In [1]:

```
from google.colab import drive
drive.mount('/content/drive/')
```

```
Drive already mounted at /content/drive/; to attempt to forcibly remount, call drive.mount(
"/content/drive/", force_remount=True).
```

In [2]:

```
import nltk
import pandas as pd
```

In [3]:

```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[3]:

```
True
```

In [4]:

```
from nltk.tokenize import word_tokenize,sent_tokenize
filepath = '/content/drive/My Drive/English-Telugu/train.en'
corpus = open(filepath, 'r').read()
words = nltk.word_tokenize(corpus)
print("The number of tokens is", len(words))
average_tokens = round(len(words)/75000)
print("The average number of tokens per sentence is",average_tokens)
unique_tokens = set(words)
print("The number of unique tokens are", len(unique_tokens))
```

```
The number of tokens is 1809312
The average number of tokens per sentence is 24
The number of unique tokens are 21095
```

In [5]:

```
from nltk.tokenize import word_tokenize,sent_tokenize
filepath = nltk.data.find('/content/drive/My Drive/English-Telugu/train.te')
corpus = open(filepath, 'r').read()
words = nltk.word_tokenize(corpus)
print("The number of tokens is", len(words))
average_tokens = round(len(words)/75000)
print("The average number of tokens per sentence is",average_tokens)
unique_tokens = set(words)
print("The number of unique tokens are", len(unique_tokens))
```

```
The number of tokens is 1030924
The average number of tokens per sentence is 14
The number of unique tokens are 106016
```

In [6]:

```
f= open("/content/drive/My Drive/English-Telugu/train.en")
en=f.readlines()
len(en)
```

Out[6]:

```
75000
```

In [7]:

```
en[:6]
```

Out[7]:

```
['we just party , and we can do whatever we want .\n',
 'and ziza the son of shiphi , the son of allon , the son of jedaiah , the son of shimri ,
the son of shemaiah;\n',
 'now a poor wise man was found in it , and he by his wisdom delivered the city; yet no man
remembered that same poor man .\n',
 "and the child grew , and she brought him unto pharaoh's daughter , and he became her son
. and she called his name moses: and she said , because i drew him out of the water .\n",
 '- i was at the jeweler and nobody came .\n',
 'and the consecrated things were six hundred oxen and three thousand sheep .\n']
```

In [8]:

```
f= open("/content/drive/My Drive/English-Telugu/train.te")
te=f.readlines()
len(te)
```

Out[8]:

```
75000
```

In [9]:

```
import re
def remove_punc(x):
  return re.sub('[!#?,.:";\n]', '', x)
```

In [10]:

```
for i in range(len(te)):
  te[i]=remove_punc(te[i])
```

In [11]:

```
te[:6]
```

Out[11]:

```
['[illegible] ',
 '[illegible] ',
 '[illegible] ',
 '[illegible] ',
 '[illegible] ',
 '[illegible] ']
```

In [12]:

```
for i in range(len(en)):
  en[i]=remove_punc(en[i])
```

In [13]:

```
en[:6]
```

Out[13]:

```
['we just party  and we can do whatever we want ',
 'and ziza the son of shiphi  the son of allon  the son of jedaiah  the son of shimri  the
son of shemaiah',
 'now a poor wise man was found in it  and he by his wisdom delivered the city yet no man r
emembered that same poor man ',
 "and the child grew  and she brought him unto pharaoh's daughter  and he became her son  a
nd she called his name moses and she said  because i drew him out of the water ",
 '- i was at the jeweler and nobody came ',
 'and the consecrated things were six hundred oxen and three thousand sheep ']
```

In [14]:

```
for i in en:
  for word in i.split():
    words_en.append(word)

words_en[:10]
```

Out[14]:

```
['we', 'just', 'party', 'and', 'we', 'can', 'do', 'whatever', 'we', 'want']
```

In [15]:

```
words_te = []
for i in te:
  for word in i.split():
    words_te.append(word)

words_te[:10]
```

Out[15]:

```
['■■■■■',
 '■■■■',
 '■■■■■',
 '■■■■',
 '■■■■■■',
 '■■■■■■■■■■',
 '■■■■■■■',
 '■■■■■■■',
 '■■■■',
 '■■■■■■■■■■']
```

In [16]:

```
from collections import Counter
english_words_counts = Counter(words_en)
telugu_words_counts = Counter(words_te)
```

In [17]:

```
import operator
english_words_counts = sorted(english_words_counts.items(), key = operator.itemgetter(1),
reverse = True)
telugu_words_counts = sorted(telugu_words_counts.items(), key = operator.itemgetter(1), rev
erse = True)
```

In [18]:

```
english_words_counts[:10]
```

Out[18]:

```
[('the', 113593),
 ('and', 80265),
 ('of', 62955),
 ('to', 34296),
 ('in', 24579),
 ('that', 21052),
 ('you', 20618),
 ('he', 19632),
 ('i', 19186),
 ('a', 17782)]
```

In [19]:

```
telugu_words_counts[:10]
```

Out[19]:

```
[('■■■■', 12085),
 ('■■■■■■', 9127),
 ('■■', 9006),
 ('■■', 8842),
 ('■', 8681),
```

```
 ('[redacted]', 8569),
 ('[redacted]', 6680),
 ('[redacted]', 5784),
 ('[redacted]', 5701),
 ('[redacted]', 5691)]
```

In [20]:

```
maxlen_english = -1
for doc in en:
    tokens = nltk.word_tokenize(doc)
    if(maxlen_english < len(tokens)):
        maxlen_english = len(tokens)
print("The maximum number of words in any document = ", maxlen_english)
```

```
The maximum number of words in any document =  74
```

In [21]:

```
maxlen_telugu = -1
for doc in te:
    tokens = nltk.word_tokenize(doc)
    if(maxlen_telugu < len(tokens)):
        maxlen_telugu = len(tokens)
print("The maximum number of words in any document = ", maxlen_telugu)
```

```
The maximum number of words in any document =  47
```

In [22]:

```
def tokenize_and_pad(x, maxlen):
  #  a tokenier to tokenize the words and create sequences of tokenized words
  tokenizer = Tokenizer(char_level = False)
  tokenizer.fit_on_texts(x)
  sequences = tokenizer.texts_to_sequences(x)
  padded = pad_sequences(sequences, maxlen = maxlen, padding = 'post')
  return tokenizer, sequences, padded
```

In [23]:

```
from tensorflow.keras.preprocessing.text import one_hot, Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
x_tokenizer, x_sequences, x_padded = tokenize_and_pad(en, maxlen_english)
y_tokenizer, y_sequences, y_padded = tokenize_and_pad(te,  maxlen_telugu)
```

In [24]:

```
[print("The tokenized version for document\n", en[-1:][0],"\n", x_padded[-1:][0])]
```

```
The tokenized version for document
 how could you stand for it
 [149 391   7 316  11  16   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   0   0]
```

Out[24]:

```
[None]
```

In [25]:

```
print("The tokenized version for document\n", te[-1:][0],"\n ", y_padded[-1:][0])
```

```
The tokenized version for document
 [redacted]
  [   6  296 8102  887    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0    0    0    0    0    0    0    0    0    0
     0    0    0    0    0]
```

In [26]:

In [26]:

```
from nltk.translate.bleu_score import sentence_bleu
```

In [27]:

```
from tensorflow.keras import models
```

In [28]:

```
NMT1=models.load_model('/content/drive/My Drive/NMT_models/NMT1.h5')
NMT1.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 74, 256)           5810688
_________________________________________________________________
lstm_2 (LSTM)                (None, 256)               525312
_________________________________________________________________
repeat_vector_1 (RepeatVecto (None, 47, 256)           0
_________________________________________________________________
lstm_3 (LSTM)                (None, 47, 256)           525312
_________________________________________________________________
time_distributed_1 (TimeDist (None, 47, 106518)        27375126
=================================================================
Total params: 34,236,438
Trainable params: 34,236,438
Non-trainable params: 0
_________________________________________________________________
```

In [29]:

```
NMT2=models.load_model('/content/drive/My Drive/NMT_models/NMT2.h5')
NMT2.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 74, 256)           5810688
_________________________________________________________________
lstm (LSTM)                  (None, 74, 256)           525312
_________________________________________________________________
lstm_1 (LSTM)                (None, 128)               197120
_________________________________________________________________
repeat_vector (RepeatVector) (None, 47, 128)           0
_________________________________________________________________
lstm_2 (LSTM)                (None, 47, 256)           394240
_________________________________________________________________
lstm_3 (LSTM)                (None, 47, 128)           197120
_________________________________________________________________
time_distributed (TimeDistri (None, 47, 106518)        13740822
=================================================================
Total params: 20,865,302
Trainable params: 20,865,302
Non-trainable params: 0
_________________________________________________________________
```

In [30]:

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x_padded, y_padded, test_size = 0.1)
import numpy as np
y_train = np.expand_dims(y_train, axis = 2)
y_train.shape
y_test = np.expand_dims(y_test, axis = 2)
y_test.shape
```

Out[30]:

```
(7500, 47, 1)
```

In [31]:

```
def pad_to_text(padded, tokenizer):

    id_to_word = {id: word for word, id in tokenizer.word_index.items()}
    id_to_word[0] = ''
    return ' '.join([id_to_word[j] for j in padded])
```

In [32]:

```
# function to make prediction
def predictionNMT1(x, x_tokenizer = x_tokenizer, y_tokenizer = y_tokenizer):
    predictions = NMT1.predict(x)[0]
    id_to_word = {id: word for word, id in y_tokenizer.word_index.items()}
    id_to_word[0] = ''
    return ' '.join([id_to_word[j] for j in np.argmax(predictions,1)])
# function to make prediction
def predictionNMT2(x, x_tokenizer = x_tokenizer, y_tokenizer = y_tokenizer):
    predictions = NMT2.predict(x)[0]
    id_to_word = {id: word for word, id in y_tokenizer.word_index.items()}
    id_to_word[0] = ''
    return ' '.join([id_to_word[j] for j in np.argmax(predictions,1)])
```

In [33]:

```
y_test=y_test.reshape(y_test.shape[0],y_test.shape[1])
```

In [34]:

```
yp1=[]
yp2=[]
yt=[]
for i in range(len(x_test)):
  yt.append(pad_to_text(y_test[i], y_tokenizer))
  yp1.append(predictionNMT1(x_test[i:i+1]))
  yp2.append(predictionNMT2(x_test[i:i+1]))
```

In [35]:

```
BLEUscore1 = nltk.translate.bleu_score.corpus_bleu(yp1, yt)
print(BLEUscore1)
```

```
0.4779748973490351
```

```
/usr/local/lib/python3.7/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:
Corpus/Sentence contains 0 counts of 2-gram overlaps.
BLEU scores might be undesirable; use SmoothingFunction().
  warnings.warn(_msg)
```

In [36]:

```
BLEUscore2 = nltk.translate.bleu_score.corpus_bleu(yp2, yt)
print(BLEUscore2)
```

```
0.48104068819490353
```

```
/usr/local/lib/python3.7/dist-packages/nltk/translate/bleu_score.py:490: UserWarning:
Corpus/Sentence contains 0 counts of 2-gram overlaps.
BLEU scores might be undesirable; use SmoothingFunction().
  warnings.warn(_msg)
```