About:

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

For example:

Input: dance like i

Output: i like to dance

In [1]:

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

In [2]:

```
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 [4]:

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

In [5]:

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

Out[5]:

	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 [6]:

```
data.shape
```

Out[6]:

```
(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 [7]:

```
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 [8]:

```
import pickle
[train,test, validation]=pickle.load(open('main_data_1.pkl','rb'))
```

In [9]:

```
train
```

Out[9]:

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

249523 rows × 2 columns

In [10]:

```
def reverse(sent):
    reverse_sent=sent.split(' ')[::-1]
    return ' '.join(reverse_sent)

train['incorrect']=train['incorrect'].apply(reverse)
validation['incorrect']=validation['incorrect'].apply(reverse)
```

In [11]:

```
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[11]:

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

In [12]:

```
validation
```

Out[12]:

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

62381 rows × 2 columns

In [13]:

```
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[13]:

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

In [14]:

```
# 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]= str(train.iloc[0]['correct_inp'])+' <end>'
train['correct_out'].iloc[0]= str(train.iloc[0]['correct_out'])+' <end>'
```

In [15]:

```
train
```

Out[15]:

	incorrect	correct_inp	correct_out
337916	them between competitions many are also there but	<start> but there are also many competitions b...	but there are also many competitions between t...
458022	night last loach ate i	<start> i had loach for dinner last night	i had loach for dinner last night <end>
356023	movie a watching started and bit little a chat...	<start> after chatting a little bit we started...	after chatting a little bit we started watchin...
209263	off day stright three	<start> three straight days off	three straight days off <end>
413285	there at planning urban study will i	<start> i will study urban planning there	i will study urban planning there <end>
...	...	...	...
107088	me inside put everything would i	<start> i have everything they put inside me	i have everything they put inside me <end>
369517	limited is life school my but busy am i recently	<start> recently i have been busy and my schoo...	recently i have been busy and my school life i...
14271	summer? enjoy to plan some have you do	<start> do you have some plans to enjoy summer?	do you have some plans to enjoy summer? <end>
305712	security! tight a what	<start> what tight security!	what tight security! <end>
277876	river the in get not should you so fishes dang...	<start> in this river there are dangerous fish...	in this river there are dangerous fishes so yo...

249523 rows × 3 columns

In [16]:

```
[vocab_size_correct,vocab_size_incorrect,correct_tk,incorrect_tk]=pickle.load(open('toker
```

In [17]:

```
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 [18]:

```
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 [19]:

```
pickle.dump([train,test, validation],open('main_data_2_reverse.pkl','wb'))
```