**Dataset**: I am working on HOTPOTQA which is a dataset for diverse, explainable multi-hop question answering. It is a question-answering dataset featuring natural, multi-hop questions, with strong supervision for supporting facts to enable more explainable question answering systems. Each question has supporting\_facts, level, question, context, answer, id, type.

**STEPS for Dataset Creation:**

Handled data preprocessing cleaning, data parsing, data preprocessing for HOTPOT QA.

* **Data Preprocessing & Cleaning phase:**

The JSON file with the number of questions indicated by start and end is fed as input to the data frame generation function, the key columns question-id,passage-id, question, passage, sentence,passage\_length,sia\_score from the JSON data are extracted. Next, the sentences from every candidate passage of the context are extracted, while extracting the sentences from candidate passage, all the sentences with an empty string or less than two characters are considered as special cases and are not included as they are producing invalid scores in the score prediction phase.

* **Score Prediction Phase**: Researched, tested, verified pretrained model ‘distilbert-base-nli-mean-tokens' & trained models like Bert, Roberta, etc. Among all, pretrained, fine-tuned Web BERT-based semantic text similarity model trained on the STS-B dataset is producing good results for the HOTPOT QA dataset, for the STSbenchmark, the sentence pairs annotated together with a score range [0-5] indicating the similarity. This model modifies PyTorch transformers by abstracting away all the research benchmarking code for ease of real-world applicability. In the score generation phase, the data frame with “question+answer”, “sentence” is fed as input to the Web BERT-based semantic text similarity model to generate scores.
* **Data Processing Phase:** After generating the scores for each sentence of every candidate passage, processed the data to generate an SIA score for candidate passage in the context for every question by taking the max of SIA scores of all sentences, saved all the results in .csv format

**Phase1**: I have created the following data frame from the HOTPOT QA dataset in the first phase

|  |  |  |  |
| --- | --- | --- | --- |
| question-id | passage-id | question | passage |
| sentence | passage\_length | sia\_score |  |

Question-id: unique id for each question

Passage-id: candidate passage id of each question

Question: question+exact answer

Passage: candidate passage

Sentence: each sentence from candidate passage

Passage\_length: length of the candidate passage, required for calculating the average SIA score for the candidate passage

Sia\_score - SIA score between question and each sentence of the candidate passage

**Phase2**: I have computed the average SIA score for each candidate passage by representing them in the following column format in a data frame

|  |  |  |
| --- | --- | --- |
| question | sentence | sia\_score |

Question: question+exact answer

Sentence: candidate passage

sia\_score: average SIA score of the candidate passage

Consider the following sample data:

[{"supporting\_facts": [["Arthur's Magazine", 0], ["First for Women", 0]],

"level": "medium",

"question": "Which magazine was started first Arthur's Magazine or First for Women?",

"context": [ ["Radio City (Indian radio station)", ["Radio City is India's first private FM radio station and was started on 3 July 2001.", " It broadcasts on 91.1 (earlier 91.0 in most cities) megahertz from Mumbai (where it was started in 2004), Bengaluru (started first in 2001), Lucknow and New Delhi (since 2003).", " It plays Hindi, English and regional songs.", " It was launched in Hyderabad in March 2006, in Chennai on 7 July 2006 and in Visakhapatnam October 2007.", " Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music-related news, videos, songs, and other music-related features.", " The Radio station currently plays a mix of Hindi and Regional music.", " Abraham Thomas is the CEO of the company."] ],

………………..

……………….

"answer": "Arthur's Magazine",

"\_id": "5a7a06935542990198eaf050",

"type": "comparison"}

From this, I have extracted question, answer, set of candidate passages from the context, again from the context, I have extracted sentences and computed STS scores between <Q+A> and every <sentence> of each candidate passage, finally calculated the SIA score ranging from 0 to 5 for each candidate passage in the context by taking the maximum of sentence SIA scores.

**The final result format:**

|  |  |
| --- | --- |
|  |  |

**Instructions to run the code:**

All the executions are aligned sequentially in Jupyter notebook

1. Run the Necessary Installations section
2. Run the data preprocessing section
3. Run the scores generation section
4. Run the data processing section
5. Please ignore the commented code

**User Controls:**

* Look for the variable ‘FOLDER’, update the value to the main folder path where the input\_data.json file exists
* Change the variables ‘start’ and ‘end’ to control data size
* The output CSV files will be generated in the same ‘FOLDER’

**Analysis and Experiments:**

**Experiment1**: In the beginning, I computed the score for the candidate passage by taking the average SIA scores of all the sentences but it was not giving good scores overall. Hence used the maximum SIA score of all sentences for candidate passage.



**Experiment1 with average SIA score of all sentences for a candidate passage**

**Experiment2**: I tried to use question + top 3 key phrases from the supporting facts + the exact answer. It was not giving good results overall.

|  |  |
| --- | --- |
|  |  |

**Experiment2 with keyphrases.**

**Experiment3: Without keyphrases- clinical model**

**In this experiment, I tried using the clinical model, it was giving little higher scores for most of the clinical data for the considered samples.**

|  |  |
| --- | --- |
|  |  |

**Experiment:3 Clinical model without keyphrases**

Also uploaded train, test, dev data and the models used while experimenting and training is available at

[**https://drive.google.com/drive/folders/17jukmmppKi\_9HhU8ehICbpbNPDkSL8Oc?usp=sharing**](https://drive.google.com/drive/folders/17jukmmppKi_9HhU8ehICbpbNPDkSL8Oc?usp=sharing)

and the experimental code is commented

The combined dataset is at <https://github.com/JainSahit/NLP576-SIA>

**References for HOTPOTQA**

Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein et al. "Natural questions: a benchmark for question answering research." *Transactions of the Association for Computational Linguistics* 7 (2019): 453-466.

**Model** :

1) <https://github.com/AndriyMulyar/semantic-text-similarity/tree/master/semantic_text_similarity>

2) 'Distilbert-base-nli-mean-tokens' model

<https://www.sbert.net/docs/usage/semantic_textual_similarity.html>

<https://github.com/UKPLab/sentence-transformers>