NLP PROJECT

Multilingual News Article Similarity

Group Name : MocktaiL

Ravi Karthik(18EC3AI19)
Nitin Bisht(21MM61R08)
Dharmana Jaswanth(18CS10014)
Ananya Das(20EC10103)

TASK DESCRIPTION:

The task is to find out if a given pair of news articles cover the same story or not regardless of the political spin, tone or the style of writing.The pair of documents could be multilingual i.e; one document could be in one language and other document could be in another language.It is a document level similarity task in the domain of new articles and we needed to label the articles on the similarity scale of '0' to '4'.Here '0' means least similar and '4' means highly similar.

INDIVIDUAL CONTRIBUTIONS:

We all contributed to project in all the below mentioned parts.The project included the following parts:

1)Scraping the Data from the json files

2)Cleaning the data obtained from the json files

3) Searched and Trained a suitable BERT model to generate embeddings for our data

4)Tried and implemented various Machine Learning models to predict the similarity between two articles.

BASELINE APPROACHES:

Since our data was in json format we had to write a program inorder to extract data from it.To do that we imported the modules **os,json** and used them.During this process we made sure that the data is properly cleaned.i.e; to make sure that there isn't any sample where there is no file corresponding to a id,if there is no text data in some samples etc.

For implementing models we used **mBERT** model[3]. Here 'm' stands for multilingual.We need this multilingual model because our dataset contains multiple languages.More specifically we used **'bert-base-multilingual-cased'** model.We can get this model from the **sentence_transformers** module offered by **Hugging Face**.

This particular module truncates the text to **512** tokens.It then encodes this into a **768** feature dense embedding.Hence using this module we can encode our text data into **768** feature dense embeddings.This encoding process took us around **50 min**.

Following this we implemented various machine learning techniques to make predictions on the similarity between the 2 articles.If the prediction value is greater than 4 we truncated it to 4 and if it less than 0 we truncated it to 0(Because the value should be in between [0,4]).

DISCUSSION NEXT STEPS:

Below discussed are the possible alternatives that we are planning to implement to increase the performance:

1) mBERT is bad at mapping sentences with sentences of similar meaning to the same vector and the performance drops even further when we mix different input languages to calculate similarity.Models like mBERT predict vector values of individual tokens and not sentences. This causes sentence aggregations to misalign the vectors due to lexical differences in languages. Meaning mBERT and other pre-trained multilingual transformers are not suitable for cross-language sentence similarity .Because of which the output of the mBERT is fine-tuned to the specific language and when comparing words of different languages, their vectors are diverged.**[6]**.So we can improve our performance by using a better models like LASER,DISTILUSE-BASE-MULTILINGUAL-CASED.

2) Using BERT-large could yield better results as it is a more robust model trained on larger corpus
3) Using more efficient Machine Learning Techniques such as Gradient Boosting etc.

RESULTS OF BASELINES:

METHODS USED	PEARSON COEFFICIENT OF TRAINING DATA	PEARSON COEFFICIENT OF TEST DATA
Cosine Similarity	0.27	0.2691
Decision Tree Classifier	0.9993	0.3405
Random Forest Classifier	0.9835	0.6057

Link to final baseline model:

https://colab.research.google.com/drive/1Zvjx4uzdyheLRfWHPOzxW1jPOiP9-9cC?usp=sharing

REFERENCES

1. https://competitions.codalab.org/competitions/33835
2. https://huggingface.co/bert-base-multilingual-cased
3. https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1
4. https://towardsdatascience.com/a-complete-guide-to-transfer-learning-from-english-to-other-languages-using-sentence-embeddings-8c427f8804a9
5. https://towardsdatascience.com/multilingual-sentence-models-in-nlp-476f1f246d2f
6. https://towardsdatascience.com/cutting-edge-bert-nlp-model-bb0bfc8b7aec