1. (2%) After your model predicts the probability of answer span start/end position, what rules did you apply to determine the final start/end position? (the rules you applied must be different from the sample code)
因為 start 的 index 一定要小於 end，因此我用 if 判斷，如果 start_idx<=end_idx 才代表是答案，否則就略過。

2. (2%) Try another type of pretrained model which can be found in huggingface's Model Hub (e.g. BERT -> BERT-wwm-ext, or BERT -> RoBERTa), and describe

- the pretrained model you used
  hfl/chinese-macbert-large
- performance of the pretrained model you used
  Best public score is 0.81928
- the difference between BERT and the pretrained model you used (architecture, pretraining loss, etc.)

原始 BERT 模型使用[MASK] token 進行掩蔽，但是[MASK] token 在微調階段從未出現，這會造成預訓練任務與下游微調任務不一致；因此 macbert 使用類似的單詞來掩蔽需要被掩蔽的單詞。

	Chinese	English
Original Sentence	使用语言模型来预测下一个词的概率。	we use a language model to predict the probability of the next word.
+ CWS	使用 语言 模型 来 预测 下 一个 词 的 概率 。	-
+ BERT Tokenizer	使 用 语 言 模 型 来 预 测 下 一 个 词 的 概 率 。	we use a language **model** to **pre ##di ##ct** the **pro ##ba ##bility** of the next word .
Original Masking	使 用 语 言 **[M]** 型 来 **[M]** 测 下 一 个 词 的 概 率 。	we use a language **[M]** to **[M] ##di ##ct** the **pro [M] ##bility** of the next word .
+ WWM	使 用 语 言 **[M] [M]** 来 **[M] [M]** 下 一 个 词 的 概 率 。	we use a language **[M]** to **[M] [M] [M]** the **[M] [M] [M]** of the next word .
++ N-gram Masking	使 用 **[M] [M] [M] [M]** 来 **[M] [M]** 下 一 个 词 的 概 率 。	we use a **[M] [M]** to **[M] [M] [M]** the **[M] [M] [M] [M] [M]** next word .
+++ Mac Masking	使 用 语 法 建 模 来 预 见 下 一 个 词 的 几 率 。	we use a **text system** to **ca ##lc ##ulate** the **po ##si ##bility** of the next word .

Figure 1: Examples of different masking strategies.

模型架構比較

	BERT	ERNIE	XLNet	RoBERTa	ALBERT	ELECTRA	MacBERT
Type	AE	AE	AR	AE	AE	AE	AE
Embeddings	T/S/P	T/S/P	T/S/P	T/S/P	T/S/P	T/S/P	T/S/P
Masking	T	T/E/Ph	-	T	T	T	WWM/NM
LM Task	MLM	MLM	PLM	MLM	MLM	Gen-Dis	Mac
Paired Task	NSP	NSP	-	-	SOP	-	SOP

TRAINING DETAILS OF CHINESE PRE-TRAINED LANGUAGE MODELS.

	BERT	BERT-wwm	RoBERTa-wwm	RBT	ELECTRA	MacBERT
Word #	0.4B	5.4B	5.4B	5.4B	5.4B	5.4B
Vocab #	21,128	21,128	21,128	21,128	21,128	21,128
Hidden Activation	GeLU	GeLU	GeLU	GeLU	GeLU	GeLU
Optimizer	AdamW	LAMB	AdamW	AdamW	AdamW	LAMB
Training Steps (base/large)	?	2M	1M / 2M	1M	1M / 2M	1M / 2M
Initial Checkpoint (base)	random	BERT	BERT	RoBERTa	random	BERT

模型表現比較

DRCD	Dev		Test	
	EM	F1	EM	F1
BERT	83.1 (82.7)	89.9 (89.6)	82.2 (81.6)	89.2 (88.8)
BERT-wwm	84.3 (83.4)	90.5 (90.2)	82.8 (81.8)	89.7 (89.0)
BERT-wwm-ext	85.0 (84.5)	91.2 (90.9)	83.6 (83.0)	90.4 (89.9)
RoBERTa-wwm-ext	86.6 (85.9)	92.5 (92.2)	85.6 (85.2)	92.0 (91.7)
ELECTRA-base	87.5 (87.0)	92.5 (92.3)	86.9 (86.6)	91.8 (91.7)
MacBERT-base	**89.4 (89.2)**	**94.3 (94.1)**	**89.5 (88.7)**	**93.8 (93.5)**
ELECTRA-large	88.8 (88.7)	93.3 (93.2)	88.8 (88.2)	93.6 (93.2)
RoBERTa-wwm-ext-L	89.6 (89.1)	94.8 (94.4)	89.6 (88.9)	94.5 (94.1)
MacBERT-large	**91.2 (90.8)**	**95.6 (95.3)**	**91.7 (90.9)**	**95.6 (95.3)**

Table 5: Results on DRCD (Traditional Chinese).

ref: https://zhuanlan.zhihu.com/p/250595837

ref: https://arxiv.org/abs/2004.13922