Responsibilities of interns -2

*Study or Literature Survey of the different pre-trained language models in NLP which can be used for our project:*

There are many pretrained language models available which are simple to incorporate and don’t require much labelled data to work with, and also makes it easy for prediction, transfer learning etc.x Some of them are OpenAI’s GPT-3, Google’s BERT, Microsoft’s CODEBERT, ELMo, Facebook’s ROBERTa, ULMFiT, XLNet, Google’s ALBERT etc[1][2].

Get the email body using Email Parser

Annotate the valid trips in the emails

Train the NER model using transformers

Combine the email chain documents into a single document

Remove punctuation and hyperlinks

Remove unwanted spaces

Remove URLS, HTML tags

Pre-Processing steps

Get the travel relevant emails from the dataset

Test and evaluate the model on unlabeled emails

Prepare the labelled set into a specified format for training

1

2

3

7

6

5

4

8

Figure 1. Pipeline of our project

In our project, we are trying to do a custom Named Entity recognition(NER) on the email data set i.e., we are labelling the trips by naming the traveler, the locations travelled and the arrival and departure date and times. In Figure 1, the trips are identified and labelled in 5 and 6; in 7,we train a model with the labelled trips data. For training, we need to identify which language model best suits our project which is travel domain specific.

Since there are many language models, we need to do a survey on the different language models; understand the advantages and disadvantages, and work on them to evaluate which gives the best results.

References

1. <https://vitalflux.com/nlp-pre-trained-models-explained-with-examples/>
2. <https://analyticsindiamag.com/top-8-pre-trained-nlp-models-developers-must-know/>