Gradescope

t11902210 張一凡

1.(2%) There are some difference between fine-tuning and prompting. Beside fine-tuning, in-context learning enable pre-trained model to give correct prediction on many downstream tasks with a few examples but without gradient descent. Please describe:

A. How encoder-only model (Bert-series) determines the answer in a extractive question answering task?

B. How decoder-only model (GPT-series) determines the answer in a extractive question answering task?

Answer:

A. The input representation of the Bert model can represent a single text sentence or a pair of texts (such as [question, answer]) in a word sequence. For a given word, its input representation can be formed by adding three parts of Embedding. The Bert model uses two new unsupervised prediction tasks to pre-train BERT, namely Masked LM and Next Sentence Prediction. In the extracted question-answering task, the model needs to extract information related to the question from a text and generate an answer, which corresponds to the two models mentioned above. The former randomly mask 15% of the tokens in each sequence during the training process, rather than predicting every word like CBOW in Word2vec. MLM randomly masks some words from input, with the goal of predicting the original vocabulary of the masked words based on their context. Unlike pre-training language models from left to right, MLM targets allow for the fusion of left and right contextual representations, which allows for the pre-training of deep bidirectional transformers. The Transformer encoder does not know which words it will be required to predict, or which have been replaced by random words, so it must maintain a distributed contextual representation for each input word. In addition, since random substitution only occurs in 1.5% of all words, it does not affect the model's understanding of language; The latter randomly divide the data into two equal-sized parts, with one pair of statements in the data being context continuous and the other pair of statements in the data being context discontinuous. Then let the Transformer model identify which pairs of statements are continuous and which pairs are not. The encoder in the Bert model is obtained by combining the above two models or by training them separately. It can represent each word in the text as a vector, which contains the Semantic information of the word. At the same time, there are different measurement methods for different vocabulary

and sentence contexts. In the extracted question-answering task, the input of the Bert model includes a question and a text. The model concatenates the question and text together and then represents each word in the text as a vector through an encoder. These vectors can be used to represent the entire text. Next, the Bert model represents the problem as a vector and fuses the problem vector with the vector of each word through an attention mechanism to obtain information related to the problem in the text. Finally, the Bert model processes this information through a fully connected layer to generate answers. In summary, the Bert model's solution in handling abstract question-answering tasks is to randomly select some words to predict when inputting a sentence, and then replace them with a special symbol. Although the model will eventually see input information at all positions, the words that need to be predicted have been replaced by special symbols, so the model cannot know in advance what words are at these positions. This allows the model to learn the words that need to be filled in these places based on the given labels. The prediction process involves various vector operations, with special symbols such as masks.

Reference Paper:
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Reference link:
https://blog.csdn.net/qq_27586341/article/details/89371017?ops_request_misc=%257B%2522request%255Fid%2522%253A%25221682135383168002172640 97%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=168213538316800217264097&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduend~default-2-89371017-null-null.142^v86^insert_down1,239^v2^insert_chatgpt&utm_term=BERT%3A%20Pre-training%20of%20Deep%20Bidirectional%20Transformers%20for%20Language%20Understanding.&spm=1018.2226.3001.4187

The following is a concise version of the reference **chatgpt**:
(1) Input questions and text paragraphs, and the encoder converts each input marker into a vector representation.
(2) For each tag, the encoder calculates the attention distribution between it and all other tags in the context. This allows the model to focus on the most relevant part of the context.
(3) Use attention distribution to weighted average all tags in the context to obtain a contextual representation.
(4) Combine the problem representation with the context representation to form a new representation vector.
(5) The vector is sent to a classification layer through the softmax activation function, which predicts the answer.

B.In answering the questions in the GPT series, I would like to select GPT2, which is

relatively basic and abundant in the paper, for analysis. GPT is a 12-layer transformer, BERT has a maximum depth of 24 layers, and GPT-2 has 48 layers with a total of 1.5 billion parameters. Its training data is a dataset called WebText, which has undergone some simple data cleaning and covers a wide range of fields. Large models require more data to converge, and experimental results show that the model is still in an underfitting state. Secondly, compared with BERT, it does not use a two-way transformer but still uses a one-way transformer. Secondly, in the pre-training phase, GPT-2 uses a multi-task approach, not only learning on one task but multiple tasks. Each task must ensure that its loss function can converge. Different tasks share the principal transformer parameters. This scheme is based on the previous MT-DNN of Microsoft, This can further enhance the generalization ability of the model, so it still performs very well even without fine-tuning. Simply put, the GPT2 model and GPT model are both unidirectional language models, with the input being a piece of text and the output predicting the probability distribution of the next word. In an extracted question answering task, the GPT model needs to generate answers based on the question and text. Unlike the Bert model, the GPT model is a one-way language model that can only predict the next word based on the previous text and cannot make predictions based on the subsequent text. Therefore, in an extracted question-answering task, the input of the GPT model includes questions and text. The model first concatenates the questions and text together and then inputs them as input sequences into the GPT model in sequence. When the model receives the complete input sequence, it will generate a sequence containing the answer, which includes the model's understanding of the entire input sequence, rather than just a part of the text. At the same time, according to the GPT series model repeatedly discussed by the teacher in class, the GPT series mainly adopts a "one by one" or "one hit fatal" strategy based on probability to complete the generation of a paragraph or answer, which is different from the vector word by word generation in traditional language models. Finally, compared to Bert in the previous question, the GPT-2 and GPT series use the Transformer's Decoder module stacked together, while BERT uses the Transformer's Encoder module built. A key difference between the two is that GPT-2 adopts a traditional language model, which only outputs one word (multiple tokens) at a time.

Reference Paper:
Language Models are Unsupervised Multitask Learners

Reference link:
https://blog.csdn.net/u013602059/article/details/107280181/?ops_request_misc=&request_id=&biz_id=102&utm_term=Language%20Models%20are%20Unsupervis&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduweb~default-1-107280181.142^v86^insert_down1,239^v2^insert_chatgpt&spm=1018.2226.3001.4187

The following is a concise version of the reference **chatgpt**:
(1) Input the problem and encode it into a vector representation.
(2) Enter the beginning of a text paragraph and use a decoder to generate a sequence that

gradually expands until the end of the generated text paragraph. Each tag in the generated sequence is based on the previous tag.
(3) At each generated tag, the softmax activation function is used to predict the next possible tag. When making predictions, the model takes into account all previously generated tags and problem representations.
(4) During the generation process, the problem representation is embedded into each generated marker vector.
(5) Once a complete text paragraph is generated, the decoder embeds the question representation into all tag vectors and predicts the answer through the softmax activation function.

2. (2%) The goal of this homework is to fine-tune a QA model. In this question, We wish you to try In-context learning on the same task and compare the difference between the prediction of in-context learning and fine-tuning.

A. Try to Instruct the model with different prompts, and describe your observation on at least 3 pairs of prompts comparison.

B. Paste the screenshot of your prediction file in A. to Gradescope. (There are at least 6 screenshots)

Answer:

A. Here are my three sets of modifications and specific description analysis.
B. A screenshot accompanied by a comparison of the corresponding results.

（1）Prompts with different languages

I chose language modification in the first prompt modification, and I used the sentence "Please find the answer to the last question from the last article" in Chinese, English, and French to compare the observation results. The following are the parts of the code that have been modified.

```
for idx, qa in enumerate(test["questions"]):
    # You can try different prompts
    prompt = "Trouvez la réponse à votre dernière question à partir du dernier article\n" #French
    #prompt = "Please find the answer to the last question from the last article\n" #English
    #prompt = "請從最後一篇的文章中找出最後一個問題的答案\n" #Chinese
    exist_question_indexs = [question_ids.index(qa["id"])]
```

The following is the image result display for problem B.

I.Chinese

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣
```

II.English

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2014年6月
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,廣州的歷史與歷史與歷史與歷史與歷史與歷史與歷史與歷史與歷史
10,7月,2010年8月1日
11,2008年,廣州获得「創建國家健康城市」獎項的時候
12,從化市,廣州的
13,80,80%
14,非洲黑人,非洲黑人
15,普通話,廣州的廣播和普通話是主要交流語言。廣州的廣播和
```

III.French

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。文章:廣州是中國最早的工業區之一。在20世紀50
3,2015,1990
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,越秀區
8,香港,香港
9,1921,2015
10,7月,2010年7月
11,2008年,廣州在1990年以前已經获得了從「創建國家健康城市」获得的獎
12,從化市,廣州的人口密度是廣州市的百倍分之一以下。
13,80,非洲黑人
14,非洲黑人,非洲黑人
15,普通話,廣州地區的歷史與文化與廣東地區的歷史與文化有密切的关系。
```

Regarding the first type of prompt modification, my observation is as follows: among the prompt modifications made in three languages, the accuracy is high in many individual word Q&A questions, but the accuracy is different in the four questions of 0, 9, 12, and 15. After analyzing the differences and representativeness of each question, I combined relevant information to draw the following conclusion.

The impact of observing prompts in different languages on contextual models may be due to the following reasons: different language structures: Chinese, English, and French

have different language structures, which may affect the model's ability to understand and infer text. Therefore, different language prompts may lead to different prediction results. Semantic differences: There are significant semantic differences between different languages, and even when expressing the same meaning, different words and phrases may be used. Therefore, using the same prompts in different languages may lead to different prediction results. In addition, insufficient training data may also have a certain impact. If the model has less training data in a certain language, its performance in that language may be affected. Therefore, using prompts in that language may lead to a decrease in prediction accuracy. In summary, based on factors such as language structure, semantic differences, and the quantity and quality of training data, using prompts in different languages may have different impacts on the predictive accuracy of the context model. Therefore, in order to obtain the best prediction results, the model should be tested and adjusted in different languages to find the most suitable prompts for a specific language. In this test, Chinese is still more accurate.

（2）Prompts with different formats

For the change of the second prompt, I chose to change the format of the prompt statement, which is to change the format of the sentence. I have decided to switch to three different methods. The first type is the conventional sentence structure given for the question; The second type is to provide an example of an answer, which is to describe the question and answer in a specific position in the prompt in comparison; The third type is the sentence structure proposed later with the condition changed. I believe that changes in sentence structure will have an impact on the final inconext result. The specific program design is shown in the picture below.

```
for idx, qa in enumerate(test["questions"]):
    # You can try different prompts
    #prompt = "請從最後一篇的文章中找出最後一個問題的答案\n" #normal format
    prompt = "在第一篇文章的問題\"從哪一天開始在廣州市內騎機車會被沒收？\"中答案為\"2007年1月16日\"\n" #example format
    #prompt = "我們需要最後一個問題的答案，依據是最後一篇文章\n" #Swap word order
    exist_question_indexs = [question_ids.index(qa["id"])]
```

The following is the image result display for problem B.

I.normal format

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣
18
```

II. example format

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日文章:廣州爆發了幾次武裝起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣
```

III. Swap word order

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年8月1日
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣
```

Based on the deformation results of the three questions presented above, I found that the deformation in the form of the problem has a smaller impact on the recognition and output of the incontext model compared to language changes. However, for answers similar to question 7, it can be seen through comparison that changes in these sentence structures will affect the judgment of the relevant sentence content. In the second and third types of deformation examples, the word order is different from the original sentence structure, I personally believe that imitating the form of 'in P, the answer to Q is A' is more conducive to the output of the target answer.

In the context model, different formats of prompts will have different impacts on the prediction accuracy of the model. This is because the model learns different semantic and syntactic information in different prompts. For example, the original sentence structure may contain more details and context information, which can help the model better understand the problem and improve its accuracy; Changing the sentence structure of the grammatical order may make the model more difficult to understand the problem, thus reducing the accuracy of the model. In addition, the prompt format of "In P, the answer to Q is A" may make the model easier to understand the problem, as it clarifies the key information of the problem, thereby improving the accuracy of the model. Therefore, when evaluating and analyzing prompts in different formats, we should consider their impact on the model and choose the most suitable prompt format based on the specific situation to improve the

accuracy of the model. This also emphasizes that selecting appropriate prompt formats is crucial for improving the performance of QA models when training and using them. For example, the results achieved by using the example method here are more accurate than the original prompts, and changing the sentence order can have an impact, but it is not too obvious.

（3）prompt that comparison with limited conditions

For the variation of the third prompt, I chose to use a prompt with limited conditions. In the two additional variations, I used restrictions on answer sentence structure and word count, which intuitively affect the accuracy of the final answer. The specific code modifications are shown below.

```
for idx, qa in enumerate(test["questions"]):
    # You can try different prompts
    #prompt = "請從最後一篇的文章中找出最後一個問題的答案\n" #normal prompt
    prompt = "答案限制在兩句話內，請根據最後一篇文章回答最後一個問題\n"  # Sentence structure restriction
    #prompt = "答案限制在10個字內，請根據最後一篇文章回答最後一個問題\n"  # WORD LIMIT
    exist_question_indexs = [question_ids.index(qa["id"])]
```

The following is the image result display for problem B.

I.normal

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣
```

II. Sentence structure restriction

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港和香港地區
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,穩定
14,非洲黑人,非洲黑人
15,普通話,廣州的公共场所,例如,廣州的公共交通工具,廣州的公共
```

III. WORD LIMIT

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,穩定
14,非洲黑人,非洲黑人
15,普通話,廣州的公共场所,例如,廣州的市廳,廣州的市廳

```

For this third type of prompt deformation, by observing three different sets of outputs, it can be found that it has a smaller impact on the result output compared to the first two deformations, mainly concentrated in question 15. The output length and sentence breaks are different, but the overall effect is not as good as adding corresponding restriction requirements directly to the question. However, the modification of such additional restriction sentences also has a slight improvement effect on the final result. The specific advantages and disadvantages are analyzed as follows.

In summary, in the context model, adding conditional constraints to different prompts can have an impact on the prediction accuracy of the model. For example, if the answer sentence structure is limited, such as only accepting a specific answer sentence structure, the model will be more accurate in predicting the answer, because the model has learned the semantic representation of a specific sentence structure. However, this restriction can also bring some problems, such as if the restriction is too strict, the model may miss some reasonable answers, such as the word limit I added in the second change, which may perform well on this issue. However, for some large text predictions, it may be necessary to have prior knowledge of statistical data. Similarly, limiting the number of response words can also affect the predictive accuracy of the model. If the upper limit of the number of words in the answer is limited, the model may prefer to choose shorter answers because they better meet the limiting conditions. However, this may also lead to the model ignoring some important information, thereby affecting the accuracy of the prediction. Therefore, when selecting prompts, it is necessary to weigh various constraints based on specific circumstances to obtain more accurate prediction results. However, compared to the previous two prompt word modifications, this item is not as outstanding and will improve the results, but it is not too obvious.

B.Here, I will also answer the texture question of the second question separately. This section is only for the convenience of displaying the texture of the second question. For specific analysis, please refer to the first question. In the first question, I also presented all the textures for comparison and observation.

（1）Prompts with different languages

I.Chinese

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣

```

II.English

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2014年6月
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,廣州的歷史與歷史與歷史與歷史與歷史與歷史與歷史與歷史與歷史
10,7月,2010年8月1日
11,2008年,廣州获得「創建國家健康城市」獎項的時候
12,從化市,廣州的
13,80,80%
14,非洲黑人,非洲黑人
15,普通話,廣州的廣播和普通話是主要交流語言。廣州的廣播和

```

III.French

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。文章:廣州是中國最早的工業區之一。在20世紀50
3,2015,1990
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,越秀區
8,香港,香港
9,1921,2015
10,7月,2010年7月
11,2008年,廣州在1990年以前已經获得了從「創建國家健康城市」获得的獎
12,從化市,廣州的人口密度是廣州市的百倍分之一以下。
13,80,非洲黑人
14,非洲黑人,非洲黑人
15,普通話,廣州地區的歷史與文化與廣東地區的歷史與文化有密切的关系。

```

(2) Prompts with different formats

I.normal format

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣

```

II. example format

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日文章:廣州爆發了幾次武裝起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣
```

III. Swap word order

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年8月1日
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣

```

(3) prompt that comparison with limited conditions

I.normal

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,首都。
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,廣州的社會環境
14,非洲黑人,非洲黑人
15,普通話,廣州的廣東人,廣東人,廣東人,廣東人,廣
```

II. Sentence structure restriction

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港和香港地區
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,穩定
14,非洲黑人,非洲黑人
15,普通話,廣州的公共场所,例如,廣州的公共交通工具,廣州的公共
```

III. WORD LIMIT

```
ID,Ground-Truth,Prediction
0,2007年1月16日,2007年1月16日
1,200公里,200公里
2,廣州,上海
3,2015,1990年
4,文化大革命,文化大革命
5,廣州,廣州
6,廣州,廣州
7,武昌起義,1911年10月10日武昌起義
8,香港,香港
9,1921,公元前214年,秦始皇出兵徵兵南,南海縣任
10,7月,2010年7月
11,2008年,廣州获得「創建國家健康城市」獎項
12,從化市,越秀區
13,80,穩定
14,非洲黑人,非洲黑人
15,普通話,廣州的公共场所,例如,廣州的市廳,廣州的市廳
```