyamadi ama suphelendi
nicin gitmemeli	nicin gitmemeli	nicin gitmemeli
ne nezahet bu hocam hayranim	ne incelik bu hocam hayranim	ne incelik bu hocam hayranim
beye bir sey soyler diye	beye bir sey soyler diye	kapici bir kadar soyler
gelsin a be	gelsin a be	bey bey mu
cok oturacak misiniz	cok oturacak misiniz	cok oturacak misiniz
ona guzel diyor	ona guzel diyor	ona bey ki
geceleri galiba bekci geziyor	geceleri galiba bekci geziyor	geceleri galiba bekci geziyor
bana sadece margarita deyin	bana sadece margarita deyin	bana sadece margarita deyin
y rab o ne encumen ne alem	ya rab o ne meclistir ne cm	ya rab o ne meclistir ne cm
hangi arkadasindaydin	hangi arkadasindaydin	hangi arkadasindaydin
nihayet	nihayet	nihayet
ben nazim beyi kabul etmedim	ben nazim beyi kabul etmedim	ben nazim beyi kabul etmedim
daha neler	daha neler	daha neler
yerden goge kadar hak etmisti	yerden goge kadar hak etmisti	yerden goge kadar hak etmisti
gunler geciyor	gunler geciyor	gunler geciyor

Result on the Test Set

Result on the Test Set

OTTOMAN (SOURCE)	TURKISH (TARGET)	AUTOMATIC TRANSLATION IN TURKISH
yalnizliktan canim sikiliyor	yalnizliktan canim sikiliyor	canim canim
ihtiyarladim hanimefendi	ihtiyarladim hanimefendi	hanimefendi
alisverisin ne	alisverisin ne	ne
daha neler daha neler	daha neler daha neler	daha daha zavalli neler
hukumeti de ayni halde	hukumeti de ayni durumda	hukumeti de durumda durumda
siz nereden biliyorsunuz	siz nereden biliyorsunuz	pasa miyiz biliyorsunuz
hangi siir	hangi siir	hangi siir
bir degil	bir degil	bir degil
haklisin teyze	haklisin teyze	haklisin teyze
silahimiz var midir	silahimiz var midir	olur olur olur
erzurumdan k o	erzurumdan k o	erzurumdan k o
kalirsiniz oyle mi	kalirsiniz oyle mi	oyle mi
kesfettigim gibi	kesfettigim gibi	saika gibi
bu sabah annesi geldi	bu sabah annesi geldi	bu bu ankaraya yaptim
yazan biraderimdir	yazan agabeyimdir	is
evet	evet	evet
acaba hata mi ettim	acaba yanlis mi yaptim	acaba yanlis mi yaptim
aman sor bana merak oldu	aman sor bana merak oldu	onu onu zavalli oldu oldu
hafta icinde biter	hafta icinde biter	kadin icinde icinde
salincakci	salincakci	salincakci
sirri sezai korkak korkak	sirri sezai korkak korkak	sirri sezai gogus gecirdi
telasla haykirdilar	telasla haykirdilar	saika ve
tatmin edilmege muhtactir	yatistirilmasi gerekir	ikinci tanidiklardan seviyor
melahat burhanettini sevmiyor	melahat burhanettini sevmiyor	siz da seviyor
kaynatmis	kaynatmis	bakistilar
dervis ahmed	dervis ahmed	dervis ahmed

Blue Score

Blue Score function

```python
def bleu_score(model, tokenizer, sources, raw_dataset):
    actual, predicted = [], []
    for i, source in enumerate(sources):
        source = source.reshape((1, source.shape[0]))
        translation = predict_seq(model, tar_tokenizer, source)
        raw_target, raw_src = raw_dataset[i]
        actual.append([raw_target.split()])
        predicted.append(translation.split())

    bleu_dic = {}
    bleu_dic['1-grams'] = corpus_bleu(actual, predicted, weights=(1.0, 0, 0, 0))
    bleu_dic['1-2-grams'] = corpus_bleu(actual, predicted, weights=(0.5, 0.5, 0, 0))
    bleu_dic['1-3-grams'] = corpus_bleu(actual, predicted, weights=(0.3, 0.3, 0.3, 0))
    bleu_dic['1-4-grams'] = corpus_bleu(actual, predicted, weights=(0.25, 0.25, 0.25, 0.25))

    return bleu_dic

bleu_train = bleu_score(model, tar_tokenizer, trainX, train)
bleu_test = bleu_score(model, tar_tokenizer, testX, test)
```

Blue Score Results

```
[27] bleu_train

{'1-grams': 0.9063629236970987,
 '1-2-grams': 0.8653306887308325,
 '1-3-grams': 0.8023825817067682,
 '1-4-grams': 0.6401645192328522}

bleu_test

{'1-grams': 0.4469939190143961,
 '1-2-grams': 0.35591626789313996,
 '1-3-grams': 0.3279145847813834,
 '1-4-grams': 0.22312765673187918}
```

Resources

1- GitHub
2- www.spicework.com
3- Stackoverflow
4- Wikipedia
5- kaggle.com
6- chatgpt
7- https://towardsdatascience.com/