Team name: Strivers
Team members:
Mahesh Sathvika - 2020101087
Mudhireddy Nandini Reddy - 2020101038
Dhulipati Lakshmi Girija - 2020101027

Interim submission

INTRODUCTION

Machine translation, question-answering, and paraphrasing are just a few of the linguistic and semantic applications that use natural language processing (NLP). Since most of these systems were created for languages with high resource requirements, the focus must be on low-resource languages. The majority of studies have focused on the rephrasing of high-resource languages, with only a limited amount of research on low-resource languages. One barrier to solving this issue is the lack of corpora in languages with scarce resources. Thorough study of low-resource languages is more appealing due to the possible benefit it may have on the population of people who find high-resource languages complex. As a result, we attempt to create a paraphrase model for Indian languages in this study employing recurrent neural networks of the Long Short Term Memory (LSTM) and the Gated Recurrent Unit(GRU), machine translation. We are mainly focused on Hindi and Telugu languages.

How we extracted datasets for modelling

1st dataset:

There aren't any high-quality datasets that are suitable for our project because Hindi and Telugu are low-resource languages. As a result, we are using the MSCOCO(Microsoft Common Objects in Context) dataset, which contains 4,5 sentences, to describe the photos in English.
By presuming that the translations produced by Google Translate API are accurate and adequate for ground truth for training our model for paraphrasing in Hindi and Telugu, we are translating these descriptions into Hindi and Telugu utilising that service.

For translations, code

```
import json
from operator import itemgetter
from itertools import groupby
from googletrans import Translator
f = open('/content/drive/MyDrive/annotations/captions_train2014.json',
'r')
x = f.read()
x = json.loads(x)
sortkeyfn = itemgetter('image_id')
x['annotations'].sort(key=sortkeyfn)
captions = []
for key,valuesiter in groupby(x['annotations'], key=sortkeyfn):
   captions.append(dict(type=key, items=list(v['caption'] for v in
valuesiter)))

g = open('/content/drive/MyDrive/caption2014.txt', 'a')
translator = Translator()
for k in range(int(len(captions)/10)):
 for i in range(10):
   for j in range(len(captions[k*10+i]['items'])):
     translations = translator.translate(captions[k*10+i]['items'][j],
dest = 'te')
     captions[k*10+i]['items'][j] = translations.text
 print(captions[k+10+i]['items'])

g.write(captions)
```

The MSCOCO dataset link is as follows: https://cocodataset.org/#download
After translating these English sentences into Hindi and Telugu using google translate api, we put them in JSON files.
In Hindi, there are 81k different types of sentences with numerous phrases having the same meaning in the train data and 21k in the validation data. This collection is divided into the proper train, validation, and test splits.
Link to those files is:**Google Drive**
https://drive.google.com/drive/folders/1_IGx4O_TtoR0plDGFUG0fjNH4w__KwnY?usp=share_link

```
    {'type': 30,
     'items': ['A flower vase is sitting on a porch stand.',
      'White vase with different colored flowers sitting inside of it. ',
      'a white vase with many flowers on a stage',
      'A white vase filled with different colored flowers.',
      'A vase with red and white flowers outside on a sunny day.']},
```

2nd dataset

At Kaggle, we are utilising the Telugu NLP dataset.
We translate news headlines about various topics like weather, movies, and health into English and then into Telugu using the Google Translate API.
The models are then trained using these two phrases, which are equivalent in both languages.This datasets also extracted same way as above method by doing two times.
Kaggle data set link
https://www.kaggle.com/datasets/sudalairajkumar/telugu-nlp
Translation datasets txt files are present in google drive.
Example:

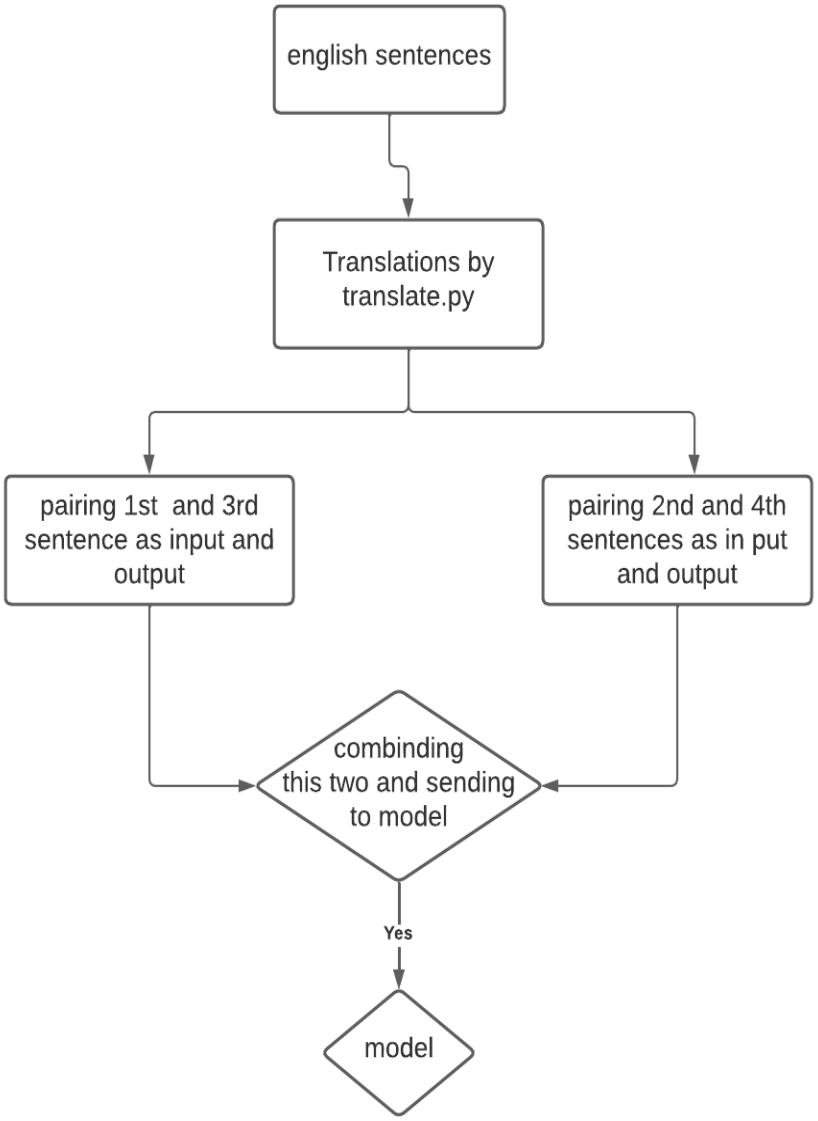
```
RBI Nazar on IDBI
Jaitley meeting with banking chiefs today
Jadeja who took a key wicket ..
Pakistan landing for another provocative action
Allu Arjun swimming with son in Goa!
Remove office bearers ..!
Indians in America do not speak in the mother tongue
Super Sunday!
ATM Working Review
Lakhs: Sushma
Sehwag's passionate tweet about Yuvi
```

Work completed:

- The MSCOCO dataset has been preprocessed, and sentences have been transformed to Hindi and Telugu. Those files can be found in the aforementioned Google Drive URL.
- Preprocessing of the Telugu Kaggle dataset is completed, and sentences with the same meaning are extracted by translating from Telugu to English and back again.
- The Hindi LSTM model was trained using MSCOCO-translated sentences. Model is saved in Google Drive as model.pth.

Workflow until now

We have worked on translating the english dataset to telugu and hindi using GoogleTrans. We have 5 MSCOCO Hindi captions of which we are discarding the 5th caption and are using 1 and 3 as inputs and 2 and 4 as outputs respectively. The data is then being tokenized to make tensor inputs and outputs. The inputs and outputs are being sent to the LSTM model to train the data.



Research papers followed:
During the course of the project we expect to follow the above papers:
https://kalaharijournals.com/resources/DEC_544.pdf
https://arxiv.org/pdf/1709.05074.pdf

Datasets
The datasets we are using are:
https://cocodataset.org/#download
The MSCOCO dataset in English which is being converted to Hindi and Telugu NLP:
https://www.kaggle.com/datasets/sudalairajkumar/telugu-nlp

This dataset is being used to convert to English and then back to telugu again to get the paraphrases.

Scope and limitations of Dataset:

Telugu NLP dataset:

The telugu NLP dataset is only limited to headlines and hence we can not form coherent sentences but we can extract another headline from the data. The model will not learn accurate punctuations or continuation in sentences but will be able to paraphrase sentence to sentence. Here we are assuming that sentences obtained after the double translation are different.

MSCOCO dataset:

Just like the previous dataset we will not be able to learn a sentence in continuation to paraphrase but we can paraphrase from sentence to sentence. This dataset is diverse and we will be able to produce better results through it. It also contains more data to accurately train the model.

All the python files for modelling LSTM and translations are uploaded to moodle.