Final Project Write-up

CMSC470 - Natural Language Processing, Spring 2019

Jake Bremerman, Megane Crenshaw, Samuel Gollob, and Asher Moldwin

Submission Folder: Submission

Introduction

For our project we wanted to see if trying to imitate human behavior would improve the performance of question answering models. We noticed that during the expo matches against humans, the machine struggled more on certain types of questions, such as common link questions. They also often had guesses that were of a completely different category than the answer, which a human would never guess. We believed that by imitating human behavior our model would do better on these types of questions. At the very least we hoped it would nudge the model in the right direction, and it would guess answers that were closer to the right answer than a model that didn't incorporate human behavior.

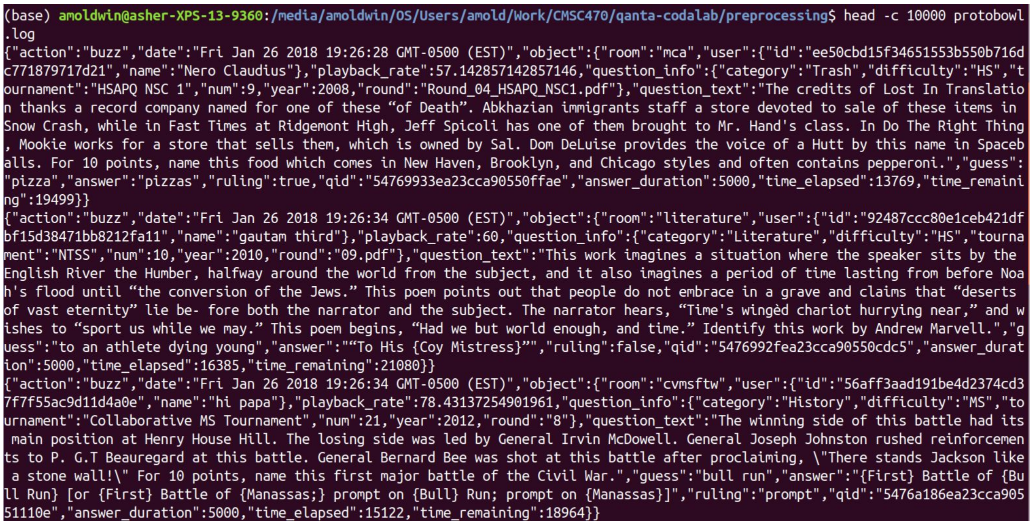
In order to get our model to imitate humans, we trained using data that contained human answers from Protobowl, since Qanta data doesn't include human responses. We also modified our loss function in order to punish our model less when it got a wrong answer that was also an answer given by humans.

Motivation: Noise Injection Techniques from Robotics

As far as we could tell, this kind of intentional use of "poorly annotated" data (in our case the annotations would be non-expert Quizbowl guesses) has not been tried for NLP or question-answering tasks. However, we found an article titled "*DART: Noise Injection for Robust Imitation Learning*" which described a similar technique in the field of robotic imitation learning. The authors of this paper describe the challenge of correcting machine-specific errors, due to the fact that there is often a lack of relevant training data. To guide the model toward better performance in these situations, specific forms of noise can be injected into the training data.

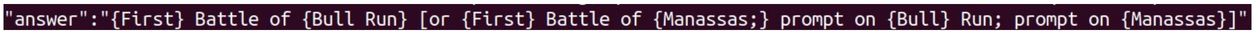
Preprocessing

In order to implement this idea, we needed to preprocess the protobowl log file to access the human guesses when training our model. This was initially challenging because the protobowl file included separate entries for each person who buzzed in on every question, making the file over 5gb. In addition, the human submitted guesses were frequently misspelled and almost never matched the official answer classes exactly:



```
(base) amoldwin@asher-XPS-13-9360:/media/amoldwin/OS/Users/amold/Work/CMSC470/qanta-codalab/preprocessing$ head -c 10000 protobowl
.log
{"action":"buzz","date":"Fri Jan 26 2018 19:26:28 GMT-0500 (EST)","object":{"room":"mca","user":{"id":"ee50cbd15f34651553b550b716d
c771879717d21","name":"Nero Claudius"},"playback_rate":57.142857142857146,"question_info":{"category":"Trash","difficulty":"HS","t
ournament":"HSAPQ NSC 1","num":9,"year":2008,"round":"Round_04_HSAPQ_NSC1.pdf"},"question_text":"The credits of Lost In Translatio
n thanks a record company named for one of these “of Death”. Abkhazian immigrants staff a store devoted to sale of these items in
Snow Crash, while in Fast Times at Ridgemont High, Jeff Spicoli has one of them brought to Mr. Hand's class. In Do The Right Thing
, Mookie works for a store that sells them, which is owned by Sal. Dom DeLuise provides the voice of a Hutt by this name in Spaceb
alls. For 10 points, name this food which comes in New Haven, Brooklyn, and Chicago styles and often contains pepperoni.","guess":
"pizza","answer":"pizzas","ruling":true,"qid":"54769933ea23cca90550ffae","answer_duration":5000,"time_elapsed":13769,"time_remaini
ng":19499}}
{"action":"buzz","date":"Fri Jan 26 2018 19:26:34 GMT-0500 (EST)","object":{"room":"literature","user":{"id":"92487ccc80e1ceb421df
bf15d38471bb8212fa11","name":"gautam third"},"playback_rate":60,"question_info":{"category":"Literature","difficulty":"HS","tourna
ment":"NTSS","num":10,"year":2010,"round":"09.pdf"},"question_text":"This work imagines a situation where the speaker sits by the
English River the Humber, halfway around the world from the subject, and it also imagines a period of time lasting from before Noa
h's flood until “the conversion of the Jews.” This poem points out that people do not embrace in a grave and claims that “deserts
of vast eternity” lie be- fore both the narrator and the subject. The narrator hears, “Time's wingèd chariot hurrying near,” and w
ishes to “sport us while we may.” This poem begins, “Had we but world enough, and time.” Identify this work by Andrew Marvell.","g
uess":"to an athlete dying young","answer":"“To His {Coy Mistress}”","ruling":false,"qid":"5476992fea23cca90550cdc5","answer_durat
ion":5000,"time_elapsed":16385,"time_remaining":21080}}
{"action":"buzz","date":"Fri Jan 26 2018 19:26:34 GMT-0500 (EST)","object":{"room":"cvmsftw","user":{"id":"56aff3aad191be4d2374cd3
7f7f55ac9d11d4a0e","name":"hi papa"},"playback_rate":78.43137254901961,"question_info":{"category":"History","difficulty":"MS","to
urnament":"Collaborative MS Tournament","num":21,"year":2012,"round":"8"},"question_text":"The winning side of this battle had its
 main position at Henry House Hill. The losing side was led by General Irvin McDowell. General Joseph Johnston rushed reinforcemen
ts to P. G.T Beauregard at this battle. General Bernard Bee was shot at this battle after proclaiming, \"There stands Jackson like
 a stone wall!\" For 10 points, name this first major battle of the Civil War.","guess":"bull run","answer":"{First} Battle of {Bu
ll Run} [or {First} Battle of {Manassas;} prompt on {Bull} Run; prompt on {Manassas}]","ruling":"prompt","qid":"5476a186ea23cca905
51110e","answer_duration":5000,"time_elapsed":15122,"time_remaining":18964}}
```

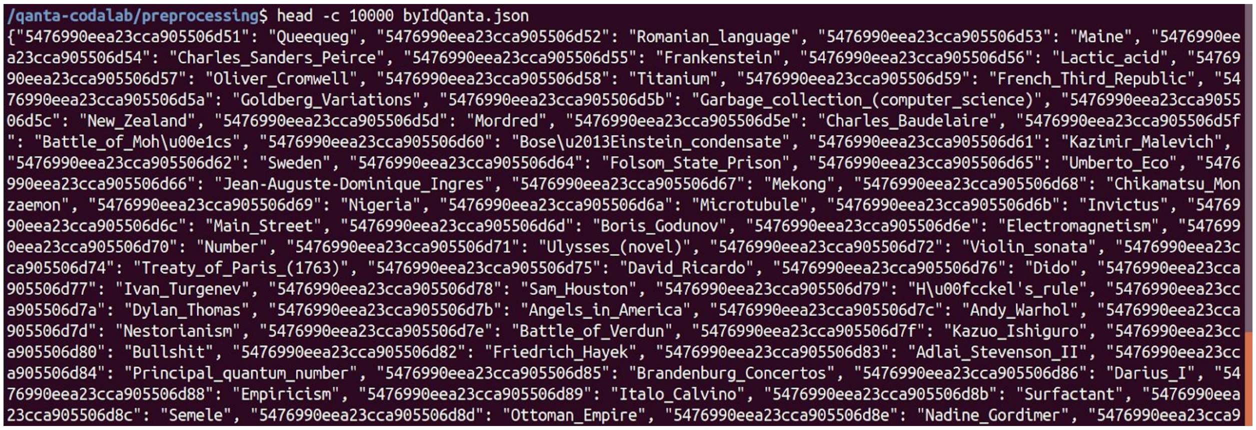
The “official” answers in the protobowl file also were in a different form than we needed, with brackets and semicolons seemingly indicating what should be marked correct or prompted, but not in a code that was intelligible to us:



```
"answer":"{First} Battle of {Bull Run} [or {First} Battle of {Manassas;} prompt on {Bull} Run; prompt on {Manassas}]"
```

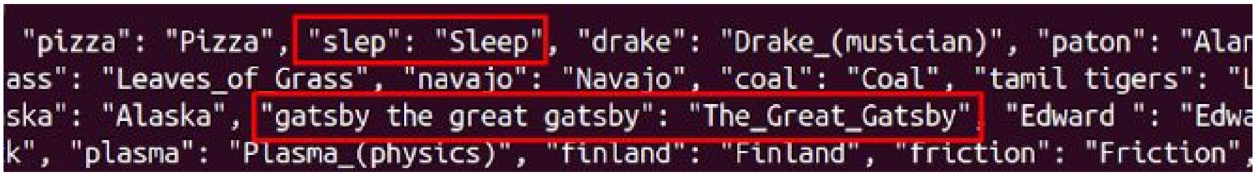
We attempted to search online to see if we could access a function that would decode these answer-matching expressions for us and tell us whether a given human-input could be considered as correct for a given question, but were unable to find anything. We considered writing a code that would match these answers using RegEx but ultimately did not pursue this.

To cross-reference the protobowl data with the qanta dataset to see the answers as Wikipedia page titles, we created a json dictionary from the qanta.train dataset mapping each question’s Protobowl ID to the Wikipedia page for that question’s answer:



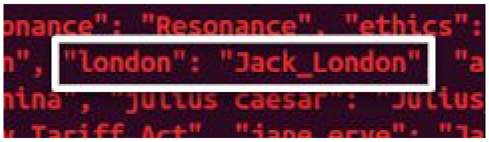
```
/qanta-codalab/preprocessing$ head -c 10000 byIdQanta.json
{"5476990eea23cca905506d51": "Queequeg", "5476990eea23cca905506d52": "Romanian_language", "5476990eea23cca905506d53": "Maine", "5476990ee
a23cca905506d54": "Charles_Sanders_Peirce", "5476990eea23cca905506d55": "Frankenstein", "5476990eea23cca905506d56": "Lactic_acid", "54769
90eea23cca905506d57": "Oliver_Cromwell", "5476990eea23cca905506d58": "Titanium", "5476990eea23cca905506d59": "French_Third_Republic", "54
76990eea23cca905506d5a": "Goldberg_Variations", "5476990eea23cca905506d5b": "Garbage_collection_(computer_science)", "5476990eea23cca9055
06d5c": "New_Zealand", "5476990eea23cca905506d5d": "Mordred", "5476990eea23cca905506d5e": "Charles_Baudelaire", "5476990eea23cca905506d5f
": "Battle_of_Moh\u00e1cs", "5476990eea23cca905506d60": "Bose\u2013Einstein_condensate", "5476990eea23cca905506d61": "Kazimir_Malevich",
"5476990eea23cca905506d62": "Sweden", "5476990eea23cca905506d64": "Folsom_State_Prison", "5476990eea23cca905506d65": "Umberto_Eco", "5476
990eea23cca905506d66": "Jean-Auguste-Dominique_Ingres", "5476990eea23cca905506d67": "Mekong", "5476990eea23cca905506d68": "Chikamatsu_Mon
zaemon", "5476990eea23cca905506d69": "Nigeria", "5476990eea23cca905506d6a": "Microtubule", "5476990eea23cca905506d6b": "Invictus", "54769
90eea23cca905506d6c": "Main_Street", "5476990eea23cca905506d6d": "Boris_Godunov", "5476990eea23cca905506d6e": "Electromagnetism", "547699
0eea23cca905506d70": "Number", "5476990eea23cca905506d71": "Ulysses_(novel)", "5476990eea23cca905506d72": "Violin_sonata", "5476990eea23c
ca905506d74": "Treaty_of_Paris_(1763)", "5476990eea23cca905506d75": "David_Ricardo", "5476990eea23cca905506d76": "Dido", "5476990eea23cca
905506d77": "Ivan_Turgenev", "5476990eea23cca905506d78": "Sam_Houston", "5476990eea23cca905506d79": "H\u00fcckel's_rule", "5476990eea23cc
a905506d7a": "Dylan_Thomas", "5476990eea23cca905506d7b": "Angels_in_America", "5476990eea23cca905506d7c": "Andy_Warhol", "5476990eea23cca
905506d7d": "Nestorianism", "5476990eea23cca905506d7e": "Battle_of_Verdun", "5476990eea23cca905506d7f": "Kazuo_Ishiguro", "5476990eea23cc
a905506d80": "Bullshit", "5476990eea23cca905506d82": "Friedrich_Hayek", "5476990eea23cca905506d83": "Adlai_Stevenson_II", "5476990eea23cc
a905506d84": "Principal_quantum_number", "5476990eea23cca905506d85": "Brandenburg_Concertos", "5476990eea23cca905506d86": "Darius_I", "54
76990eea23cca905506d88": "Empiricism", "5476990eea23cca905506d89": "Italo_Calvino", "5476990eea23cca905506d8b": "Surfactant", "5476990eea
23cca905506d8c": "Semele", "5476990eea23cca905506d8d": "Ottoman_Empire", "5476990eea23cca905506d8e": "Nadine_Gordimer", "5476990eea23cca9
```

Next, we mapped the human-submitted guesses to actual class labels, by mapping every guess which was ruled “correct” to the “true answer” for that question:



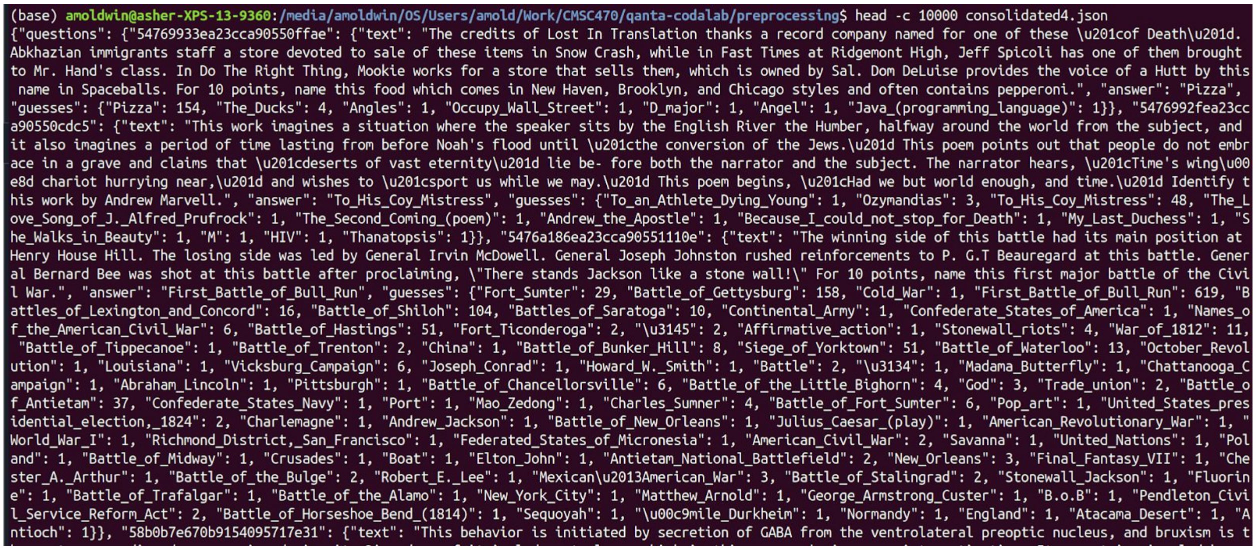
```
 "pizza": "Pizza", "slep": "Sleep", "drake": "Drake_(musician)", "paton": "Alan
ass": "Leaves_of_Grass", "navajo": "Navajo", "coal": "Coal", "tamil tigers": "L
ska": "Alaska", "gatsby the great gatsby": "The_Great_Gatsby", "Edward ": "Edwa
k", "plasma": "Plasma_(physics)", "finland": "Finland", "friction": "Friction",
```

This was a one-to-one mapping, so whatever a guess was interpreted as last became the answer class it was mapped to. This led to some problems when a human input could refer to more than one real answer:



```
onance": "Resonance", "ethics":
n", "london": "Jack_London" "a
ina", "julius caesar": "Julius
 Tariff Act", "jane eyre": "Ja
```

Finally, we created a consolidated version of our training data which included the guesses for each question as well as counts for each guess:



```
(base) amoldwin@asher-XPS-13-9360:/media/amoldwin/OS/Users/amold/Work/CMSC470/qanta-codalab/preprocessing$ head -c 10000 consolidated4.json
{"questions": {"54769933ea23cca90550ffae": {"text": "The credits of Lost In Translation thanks a record company named for one of these \u201cof Death\u201d.
Abkhazian immigrants staff a store devoted to sale of these items in Snow Crash, while in Fast Times at Ridgemont High, Jeff Spicoli has one of them brought
to Mr. Hand's class. In Do The Right Thing, Mookie works for a store that sells them, which is owned by Sal. Dom DeLuise provides the voice of a Hutt by this
 name in Spaceballs. For 10 points, name this food which comes in New Haven, Brooklyn, and Chicago styles and often contains pepperoni.", "answer": "Pizza",
"guesses": {"Pizza": 154, "The_Ducks": 4, "Angles": 1, "Occupy_Wall_Street": 1, "D_major": 1, "Angel": 1, "Java_(programming_language)": 1}}, "5476992fea23cc
a90550cdc5": {"text": "This work imagines a situation where the speaker sits by the English River the Humber, halfway around the world from the subject, and
it also imagines a period of time lasting from before Noah's flood until \u201cthe conversion of the Jews.\u201d This poem points out that people do not embr
ace in a grave and claims that \u201cdeserts of vast eternity\u201d lie be- fore both the narrator and the subject. The narrator hears, \u201cTime's wing\u00
e8d chariot hurrying near,\u201d and wishes to \u201csport us while we may.\u201d This poem begins, \u201cHad we but world enough, and time.\u201d Identify t
his work by Andrew Marvell.", "answer": "To_His_Coy_Mistress", "guesses": {"To_an_Athlete_Dying_Young": 1, "Ozymandias": 3, "To_His_Coy_Mistress": 48, "The_L
ove_Song_of_J._Alfred_Prufrock": 1, "The_Second_Coming_(poem)": 1, "Andrew_the_Apostle": 1, "Because_I_could_not_stop_for_Death": 1, "My_Last_Duchess": 1, "S
he_Walks_in_Beauty": 1, "M": 1, "HIV": 1, "Thanatopsis": 1}}, "5476a186ea23cca90551110e": {"text": "The winning side of this battle had its main position at
Henry House Hill. The losing side was led by General Irvin McDowell. General Joseph Johnston rushed reinforcements to P. G.T Beauregard at this battle. Gener
al Bernard Bee was shot at this battle after proclaiming, \"There stands Jackson like a stone wall!\" For 10 points, name this first major battle of the Civi
l War.", "answer": "First_Battle_of_Bull_Run", "guesses": {"Fort_Sumter": 29, "Battle_of_Gettysburg": 158, "Cold_War": 1, "First_Battle_of_Bull_Run": 619, "B
attles_of_Lexington_and_Concord": 16, "Battle_of_Shiloh": 104, "Battles_of_Saratoga": 10, "Continental_Army": 1, "Confederate_States_of_America": 1, "Names_o
f_the_American_Civil_War": 6, "Battle_of_Hastings": 51, "Fort_Ticonderoga": 2, "\u3145": 2, "Affirmative_action": 1, "Stonewall_riots": 4, "War_of_1812": 11,
 "Battle_of_Tippecanoe": 1, "Battle_of_Trenton": 2, "China": 1, "Battle_of_Bunker_Hill": 8, "Siege_of_Yorktown": 51, "Battle_of_Waterloo": 13, "October_Revol
ution": 1, "Louisiana": 1, "Vicksburg_Campaign": 6, "Joseph_Conrad": 1, "Howard_W._Smith": 1, "Battle": 2, "\u3134": 1, "Madama_Butterfly": 1, "Chattanooga_C
ampaign": 1, "Abraham_Lincoln": 1, "Pittsburgh": 1, "Battle_of_Chancellorsville": 6, "Battle_of_the_Little_Bighorn": 4, "God": 3, "Trade_union": 2, "Battle_o
f_Antietam": 37, "Confederate_States_Navy": 1, "Port": 1, "Mao_Zedong": 1, "Charles_Sumner": 4, "Battle_of_Fort_Sumter": 6, "Pop_art": 1, "United_States_pres
idential_election,_1824": 2, "Charlemagne": 1, "Andrew_Jackson": 1, "Battle_of_New_Orleans": 1, "Julius_Caesar_(play)": 1, "American_Revolutionary_War": 1, "
World_War_I": 1, "Richmond_District,_San_Francisco": 1, "Federated_States_of_Micronesia": 1, "American_Civil_War": 2, "Savanna": 1, "United_Nations": 1, "Pol
and": 1, "Battle_of_Midway": 1, "Crusades": 1, "Boat": 1, "Elton_John": 1, "Antietam_National_Battlefield": 2, "New_Orleans": 3, "Final_Fantasy_VII": 1, "Che
ster_A._Arthur": 1, "Battle_of_the_Bulge": 2, "Robert_E._Lee": 1, "Mexican\u2013American_War": 3, "Battle_of_Stalingrad": 2, "Stonewall_Jackson": 1, "Fluorin
e": 1, "Battle_of_Trafalgar": 1, "Battle_of_the_Alamo": 1, "New_York_City": 1, "Matthew_Arnold": 1, "George_Armstrong_Custer": 1, "B.o.B": 1, "Pendleton_Civi
l_Service_Reform_Act": 2, "Battle_of_Horseshoe_Bend_(1814)": 1, "Sequoyah": 1, "\u00c9mile_Durkheim": 1, "Normandy": 1, "England": 1, "Atacama_Desert": 1, "A
ntioch": 1}}, "58b0b7e670b9154095717e31": {"text": "This behavior is initiated by secretion of GABA from the ventrolateral preoptic nucleus, and bruxism is t
```

Our training data looked like above sample, except with the counts normalized as probabilities. We experimented with several versions of this, including a version where we had a threshold for a guess's frequency (as a percentage of the total answers) to be included in the dataset. We also used a version where correct guesses were ignored and left out of the normalization calculation.

Model Design

We needed to work with a neural network model to allow for a loss function to be incorporated so we could reward the system for guessing human guesses. We thought that an LSTM might be a good idea because of its linguistic relevance and way that it decides to remember and forget parts of a passage just like a human would, and human imitation was the driving force of our project. However, we ended up deciding to start with a DAN because we had a better foundation with the DAN based on previous work we did.

We had to alter various parts of our DAN to get it to work with our new data. The DAN normally only expected question text and "labels", which were just a single answer for each question text. We needed various sections of the program to work with our new data. We needed to change the batchify function to make sure that our human guess data was getting passed into other functions like evaluate and train. However, we couldn't just change the model entirely since it needed to run as before. For example, it needed to run the same as when it would run the development or test data within the same functions as the new training data like evaluate.

The most important change we needed to implement was the loss function. Normally the loss function simply looks at the correct answer and sees what level of certainty the model had for that one class. However, we needed a way to alter the loss function so that the model would be rewarded for human guesses. Once we were able to make sure the human guess data could be passed into all the necessary functions, we could use it to modify the loss function.

The way we did this was by performing a normal cross entropy loss on the model and the true answer labels and then multiplying it by one minus the frequency in human guesses of that model's output:

$$Loss = CrossEntropyLoss(model, answer) * (1 - HumanGuessFrequency)$$

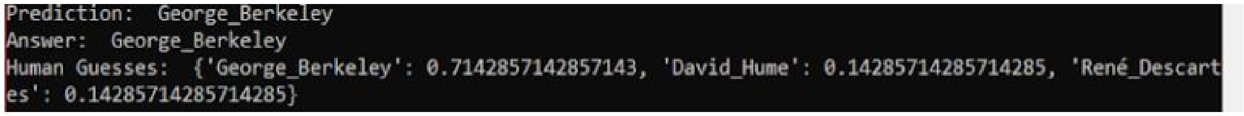
The HumanGuessFrequency variable was a value represented by the average human guess frequency across all of the examples for the specific batch. For example, in a batch of two questions, if the model outputs "Pizza", which was guessed by humans 90% of the time for that question and "George_Washington" for the second question, which was guessed by humans 10% of the time for that question, the HumanGuessFrequency for that batch would be 0.5 since the average of .9 and .1 over 2 questions is 0.5. The final loss value then would be half of what it would be in a normal DAN that only uses gold labels.

Another change we needed to make to the model involved modifying it to work on a GPU. We knew we would be training the model often under different conditions, so we wanted to speed up the process by using a GPU. We had to modify some of the code to work on GPUs. This was not an issue, but it along with some other things became an issue related to saving our model and running it on Codalab.

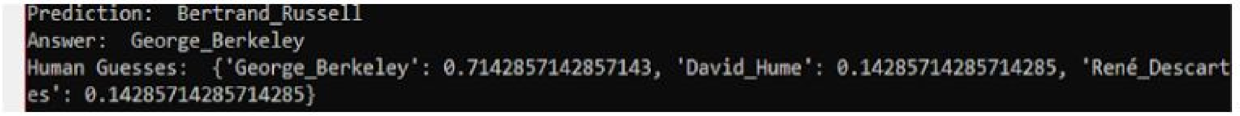
We had to save other aspects of the system besides just the model to run it on Codalab. We needed to save our dictionaries that related the int values for the words and classes to the associated strings. We also needed to make sure the system would tokenize and split up the question text the same way on Codalab as it was in our model. Both of these things together caused issues because the GPU tensors from the model could not work on systems without GPUs. Also, Codalab seemed to split up text differently and we weren't able to incorporate our version of text tokenization.

Error Analysis

For our error analysis, we decided to not only look at our model's responses to questions but also a version of the DAN that was the same in all aspects (including the training development and test data, hyperparameters, etc.) except for the inclusion or omission of human feedback. In the following examples, the "top" box represents our model, and the "bottom" box is the non-human-feedback DAN.

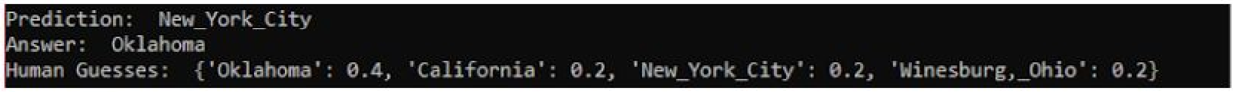


```
Prediction:  George_Berkeley
Answer:  George_Berkeley
Human Guesses:  {'George_Berkeley': 0.7142857142857143, 'David_Hume': 0.14285714285714285, 'René_Descart
es': 0.14285714285714285}
```

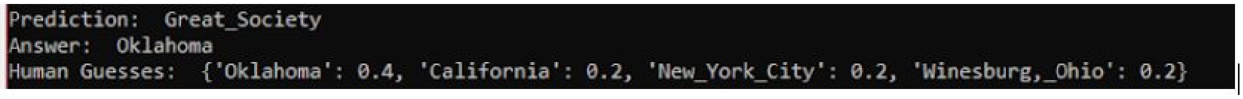


```
Prediction:  Bertrand_Russell
Answer:  George_Berkeley
Human Guesses:  {'George_Berkeley': 0.7142857142857143, 'David_Hume': 0.14285714285714285, 'René_Descart
es': 0.14285714285714285}
```

Our model was able to correctly predict George Berkeley compared to the original DAN. The fact that humans had also guessed this answer to the question may imply that humans guessed this answer to other questions, which the model was able to learn.

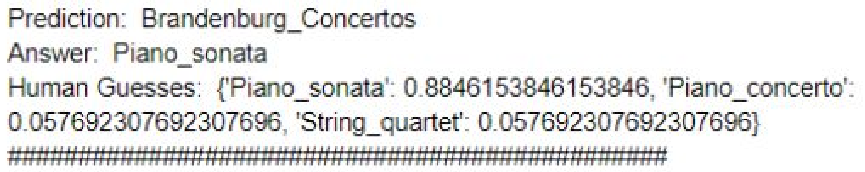


```
Prediction:  New_York_City
Answer:  Oklahoma
Human Guesses:  {'Oklahoma': 0.4, 'California': 0.2, 'New_York_City': 0.2, 'Winesburg,_Ohio': 0.2}
```

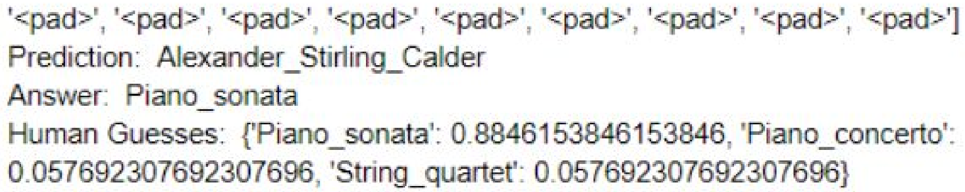


```
Prediction:  Great_Society
Answer:  Oklahoma
Human Guesses:  {'Oklahoma': 0.4, 'California': 0.2, 'New_York_City': 0.2, 'Winesburg,_Ohio': 0.2}
```

Our model, while guessing the wrong answer, is able to make predictions within the realm of the correct answer, much as a human would. Our model guessed a specific location within the U.S. compared to the original model.



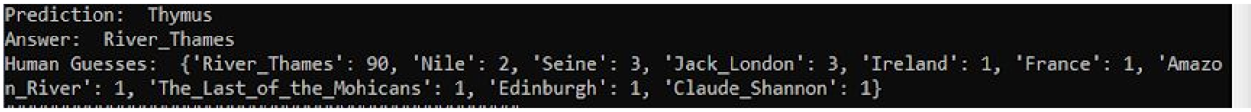
```
Prediction:  Brandenburg_Concertos
Answer:  Piano_sonata
Human Guesses:  {'Piano_sonata': 0.8846153846153846, 'Piano_concerto':
0.057692307692307696, 'String_quartet': 0.057692307692307696}
####################################################
```



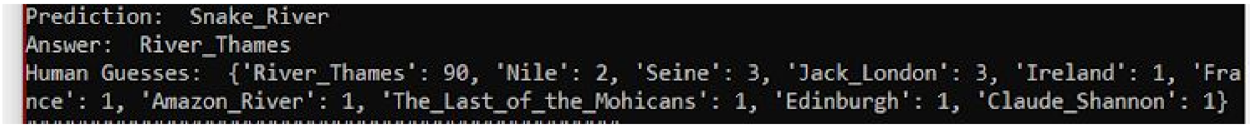
```
'<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>']
Prediction:  Alexander_Stirling_Calder
Answer:  Piano_sonata
Human Guesses:  {'Piano_sonata': 0.8846153846153846, 'Piano_concerto':
0.057692307692307696, 'String_quartet': 0.057692307692307696}
```

Our model was able to guess a piece of music, which the question was looking for, compared to the other model's guess, which was a person. We hypothesize that rewarding the model for human answers helped it to at least predict answers in the same category as the true question like a real human would whereas the other model seems to guess more randomly.

We also wanted to look at specific examples of where our model guessed very badly and try to understand what might have been going wrong.

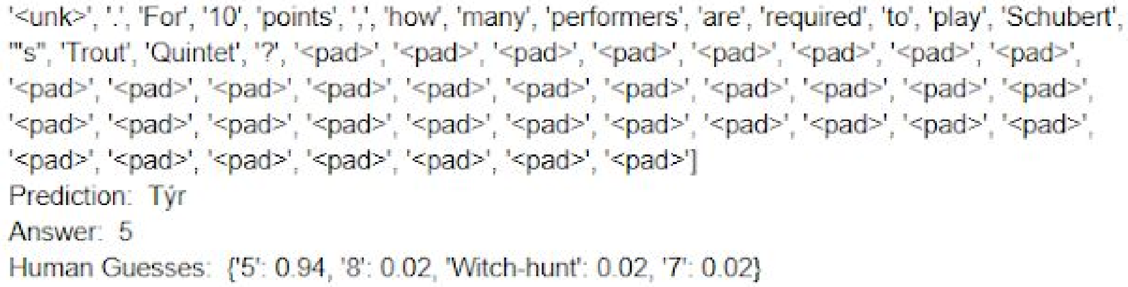


```
Prediction:  Thymus
Answer:  River_Thames
Human Guesses:  {'River_Thames': 90, 'Nile': 2, 'Seine': 3, 'Jack_London': 3, 'Ireland': 1, 'France': 1, 'Amazo
n_River': 1, 'The_Last_of_the_Mohicans': 1, 'Edinburgh': 1, 'Claude_Shannon': 1}
```

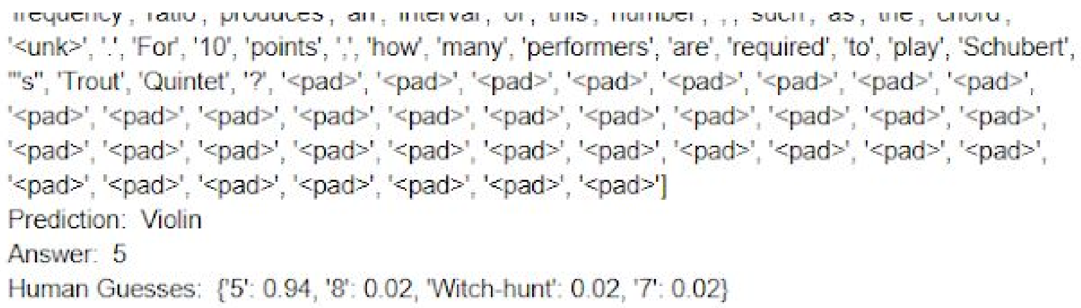


```
Prediction:  Snake_River
Answer:  River_Thames
Human Guesses:  {'River_Thames': 90, 'Nile': 2, 'Seine': 3, 'Jack_London': 3, 'Ireland': 1, 'Fra
nce': 1, 'Amazon_River': 1, 'The_Last_of_the_Mohicans': 1, 'Edinburgh': 1, 'Claude_Shannon': 1}
```

The original model actually guessed closer to the real answer. However, we think this may be because our mapping process mapped "Thames" to "Thymus" since the spelling is similar. We could potentially fix this if we could improve the mapping process to avoid these errors, perhaps by mapping based on frequency of the answer in other questions or maybe even a learned mapping using neural networks.



```
'<unk>', '.', 'For', '10', 'points', ',', 'how', 'many', 'performers', 'are', 'required', 'to', 'play', 'Schubert',
"'s", 'Trout', 'Quintet', '?', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>',
'<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>',
'<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>',
'<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>']
Prediction:  Týr
Answer:  5
Human Guesses:  {'5': 0.94, '8': 0.02, 'Witch-hunt': 0.02, '7': 0.02}
```



```
'frequency', 'ratio', 'produces', 'an', 'interval', 'of', 'this', 'number', ',', 'such', 'as', 'the', 'chord',
'<unk>', '.', 'For', '10', 'points', ',', 'how', 'many', 'performers', 'are', 'required', 'to', 'play', 'Schubert',
"'s", 'Trout', 'Quintet', '?', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>',
'<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>',
'<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>',
'<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>']
Prediction:  Violin
Answer:  5
Human Guesses:  {'5': 0.94, '8': 0.02, 'Witch-hunt': 0.02, '7': 0.02}
```

Both models were quite off with their answers (the original DAN may be arguably closer). This questions was very difficult because it was a common link question. We do not necessarily have a clear reason for why our model was unable to get this question correct other than the DAN model and other QA models being generally bad at common-link questions like this one.

Results

To assess the performance of our model, we ran 100 epochs of training on two iterations of a standard DAN (without the custom loss function), as well as 100 epochs in two different human feedback (HF) models.

The first (Answer Frequency) the HumanGuessFrequecy variable included the correct answers in its frequency distribution. This rewarded the model more for guessing the correct answer when compared to incorrect guesses. For example, an easy question might have had a distribution (0.8, 0.1, 0.05, 0.05) where the correct answer was guessed 80% of the time, while a more difficult or ambiguous question might have had something like (0.6, 0.3, 0.1). In that second case, the model would receive a smaller reward for the correct answer and more significant rewards for guessing other incorrect human guesses. We hypothesized this property would help capture more accurately the human “thought process” when approaching a question.

The second HF model (Guess Frequency), used only the incorrect human guesses when calculating the frequencies, guaranteeing a more substantial reward for incorrect human guesses while giving no additional reward for a correct guess (aside from the naturally lower loss). This came from our observation that many questions had fairly low values for incorrect guess frequencies, relative to the correct guesses. This risked a higher reward for incorrect human guesses than the correct guesses, but as will be shown, that did not dominate the trend.

Figure 1 displays the performance of the two standard DAN models and the two HF models over the epochs, and though they perform similarly, it is also quite visible that the HF models have a slight head up (final accuracies around 5% higher than the standard DANs). Note that the accuracies are still quite low; more epochs were not run due to the time constraints of running the four models.

Especially observing the Guess Frequency HF (GFHF) model, note that it is above the standard DAN models in accuracy throughout the training. Comparing the GFHF with the Answer Frequency (AFHF) model, it is not clear that there is a benefit between the two. It is interesting to note, however, that the AFHF model looked more like DAN1,2 during the earlier epochs, and later rose in accuracy, closer to the GFHF. This may be related to the randomness in the training data, or it is a sign that the AFHF model took longer to absorb the “conceptual understanding” that the loss function tried to represent.

Figure 2 looks at the accuracy at each step for the GFHF and DAN1 models, mostly to show the difference in their learning trajectories. With DAN 1 (as with DAN 2), the model seemed to get stuck in certain plateaus, where its accuracy had a noticeable dip (the one around step 800 being a good example), probably as its parameters went down an unproductive route. The GFHF (as well as the AFHF) seemed to have a more consistently “smooth” trajectory throughout the training, not showing any major dips in its accuracy. This may be a sign that the modified loss function gave its parameter adjustments a “safety net,” as it was able to adjust conceptually-similar parameters more evenly. This allowed it to more quickly “go in the right direction” when improving its parameters. Finally, it should be noted that the GFHF’s performance is a sign that our fear of it favoring incorrect human guesses did not overtake its tendency towards the correct answer.

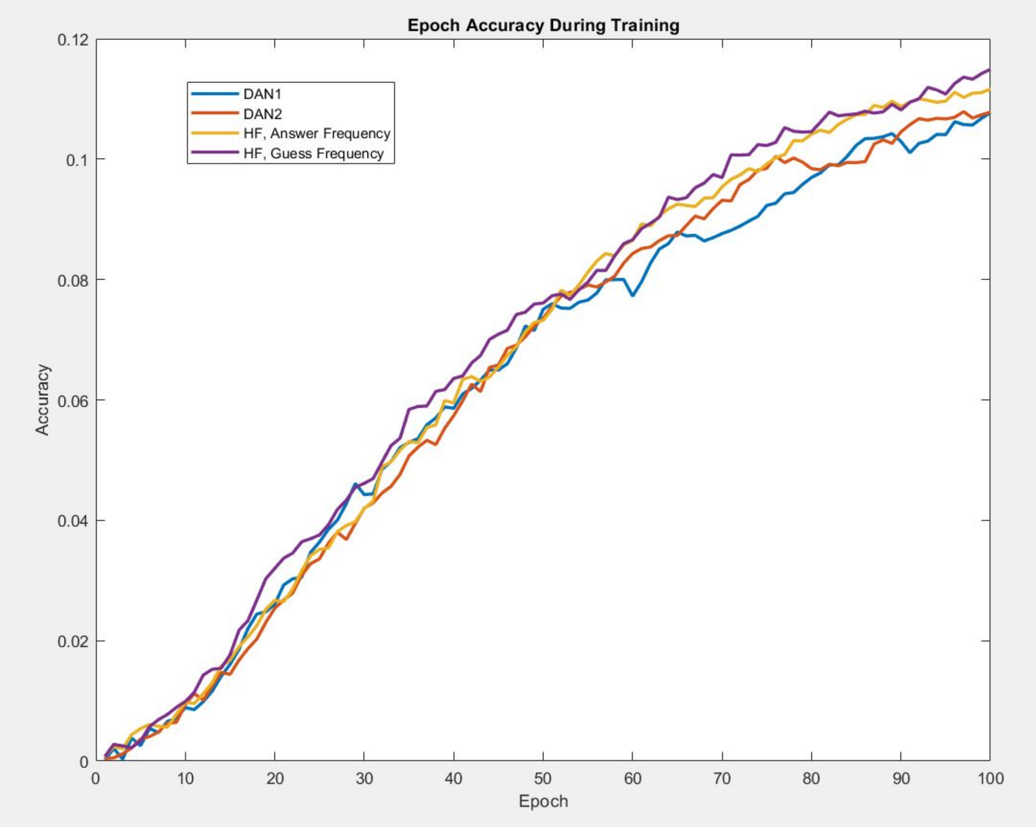


Figure 1 - Averaged epoch accuracy for four DAN models during training

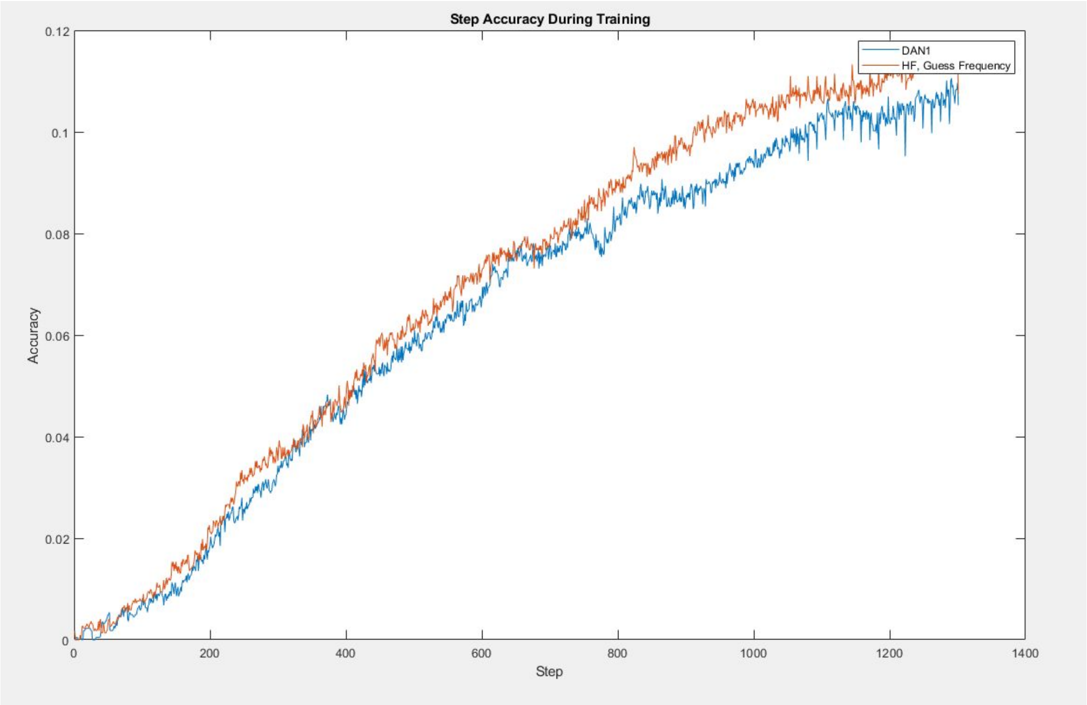


Figure 2 - Step-by-step accuracy for a standard DAN and the GFHF during training

Conclusion

While our model did do slightly better than a regular DAN and we learned a lot in the process, there are several things we could improve upon. In the future we would spend more time preprocessing and devising our loss function, since those are the key components to our model. In addition, we would spend more time tuning hyperparameters, such as the number of epochs we train for and the batch size. In addition, if we stuck to a DAN in the future, we would also experiment with the number of layers and the number of hidden nodes per layer. However, we were also interested in doing a LSTM instead of a DAN. If we had more time, we wanted to also include categorization in our algorithm, since we believe this, in combination with using human guesses, would further help our model with common link questions. We also didn’t have enough time to do a more quantitative error analysis, so this is another area we would focus on in the future.