SI630 Project Report: Answering Multiple Choice Questions from Gaokao English Exam

Jiaoyang Hu
SI630
University of Michigan, School of Information
ivanhujy@umich.edu

Abstract

This project attempts to answer multiple choice questions from Chinese Gaokao English exam papers. While the topic of multiple choice question answering has been widely studied, there has been no research effort into Gaokao English multiple choice questions. We fine-tuned GPT2 model with the RACE dataset(Lai et al., 2017), FitBert model to will be implemented to predict the correct answer, and their accuracy will be benchmarked against a random guess baseline, which has 25% accuracy. Here we show that the fine-tuned GPT2 model achieved 68.3% accuracy,higher than the untuned GPT2 model, which has 66.9% accuracy. The FitBert model achieved an accuracy of 73.5%.

1 Introduction

This projects aims to fine-tune pretrained language models to answer multiple choice questions from Chinese Gaokao English exam papers.The multiple choice questions requires the student to pick from 1 out of 4 choices to fill in the blanks in the question text.

Previous work on this has been focusing on reading comprehension multiple choice questions, which have contexts of hundreds of tokens in length, and their models are usually complex. The questions that these models are answering are reading comprehension questions as opposed to simple fill-in-the-blank questions that this project aims to do.

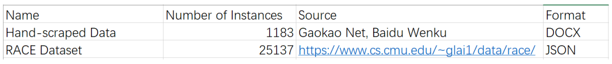
I scraped multiple choice questions from multiple online sources(see Data section for details), and my goal is to fine-tune models to achieve a high accuracy in these Gaokao questions. I fine-tuned GPT2 model with the RACE dataset (Lai et al., 2017). A dataset that contains passages from Chinese English exam papers. Another model that I used was the FitBert modelhttps://github.com/Qordobacode/fitbert, a model that can rank options to fill in the blanks of a sentence. The fine-tuned GPT2 model with achieved 68.3% accuracy,higher than the untuned GPT2 model, which has 66.9% accuracy. The FitBert model achieved an accuracy of 73.5%. All models outperformed the random guess baseline, which has 25% accuracy.

How models comprehend and answer Gaokao English questions will be of great benefit to the field of Muiltiple Choice Question Answering. Future NLP research into Gaokao Exam questions or even generate these questions will also find the results of this project useful. As an eye to the future, hopefully this project, and more work in the same field will free teachers from the repetitive task of creating questions and answers themselves, as well as provide high-quality practices for students.

The main contribution of this project is the fine-tuned GPT2 model that can achieve 68.3% accuracy, and the application of FitBert model to answer Gaokao English multiple choioces questions, which has an accuracy of 73.5%. The superior performance of a masked model like FitBert over the GPT2 model suggests that being able to use contexts both before and after the blanks would greatly improve the prediction accuracy.

2 Data

A summary of the data used for this project is listed below, please refer to corresponding parts in this section for details.

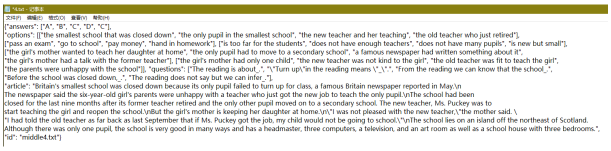


Name	Number of Instances	Source	Format
Hand-scraped Data	1183	Gaokao Net, Baidu Wenku	DOCX
RACE Dataset	25137	https://www.cs.cmu.edu/~glai1/data/race/	JSON

Figure 1: Data Summary Table

2.1 RACE Dataset

The RACE dataset is developed by Guokun Lai et al. and contains passages from Chinese English exam papers, which is valuable training data for my models. The RACE dataset is composed of txt files containing one instance per txt file in json format. The json includes ”answers”,”options”,”article” and ”Id”. For convenience, only the ”article” part of the dataset is used to fine-tune my model.

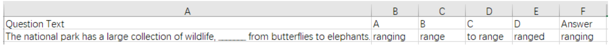


```
{"answers": ["A", "B", "C", "D", "C"],
"options": [["the smallest school that was closed down", "the only pupil in the smallest school", "the new teacher and her teaching", "the old teacher who just retired"],
["pass an exam", "go to school", "pay money", "hand in homework"], ["is too far for the students", "does not have enough teachers", "does not have many pupils", "is new but small"],
["the girl's mother wanted to teach her daughter at home", "the only pupil had to move to a secondary school", "a famous newspaper had written something about it",
"the girl's mother had a talk with the former teacher"], ["the girl's mother had only one child", "the new teacher was not kind to the girl", "the old teacher was fit to teach the girl",
"the parents were unhappy with the school"]], "questions": ["The reading is about_.", "\"Turn up\"in the reading means \"_\".", "From the reading we can know that the school_.",
"Before the school was closed down,_.", "The reading does not say but we can infer_."],
"article": "Britain's smallest school was closed down because its only pupil failed to turn up for class, a famous Britain newspaper reported in May.\n
The newspaper said the six-year-old girl's parents were unhappy with a teacher who just got the new job to teach the only pupil.\nThe school had been
closed for the last nine months after its former teacher retired and the only other pupil moved on to a secondary school. The new teacher, Ms. Puckey was to
start teaching the girl and reopen the school.\nBut the girl's mother is keeping her daughter at home.\n\"I was not pleased with the new teacher,\"the mother said. \
"I had told the old teacher as far back as last September that if Ms. Puckey got the job, my child would not be going to school.\"\nThe school lies on an island off the northeast of Scotland.
Although there was only one pupil, the school is very good in many ways and has a headmaster, three computers, a television, and an art room as well as a school house with three bedrooms.",
"id": "middle4.txt"}
```

Figure 2: RACE Dataset Format

2.2 Hand-scraped Multiple Choice Questions

The data used in this project will consist of multiple choice questions from Gaokao English exam papers. As there is no available database for these questions, these data will be scraped by hand from Gaokao English exam papers or practice questions of the same form from online sources. The data will be entered into a csv file with six columns, starting with ’Question Text’, followed by four choices, and the correct answer. The models will read in the inputs and generate answers in a following column ’Generated Answer’.

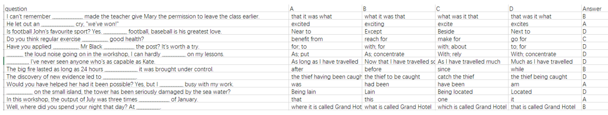


A	B	C	D	E	F
Question Text	A	B	C	D	Answer
The national park has a large collection of wildlife, _______ from butterflies to elephants.	ranging	range	to range	ranged	ranging

Figure 3: Sample Data Input Format

2.2.1 Data Analysis

As there is no available database for gaokao English exam questions, we collected gaokao english exam practice questions from Gaokao Net(http://www.gaokao.com/e/20110512/4dcbca16a78cf.shtml), and Baidu Wenku(https://wenku.baidu.com/view/f550beff32d4b14e852458fb770bf78a64293a54.html). While the original sources was relatively clean, the format of the original resources was in Microsoft doc, contained Chinese punctuations and Chinese characters. We manually editted out the Chinese characters and punctuations, as those cannot be recognized with UTF-8 encoding, and then we preprocessed the data into a clean, csv format.This data is only used for evaluation. To evaluate the GPT2 model, we fill in the blanks with the 4 options, which creates 4 sentences. The 4 sentences are then fed to the GPT2 model, which will calculate the perplexity of each and pick the one with the lowest perplexity. To evaluate the FitBert model, the blank part in the question text is masked, and the masked question text and the options are fed to the FitBert model, which will rank the four options to fill in the blank. The code for preprocessing data can be viewed in my Github(https://github.com/PepperinoHu/gaokao_mcq_answering/blob/master/basic_data_preprocess.ipynb). With data processed,



question	A	B	C	D	Answer
I can't remember __________ made the teacher give Mary the permission to leave the class earlier.	that it was what	what it was that	what was it that	that was it what	B
He let out an ___________ cry, "we've won!"	excited	exciting	excite	excites	A
Is football John's favourite sport? Yes. ________ football, baseball is his greatest love.	Near to	Except	Beside	Next to	D
Do you think regular exercise _________ good health?	benefit from	reach for	make for	go for	C
Have you applied _________ Mr Black __________ the post? It's worth a try.	for; to	with; for	with; about	to; for	D
______ the loud noise going on in the workshop, I can hardly ________ on my lessons.	As; put	As; concentrate	With; rely	With; concentrate	D
________, I've never seen anyone who's as capable as Kate.	As long as I have travelled	Now that I have travelled sc	As I have travelled much	Much as I have travelled	D
The big fire lasted as long as 24 hours ___________ it was brought under control.	after	before	since	while	B
The discovery of new evidence led to ___________.	the thief having been caugh	the thief to be caught	catch the thief	the thief being caught	D
Would you have helped her had it been possible? Yes, but I ________ busy with my work.	was	had been	have been	am	A
_________ on the small island, the tower has been seriously damaged by the sea water?	Being lain	Lain	Being located	Located	D
In this workshop, the output of July was three times __________ of January.	that	this	one	it	A
Well, where did you spend your night that day? At ________.	where it is called Grand Hot	what is called Grand Hotel	which is called Grand Hotel	that is called Grand Hotel	B

Figure 4: Data CSV File Format

we continue to analyze the data. The code for analysis is on my Github(https://github.com/PepperinoHu/gaokao_mcq_answering/blob/master/basic_data_analysis.ipynb) We can see that there are a total of 1147 total questions in my dataset. The correct answers are not equally distributed as there are more Bs and Cs than As and Ds. As for length of questions, we can see that

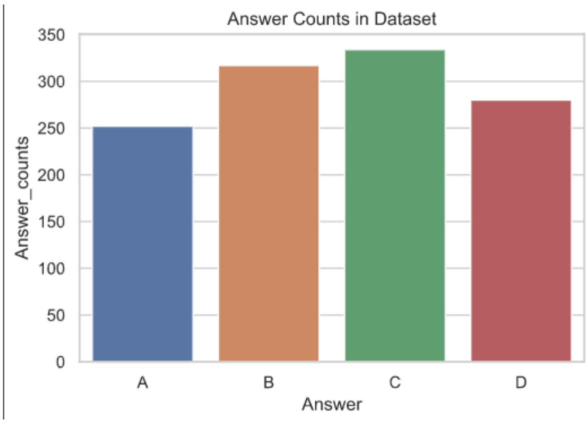


Figure 5: Answer Counts

questions has a mean length of 83.6, with the max at 193. This means that we are dealing with short texts ranging from one to three, four sentences. Multiple question format are also present in the dataset. The first is multiple blanks. The questions can have up to three blanks to fill in, bringing additional challenge to prediction. There is a total of 160 questions in my dataset with multiple blanks. There is also variety of answers, some answers are displayed as ” / ”, indicating a no-fill, while some answers are in the format of ”Both A and B”, suggesting that two of the other options are also correct. After examinatino, there are 52 questions with no-fill and 6 questions with ’both’ form. All of these problems will add complications to my model.

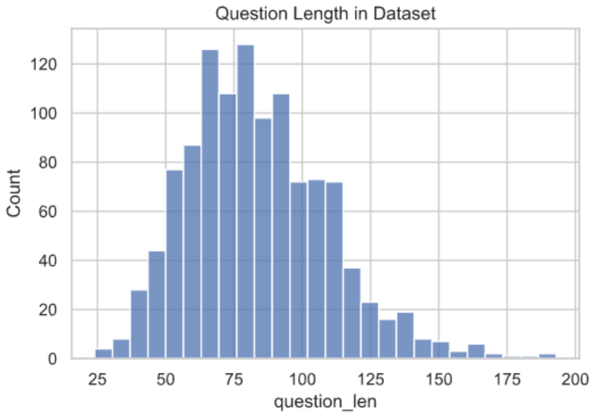


Figure 6: Question Length

3 Related Work

3.1 Annotated Related Work in Proposal

1. In the paper *CNN for Text-Based Multiple Choice Question Answering* , the authors proposed a Convolutional Neural Network (CNN) model for textbased multiple choice question answering where questions are based on a particular article. While my project only deal with questions, their model also have to comprehend an article. Their model assigns a score to each question-option tuple and chooses the final option accordingly. Their model was tested on Textbook Question Answering (TQA) and SciQ dataset,and it ended up outperforming several LSTM-based baseline models on the two datasets(Chaturvedi et al., 2018).

2. In the paper *ElimiNet: A Model for Eliminating Options for Reading Comprehension with Multiple Choice Questions*, the authors proposed a neural network-based model which tries to mimic this process. The model makes multiple rounds of partial elimination to refine the passage representation and finally uses a selection module to pick the best option(Parikh et al., 2019). Their approach was innovative and highly useful in more complicated reading comprehension. However, the complexity of Gaokao English exam is much lower. A more direct approach like what I proposed would be much faster and complete the question with much less rounds of operation.

3. In the paper *CMU Multiple-choice Question Answering System at NTCIR-11 QA-Lab*, the authors created a model for answering multiple choice English questions for the world history entrance exam. Questions are preceded by short descriptions providing a historical context. Given the context and question-specific instructions, the model generates verifiable assertions for each answer choice. These assertions are evaluated using several evidencing modules, which assign a plausibility score to each assertion. These scores are then aggregated to produce the most plausible answer choice(Wang et al., 2014). Their model, when answering questions, will write queries to retrieve information from Wikipedia. I do not have the information retrieval skills required to do so, and the Engilsh textbooks used in China cannot be easily accessed in such a way.

3.2 Additional Annotated Related Work in Project Update

1. In the previous course project done by Zhang Dehao, he did a very similar project compared to mine. He built a word2vec model, transformers model with random guess as a baseline. He discovered that his model(SpaCy representation with weighted average and duplicates removed) was able to handle numerous questions that are short and object-oriented (“What/Which” type of questions)(Zhang, 2020). When the question descriptions and/or answer choices become long and require reasoning (“Why/How” type of questions), most of the model’s predictions were incorrect(Zhang, 2020). My workflow will be very similar to his as we are all trying to get high accuracy on multiple choice questions. However, the nature of our task is still a bit different because Gaokao English questions is more like fill-in-the-blank and my approach of picking choices with perplexity as criterion, or using a masked model like FitBert will not work well in his project.

3.3 Additional Annotated Related Work in Project Report

1. In the paper *Enabling Language Models to Fill in the Blanks*, the authors proposed a simple approach for text infilling, which is the task of predicting spans of text at any position in a document. The authors claim that their trained GPT2 model can fill in the blanks and generate sentences that is hard

for a human to distinguish if it is machine-generated(Donahue et al., 2020). The potential of this model on multiple choice question answering could be huge, and could lead to superior performance compared to mine. They managed to use masking with a GPT2 model, which is something that I am unawared GPT2 is capable of. I shall look into it and see how far masking can get the GPT2 model to improve in accuracy.

4 Methods

This project will first implement a random guess method as a baseline. The two models this project used are a GPT2 model fine-tuned with the RACE dataset, as well as the FitBert model.

4.1 Implementation of Random Guess Baseline

The code for implementation of random guess is included in my github(https://github.com/PepperinoHu/gaokao_mcq_answering/blob/master/basic_data_analysis.ipynb). The code chooses one out of four correct choices randomly. And consistent with our intuition, we achive about 0.2502 accuracy, close to theoretical value of 0.25, in 1000 trial runs.

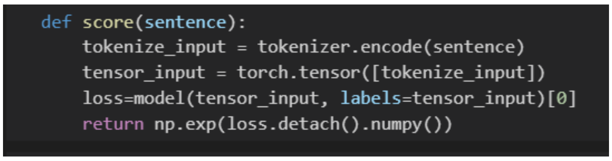
4.2 Fine-tuned GPT2 Model

The GPT2 model that this project fine-tuned comes from Huggingface https://huggingface.co/transformers/model_doc/gpt2.html. The RACE dataset is developed by Guokun Lai et al. and contains passages from Chinese English exam papers, which is valuable training data for my models.

To answer the quesitions, my GPT2 models would fill in the blanks with each of the 4 available choices, resulting in four sentences. Perplexity is a measure of how surprised the model is about the next word. This means that an incorrect choice would lead to a sentence that is more surprising for the model. The model calculates the perplexity of the 4 sentences and pick the sentence with the lowest perplexity.

4.3 FitBert Model

The FitBert Model that this project applied to answering multiple choice questions from Gaokao English Exams comes from https://github.com/Qordobacode/fitbert. This project only applied the model in answering and did not do any additional fine-tuning due to the time constraints of the project.



```
def score(sentence):
    tokenize_input = tokenizer.encode(sentence)
    tensor_input = torch.tensor([tokenize_input])
    loss=model(tensor_input, labels=tensor_input)[0]
    return np.exp(loss.detach().numpy())
```

Figure 7: Perplexity Calculation

The FitBert model accepts masked strings, and a list of options. It would rank the options based on probability calculated from the contexts before and after the blanks. The first option returned in its list of ranked options is what the model considers to be the answer to the multiple choice question.

5 Evaluation and Results

The sold evaluation method is the calculate the accuracy of model predicitions. Which is:
$accuracy = num\_of\_correct\_answers/num\_of\_questions$
We feed our hand-scraped Gaokao English Exam Questions to our models and let it make the predictions. The GPT2 models fill the 4 options in the blank and generate 4 sentences. They calculate perplexity of each sentence and return the option with the lowest perplexity. The FitBert model masks the blanks and rank the options to fill in the blank. The model returns the favorite option as the answer to the questions. With each returned choice, we compare the model's choice with the answer provided by our source, and record 1 for correct answer and 0 for incorrect. When the model finishes answering, we calculate the accuracy.

Results show that the fine-tuned GPT2 model achieved 68.3% accuracy,higher than the untuned GPT2 model, which has 66.9% accuracy. The FitBert model achieved an accuracy of 73.5%. All models outperformed the random guess baseline, which has 25% accuracy.

6 Discussion

In the process of this project, we discovered that fine-tuning with the RACE dataset, which is readings from Chinese English exams, can improve the performance of the GPT2 model. Training for more steps and more epoches also lead to a small accuracy increase. The biggest improvement comes from using a masked model like FitBert rather than a causal language model

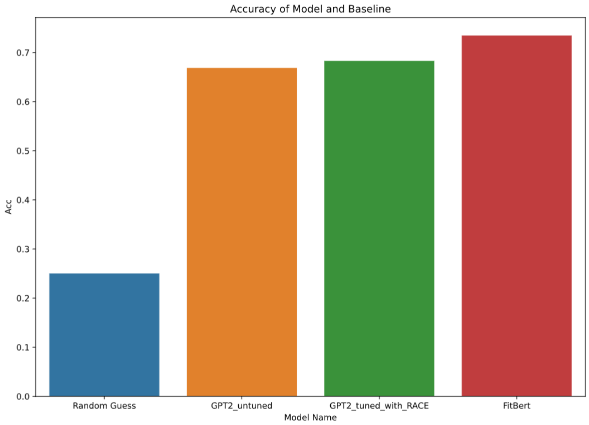


Figure 8: Model and Baseline Accuracy Summary

like the GPT2. The FitBert model was able to make use of contexts from both before the blanks and after the blanks, which leads to more accurate predicitions.

The performance of my model could be satisfactory to some, as it manages to capture some features from the RACE dataset, and seems to understand the pattern of text in Gaokao a little better after my fine-tuning. And the models are definitely better than random guessing baseline. This would mean that my models could have potential in future studies regrading Gaokao multiple choice questions. But since a well-trained student can consistently get 90% or even 100% accuracy on these multiple choice questions, I would say that our model still has a long way to go.

Here are some thoughts as to why my models could not match the accuracies of a human. The GPT2 models are limited to contexts before the blank, which means that they receive less information than a student when doing the questions. The FitBert model, while it could take contexts both before and after the blanks into account, it is limited by the order in which it has to predict. If there are three blanks to fill in, a student could decide on the second blank first, eliminate some options, and then decide on the rest. The FitBert model can only do the predictions in one linear order. Finally, there is domain knowledge. A well-trained student sees enough questions that they can capture patterns and traps inside the questions. They know when the question text is loring them to pick a seemingly right but incorrect answer, while my models could fall in the traps.

7 Conclusion

This project attempted to answer multiple choice questions from Chinese Gaokao English exam papers. We used two models, a fine-tuned GPT2 model with the RACE dataset(Lai et al., 2017), and the FitBert model to predict the correct answer. The fine-tuned GPT2 model achieved 68.3% accuracy,higher than the untuned GPT2 model, which has 66.9% accuracy. The FitBert model achieved an accuracy of 73.5%. Github link to this project is `https://github.com/PepperinoHu/gaokao_mcq_answering`.

8 Other Things We Tried

I also tried fine-tuning the GPT2 model with some of the correct sentences from my hand-scraped dataset. Unfortunately, the prediction accuracy of the resulting model actually dropped by 0.8% compared to the untuned GPT2 version. Therefore, I decided to abandon this idea and use the RACE dataset to fine-tune.

9 What I Would Have Done Differently

Looking back, what I would do differently is to just try more things with my models. Fine-tune the FitBert model, try more hyperparameters in fine-tuning etc. I would also try harded in getting better-quality exam questions and answers.

References

Akshay Chaturvedi, Onkar Pandit, and Utpal Garain. 2018. Cnn for text-based multiple choice question answering.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Soham Parikh, Ananya B Sai, Preksha Nema, and Mitesh M Khapra. 2019. Eliminet: A model for eliminating options for reading comprehension with multiple choice questions. *arXiv preprint arXiv:1904.02651*.

Di Wang, Leonid Boytsov, Jun Araki, Alkesh Patel, Jeff Gee, Zhengzhong Liu, Eric Nyberg, and Teruko Mitamura. 2014. Cmu multiple-choice question answering system at ntcir-11 qa-lab. In *NTCIR*.

Dehao Zhang. 2020. Multiple-choice question answering through semantic representation space.