Dictionary Augmented Transformers

Proposal for Code4GovTech

Mentoring Organization

Samagra-Development

Harsh Goyal

June 11, 2023

Abstract:

Code4GovTech is an organization which helps passionate developers to interact with each other to solve various open source issues. Samagra-Development/ai-tools involves the project “Dictionary Augmented Transformesr”. It has been observed that people do use odia words in various platforms which are not translated with good accuracy. Main objective of the project “Dictionary Augmented Transformers” is to increase the current translation transformer accuracy by augmenting the dictionary.

Contents

Why Code4GovTech?

1. Code4GovTech helps me gain knowledge about how to contribute to any open source.
2. They provide good connections with people from various fields and departments.
3. Projects are well organized and described by mentors. They are dedicated and passionate to help.
4. Working with Code4GovTech provides a good platform to improve yourself for future opportunities.

Project Details

Overview:

Document Augmented Translation Models aims to enhance the translation accuracy by incorporating a dictionary of correct translation.The goal is to ensure that the translated output contains the translated words from a supplied dictionary, especially for words where the translation is known with certainty. Skills required for this project include Python, NLP.

Current Problems:

1. Data preparation involves parsing an odia dictionary. Since some dictionaries involve non-copyable text. Many available free OCR tools do not support many languages. Some supporting OCR tools like Tesseract do not yield good accuracy with text containing multiple languages. These difficulties harden the data preparation task.
2. Augmenting the dictionary with a text translation model involves identifying the words or phrases from input text and replacing them using the dictionary into an initial translation output. Identifying the words and phrases in the input text requires breaking the input into a word or set of words such that their correct translation is present in the dictionary.

Possible Solutions:

1. Tesseract works well when provided with a single language to read from an image. Given a document we can achieve word level segmentation using the tesseract inbuilt function (like tesseract.images_to_boxes) which assures us the output image contains only a single language. Next task is to identify the language of the output image.
2. To Identify the language in an image, We can rely on certain hacks. For example, if we can assume the unicode range of two languages do not overlap, converting an image into text with language different from the language of the image produces an arbitrary sequence of letters in that language. We can identify this sequence, and identify the correct language of image. This works only when an image contains one of the two possible languages.
3. There are more reliable approaches which involve manual parsing of some pages. For example, we can build a CNN model which classifies the language of an image. To train this model we require to parse some pages manually or using some script.
4. This manual parsing can be automated using unsupervised learning like K-mean but still involves some manual work to check the classified data from unsupervised models.

Milestones

Milestone 1:

Automating word level segmentation and analyzing the structure of pdf to generate structured dataset. Location of every word on a page provides information about the structure of the document. For example, if a line is continued with some margin, it indicates it is continuation with the previous line,or identification of horizontal and vertical lines on page helps in generating columns, or identifying same font size images to filter unnecessary words etc.

Milestone 2:

Identifying the language of image using unsupervised learning and Convolutional Neural Network.

Milestone 3:

Producing a high accuracy dataset by reducing errors through some hacks. Hacks like dictionary contain data in lexicographical order or any word translation should not contain special characters like “$”, “#”, “@” etc.

Milestone 4:

Translating the input text using dictionary augmentation. It involves segmenting the input text such that segmented words can be translated by parsed dictionaries.

Timelines

11 June-15 June	***Milestone 1***
16 June- 5 July	***Learning and testing various unsupervised models***
6 July-10 July	***Milestone 2***
11 July- 17 July	***Milestone 3***
18 July- 25 July	***Buffer Time (Completing left over tasks)***
25 july- 5 August	***Learning transformers, translation model and other relevant skills for augmentation***
6 August- 15 August	***Milestone 4***
15 August- 20 August	***Evaluation time***

Availability

Ques: When do your classes and exams finish?

Ans: Classes will start from August and there will be no exams in August. Generally, there are 2-3 classes per day on average (28 hours per week). I will be able to manage my schedule .

Ques: How many hours can you contribute to this project?

Ans: I can contribute approximately 5 hours per day (30 hours per week).

Ques: Do you have a full time or part time internship planned this summer ?

Ans: No, I am not having any full or part time internship this summer. I can dedicate all my attention to this project.

Ques: Are you traveling during the summer?

Ans: No, I won't be traveling anywhere this summer.

Personal details

About Me

Name	**Harsh Goyal**
Email	**harsh1513088@gmail.com**
Contact No.	**+91 8448673532**
Github	**Github link**
LinkedIn	**LinkedIn Profile**
Resume	**My Resume**
Country/Time Zone	**India/ IST(GMT + 5:30)**
Location	**Tagore Garden, West Delhi, Delhi, India**

Education

University	**Indian Institute of Technology (IIT), Hyderabad**
Major	**Computer Science and Engineering**
Year	**3rd Year**
Degree	**Bachelor of Technology (B. Tech.)**

Why me?

1. I am passionate and dedicated about contributing to my first open source project. I assure you to yield the best out of me.
2. I am a Team Worker. I have done many successful projects on the college level which involve team management.
3. I want to leave my comfort zone and explore my limits. I like problem solving.
4. I believe in sharing knowledge. This helps me to manage the communication among teammates and also increases the possibility of better outcome.

Thanking note

1. Code4GoveTech helps me learn about recent technologies for NLP. It helps me to get out of my comfort zone.
2. I will look forward to any feedback from the organization reviewing this document and would be grateful to discuss/change accordingly.