CSE 576 Natural Language Processing
Synthetic Dataset Creation

Rajasree Chennupati – 1218519276 – rchennu1@asu.edu

Dataset:

The dataset I am working in this is Open Book QA dataset. The dataset QA is provided by Allen AI has a collection of labeled pairs of Question Answers. The data is encoded in JSON lines format. More details about this dataset is present at https://allenai.org/data/open-book-qa.

The dataset includes 3 jsonl files namely dev_complete.jsonl, dev_complete.jsonl, dev_complete.jsonl and 1 txt file namely crowdsourced-facts.txt file.

The columns in my dataset includes:

a. Question_Id
b. Question_stem
c. Question_text (There are 4 text for each choice)
d. Fact
e. humanScore
f. clarity
g. turkIdAnonymized
h. answerKey

There are total of 4957 questions in the dataset. Each question has 4 choices associated with it. Each question has an exact answer and a fact which is helpful in answering the question. The questions were created via a multi-stage crowdsourcing and partial expert filtering process

Preprocessing the data:

Initially the data is present in jsonl file and one example of data is given below as :

{"id": "7-980", "question":
{"stem": "The sun is responsible for", "choices":
[{"text": "puppies learning new tricks", "label": "A"},
{"text": "children growing up and getting old", "label": "B"},
{"text": "flowers wilting in a vase", "label": "C"},
{"text": "plants sprouting, blooming and wilting", "label": "D"}]},
"fact1": "the sun is the source of energy for physical cycles on Earth",
"humanScore": "1.00", "clarity": "2.00", "turkIdAnonymized": "b356d338b7",
"answerKey": "D"
}

I loaded the jsonl file and loaded the data into corresponding lists. With the help of these lists I generated exact_answercandidates list which has the question and the corresponding exact answer. To generate more sentences, I used BM25 Model which is useful in generating the text from corpus. manner I generated top2 relevant choices for each question in the dataset by mapping with the corpus. I formatted all the data from questions and BM25 generated facts such that each question will have 9 facts associated with it. Top 2 facts for each choice in the question with BM25 model with including the fact given in the question with the dataset.

SIA score Generation:

To generate SIA score I used approach 2 by using Semantic textual similarity between Question + Exact Answer with the facts.

To compute semantic textual Similarity, I tried with multiple models like BERT Model (Mihir generated this model with LR), RobertA Model and a pretrained model Web STS Bert (pretrained with STS-B) and Clinical based Web STS model. I passed the Question + Exact Answer, Answer as input to the model and tried to generate the SIA score. I additionally tried using Question + Fact, Answer also an input to the model. Among all the models I tried I got better results with Web STS Bert Model.

After passing the input as Question +Exact Answer, Sentence as input to the model, the following are the columns present in my dataset after generating the SIA score:

a. Sentence_1: Question in the dataset
b. Sentence_2: Facts corresponding to the question
c. Score: SIA score between Sentence_1 and Sentence_2

Manual Evaluation:

Question	Answer	Exact Answer	Key phrase1	Alignment1	Text1	Score1	Key Phrase2	Alignment2	Text2	Score2	Avg_score
The sun is responsible for	strawberries are plants and plants are producers	plants sprouting, blooming and wilting	sun	producer	SIMI	1	responsible	plants	SIMI	2	1.5
The sun is responsible for	grass and plants get wet with dew	plants sprouting, blooming and wilting	sun		NOALI	0	responsible	plants	SIMI	2	1
The sun is responsible for	the sun is the source of energy for physical cycles on Earth	plants sprouting, blooming and wilting	sun	sun	EQUI	4	responsible		NOALI	0	2
Stars are	stars are billions of miles away	great balls of gas burning billions of miles away	stars	stars	EQUI	4	are	billions	ANS	2	3
Stars are	balloons and balls are gas filled	great balls of gas burning billions of miles away	stars				are	balloons and gas	ANS	3.5	1.75
Stars are	the sky is bright in sunny weather	great balls of gas burning billions of miles away	stars	sky	SIMI	2	are		NOALI	0	1
Stars are	the yellow dwarf in the sky is our sun	great balls of gas burning billions of miles away	stars	sun	EQUI	2	are		NOALI	0	1
Stars are	a star is made of gases	great balls of gas burning billions of miles away	stars	stars	EQUI	4	are	gas	ANS	2	3
an inherited characteristic found on all mammals is	an animal's fur forms their coat	fur	mammal	animal	SIMI	2	characteristic	fur	ANS	4	3
an inherited characteristic found on all mammals is	some hares grow longer fur for winter	fur	mammal	hare	SIMI	1	characteristic	fur	ANS	4	2.5
an inherited characteristic found on all mammals is	the colors of the parts of an organism are inherited characteristics	fur	mammal		NOALI	0	characteristic	characteristic	EQUI	3	1.5

Instructions to run the code:

The code to the dataset generation is included in the repo here:
https://colab.research.google.com/drive/1lL0s8TeG2IWSrG22cIA46vjBnXFIPb_s?authuser=1#scrollTo=ueJkcg1fYaYv

The first section consists of all necessary libraries required to run the code.
In the second section I performed the data preprocessing taking the jsonl format of the questions and got the data in the format of 'Sentence_1, Sentence_2'.
In the last section I used Web Bert Model and passed the data to generate the SIA scores.

Also I tried with Roberta-Model and BERT-Model which are included in the next sections.The link to these models are present here.
https://drive.google.com/drive/u/1/folders/1-ECcrDRqkziVcq_JiyQXB5ETRTXvdSWD
https://drive.google.com/drive/u/1/folders/1-1bwDIK2rZ0BERCLIGJ1VbsrZxY5-xdI

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