NLP Application on Taking Exam

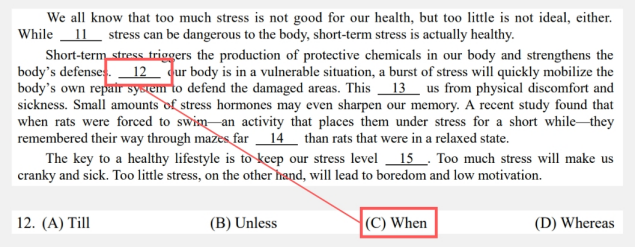
TEAM36 109022136陳致寧　109022127徐偉晉　108022138楊宗諺　108022108紀宇恆　109030039廖奕愷

***Abstract*—Here is the final project by TEAM36. Our goal is to finish all the multiple choice questions in the exam. There are five sections in the exam, and each has different types of questions. The models we chose to solve the questions are BERT and SBERT, which are currently the best choice for these NLP problems.**

I. Introduction

In the beginning of this report, we will explain the exam paper and how we are dealing with it. We focus on what problems or questions we conflicted with and the preparation to solve various types of questions.

A. Cloze


We all know that too much stress is not good for our health, but too little is not ideal, either. While __11__ stress can be dangerous to the body, short-term stress is actually healthy.

Short-term stress triggers the production of protective chemicals in our body and strengthens the body's defenses. __12__ our body is in a vulnerable situation, a burst of stress will quickly mobilize the body's own repair system to defend the damaged areas. This __13__ us from physical discomfort and sickness. Small amounts of stress hormones may even sharpen our memory. A recent study found that when rats were forced to swim—an activity that places them under stress for a short while—they remembered their way through mazes far __14__ than rats that were in a relaxed state.

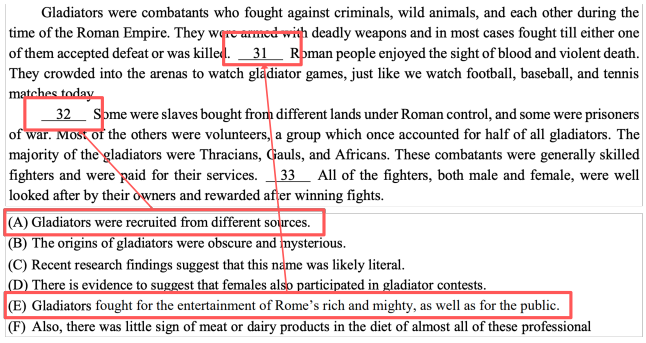
The key to a healthy lifestyle is to keep our stress level __15__. Too much stress will make us cranky and sick. Too little stress, on the other hand, will lead to boredom and low motivation.

12. (A) Till　(B) Unless　(C) When　(D) Whereas

Fig. 1. The schematic figure for cloze.

For the first category, sections 1 to 3 are included. Since these questions are all with cloze question type, there is no need to consider the case that the options of question are sentences. The strategy for solving this type of question is extremely straightforward. What we need to do is to scan for the options and then choose the most suitable one.

B. Sentences Reordering


Gladiators were combatants who fought against criminals, wild animals, and each other during the time of the Roman Empire. They were armed with deadly weapons and in most cases fought till either one of them accepted defeat or was killed. __31__ Roman people enjoyed the sight of blood and violent death. They crowded into the arenas to watch gladiator games, just like we watch football, baseball, and tennis matches today.

__32__ Some were slaves bought from different lands under Roman control, and some were prisoners of war. Most of the others were volunteers, a group which once accounted for half of all gladiators. The majority of the gladiators were Thracians, Gauls, and Africans. These combatants were generally skilled fighters and were paid for their services. __33__ All of the fighters, both male and female, were well looked after by their owners and rewarded after winning fights.

(A) Gladiators were recruited from different sources.
(B) The origins of gladiators were obscure and mysterious.
(C) Recent research findings suggest that this name was likely literal.
(D) There is evidence to suggest that females also participated in gladiator contests.
(E) Gladiators fought for the entertainment of Rome's rich and mighty, as well as for the public.
(F) Also, there was little sign of meat or dairy products in the diet of almost all of these professional

Fig. 2. The schematic figure for sentences reordering.

This section deals with discourse structure. Unlike the first category, we must consider the entire sentence and fill the correct sentence into the blank. Therefore, we cannot use Mask LM but must find an alternative strategy.

C. Reading Question

Located in Black Canyon straddling the border between Nevada and Arizona in the southwestern region of the United States, Hoover Dam is named one of the Top 10 Construction Achievements of the 20th century. The dam, constructed between 1931 and 1936, was the largest of its kind at the time. Its construction was the result of a massive effort involving thousands of workers and cost over one hundred lives.

Since about 1900, the Black Canyon and nearby Boulder Canyon had been investigated for their potential to support a dam that would control floods, provide irrigation water, and produce hydroelectric power. In 1928, the US Congress authorized the project. The winning bid to build the dam was submitted by Six Companies, Inc. However, such a large concrete structure had never been built before, and some of the techniques were unproven. The extreme summer heat and lack of facilities near the site also presented tremendous difficulties. Nevertheless, Six Companies turned over the finished dam to the federal government on March 1, 1936, more than two years ahead of schedule.

The initial design of the dam, which was more concerned with the dam's functionality than its exterior, was criticized by many as being too plain and unremarkable for a project of such immense scale. So Gordon B. Kaufmann, the architect who was brought in to redesign the exterior, greatly streamlined the design and applied an elegant Art Deco style to the entire project. Allen Tupper True, an American illustrator, was also hired to handle the design and decoration of the walls and floors of the new dam. He integrated into his design the images and colors based on Native American visions of rain, lightning, clouds, and animals, thereby creating symbolic patterns which appear both ancient and modern.

Today, Hoover Dam has become a national historic landmark. Standing at more than 725 feet above the Colorado River, the highest concrete dam in the Western Hemisphere continues to draw crowds 85 years after its creation, attracting more than a million visitors a year.

36. Which of the following is **NOT** mentioned as a reason for building the dam in the beginning?
(A) To promote tourism.　(B) To support agriculture.
(C) To generate electricity.　(D) To prevent natural disasters.

37. Which of the following statements is true about Hoover Dam?
(A) Its construction lasted for more than a decade.
(B) It is strong in functionality, but plain in design.
(C) Its site stretches over two states in the United States.
(D) It became famous because it led to the discovery of Black Canyon.

Fig. 3. An example of the reading question.

Reading question is the most important and challenging section in the project. Here is an example of the reading question. As the figure 3 shows, there will be a paragraph and it should be read by the reader(or model) first, which we can also call a hypothesis. And below the paragraph, there will be some questions, each includes four choices with a single correct answer.

Next, we combine the question and a single choice respectively, which we call a statement. For a correct statement, we give it a simple label True and otherwise False. Our model will be trained by these hypotheses and labeled statements, and we'll give test questions with hypotheses into the model, and the model will give each choice a logistic score. The choice with the highest score will be the final answer for a question.

II. Methods

To achieve our goal, we mainly divide the whole exam questions into three categories. For each category, we use different data preprocessing techniques and different NLP model structures to analyze and then generate the answers.

A. Cloze

To beat this kind of monster, the workflow we adapted is shown in Fig. 4.

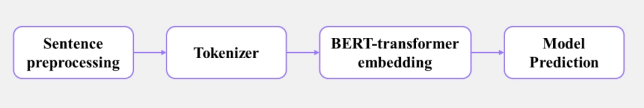


Fig. 4. The workflow diagram.

For sentence preprocessing, we initially extract the sentences with blank from the passage given. Then, we replace the blanks with masks like the figure shown below.

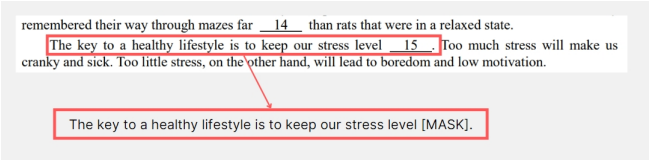


Fig. 5. The schematic figure for sentence preprocessing in the cloze section.

Subsequently, we input the preprocessed sentences into a pre-trained BERT Mask Language model (Mask LM) one at a time. Mask LM would generate a word vector for the mask and perform an inner product of this vector with all the word vectors that have already been stored in the model to obtain its similarity with each word.

Finally, we judge whether the option is the correct answer based on its similarity to the word vector of the mask.

B. Sentences Reordering

For sentence preprocessing, we use blanks to divide the article into sentences as shown in the following figure. This process is important because we want to concatenate the blank with previous and following sentences, and fill in each answer to find the best prediction for the blank.

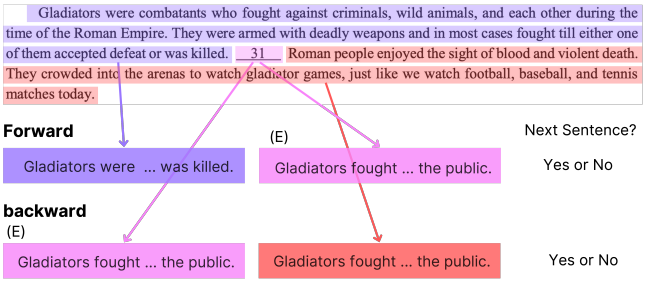


Fig. 6. The schematic figure for sentence preprocessing in the sentences reordering section.

After sentence preprocessing, we can divide the process into two parts. First, we need to assign a score to each answer in different blanks. To do this, we input the preceding sentence and the candidate answer into a pretrained Bert Next Sentence Prediction model, and obtain the logits value for the prediction. We repeat the same process for the following sentence, but switch the positions. In this step, we find the score of each answer by averaging the evaluation of the preceding and following sentences. We do this because in the next step, we need to exhaust all the combinations of answers and calculate all the scores in order to avoid repeated calculations. Last,

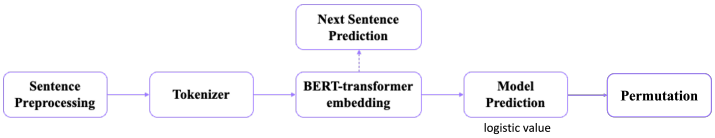


Fig. 7. The workflow diagram of the model of the sentences reordering.

we permute the answers to generate all possible answers. For example, in an exam in 2021, we can get $P_5^6 = 720$ possible answer lists. We then calculate the sum of the scores for each answer list and the list that obtains the highest score is the final answer of this section.

C. Reading Questions

1) DataSets and Data Preprocess: The Datasets we used are the reading question part of GSAT and Advanced Subjects Test in the past twenty years. There are totally about 605 questions and 2410 choices with the label of T/F(544 for train, 29 for validation, 32 for testing).

Subsequently, data preprocess is a key point in our project. We first filter out only the reading question part from the original text of the exam paper. To effectively access the information, we use the module docx and PyPDF2 to load the exam paper. Next, we use a powerful tool—regular expression, to clearly tidy up all of the paragraph and the question.

To obtain our goals, we first clear out the hierarchy between paragraph and question. Originally, a paragraph corresponds to four questions. But since each question is independent, we convert it into a question corresponding to a paragraph, still with the four choices. Sure, we also record the correct answer of the question to training. Last, we do question classification to help us analyze the results. The detail of question classification will be mentioned in the method part.

2) The Structure of the Model: Here is the structure of our model. Each of the paragraphs and choices will be packed and go through a tokenizer. Next, they will be encoded into high-dimensional vectors by the SBERT model. After encoded, we can simply make a prediction by comparing similarity between the vector of the paragraph and each statement.

Then, we transfer them into embedding vectors, and finally go through a Neural Network to make predictions for each question.

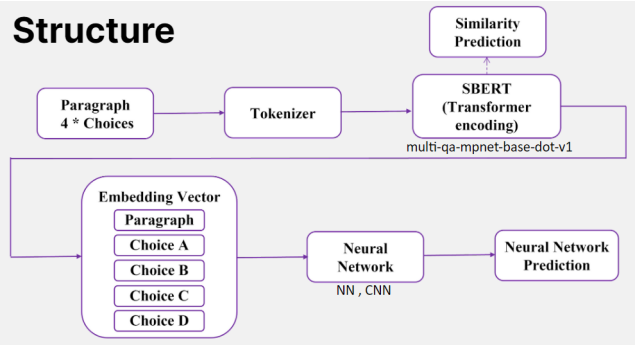


Fig. 8. The structure of the model in flow chart in reading question section.

3) Question Classification: The question classification can help us analyze the results and try to see the performance of the model. The datasets of the question classification is from the "Experimental Data for Question Classification". In these datasets, the question is classified in six classes: abbreviation, entity, description, human, location, numeric.

The size of the datasets is about 5452. We split it into 5000 training datasets and 452 validation datasets. We train the training datasets with BERT model(I choose for 'distilbert-base-uncased' model), and the below confusion matrix is the result of validation datasets for question classification model in six classes:

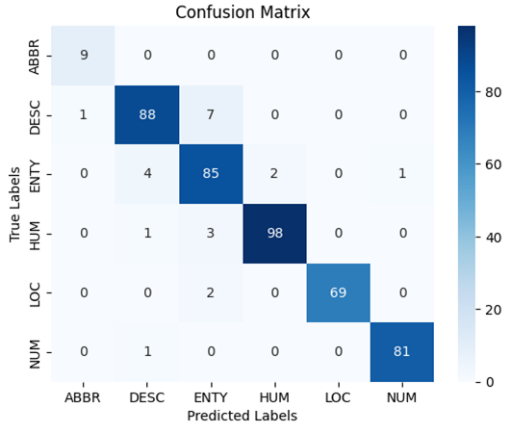


Fig. 9. The confusion matrix of question classification with six classes.

The accuracy of the question classification model is about 95.13%. From this result we can see that we can trust this model being a simple classifier, helping us with the remaining work.

4) SBERT Transformer and Embedding: The pretrained model we choose is "multi-qa-mpnet-base-dot-v1" in SBERT. Its max sequence length is 512, which is required in this reading test answering, where the total number of words of a single paragraph in this dataset may be up to around 490. Also, the performance is important. Among all the models, this one performs the best, in using either similarity predicting or neural networks.

As mentioned above, the paragraph in the dataset won't have more than 512 words, which is the limitation of SBERT, so the whole paragraph can be fed into the model entirely and encoded into an embedding vector. However, only using a single embedding vector to represent the whole paragraph may lose too much information during the pooling process. Therefore, the paragraph is first splitted into multiple parts to prevent losing too much information.

Here, the paragraph is splitted into 8 parts, each part may contain two or three sentences depending on the paragraph. Then, each part is fed into SBERT to create the embedding vectors. So the whole encoding result of a paragraph is 8 embedding vectors with 768-dimensions.

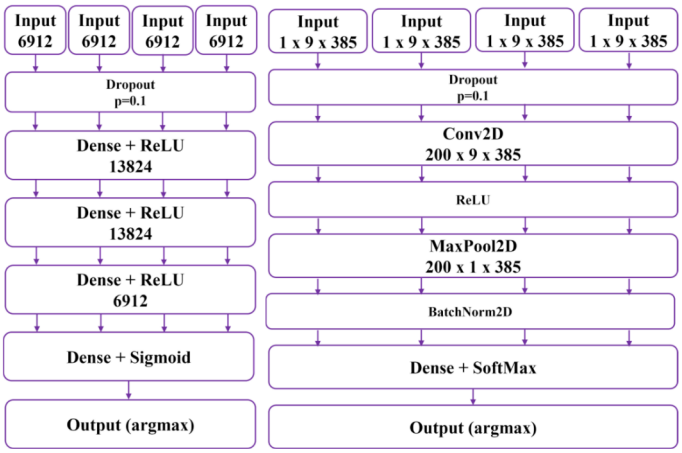


Fig. 10. The architecture of our NN(left) & CNN(right).

5) Neural Network: In the last part of our model structure, we try two types of Neural Network, common Neural Network(NN) and Convolutional Neural Network(CNN). Figure 4 shows the architecture of our NN and CNN model.

In the first one, which is NN, each statement's embedding vector is concatenated with the paragraph embedding vector, flatten, and fed into NN. Each one creates a score, and the one with the highest score is the predicted choice.

In the second one, which is CNN, the process is the same as NN, but without flattening. Each choice has a probability, and the highest probability one is the predicted one.

III. Results

A. Cloze

In this part, we achieved a correct rate of 23/30 in the Advanced Subjects Test in 2021. Especially the model shows better performance on section 1 since the questions given in this section are all sentences rather than a passage. There is no dependence on other sentences. All of the helpful clues can be found in the extracted sentence itself. Therefore, it reaches the highest correct rate compared to section 2 and 3.

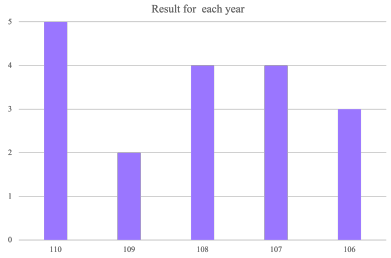


Fig. 11. The number of correct answer in the exam of each year. There are 5 questions in the sentences reordering section.

B. Sentences Reordering

In this section, we achieved a perfect score of 5/5 in the Advanced Subjects Test in 2021. This is a remarkable achievement, as the sentence prediction task is much more difficult than word prediction. To ensure stability, we additionally tested exams from 109 to 106 and found an average score of 3.6/5.

C. Reading Question

1) Training Curve: The validation loss of NN has never dropped from the beginning of the training, meaning that NN overfit the data easily, so the training takes only 30 epochs, otherwise the result is bad. For the CNN part, there exists a lot of fluctuation in the loss curve, but the result does get better.

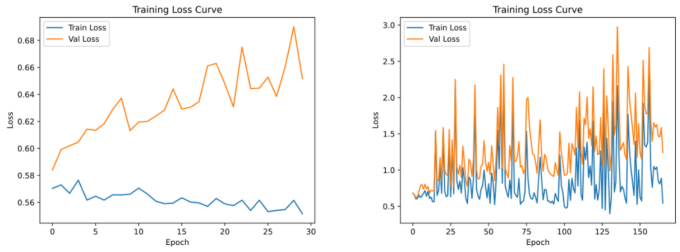


Fig. 12. The training loss curve of NN(left) and CNN(right).

2) Accuracy of NN prediction and similarity prediction: All of the above is our method and details of the model. Totally, we made three types of predictions—-by similarity, NN, CNN. For a simple presentation, we show the results of only the Advanced Subjects Test in 2021.

The accuracy of all of them is 7/16(44%) in similarity prediction, 7/16(44%) in NN prediction, and 10/16(63%) in CNN prediction.We can see that the result of CNN is relatively good(also in a more test datasets we tried).

We deduce that the reason is CNN is good at filtering out the important information, which is crucial in solving the reading question. It's making sense for us since it's often to determine the answer of the question in a key sentence or piece of paragraph. Prediction by similarity or common NN may be more difficult to find the correlation of paragraph(hypotheses) and statements.

ID	answer	nn_prediction	cnn_prediction	sim_prediction	question_class
36	A	A	A	D	DESC
37	C	B	A	A	DESC
38	B	B	B	D	HUM
39	A	A	A	D	HUM
40	B	D	C	B	LOC
41	A	D	D	A	DESC
42	C	D	C	C	DESC
43	D	D	D	A	DESC
44	D	B	D	D	DESC
45	B	D	D	B	DESC
46	D	D	D	C	ENTY
47	D	B	B	D	DESC
48	C	C	C	B	ENTY
49	D	D	D	C	DESC
50	A	D	A	C	DESC
51	A	A	D	A	ENTY

Fig. 13. The result of predictions and answers of the Advanced Subjects Test in 2021 in the reading question section.

Although our best results, CNN prediction, may not be good, it also helps us filter two choices in a four choices question on average. Additionally, with a big difference of 25%(by directly guessing), the design direction is correct to solve the reading question in a limited training datasets.

IV. Discussion

A. Cloze

In this part, there's a chance we get a wrong answer directly from the model if there is a multiple-word option such as phrases.

The original model is not able to fit a two-word token into the MASK (the multiple-word tokens will all get the same token ID, meaning it couldn't handle phrases). In order to deal with this problem, we applied another NLP model Phrase-Bert. This model is trained for producing powerful phrase embedding, which can be extremely useful for our problem. After we generate the word vector for the mask from the original model, for each option we simply calculate its similarity with the mask based on the embedding results of Phrase-Bert and choose the one with the highest similarity as our final answer.

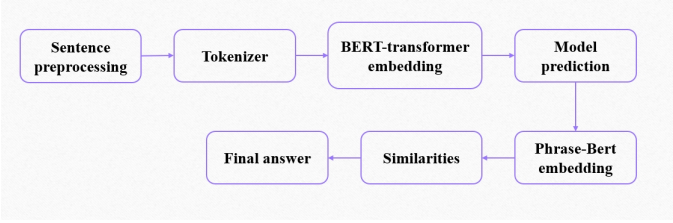
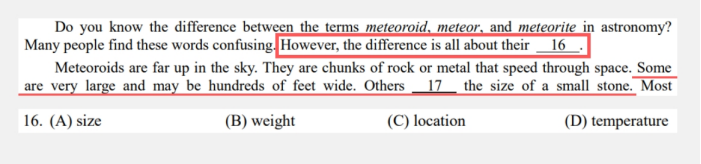


Fig. 14. The updated workflow in the cloze section.

Although there are mostly similar types of questions in one section, some of the questions cannot still be answered correctly.

Take question 16 for instance, the answer actually depends on the sentences located in the next paragraph. As we know,

the calculated vector of the MASK should be mostly dependent on other tokens nearby, the model couldn't generate the best answer based on the given input since the distance between the MASK position and the clue is too far away.


Do you know the difference between the terms *meteoroid*, *meteor*, and *meteorite* in astronomy? Many people find these words confusing. However, the difference is all about their __16__.

Meteoroids are far up in the sky. They are chunks of rock or metal that speed through space. Some are very large and may be hundreds of feet wide. Others __17__ the size of a small stone. Most

16. (A) size (B) weight (C) location (D) temperature

Fig. 15. Example of question 16.

Moreover, another type of questions that we frequently get wrong answers are grammar related problems. Since different tenses of a verb can share the same or similar word vector values, the result obtained from the model could not be concrete. Therefore, another model for grammar checking should be applied in this case. Classifying these special types of questions could take much effort. Since we haven't thought of an ideal way to detect this question type and we mainly focus on the section of reading question for this project, we've decided to neglect the minor special cases in these sections.

B. Sentences Reordering

In this section, we discovered trap answer sentences, where the scores in each question blank are quite high, and the tone of the sentences fit perfectly in each blank. This is likely to be a design by the exam center, with the intention of leading students to fill in coherent but unified answers. Initially, we did not obtain the best solution by arranging and combining all possible answers, but instead calculated the best answer for each question blank. This was successful in the exam in 2021, however, in the exam in 2019, it was obviously not effective due to the existence of the aforementioned trap answers. This would lead to each question's answer being misleading and meaningless. Arranging and combining all possible answers can avoid this situation, because we calculate the overall score rather than a single score, thus successfully avoiding being disturbed by traps.

We have previously tried to normalize the scores to reduce the impact of trap questions, but the results were not significant. We also tried multiplying the scores with the standard deviation to emphasize the discrete effect of the answers, but this method was not stable. Therefore, we ultimately decided to retain the simple scores without normalization or other adjustments.

C. Reading Question

1) About the Datasets: The size of datasets may not be big enough to help us solve diverse types of reading questions. The reason is that we need to collect the question and tidy it up by ourselves before using it, and this is a time consuming process. If we have a larger dataset, the training can last longer before it overfits the data, thus having a higher accuracy.

2) The effect of question classification: In our results, the different types of questions didn't have a big difference in accuracy. But ideally we have to choose different models for different types of questions. The six classes we classified may not be a good way to classify the question type of reading question.(but it's also hard to find the labeled question sets). In conclusion, we found it is crucial to classify the question, and we tried it but we didn't do it well(may also be affected by the size of datasets).

V. Conclusion

After all procedures, we go through three types of problems and solve them in different ways in NLP. To see the total performance of this project, we finally took the Advanced Subjects Test in 2021 by our machine. It got 52 points in multiple questions(72 points in total). Assuming that the machine can get 23 points in the writing part, totally it got 75 points in this exam(100 points in total).

Unfortunately, its strongest opponent, ChatGPT, got 90 points in the test. Here we can see that there is a room of improvement in our machine, and it has great potential to solve various problems. However, the performance of our machine is better than 70% of people in this exam, which shows the power and effort during the research of this project. In conclusion, it's a good application to solve the problem in exam by NLP and the knowledge we learned in the course.

VI. Author Contribution Statements

109022136 陳致寧: 22% in total, collecting data in past 20 years of exam paper, plan and solve the reading question section, lead the discussion of the whole team.

109022127 徐偉晉: 22% in total, plan and solve the reading question section, focus on the model structure and parameter, help with every issue in each section.

108022138 楊宗諺: 19% in total, plan and solve the cloze section, the main role in integrating and solving the cloze problems, help with arrange the ppt of final presentation.

108022108 紀宇恆: 19% in total, plan and solve the sentences reordering section, a lots effort in design the ppt of final presentation, the main and only speaker of final presentation.

109030039 廖奕愷: 19% in total, plan and solve the cloze section, construct different model in solving various cloze problems.

References

[1] Experimental Data for Question Classification https://cogcomp.seas.upenn.edu/Data/QA/QC/

[2] SentenceTransformers Documentation https://www.sbert.net/

[3] College Entrance Examination Center (data source) https://www.ceec.edu.tw/

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," CoRR, vol. abs/1706.03762, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.0480

[6] Nils Reimers, Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084