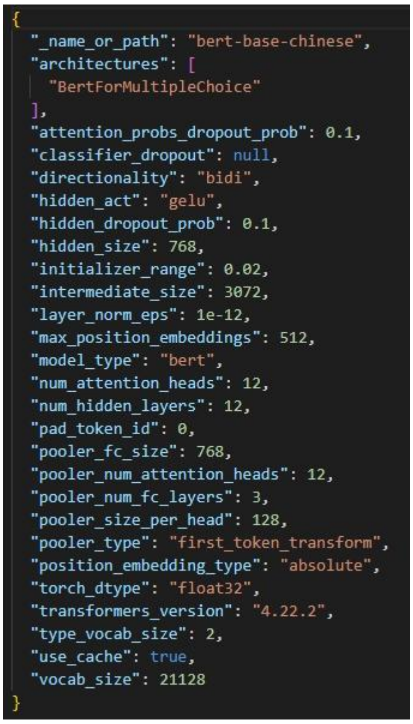
Q1 : Data processing

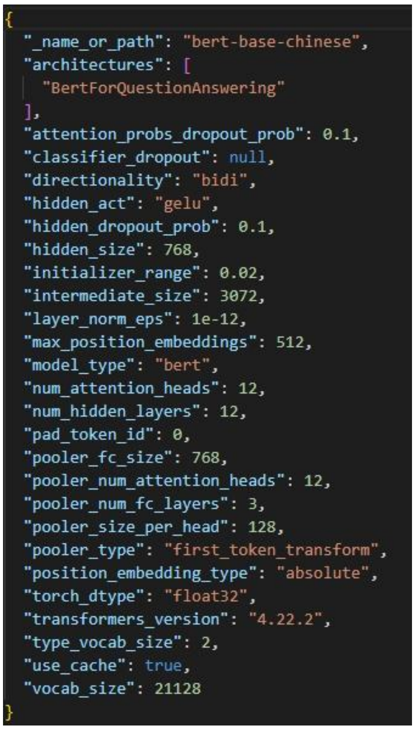
1. Tokenizer
    a. Bert tokenizer 是一種類似byte pair encoding的一種方式, 會把word切成一塊一塊的subword來避免一些unknown的word或是相似的字互相影響。而中文與英文在定義單位的時候是不相同的, 英文是以subword的概念而中文是以字為基本單位。
    bert toknizer的步驟老師課程中有解釋為:
    Step1 : Define vocabulary size
    Step2 : 將word切成character
    Step3 : 根據前面的資料建立其language model
    Step4 : 選擇能夠進步最大likelihood的subword添加進去
    Step5 : 重複step4直到threshold
2. Answer Span
    a. 我用tokenizer裡面的return_offset_mapping, 這個方法會回傳每個token對應的(char start, char end), 因此只後只要根據這兩個當成相對應的start position跟end position.
    b. 每一組的start/end會進行一個配對機率的相乘, 然後其中會將一些不符合條件的配對進行刪除, 像是一些subsentence比原本sentence長的或是一些end position < start position的, 然後找出機率最大的就是最後我們要用的predict.

Q2 : Modeling with BERTs and their variants.

1. BERT
    a. Configuration(使用bert-base-chinese)

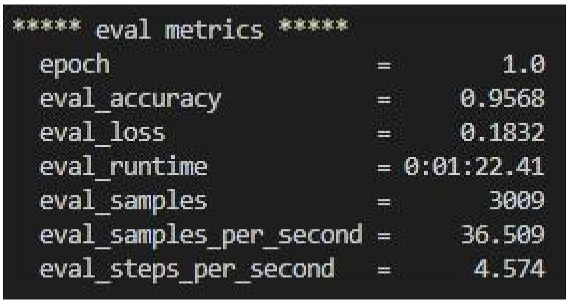


```
{
  "_name_or_path": "bert-base-chinese",
  "architectures": [
    "BertForMultipleChoice"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "directionality": "bidi",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 0,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "position_embedding_type": "absolute",
  "torch_dtype": "float32",
  "transformers_version": "4.22.2",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```



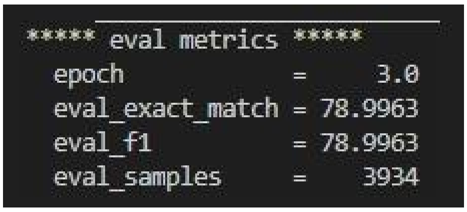
```
{
  "_name_or_path": "bert-base-chinese",
  "architectures": [
    "BertForQuestionAnswering"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "directionality": "bidi",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 0,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "position_embedding_type": "absolute",
  "torch_dtype": "float32",
  "transformers_version": "4.22.2",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```

    b. Performance
    Context selection accuracy : 0.9568



```
***** eval metrics *****
  epoch                   =        1.0
  eval_accuracy           =     0.9568
  eval_loss               =     0.1832
  eval_runtime            = 0:01:22.41
  eval_samples            =       3009
  eval_samples_per_second =     36.509
  eval_steps_per_second   =      4.574
```

Question answering EM: 0.7899
Question answering F1: 0.7899



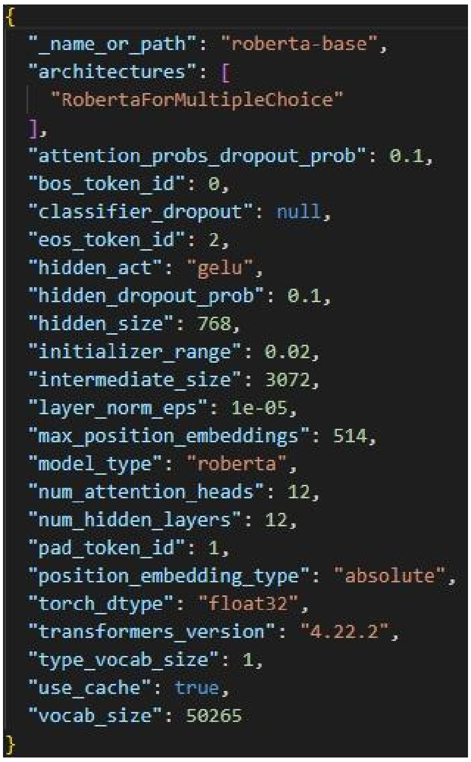
```
***** eval metrics *****
  epoch            =     3.0
  eval_exact_match = 78.9963
  eval_f1          = 78.9963
  eval_samples     =    3934
```

c. Loss function
Cross Entropy

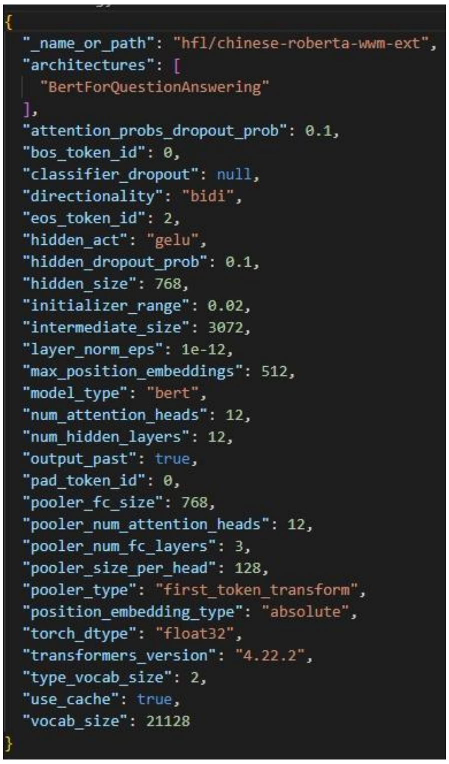
d. Training arguments
<u>Context selection</u>
optimization : adamw(lr=3e-5)
batch size:2
gradient accumulation:8
num_train_epochs:1
max_len:512
<u>Qustion answering</u>
optimization : adamw(lr=3e-5)
batch_size : 2
gradient accumulation:8
num_train_epochs:3
max_len:512

2. Roberta-wwm-ext

   a. Configuration(使用hfl/chinese-roberta-wwm-ext)



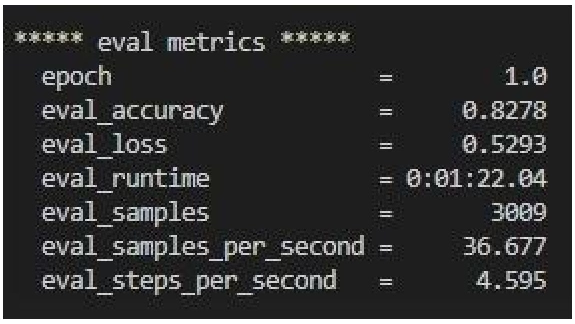
```
{
  "_name_or_path": "roberta-base",
  "architectures": [
    "RobertaForMultipleChoice"
  ],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "classifier_dropout": null,
  "eos_token_id": 2,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "roberta",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 1,
  "position_embedding_type": "absolute",
  "torch_dtype": "float32",
  "transformers_version": "4.22.2",
  "type_vocab_size": 1,
  "use_cache": true,
  "vocab_size": 50265
}
```



```
{
  "_name_or_path": "hfl/chinese-roberta-wwm-ext",
  "architectures": [
    "BertForQuestionAnswering"
  ],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "classifier_dropout": null,
  "directionality": "bidi",
  "eos_token_id": 2,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "output_past": true,
  "pad_token_id": 0,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "position_embedding_type": "absolute",
  "torch_dtype": "float32",
  "transformers_version": "4.22.2",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```

b. Performance

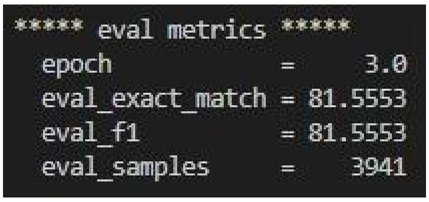
Context selection accuracy : 0.8278



```
***** eval metrics *****
  epoch                   =        1.0
  eval_accuracy           =     0.8278
  eval_loss               =     0.5293
  eval_runtime            = 0:01:22.04
  eval_samples            =       3009
  eval_samples_per_second =     36.677
  eval_steps_per_second   =      4.595
```

Question answering EM: 0.8155

Question answering F1: 0.8155



```
***** eval metrics *****
  epoch            =     3.0
  eval_exact_match = 81.5553
  eval_f1          = 81.5553
  eval_samples     =    3941
```

c. Diffierence

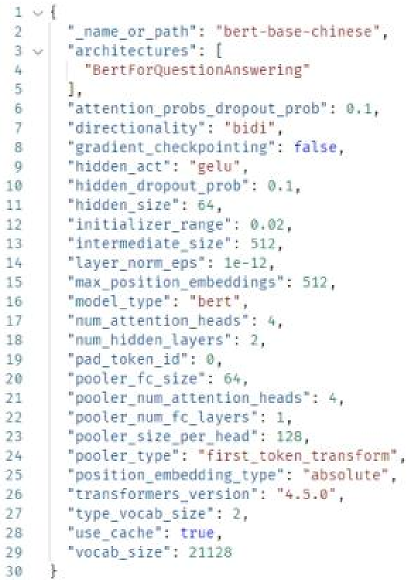
roberta與bert不一樣的地方在於masking，roberta的mask position不同於bert會在一定的時間內動態條整位置，並且其mask的方式也不同於bert，在roberta中如果詞有部分被mask則會講整個詞都給mask掉。

Q3 : Curve

Q4 : Pre Trained vs Not Pretrained

a. Configuration

減少了layer, hidden size的數量



```json
{
  "_name_or_path": "bert-base-chinese",
  "architectures": [
    "BertForQuestionAnswering"
  ],
  "attention_probs_dropout_prob": 0.1,
  "directionality": "bidi",
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 64,
  "initializer_range": 0.02,
  "intermediate_size": 512,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 4,
  "num_hidden_layers": 2,
  "pad_token_id": 0,
  "pooler_fc_size": 64,
  "pooler_num_attention_heads": 4,
  "pooler_num_fc_layers": 1,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "position_embedding_type": "absolute",
  "transformers_version": "4.5.0",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```

b. Performance

Context selection accuracy : 0.4667

Question answering EM : 0.0509

Question answering F1: 0.0509

c. Compare

用沒有pretrained的model loss會長時間維持在一個很高的數字，且即使最後loss下降了但validation的效能卻沒有跟著上升，我想這可能是overfitting的問題，應該需要更多的訓練資料才會訓練得比較好。