CSE 576 Natural Language Processing

Project Phase 2 – Automated Data Creation

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

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**.

Structure and Details of the Dataset

Here is the link to the paper describing the Dataset https://arxiv.org/pdf/1910.11473.pdf and the dataset can be found at https://github.com/allenai/qasc .

QASC is a multi-hop question-answering dataset with a focus on sentence composition. It consists of 9,980 8-way multiple-choice questions about grade school science. It comes with a corpus of 17M sentences. It requires retrieving facts from a large corpus and composing them to answer a multiple-choice question.

It the dataset is split as:

- **Training:** 8134 questions
- **Validation:** 926 questions
- **Testing:** 920 questions

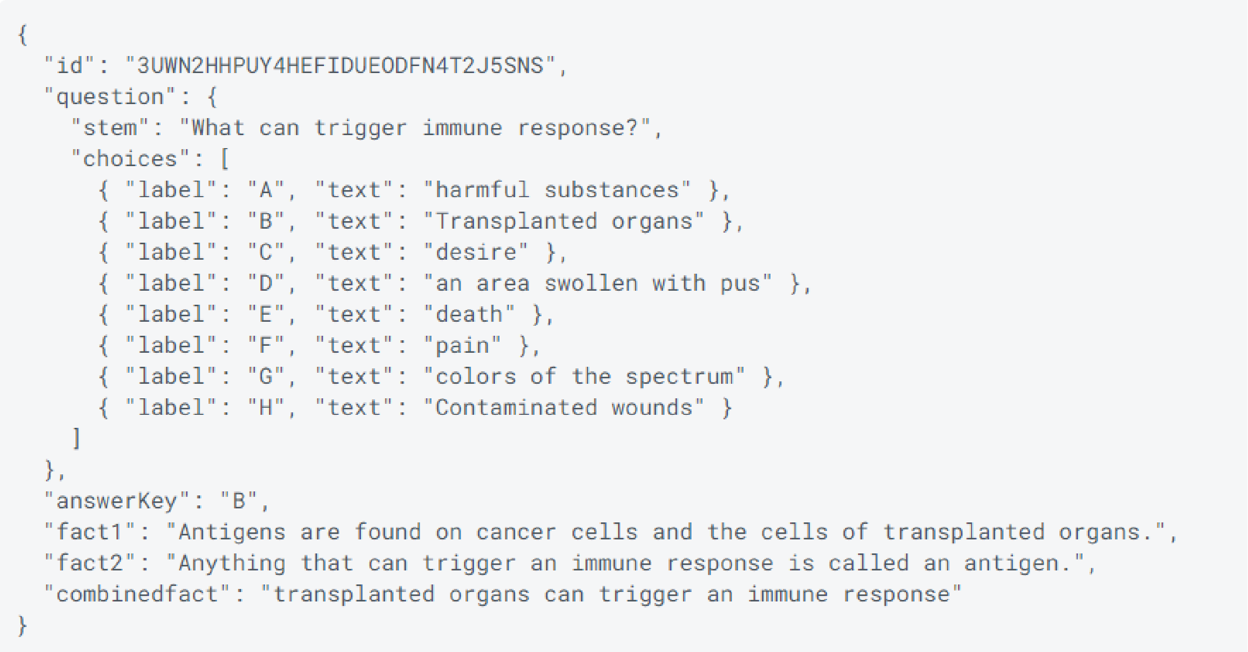
Each **train, validation example** consists of a question with **8 answers candidates** as options, followed by **two supporting facts** named as f_S and f_L and **a composed fact** $f_{C,}$ composed from f_S and f_L using the broad knowledge that is used to answer the question.

All questions in QASC are human-authored, obtained via a multi-step crowdsourcing process. To better enable development of both the reasoning and retrieval models, the pair of facts that were composed to create the question are also provided. These annotations are used to develop a novel two-step retrieval technique that uses question-relevant facts to guide a second retrieval step.

The corpus consists of 17 Million facts.

Pre-Processing the Dataset

Initially the training and the validation set, which are in ‘jsonl’ format are loaded into the script.



```
{
  "id": "3UWN2HHPUY4HEFIDUEODFN4T2J5SNS",
  "question": {
    "stem": "What can trigger immune response?",
    "choices": [
      { "label": "A", "text": "harmful substances" },
      { "label": "B", "text": "Transplanted organs" },
      { "label": "C", "text": "desire" },
      { "label": "D", "text": "an area swollen with pus" },
      { "label": "E", "text": "death" },
      { "label": "F", "text": "pain" },
      { "label": "G", "text": "colors of the spectrum" },
      { "label": "H", "text": "Contaminated wounds" }
    ]
  },
  "answerKey": "B",
  "fact1": "Antigens are found on cancer cells and the cells of transplanted organs.",
  "fact2": "Anything that can trigger an immune response is called an antigen.",
  "combinedfact": "transplanted organs can trigger an immune response"
}
```

Figure 1. Structure of a Question in training and Validation Set

Each question, answer candidate, fact1, fact2 and combined fact are extracted and are appended to corresponding list. The answer candidate corresponding to the answer key is considered as the exact answer.

For every answer candidate a fact is retrieved from the corpus using the Anserini Information Retrieval toolkit built on Lucene. Also, the fact1, fact2 and the combined fact are pushed into answer candidate, hence every Question(query) has 11 corresponding answer candidates for training and validation set.

Pyserini

Pyserini is a python interface to the Anserini IR toolkit built on Lucene. **Okapi BM25** (best matching) is a ranking function used by search engines to estimate the relevance of documents to a given search query. It is based on the probabilistic retrieval framework.

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of their proximity within the document. It is a family of scoring functions with slightly different components and parameters. One of the most prominent instantiations of the function is as follows.

Given a query Q, containing keywords q_{1},...,q_{n}, the BM25 score of a document D is:

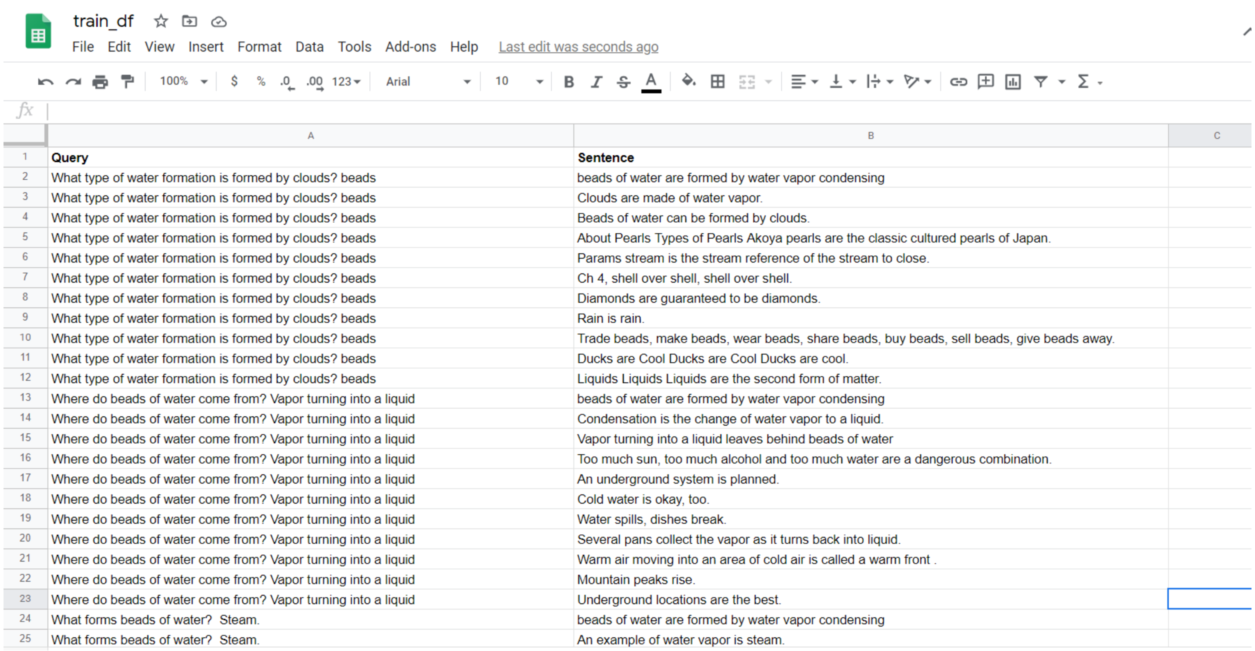
$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

where **$f(q_i,$ $D)$** is q_i's term frequency in the document *D*, **|D|** is the length of the document *D* in words, and **avgdl** is the average document length in the text collection from which documents are drawn. **k1** and **b** are free parameters, usually chosen, in absence of an advanced optimization as K1belongs to [1.2,2.0] and b-0.75. is the **IDF** (inverse document frequency) weight of the query term . It is usually computed as:

$$\mathrm{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i .

Processed Data frame for both train and validation set looks like this:



	A	B	C
1	**Query**	**Sentence**	
2	What type of water formation is formed by clouds? beads	beads of water are formed by water vapor condensing	
3	What type of water formation is formed by clouds? beads	Clouds are made of water vapor.	
4	What type of water formation is formed by clouds? beads	Beads of water can be formed by clouds.	
5	What type of water formation is formed by clouds? beads	About Pearls Types of Pearls Akoya pearls are the classic cultured pearls of Japan.	
6	What type of water formation is formed by clouds? beads	Params stream is the stream reference of the stream to close.	
7	What type of water formation is formed by clouds? beads	Ch 4, shell over shell, shell over shell.	
8	What type of water formation is formed by clouds? beads	Diamonds are guaranteed to be diamonds.	
9	What type of water formation is formed by clouds? beads	Rain is rain.	
10	What type of water formation is formed by clouds? beads	Trade beads, make beads, wear beads, share beads, buy beads, sell beads, give beads away.	
11	What type of water formation is formed by clouds? beads	Ducks are Cool Ducks are Cool Ducks are cool.	
12	What type of water formation is formed by clouds? beads	Liquids Liquids Liquids are the second form of matter.	
13	Where do beads of water come from? Vapor turning into a liquid	beads of water are formed by water vapor condensing	
14	Where do beads of water come from? Vapor turning into a liquid	Condensation is the change of water vapor to a liquid.	
15	Where do beads of water come from? Vapor turning into a liquid	Vapor turning into a liquid leaves behind beads of water	
16	Where do beads of water come from? Vapor turning into a liquid	Too much sun, too much alcohol and too much water are a dangerous combination.	
17	Where do beads of water come from? Vapor turning into a liquid	An underground system is planned.	
18	Where do beads of water come from? Vapor turning into a liquid	Cold water is okay, too.	
19	Where do beads of water come from? Vapor turning into a liquid	Water spills, dishes break.	
20	Where do beads of water come from? Vapor turning into a liquid	Several pans collect the vapor as it turns back into liquid.	
21	Where do beads of water come from? Vapor turning into a liquid	Warm air moving into an area of cold air is called a warm front .	
22	Where do beads of water come from? Vapor turning into a liquid	Mountain peaks rise.	
23	Where do beads of water come from? Vapor turning into a liquid	Underground locations are the best.	
24	What forms beads of water? Steam.	beads of water are formed by water vapor condensing	
25	What forms beads of water? Steam.	An example of water vapor is steam.	

Figure 2. Data frame for training set after extracting facts from corpus

Where each query is a combination Question and Exact Answer <Q+E_A>, Sentence are nothing but the fact1, fact2, combined fact and the fact retrieved from the corpus corresponding to each answer candidate from 8 options. Hence every question has 11 answer candidates.

Semantic Information Availability (SIA) Score Generation

Since there are not any existing Natural Language Processing models that can directly generate SIA scores, hence I had suggested to utilize existing State of the art (SOTA) Semantic Textual Similarity (STS) model to generate the STS scores.

Now as STS scores are generally in range [0, 5] however SIA scores fall in range [0, 4], the generated STS scores are converted to the range [0, 4] and are labelled as the SIA scores (gold Label) for the dataset.

Web STS Bert

Web STS Bert is an easy-to-use interface to fine-tune BERT models for computing semantic similarity, it contains and interface to fine-tuned BERT based semantic text similarity models. It modifies the pytorch-transformers by abstracting away all the research benchmarking code for ease of real-world applicability.

Web STS BERT was pretrained on the STS-B dataset and has a Pearson correlation of 0.893.

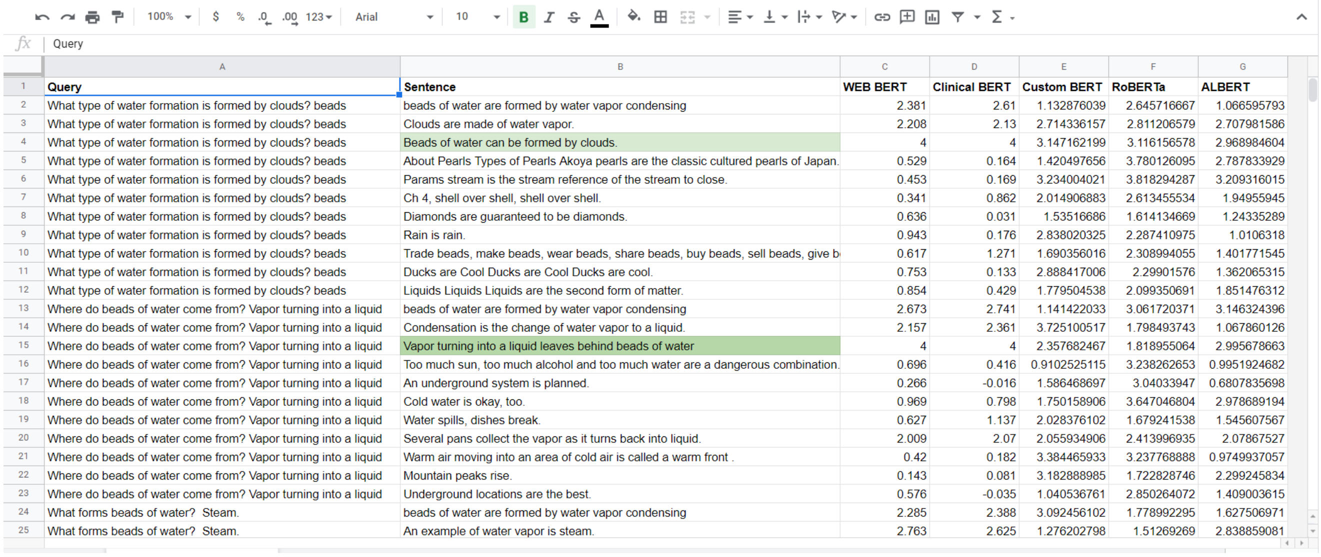
The generated scores for each Query, Sentence pair from the training and the validation set is in the range [0, 5] they converted to [0,4].

Experimenting with Commercially available STS Models

The following models were taken into consideration before sticking to Web Bert Model, to identify the best performing model for QASC Dataset.

1. ALBERT
2. RoBERTa
3. Custom BERT model pre trained on STS-D dataset with a Linear output Layer to predict STS score.
4. Clinical STS BERT
5. Web STS BERT

Comparison of Scores Generated by various considered Model



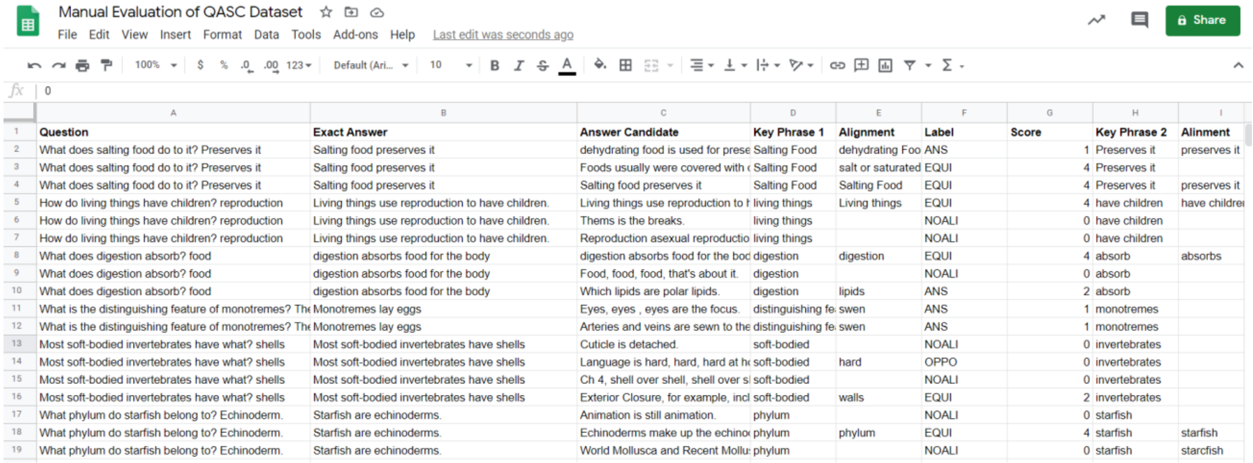
	A	B	C	D	E	F	G
1	**Query**	**Sentence**	**WEB BERT**	**Clinical BERT**	**Custom BERT**	**RoBERTa**	**ALBERT**
2	What type of water formation is formed by clouds? beads	beads of water are formed by water vapor condensing	2.381	2.61	1.132876039	2.645716667	1.066595793
3	What type of water formation is formed by clouds? beads	Clouds are made of water vapor.	2.208	2.13	2.714336157	2.811206579	2.707981586
4	What type of water formation is formed by clouds? beads	Beads of water can be formed by clouds.	4	4	3.147162199	3.116156578	2.968984604
5	What type of water formation is formed by clouds? beads	About Pearls Types of Pearls Akoya pearls are the classic cultured pearls of Japan.	0.529	0.164	1.420497656	3.780126095	2.787833929
6	What type of water formation is formed by clouds? beads	Params stream is the stream reference of the stream to close.	0.453	0.169	3.234004021	3.818294287	3.209316015
7	What type of water formation is formed by clouds? beads	Ch 4, shell over shell, shell over shell.	0.341	0.862	2.014906883	2.613455534	1.94955945
8	What type of water formation is formed by clouds? beads	Diamonds are guaranteed to be diamonds.	0.636	0.031	1.53516686	1.614134669	1.24335289
9	What type of water formation is formed by clouds? beads	Rain is rain.	0.943	0.176	2.838020325	2.287410975	1.0106318
10	What type of water formation is formed by clouds? beads	Trade beads, make beads, wear beads, share beads, buy beads, sell beads, give b	0.617	1.271	1.690356016	2.308994055	1.401771545
11	What type of water formation is formed by clouds? beads	Ducks are Cool Ducks are Cool Ducks are cool.	0.753	0.133	2.888417006	2.29901576	1.362065315
12	What type of water formation is formed by clouds? beads	Liquids Liquids Liquids are the second form of matter.	0.854	0.429	1.779504538	2.099350691	1.851476312
13	Where do beads of water come from? Vapor turning into a liquid	beads of water are formed by water vapor condensing	2.673	2.741	1.141422033	3.061720371	3.146324396
14	Where do beads of water come from? Vapor turning into a liquid	Condensation is the change of water vapor to a liquid.	2.157	2.361	3.725100517	1.798493743	1.067860126
15	Where do beads of water come from? Vapor turning into a liquid	Vapor turning into a liquid leaves behind beads of water	4	4	2.357682467	1.818955064	2.995678663
16	Where do beads of water come from? Vapor turning into a liquid	Too much sun, too much alcohol and too much water are a dangerous combination.	0.696	0.416	0.9102525115	3.238262653	0.9951924682
17	Where do beads of water come from? Vapor turning into a liquid	An underground system is planned.	0.266	-0.016	1.586468697	3.04033947	0.6807835698
18	Where do beads of water come from? Vapor turning into a liquid	Cold water is okay, too.	0.969	0.798	1.750158906	3.647046804	2.978689194
19	Where do beads of water come from? Vapor turning into a liquid	Water spills, dishes break.	0.627	1.137	2.028376102	1.679241538	1.545607567
20	Where do beads of water come from? Vapor turning into a liquid	Several pans collect the vapor as it turns back into liquid.	2.009	2.07	2.055934906	2.413996935	2.07867527
21	Where do beads of water come from? Vapor turning into a liquid	Warm air moving into an area of cold air is called a warm front .	0.42	0.182	3.384465933	3.237768888	0.9749937057
22	Where do beads of water come from? Vapor turning into a liquid	Mountain peaks rise.	0.143	0.081	3.182888985	1.722828746	2.299245834
23	Where do beads of water come from? Vapor turning into a liquid	Underground locations are the best.	0.576	-0.035	1.040536761	2.850264072	1.409003615
24	What forms beads of water? Steam.	beads of water are formed by water vapor condensing	2.285	2.388	3.092456102	1.778992295	1.627506971
25	What forms beads of water? Steam.	An example of water vapor is steam.	2.763	2.625	1.276202798	1.51269269	2.838859081

Figure 3. Comparison of scores predicted by different Models

Upon close observation I could refer that **Web BERT and Clinical BERT** were performing for more better than all other models, however there where instances where Clinical BERT was calculating negative scores. **Web BERT Model** showed consistency over then entire Dataset.

Manual Evaluation of the Predicted Scores

20 randomly picked query and sentence pair from the train set, and were manually evaluated to calculate the STS score by extracting the key phrase in the question and corresponding terms were identified in the sentence (Answer Candidate). The predicted scores were similar to that of the scores predicted by the Web STS model.


Manual Evaluation of QASC Dataset
File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago
Share
100%
Default (Ari...
10
fx 0

	A	B	C	D	E	F	G	H	I
1	**Question**	**Exact Answer**	**Answer Candidate**	**Key Phrase 1**	**Alignment**	**Label**	**Score**	**Key Phrase 2**	**Alinment**
2	What does salting food do to it? Preserves it	Salting food preserves it	dehydrating food is used for prese	Salting Food	dehydrating Foo	ANS	1	Preserves it	preserves it
3	What does salting food do to it? Preserves it	Salting food preserves it	Foods usually were covered with (	Salting Food	salt or saturated	EQUI	4	Preserves it	
4	What does salting food do to it? Preserves it	Salting food preserves it	Salting food preserves it	Salting Food	Salting Food	EQUI	4	Preserves it	preserves it
5	How do living things have children? reproduction	Living things use reproduction to have children.	Living things use reproduction to h	living things	Living things	EQUI	4	have children	have childre
6	How do living things have children? reproduction	Living things use reproduction to have children.	Thems is the breaks.	living things		NOALI	0	have children	
7	How do living things have children? reproduction	Living things use reproduction to have children.	Reproduction asexual reproductio	living things		NOALI	0	have children	
8	What does digestion absorb? food	digestion absorbs food for the body	digestion absorbs food for the bod	digestion	digestion	EQUI	4	absorb	absorbs
9	What does digestion absorb? food	digestion absorbs food for the body	Food, food, food, that's about it.	digestion		NOALI	0	absorb	
10	What does digestion absorb? food	digestion absorbs food for the body	Which lipids are polar lipids.	digestion	lipids	ANS	2	absorb	
11	What is the distinguishing feature of monotremes? The	Monotremes lay eggs	Eyes, eyes , eyes are the focus.	distinguishing fe	swen	ANS	1	monotremes	
12	What is the distinguishing feature of monotremes? The	Monotremes lay eggs	Arteries and veins are sewn to the	distinguishing fe	swen	ANS	1	monotremes	
13	Most soft-bodied invertebrates have what? shells	Most soft-bodied invertebrates have shells	Cuticle is detached.	soft-bodied		NOALI	0	invertebrates	
14	Most soft-bodied invertebrates have what? shells	Most soft-bodied invertebrates have shells	Language is hard, hard, hard at h	soft-bodied	hard	OPPO	0	invertebrates	
15	Most soft-bodied invertebrates have what? shells	Most soft-bodied invertebrates have shells	Ch 4, shell over shell, shell over sl	soft-bodied		NOALI	0	invertebrates	
16	Most soft-bodied invertebrates have what? shells	Most soft-bodied invertebrates have shells	Exterior Closure, for example, incl	soft-bodied	walls	EQUI	2	invertebrates	
17	What phylum do starfish belong to? Echinoderm.	Starfish are echinoderms.	Animation is still animation.	phylum		NOALI	0	starfish	
18	What phylum do starfish belong to? Echinoderm.	Starfish are echinoderms.	Echinoderms make up the echino	phylum	phylum	EQUI	4	starfish	starfish
19	What phylum do starfish belong to? Echinoderm.	Starfish are echinoderms.	World Mollusca and Recent Mollu	phylum		NOALI	0	starfish	starcfish

Figure 4. Manual Evaluation

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. Pyserini_and_Data_PreProcessing.ipynb

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

2. SIA-Scores-Generation-Using-WEB-BERTandClinical-BERT.ipynb

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

3. Generating-Dataset-Using-Preprocessed-Huggingface-Models.ipynb

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