About:

- In this notebook, I have processed the files needed for me to train the model which will be taking the sentence in a left to right fashion and output will also be from left to right.

For example:

Input: i like dance

Output: i like to dance

In [1]:

```
import os
os.environ['TF_CPP_MIN_LOG_LEVEL'] = '3'
```

In [49]:

```
import matplotlib.pyplot as plt
%matplotlib inline
# import seaborn as sns
import pandas as pd
import re
import tensorflow as tf
from tensorflow.keras.layers import Embedding, LSTM, Dense
from tensorflow.keras.models import Model
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
import numpy as np
import seaborn as sns
```

In [50]:

```
data=pd.read_csv('processed_sentence_pairs_unique.csv')
```

In [51]:

```
data.drop('are_same', axis=1,inplace=True)
data.head()
```

Out[51]:

	incorrect	correct
0	and he took in my favorite subject like soccer	and he took in my favorite subjects like soccer
1	actually who let me know about lang was him	actually he was the one who let me know about ...
2	his kanji is ability is much better than me	his kanji ability is much better than mine
3	we have known each other for only half a year ...	we have known each other for only half a year ...
4	i heard a sentence last night when i watched tv	i heard a sentence last night when i was watch...

In [52]:

```
data.shape
```

Out[52]:

```
(495873, 2)
```

Note:

- From the EDA we can see that both the Correct & Incorrect sentence have a maximum sentence length of 68 after we have removed the outliers. But we do not have the computational resource to work on these kind of sentences, hence we are limiting the sentence length to 16, and then we will be working on them.

In [53]:

```
data['length']=data['correct'].astype(str).apply(lambda i:len(i.split(' ')))

data=data[data['length']<=16]

data.drop('length',axis=1, inplace=True)
```

Splitting the Data

In [54]:

```
from sklearn.model_selection import train_test_split
train_1, test = train_test_split(data, test_size=0.2)
train, validation = train_test_split(train_1, test_size=0.2)
```

In [55]:

```
import pickle

pickle.dump([train,test, validation],open('main_data_1.pkl','wb'))
```

In [56]:

```
train['correct_inp'] = '<start> ' + train['correct'].astype(str)
train['correct_out'] = train['correct'].astype(str) + ' <end>'

train = train.drop(['correct'], axis=1)
# only for the first sentance add a toke <end> so that we will have <end> in tokenizer
train.head()
```

Out[56]:

	incorrect	correct_inp	correct_out
337916	but there also are many competitions between them	<start> but there are also many competitions b...	but there are also many competitions between t...
458022	i ate loach last night	<start> i had loach for dinner last night	i had loach for dinner last night <end>
356023	and we chatted a little bit and started watchi...	<start> after chatting a little bit we started...	after chatting a little bit we started watchin...
209263	three stright day off	<start> three straight days off	three straight days off <end>
413285	i will study urban planning at there	<start> i will study urban planning there	i will study urban planning there <end>

In [57]:

```
validation
```

Out[57]:

	incorrect	correct
264288	they were so beautiful song	they were such beautiful songs
386352	therefor i can not easy to sleep	so i can not sleep very well
126621	i never gone to china but i totally agree with...	i have never gone to china but i totally agree...
121533	deep breath help me	deep breath helps me
359697	i admitted japanese people got high scores in ...	i admit that japanese people received high sco...
...	...	...
147342	you must finish this work	this work must be finished by you
197930	i went to obi festival to take pictures	i went to the obi festival to take pictures
77006	many customer came to my day	many customers came today
165890	my friends said to me decoration versions are ...	my friends said that the decorated versions ar...
264329	that sounds interesting is not that?	that sounds interesting does not it?

62381 rows × 2 columns

In [58]:

```
validation['correct_inp'] = '<start> ' + validation['correct'].astype(str)
validation['correct_out'] = validation['correct'].astype(str) + ' <end>'

validation = validation.drop(['correct'], axis=1)
# only for the first sentance add a toke <end> so that we will have <end> in tokenizer
validation.head()
```

Out[58]:

	incorrect	correct_inp	correct_out
264288	they were so beautiful song	<start> they were such beautiful songs	they were such beautiful songs <end>
386352	therefor i can not easy to sleep	<start> so i can not sleep very well	so i can not sleep very well <end>
126621	i never gone to china but i totally agree with...	<start> i have never gone to china but i total...	i have never gone to china but i totally agree...
121533	deep breath help me	<start> deep breath helps me	deep breath helps me <end>
359697	i admitted japanese people got high scores in ...	<start> i admit that japanese people received ...	i admit that japanese people received high sco...

In [59]:

```
# for one sentence we will be adding <end> token so that the tokanizer learns the word <e
# with this we can use only one tokenizer for both encoder output and decoder output
train['correct_inp'].iloc[0]= train.iloc[0]['correct_inp']+' <end>'
train['correct_out'].iloc[0]= train.iloc[0]['correct_out']+' <end>'
```

In [60]:

```
train['correct_inp'].iloc[0]
```

Out[60]:

```
'<start> but there are also many competitions between them <end>'
```

In [61]:

```
train
```

Out[61]:

	incorrect	correct_inp	correct_out
337916	but there also are many competitions between them	<start> but there are also many competitions b...	but there are also many competitions between t...
458022	i ate loach last night	<start> i had loach for dinner last night	i had loach for dinner last night <end>
356023	and we chatted a little bit and started watchi...	<start> after chatting a little bit we started...	after chatting a little bit we started watchin...
209263	three stright day off	<start> three straight days off	three straight days off <end>
413285	i will study urban planning at there	<start> i will study urban planning there	i will study urban planning there <end>
...	...	...	...
107088	i would everything put inside me	<start> i have everything they put inside me	i have everything they put inside me <end>
369517	recently i am busy but my school life is limited	<start> recently i have been busy and my schoo...	recently i have been busy and my school life i...
14271	do you have some plan to enjoy summer?	<start> do you have some plans to enjoy summer?	do you have some plans to enjoy summer? <end>
305712	what a tight security!	<start> what tight security!	what tight security! <end>
277876	in these fishes there are dangerous fishes so ...	<start> in this river there are dangerous fish...	in this river there are dangerous fishes so yo...

249523 rows × 3 columns

In [62]:

```
#Correct Sentence tokenizer
#We will be considering . and , and ; as tokens
correct_tk = Tokenizer(filters='!"#$%&()*+-/;=?@[\\]^_`{|}~\t\n,.:',oov_token='<UNK>')
correct_tk.fit_on_texts(train['correct_inp'].values)

#Incorrect Sentence Tokenizer
incorrect_tk = Tokenizer(filters='!"#$%&()*+-/;=?@[\\]^_`{|}~\t\n.,:',oov_token='<UNK>')
incorrect_tk.fit_on_texts(train['incorrect'].values)
```

In [63]:

```
print(correct_tk.word_index.get('<start>'))
print(correct_tk.word_index.get('<end>'))
print(correct_tk.word_index.get('<UNK>'))
```

```
2
20421
1
```

In [64]:

```
vocab_size_correct=max(correct_tk.word_index.values())
print(vocab_size_correct)
vocab_size_incorrect=max(incorrect_tk.word_index.values())
print(vocab_size_incorrect)
```

```
40176
52192
```

In [65]:

```
pickle.dump([vocab_size_correct,vocab_size_incorrect,correct_tk,incorrect_tk],open('toker
pickle.dump([train,test, validation],open('main_data_2.pkl','wb'))
```