Telugu LLMs

Report: May 4, 2024 - May 7, 2024

Work Done:
In the past two to three days, significant progress has been made towards understanding the landscape of Telugu Language Models (LLMs) and exploring existing resources. Here's a summary of the work completed:

Research on Existing Telugu LLMs:
We have searched for various Telugu Language Models available, including small language models (SLMs) like Chandamama Kathalu and some deep learning-based models such as Llama-3-8b-Telugu_Romanized, POS-NER models, Comprehension Question Answering models, Summarization Models, Whisper-large-v3 by OpenAI, and Navarasa 2.0. Each model's capabilities, datasets used, and potential applications have been analysed to understand their strengths and weaknesses.

1) Chandamama Kathalu: An SLM. Around 40,000 short stories have been selected, proofread, and used for training this SLM.

Dataset:
https://code.swecha.org/panini-dhpc/chandamama-kathalu-dataset
https://huggingface.co/datasets/swechatelangana/chandamama-kathalu

Model:
https://chandamama.swecha.org/

Data collected using ocr service:
https://code.swecha.org/telugu-ai/chandamama-kathalu-ocr-service

2) Llama-3-8b-Telugu_Romanized: The model is designed specifically to handle tasks for the spoken Telugu language, and it uses the Romanized script. It even supports code-mixing (a combination of English and Telugu), which is common in informal contexts. The model can be used directly for tasks such as language generation, text completion, and question answering in the Telugu language with code-mixing.

Link: https://huggingface.co/jayasuryajsk/Llama-3-8b-Telugu-Romanized

3) POS-NER, Comprehension Question Answering and Summarization Models
Link:
https://medium.com/analytics-vidhya/language-modeling-for-%E0%B0%A4%E0%B1%86%E0%B0%B2%E0%B1%81%E0%B0%97%E0%B1%81-telugu-b590a029a565

4) Whisper-large-v3: A model by OpenAI which is mainly used for speech recognition and speech translation.
Link: https://huggingface.co/openai/whisper-large-v3
Dataset: For multilingual languages, they used the Google Fleurs dataset.
Dataset Link: https://huggingface.co/datasets/google/fleurs

5) Navarasa 2.0: A model by Telugu LLM's Lab, which is finetuned on Google Gemma and mainly used for text generation. The dataset they used was the Indic alpaca dataset. It is fine-tuned for 15 Indian languages.

Link:https://huggingface.co/collections/Telugu-LLM-Labs/navarasa-20-models-65f7c72addf0619cb0991309
Dataset Link:
https://huggingface.co/collections/Telugu-LLM-Labs/indic-alpaca-datasets-65f2a3687d5cdbce8880c581

Exploration of other multi-lingual Models:
In addition to Telugu-specific models, investigation into multi-lingual models focusing on Indic languages, such as AI for Bharat and Bloom, has been conducted.

Information about already existing multi-lingual models:

- **AI for Bharat:** These are the ones who are mainly working on Indic languages.
  Link: https://github.com/AI4Bharat/IndicBERT/tree/main?tab=readme-ov-file#indiccorp-v2
- **Bloom:** They are also working on some Indic languages, and Telugu is one of them.
  Link: https://huggingface.co/bigscience/bloom

DataSets:
- Telugu_news

Work to be Done:
Moving forward, the following tasks are planned for the next two to three days:

Model Interaction:
The immediate focus will be on interacting with some of the existing models from the Hugging Face repository. This hands-on exploration will provide practical insights into model capabilities, limitations, and potential customisation requirements.

Data Collection Strategy:
Efforts will be directed towards devising a comprehensive strategy for data collection. This involves identifying diverse sources of Telugu text data, including literature, news articles, social media content, and domain-specific texts. Moreover, considerations for data preprocessing, cleaning, and annotation will be outlined.

Things to ponder:

Parameters, number of layers, context length size, vocab size - whether it is an encoder/decoder or both.

See datasets like:
ROOTS Corpus - how big
Common Crawl

sarwan.ai - Pratyush Kumar, iit bombay - pushpak Bhattacharyya
Ola Krutrim ai - ravi jain

****

Report: May 7, 2024 - May 11, 2024

Work Done:
In the last couple of days, we've been working on finding and organising Telugu text datasets. We will also gather more data from websites using web scraping techniques. The primary objective is to gather diverse Telugu text data to support the development of a comprehensive Telugu Language Model (TLM).

Available Datasets:
Several Telugu language datasets have been identified, each varying in size and content. These datasets include:

Size of Telugu language dataset in Roots Corpus: 3 GB
Size of Telugu language dataset in AI4Bharat-IndicNLP Dataset: 674 MB - LINK
Size of Telugu language dataset in CC-100: 500 MB - LINK
Size of Telugu language dataset in OSCAR-2201: 500 MB - LINK
Size of Telugu language dataset in OSCAR-2301: 3.9 GB - LINK
Size of Telugu language dataset in Kaggle/Telugu_NEWS: 403.86 MB - LINK, SCRAPER
SOME MORE TO LOOK AT
Andra Jyothi 2015, 2016, 2017 newspaper data - Link
Telugu Translation task dataset LINK
Mahabharatam in Telugu - Link
Telugu PDfs of Books - Link
Telangana books - App(for Telugu medium books from class 1 - 10)
Bible in Telugu - LINK
One Hundred and Fifth Amendment of the Indian Constitution - LINK
Constitution of India in Telugu - LINK
Telugu books in pdf form - LINK
Legislative department - LINK
Pedda bala siksha PDF - LINK

Greater Telugu Website - LINK

IIITH LLM Lab Datasets.
Speech DATASET
IITA Dataset
Telugu Raw Speech Corpus
Telugu Raw Text Corpus
Telugu (India) General Conversation Speech Dataset

Websites to Scrape from:
To supplement the existing datasets, several websites have been identified as potential sources for Telugu text data. These include (more will be added to the list):

NPTEL COURSES TRANSLATION: https://nptel.ac.in/translation
NEWSPAPER SITES
https://www.smartial.net/smart-tools/wextractor.php

By systematically curating Telugu text datasets and supplementing them with freshly scraped data from online sources, We aim to build a robust foundation for the development of an effective Telugu Language Model.

****

Report: May 11, 2024 - May 18, 2024

Work Done:
Objective: Over the past few days, the primary goal has been to develop a system for scraping Eenadu newspapers from the Wayback Machine and converting PDF files into text to extract data.

Scraping Eenadu Newspapers:
Initiated the development of code to scrape Eenadu newspapers from the Wayback Machine.

Successfully accessed and extracted newspaper content from the archived versions on the Wayback Machine.

Utilised Python libraries such as BeautifulSoup and requests to handle the web scraping tasks.
Ensured that the extracted data was saved in a structured format for further processing and analysis.

PDF to Text Conversion:
Attempted to convert PDF files of the scraped newspapers into text format.
Employed various tools and libraries like PyPDF2 and PDFMiner for the text extraction process.

Faced challenges with accurately converting the PDF content into a readable text format. The extracted text was either incomplete or incorrectly formatted, which hindered data usability.

Challenges Encountered:
Difficulty in handling the complex structure and formatting of PDF files.

Inconsistent results from different PDF-to-text conversion libraries make standardising the extracted data challenging.

Work to be Done
Objective: The focus for the upcoming days will be to address the issues encountered with the PDF-to-text conversion and to expand the scope of data extraction from additional archive sites, particularly the Wayback Machine.

Improving PDF to Text Conversion:
Investigate alternative methods and tools for more effective PDF-to-text conversion.

Consider using OCR (Optical Character Recognition) technology with libraries like Tesseract to handle scanned or image-based PDFs.

To enhance text extraction accuracy, develop a robust pipeline to preprocess PDFs before conversion.

Test and validate the improved conversion methods with a sample set of PDFs to ensure reliability and consistency.

Expanding Data Extraction:
Explore additional archives and repositories similar to the Wayback Machine for potential sources of historical newspaper data.

Modify and enhance the existing scraping code to adapt to different archive structures and access protocols.

Implement error handling and logging to improve the resilience and maintainability of the scraping system.

Establish a workflow for periodically updating and expanding the dataset with new newspaper archives as they become available.

****

Report: May 18, 2024 - May 21, 2024

Work Done:
Over the past few days, We have scrapped from more archive sites such as sakshi.com and telugu360 news. We were also successful in figuring out how to convert PDF to text, but the main source of data remains as websites.
Data collected so far (approx figures):
Eenadu from way back - 180 MB - being a purely archive site, data collection is very slow
Telugu360 news - 120 MB - data collection is fast here, but data is less
sakshi.com - collected till now 515 MB - data collection speed is moderate, and a huge amount of data is present in the archive spanning over ten years (2024 - 2014). We can expect at least seven gigs in total.
Data collection from PDFs: This is working but not implemented at all. It needs some modification.

Total till now: 815 MB (approx)

Work To be done:
We have figured out more archive sites, which are listed below:
Newspaper archives:
1) https://www.sakshi.com/archive
2) https://www.telugu360.com/te/category/politics/
3)http://www.andhrabhoomi.net/nation
4)https://www.vaartha.com/category/telangana/page/2/
5)https://www.telugumirchi.com/telugu/movies
6)https://www.netitelugu.com/
7)https://telugumopo.com/category/political-news/telugu-political-news/
8)https://www.manatelangana.news/category/telangana-news/telangana-state-news/
9)https://telugu.webdunia.com/andhra-pradesh-news

Our aim will also be to scrap out these websites. More archives will also be appended to this list.

****

Report: May 21, 2024 - May 28, 2024

Work Done:
In the past few days, we have curated a lot of websites and scraped some of them. We have also started working on collecting data from existing popular datasets like common crawl and roots corpus.

We have collected around 15 GB (approx) till now, of which almost 9 GB consists of data from existing datasets, and we scrape the rest.

Link to sources we got data from

Work To be done:
Since we have a good amount of data, we will be focusing on its processing, such as deduplication and tokenisation. We will also be curating more data in parallel.

****

Report: May 28, 2024 - June 10, 2024

Work Done:
In the past few days, we have collected a massive amount of text data from the internet of around 150 GB, which consists of data from existing datasets, websites scraped from us and data extracted from PDFs.

Work To be Done:
We will be working on deduplicating the obtained data from now on. We are thinking of following the sim-hash method of deduplication. After deduplication, we will be working on tokenisation.

****

Mid-Term Report on Telugu Language Model (LLMs)

Introduction
The development of the Telugu Language Model (LLM) has seen significant advancements from May 4, 2024, to June 20, 2024. Our primary focus has been to understand the landscape of existing Telugu LLMs, gather diverse datasets, and lay the groundwork for building a comprehensive Telugu Language Model. This mid-term report provides an overview of the work accomplished, the challenges encountered, and the roadmap for the upcoming phases.

Research on Existing Telugu LLMs
Identification and Analysis
We began by researching the existing Telugu Language Models to understand their capabilities, datasets, and potential applications. This included models like Chandamama Kathalu, Llama-3-8b-Telugu_Romanized, POS-NER models, Comprehension Question Answering models, Summarization Models, Whisper-large-v3 by OpenAI, and Navarasa 2.0. Key insights from these models are:

Chandamama Kathalu: This Small Language Model (SLM) is trained on a dataset of around 40,000 short stories, focusing on traditional Telugu literature.
Llama-3-8b-Telugu_Romanized: Designed for spoken Telugu in Romanized script, it supports code-mixing with English, making it suitable for informal contexts.
POS-NER, Comprehension QA, Summarization Models: These models are crucial for various NLP tasks in Telugu, such as part-of-speech tagging, named entity recognition, and text summarisation.

Whisper-large-v3: An OpenAI model used primarily for speech recognition and translation, utilising the Google Fleurs dataset.
Navarasa 2.0: A text generation model fine-tuned on the Indic Alpaca dataset, supporting multiple Indian languages, including Telugu.

Exploration of Multilingual Models
We also explored multilingual models focusing on Indic languages, such as AI for Bharat and Bloom, which include Telugu among the supported languages. These models offer a broader context for developing robust and versatile LLMs for Telugu.

Data Collection and Dataset Analysis
Identified Datasets
Our data collection strategy focused on curating diverse sources of Telugu text data. We identified several datasets of varying sizes and content:

Roots Corpus: 3 GB of Telugu data.
AI4Bharat-IndicNLP Dataset: 90 GB (including synthetic data).
CC-100: 4.7 GB.
OSCAR-2201 and OSCAR-2301: 2.5 GB and 3.6 GB, respectively.
Kaggle/Telugu_NEWS: 310 MB.
etc

We also sourced data from the Mahabharatam in Telugu, Andhra Jyothi newspaper archives, Telangana books, and various other repositories.

Scraping and Conversion
We developed systems to scrape websites and convert PDFs to text, focusing on historical newspapers and other valuable text sources:

Eenadu from Wayback Machine: Collected 820 MB of data.
Telugu360 News: 220 MB.
Sakshi.com: 1.5 GB.
Challenges faced included handling the complex structure of PDFs and ensuring accurate text extraction.

Progress on Data Collection
As of June 10, 2024, we have amassed approximately 150 GB of text data, combining existing datasets, scraped data from websites, and extracted text from PDFs. This data serves as the foundation for developing a robust Telugu Language Model.

Next Steps: Data Processing and Model Development
Deduplication and Tokenization
Our immediate focus is on processing the collected data. We plan to use the sim-hash method for deduplication, which will help remove redundant data and ensure our dataset's uniqueness.

Following deduplication, tokenisation will be performed to convert the text into a format suitable for model training.

Model Interaction and Customization
Interacting with existing models from the Hugging Face repository will provide practical insights into their capabilities and limitations. This hands-on exploration will inform potential customisation requirements, enhancing the models' performance for specific Telugu language tasks.

Data Expansion and Maintenance
We will continue to explore and scrape additional websites, particularly those with rich archives like sakshi.com and telugu360 news. Maintaining a workflow for periodic updates and expansion of the dataset is crucial for keeping the model relevant and up-to-date.

Conclusion
Significant progress has been made in understanding the landscape of Telugu Language Models and gathering the necessary datasets. Despite challenges in data extraction and conversion, we have successfully compiled a substantial corpus of Telugu text data. Moving forward, our focus will be on data processing, model interaction, and continuous data expansion to develop a comprehensive and effective Telugu Language Model.

****

Report: June 10, 2024 - June 16, 2024

Work Done:
In the past few days, we tried implementing the sim-hash algorithm in three stages.

1) We have divided our 150 GB of data into nearly 85 lakh text files of 800 - 1000 words each. Now, we calculate the sim-hash for each text file.
2) We have calculated sim-hashes in step 1 and stored them across 77020 CSV files. In step 2, we will drop the rows with exact hashes (duplicates, keeping the first instance) for each CSV.
3) After step 2, we now have 81 lakh text files, and nearly 40,000 files were removed. We will now compare a row in a CSV with all other rows in all other CSVs (pair-wise comparison of 81 lakh hashes)

Challenges:
The above approach is promising, but step 3 is quite expensive and takes a lot of time (nearly years) cause we have 81 lakh pair-wise comparisons to be done.

Work to be done (Alternate approach):
Since the above method is not feasible, we will be shifting to the min-hash algorithm, which removes similarities by identifying the nearest neighbours of a particular doc instead of pair-wise comparisons between all docs. This would be a lot faster than the previous approach.

****

Report: June 16, 2024 - July 2, 2024

Work Done:
In the past few days, we have focused on cleaning our data. This involved removing vulgar words, English text, and promotional content. Additionally, we have completed tokenisation of approximately 2% of our total dataset, which amounts to around 4 GB of data. Our initial tests, which were over 1,000 sentences, yielded an average fertility score of 1.71.

Work To Be Done:
Our primary focus now is to commence training our language model. We will use the Llama architecture as a reference as we proceed with the pre-training phase.

A detailed table showing the experiments performed:
Vocab size - 32768 fixed for all experiments

S.No	1000 Sentences		5380 Sentences	
	Average	Maximum	Average	Maximum
Batch_1(4.1GB)	1.7175	5.066	1.9044	11.22
Batch_2(9.9GB)	2.784	5.5	2.891	12.33
Batch_3(15 GB)	2.784	5.5	2.891	12.33
Batch_4(25 GB)	2.784	5.5	2.891	12.33
Batch_5(38 GB)	2.784	5.5	2.891	12.33

S.N0	Frequency = 5	Frequency = 7	Frequency = 9
Batch_1	1.7175	1.7175	1.7175

****

TO BE CONTINUED……