CSE 576 Natural Language Processing

Project Phase 2 – Automated Data Creation

Jayasurya Sevalur Mahendran | ASU ID: 1217399443 | E-mail ID: jsevalur@asu.edu

Links:

Entire Team's submission can be found at: https://github.com/JainSahit/NLP576-SIA

All the files can be found at: https://drive.google.com/drive/folders/1HCH6OYs6U56eNR5C03J_pOZW1TElDOYd?usp=sharing

1. Preprocessed Dataset and Results

   https://drive.google.com/drive/folders/1HCH6OYs6U56eNR5C03J_pOZW1TElDOYd?usp=sharing

Python notebook links

1. **Generating-Dataset-Using-Preprocessed-Huggingface-Models.ipynb**

   https://colab.research.google.com/drive/1117iKWm6Vju8yVPFHq1s_nxpkVCbS0Cq?usp=sharing

2. **Pyserini_and_Data_PreProcessing.ipynb**

   https://colab.research.google.com/drive/1yKHTbOUMYdRdb0_N6fFlkoJU9h_hwPan?usp=sharing

3. **SIA-Scores-Generation-Using-WEB-BERTandClinical-BERT.ipynb**

   https://colab.research.google.com/drive/1ndFdUtDpT_Wh-H5kvhTiv0OoIh1MAp9A?usp=sharing

Task Description

Since Semantic Information Availability (SIA) does not have a dedicated dataset for itself, the task is to use the publicly available dataset to create answer candidates and assign a SIA for each answer candidate and create a diverse dataset for SIA. For this purpose, I had chosen the **multi-hop question-answering QASC dataset**.

Steps Performed:

1. Extraction of Answer candidates from Corpus using Pyserini (Anserini + Okapi BM25).
2. Utilize Web Bert Model to generate STS scores.
3. Convert generated STS scores in range [0, 5] to SIA score [0, 4].
4. Export the Result dataframe.

Detailed explanation can be found inside the main report.

The final data consists of three columns namely question, Sentence(Answer candidates), sia score.

Running the code

Above, I have attached links to the colab notebooks, I have attached the ipynb files in the submission folder.

Note: Before running any section of code kindly download the entire folder and upload it to your drive, and change the path wherever its necessary

The preprocessing of data and extracting answer candidate answer from the corpus is done in the file named: Pyserini_and_Data_PreProcessing.ipynb.

> https://colab.research.google.com/drive/1yKHTbOUMYdRdb0_N6fFlkoJU9h_hwPan?usp=sharing

Using the answer candidates and Question and exact answer pair, STS scores is generated and converted to SIA scores in the file titled: SIA-Scores-Generation-Using-WEB-BERTandClinical-BERT.ipynb

> https://colab.research.google.com/drive/1ndFdUtDpT_Wh-5kvhTiv0Oolh1MAp9A?usp=sharing

I had also experimented with various state of the art models to generate scores which can be found at

> https://colab.research.google.com/drive/1117iKWm6Vju8yVPFHq1s_nxpkVCbS0Cq?usp=sharing