A Study on Contextualized Language Modeling for Machine Reading Comprehension
(上下文語言模型化技術於閱讀理解之研究)

吳沁穎 Chin-Ying Wu
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
elain1224@gmail.com

許永昌 Yung-Chang Hsu
易晨智能股份有限公司
mic@ez-ai.com.tw

陳柏琳 Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

摘要

隨著深度學習的發展，機器閱讀理解的研究已有了長足的進步，並在許多實際應用情境上展露頭角。機器閱讀理解是一項用於評估機器對語言的理解能力的自然語言處理任務，其形式為：給定文章段落與相關問題，電腦自動根據文章段落與相關問題來進行回答。本研究嘗試使用兩種以 BERT 為基礎的預訓練語言模型：BERT-wwm 和 MacBERT，發展能夠達到更佳預測表現的機器閱讀理解方法。此外，考慮到閱讀理解中的文章類型可能對於回答模式有潛在影響，我們針對訓練資料集的文章段落進行分群，以此做作為額外資訊結合到語言模型的輸入。另一方面，我們也探索使用集成學習法來結合上述兩種預訓練語言模型，以進一步提升機器閱讀理解的表現。

Abstract

With the recent breakthrough of deep learning technologies, research on machine reading comprehension (MRC) has attracted much attention and found its versatile applications in many use cases. MRC is an important natural language processing (NLP) task aiming to assess the ability of a machine to understand natural language expressions, which is typically operationalized by first asking questions based on a given text paragraph and then receiving machine-generated answers in accordance with the given context paragraph and questions. In this paper, we leverage two novel pretrained language models built on top of Bidirectional Encoder Representations from Transformers (BERT), namely BERT-wwm and MacBERT, to develop effective MRC methods. In addition, we also seek to investigate whether additional incorporation of the categorical information about a context paragraph can benefit MRC or not, which is achieved based on performing context paragraph clustering on the training dataset. On the other hand, an ensemble learning approach is proposed to harness the synergistic power of the aforementioned two BERT-based models so as to further promote MRC performance.

關鍵字：深度學習、自然語言處理、機器閱讀理解、語言模型

Keywords: Deep Learning, Natural Language Processing, Machine Reading Comprehension, Language model

文章段落： 新北市的人口眾多，市區的交通流量十分龐大。每逢尖峰時段或假日，經常會有大量人潮、車潮流動於市區內或臺北、新北兩市之間，導致市區內各重要幹道常出現交通阻塞的情形。……
問題： 新北市的交通流量龐大的狀況與何有關？
可能回答： 人口眾多

圖 1、閱讀理解問題範例

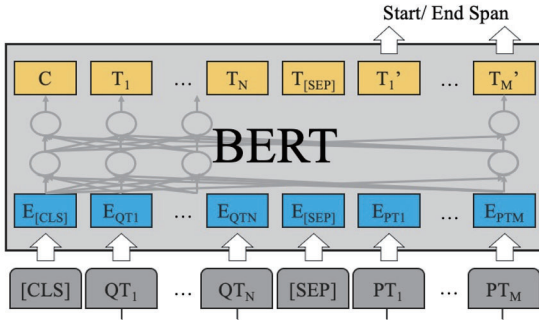

圖 2、BERT 在閱讀理解任務之應用示意圖

1 緒論

隨著各領域的文本數據大量產生，傳統的人工處理方式受限於速度、人力成本等因素使其逐漸成為產業發展的瓶頸；與此同時，能自動分析文本，並且從中抽取語意知識的機器閱讀理解 (Machine reading comprehension, MRC) 技術也漸漸開始受到關注。機器閱讀理解的主要應用的方向為：在既有的文本中查詢目標知識。例如：自動客服，可以從產品相關說明資料中找到與用戶描述相符的部分並給出詳細解答；在醫療領域，模型可以根據患者的症狀描述自動查詢大量病例與醫療論文，尋找相關的資訊與診療方式。舉凡需要分析大量文本的任務，都能夠以機器閱讀理解模型進行協助。

機器閱讀理解是一個典型的自然語言處理任務，用以評估機器對於語言的理解能力。任務的進行方法為：給定一段文章段落與一個相關的問題，機器需要根據文章進行回答。；此篇著重在段落擷取 (Span Extraction) 的類型，亦即，在任務中，模型在需要從給定的文章中擷取一個段落作為回答。範例如圖 1。

過去傳統的類神經網路架構是由多個不同功能的模組構成，研究的主軸在於如何運用注意力機制 (Attention-based) 讓模型取得更豐富的文章段落與問題之間的交互關係，例如：Attention Sum (Kadlec et al. 2016), Gated attention (Dhingra et al. 2017), Self-matching (Wang et al. 2017), Attention over Attention (Cui et al. 2017) Bi-attention (Seo et al. 2016) 。近年來，由於預訓練語言模型 (Pre-trained language model) 的出現，例如 ELMo (Peters et al. 2018) 、GPT (Radford et al. 2018) 、BERT (Devlin et al. 2019) ，使得機器閱讀理解中相當大一部分的模組功能都可以以此取代，並且因其具有更加豐富的語意資訊，使得後續模型皆以預訓練語言模型作為主幹進行研究。

本篇研究使用了兩個基於 Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) 架構的預訓練語言模型：BERT - Whole Word Masking (BERT-wwm) (Yiming Cui et al. 2019)、Masked As Correction BERT (Mac-BERT) (Yiming Cui et al. 2020)，兩者皆為針對中文語言特性提出的模型。使用的資料集為兩個機器閱讀理解資料集：簡體中文的 CMRC (Cui et al., 2019) 與繁體中文的 DRCD (Shao et al., 2018)。實驗有兩個主要的提升模型表現的方向。首先，考慮到資料集中的標題資訊可能帶有一些資訊，且可能對於回答模式有潛在影響，故利用此對所有文章進行分群 (Clustering) ，並以分群的結果作為額外資訊結合到語言模型的輸入，再以相同的方式重新訓練模型。最後一個部分，由於前述兩個語言模型都已經能達到一定的成果，故使用一個相對較簡單、快速的模型增進方法：集成學習法 (Ensemble Learning Method) ，將對於兩個模型的預測分數進行平均，以獲得更佳的結果。

2 相關研究

2.1 語言模型

由於語言模型有著可以提供更豐富的詞彙關聯性資訊、降低模型架構成本等優勢，在自

Strategies	Example
Original Sentence	使用語言模型來預測下一個詞的頻率。
+ BERT Tokenizer	使 用 語 言 模 型 來 預 測 下 一 個 詞 的 頻 率 。
+ CWS	使用 語言 模型 來 預測 下一個 詞 的 頻率 。
Original Masking	使用語言 [M] 型來[M] 測下一個詞的頻率。
+ WWM	使用語言 [M] [M] 來[M] [M] 下一個詞的頻率。
++ N-gram masking	使用[M] [M] [M] [M] 來 [M] [M] 下一個詞的頻率。
+++ Mac masking	使用語法建模來預見下一個詞的頻率。

表 1、不同切分策略與遮蓋策略示意圖。“+”代表沿用前述策略設定。

然語言處理的任務常選擇加入語言模型輔助以增進表現。近年來，預訓練的深層類神經網路語言模型，如 ELMo (Peters et al., 2018)、GPT (Radford et al., 2018)、BERT (Devlin et al., 2019) 等模型陸續被提出。這類的模型預先在其他相關的任務上以大量資料訓練，再將所學知識遷移到新的任務。如此一來，除了可以克服目標任務資料不足的問題，也能利用學到的知識提高目標模型的準確度。歸功於這些優點，預訓練的語言模型在各個自然語言處理領域都取得顯著的進展。其中，BERT 模型是最被廣泛運用在不同任務上的語言模型之一。BERT 運用 Transformer (Radford et al., 2018) 的自注意力機制 (Self-attention mechanism) 學習文本中單詞之間的上下文關係，並且由於其雙向進行的特性，使得上文與下文的資訊能更充分地被使用。BERT 在當時最知名的英文閱讀理解任務：SQuAD (Rajpurkar et al., 2016) 上展現了這方面的能力，不僅超越了當時所有的類神經網路模型，也改變了機器閱讀理解領域的研究模式，使得目前的模型大多是使用預訓練語言模型為主幹的方法。

2.2 機器閱讀理解

隨著資料集的發佈，機器閱讀理解開始受到越來越多研究者的關注。傳統類神經網路架構的機器閱讀理解模型可以分為四個核心模組，轉換文字為表徵的模組 (embedding)、特徵抽取模組 (feature extraction) 、文章段落與問題交互關係模組 (context-question interaction) 及預測答案模組 (answer prediction)。早期的研究趨勢在於文章段落與問題之間的交互關係，主要以基於注意力機制作為研究重點。研究包括 Attention Sum (Kadlec et al., 2016)、Gated attention (Dhingra et al. 2017)、Self-matching (Wang et al., 2017)、Attention over Attention (Cui et al., 2017)、Bi-attention (Seo et al., 2016) 等。近年來，預訓練語言模型陸續提出，由於其具備的豐富資訊，使得上述前三個模組的功能，都得以用一個預訓練語言模型就完全囊括。因此，預訓練的語言模型逐漸取代過去的注意力機制模型，開始作為機器閱讀理解模型的主要結構，並且在最近幾年取得了非常優秀的成果。這些預訓練語言模型包括 ELMo (Peters et al., 2018)、GPT (Radford et al., 2018)、BERT (Devlin et al., 2019)、XLNet (Yang et al., 2019)、RoBERTa (Liu et al., 2019)、ALBERT (Lan et al., 2020)、ELECTRA (Clark et al., 2020)。

3 研究方法

本研究主要探討 BERT 與兩種針對中文的語言模型：BERT-wwm (Yiming Cui et al. 2019)、MacBERT (Yiming Cui et al. 2020) 在閱讀理解任務上的表現，將分為三個部分進行。第一個部分，分別對三種語言模型進行最基本的微調。第二個部分，實驗目標是更有效利用文章的類別性質，故將資料集中的標題資訊作為分群基準，並將分群資訊加入模型的輸入端重新訓練模型。最後一個部分是集成模型，合併兩個模型的結果以提昇表現。

3.1 BERT

BERT 為由 Google 提出的預訓練語言模型，全名為 Bidirectional Encoder Representations from Transformers。主要架構為 Transformers 的編碼器 (Encoder)，使用 Masked Language model (MLM) 與 Next Sentence Prediction (NSP) 作為

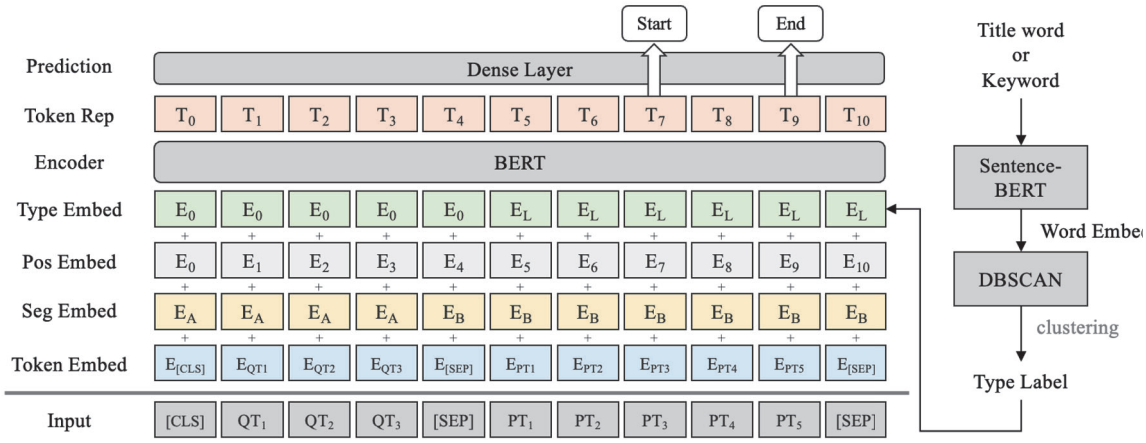

圖 3、文章標題分群與模型輸入整合之流程示意圖

訓練方式。句子在輸入模型後會切割句子，並以 tokenizer 轉換成以詞為單位的 token 序列；其中，MLM 是以特殊 token：[MASK] 隨機遮蔽 (Masking) 掉原始的 token 並進行遮蔽位置單詞的預測，目的是讓機器學習使用僅剩的上下文資訊推測目標 token 適合填入的單詞。此篇用的是同樣由 Google 提出的中文版的 BERT，使用的訓練語料為中文的維基百科文章，並且在所有文字間填入空格以利切分。切分前後狀態可參考表 1。

本實驗預計對模型進行微調，以建構成適合用於機器閱讀理解的模式。如圖 2 所示，首先將資料集中的一個問題與對應文章進行串接，前端會加入辨識輸出位置的特殊 token：[CLS]，兩段段文字之間則有 [SEP] ，用以區別文章與問題。此串 token 序列作為模型的輸入，BERT 模型會根據每個 token 的資訊生成一個對應此 token 的表徵 (representation) ，並依此算出文章中每個單詞適合作為起點和終點的機率。令 $\mathbf{T}_i$ 為 BERT 生成的第 i 個 token 的表徵，$\mathbf{S} \in \mathbb{R}^h$ 為起始點向量 (start vector)，其中 h 代表 token 表徵的大小。則起始機率可以表示為下列公式。

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \qquad (1)$$

終點機率算法概念相同。損失函數的計算為起點與終點的交叉熵損失 (cross-entropy loss) 平均值。模型最終會輸出根據起點與終點的一段文章段落。

3.2 BERT-wwm

BERT-wwm (Yiming Cui et al. 2019) 全名為 BERT - whole word masking，是考慮中文語言特性的 BERT 延伸模型。最早的 BERT 是使用英文作為訓練語料，以空格為基準所切分，切分出的英文單字是具有完整的意義的最基本單位，並以此作為遮蓋的單位；BERT 在運用到中文版本上時，會預先在所有字之間加入空格，再以字為單位進行切分，這也是訓練時的遮蓋的單位。然而，儘管字並非無意義，卻不一定會與中文使用者解讀句子的意思吻合。故以字作為單位訓練時，可能會產生一些誤差。考慮到中文的語言特性，BERT-wwm 在切分詞時使用 LTP (Che et al., 2010) 作為切分中文的工具 (Chinese Word Segmentation, CWS)，並以實際用於解讀意義的全詞作為 MLM 訓練時的遮蓋單位，用以接近中文在使用時的情境。實際使用情境與 BERT tokenizer 的比較可參考表 1 的範例。

3.3 MacBERT

MacBERT (Yiming Cuil et al. 2020) 全名為 Masked As Correction BERT，是另一個考慮中文語言特性的 BERT 延伸版本。在 MLM 的訓練任務上，除了沿用 BERT-wwm 的以全詞為單位訓練的概念之外，還修改了兩個部分的設定。首先，在選取欲被遮蓋的部分，以 N-gram masked 取代隨機遮蓋，先取得候選的 token，再從中選擇遮蓋目標。其次，更改遮

Dataset	Title	Paragraph
DRCD	函數	函數在數學中為兩集合間的一種對應關係：輸入值集合中的每項元素皆能對應唯一一項輸出值集合中的元素。氣溫的分布也能用函數表達，以時間和地點作為參量輸入…
CMRC	国际初中科学奥林匹克	国际初中科学奥林匹克（ International Junior Science Olympiad）是一项给予 15 岁或以下的学生参与的国际科学比赛。此比赛最先在 2004 年举办，然后一年举办一次。…

表 2、資料集之文章段落與對應標題範例

蔽的內容，針對目標 token 先以 word2vec (Tomas Mikolov et al., 2013) 計算相似度以獲得相似的單詞，並且用此相似詞進行遮蓋，可以減少因為 [MASK] 只在預訓練使用而不會出現在微調階段所造成的差異，也可以藉此讓模型學到更多的相似詞與上下文之間的關係。遮蓋策略比較可參考表 1 的範例。

3.4 資料集分群與輸入整合

考慮到文章的類型可能會對於文章回答的形式有潛在的影響，故決定加入文章類型資訊。由於資料集中的文章沒有人工標記類別，此處根據資料集中既有的文章標題以及從文章取出的重要單詞等兩種來源作為依據進行分群，取得類似分類的資訊，將此資訊加入到模型中以提升表現。上述流程與模型架構如圖 3 所示。加入方式為新增一個 type token，於輸入 BERT 模型前將加入到所有 token 中。

其中，keyword 的取得方法為將全文進行分詞，取得每篇文章各自的詞庫後，再以 Term Frequency-Inverse Document Frequency (TF-IDF) 計算單詞重要性，最終取分數較高者作為此處輸入的內容。TF-IDF 包含兩個部分，詞頻 (Term Frequency, TF) 與逆向文件頻率 (Inverse Document Frequency, IDF) 。其中 TF 表示單詞於單一文章中的出現頻率，IDF 表示單詞在整個資料集中出現過的文章數量，TF-IDF 則為兩者相乘。公式如下。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2)$$

$$IDF_i = \log \frac{N}{n_i} \quad (3)$$

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \quad (4)$$

	Title#	Paragraph#	Question#
DRCD	2,108	10,014	30K
CMRC	3,251	3,251	20K

表 3、資料集之標題、文章、問題數量比較

上式中 *TF* 的 $n_{i,j}$ 為單詞 i 於文章 d_j 中的出現次數，分母為文章 d_j 中所有單詞的出現次數總和。 *IDF* 中的 N 為整個資料集的文章總數量，n_i 則為文章中包含文字 i 的文章數目。

前段得出分群根據的文字 (keyword) 與標題 (title word) 的後續方法相同。首先，利用 sentence-BERT (Nils Reimers and Iryna Gurevych, 2019) 取得標題或 keyword 的詞嵌入 (Word embedding)，再以基於密度的聚類演算法方法：DBSCAN (Density-based spatial clustering) (Martin Ester et.al., 1996) 分出數個相似度較高的群集。其中，只會將有一定數量的聚類識別為一個組別，在一定範圍內數量不足以構成組別的的離群值 (outlier) 將會被標記為雜訊 (noise)。

分組的結果會做為額外資訊提供給模型，並且作為 BERT 的輸入資訊之一整合在模型中。加入的方式如圖 3 所示，在 BERT 的輸入層中添加一層 Type Embed ，並在對應到的文章上進行組別標記，其中，被標記為離群值的部分會以 0 加入，等同於不加入任何額外資訊。後續實驗步驟與一般閱讀理解任務相同，即 BERT 模型進行微調，預測出機率最高的起點與終點。可參考圖 2。

3.5 集成學習 (Ensemble Method)

集成學習指結合多個模型的結果來提升整體表現。段落擷取類型的閱讀理解模型最終會對文章中的每個詞進行兩種預測，分別為適

	Type	Title word	1 keyword	2 keywords
DRCD	Outlier	657	768	346
	Big group	1,097	5,105	8483
	Small group	10 ~ 193 (19)	10 ~ 436 (75)	10 ~ 40 (7)
CMRC	Outlier	810	678	591
	Big group	2,035	2,117	2453
	Small group	9 ~116 (14)	10 ~ 43 (26)	10 ~ 29 (12)

表 4、分群組別數量與樣本數量分佈結果，表格中數字代表樣本數，括弧為組數

合作為起點的機率以及適合作為終點的機率。每組機率值都會對應到文章內容並擷取起點與終點所對應的句子。使用單模型時取用機率最高的句子作為解答。使用兩個以上的模型結果時，將列出所有可能的預測句子，句子完全相同者對機率進行平均，不同者沿用單模型的機率作為新的句子候選。最終，再選出新的句子候選中機率最高者作為集成模型的輸出。

4 實驗結果與討論

4.1 實驗材料

本篇採用兩個公開的數據集：台達閱讀理解資料集 (Delta Reading Comprehension Dataset, DRCD) (Shao et al., 2018) 與訊飛杯中文機器閱讀理解評測 (The Third Evaluation Workshop on Chinese Machine Reading Comprehension, CMRC) (Cui et al., 2019)。兩者皆為段落擷取類型機器閱讀理解資料集，其訓練資料都來源於維基百科。其中，前者為繁體中文，包含 2,108 個主題 (Title) 的 10,014 個文章段落 (Paragraph)，以及三萬多個問題 (Query)；後者為簡體中文，包含 3251 個文章段落及兩萬多個問題。

在資料集分群的實驗，採用資料集中的文章主題名稱作為分群根據，主題與文章段落的關係可參考表 2。實驗用的兩個資料集在這部分有些差異：DRCD 的每個主題對應到多篇數量不等的文章段落；CMRC 則是一個主題只會對應到一篇文章。可參考表 3 的標題、文章以及問題的數量比較。另外，在以內容關鍵字作為分群依據的部分，則是用每篇文章的內容進行分詞與擷取，每篇文章都有對應的分群結果。

4.2 實驗設定

取得分群資訊的部分，使用 JIEBA 中文斷詞工具對目標文章進行分詞以取得 keyword，並且由於慣用語的不同，使用 JIEBA 的官方簡體中文字典與中央研究院資訊科學所繁體詞庫分別作為簡體中文與繁體中文的分詞依據；DBSCAN 分群中，title 與 keyword 的設定相同。兩點可作為鄰近點的最大距離閾值 (eps)為 3，每個群集中必須要有的最小鄰近點數量(min samples) 為 10。BERT 微調訓練的參數設定的部分，模型 learning rate 為 5e-5，訓練 batch size 為 32、training epoch 為 3；文字處理設定的部分，文章的最大長度為 384 個文字、問題的最大長度為 64 個文字、預測答案的最大長度為 30 個字。在文章與問題中，大於限制者會捨去超出字數的文字，小於限制者則會填空至此長度。

4.3 評估指標

採用 Exact match (EM) 與 F1-Score 兩種指標進行評估。EM 著重於預測回答與標準答案的完全匹配程度，有助於當正確答案為一個短句或單字時的回答精準度。例如，一個閱讀理解任務包括N個問題，每個問題對應的正確答案只有一個，此輪中回答正確的題數為 M 題。則完全匹配的回答數為 M 個，剩餘的 N-M 個為不完全匹配。不完全匹配包括與標準答案部分匹配的回答與完全無關的回答，其公式如下。

$$Exact\ Match = \frac{M}{N} \quad (5)$$

F1-Score 主要用來評估預測回答與標準答案的重合程度，相對於精準匹配值較具有彈性。

		DRCD		CMRC	
		EM (%)	**F1**	**EM (%)**	**F1**
Fine-tune	BERT	85.600	0.917	60.298	0.840
	BERT-wwm	85.686	0.921	61.479	0.844
	MacBERT	88.606	0.938	63.156	0.856
Add Clustering Info.	BERT-wwm				
	+ Title Label	85.377	0.919	61.819	0.844
	+ 1 Keyword Label	85.680	0.917	61.912	0.847
	+ 2 Keyword Labels	85.834	0.920	61.484	0.842
	MacBERT				
	+ Title Label	88.405	0.937	63.591	0.856
	+ 1 Keyword Label	88.262	0.936	63.778	0.855
	+ 2 Keyword Labels	88.892	0.940	63.156	0.858
Ensemble	MacBERT + BERT-wwm	88.663	0.938	64.461	0.864
	The two best model	**89.207**	**0.942**	**65.238**	**0.862**

表 5、使用 BERT-wwm、MacBERT、MacBERT 加入分群資訊及兩者集成的實驗結果。表中底線代表集成學習以外的方法中，表現較佳的兩個模型，亦為用於後續集成學習實驗；表中粗體代表表現最佳的分數

F1-Score 以準確率 (Accuracy) 和召回率 (Recall) 的調和平均數得出。公式如下。

$$precision = \frac{TP}{TP+FP} \quad (6)$$

$$recall = \frac{TP}{TP+FN} \quad (7)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision+recall} \quad (8)$$

圖中的 TP、TN、FP、FN 分別代表四種可能的預測情況。True Positive (TP) 為將正確為預測為正確的情況；True Negative (TN) 為將錯誤預測為錯誤的情況；False Positive (FP) 為將錯誤預測為正確的情況；False Negative (FN) 為將正確預測為錯誤的情況。

4.4 分群結果

分群結果可參考表 4。根據分群後的狀態不同，將這些分群組別分成三種類型：離群值 (outlier)、一個大組別 (Big group) 與數個小組別 (Small group)。離群值指的是和其他文章相關性較低，或是存在相關性高的文章，但是其數量總和遠不足以構成一個新組別的文章樣本；大組別和小組別的區分主要是組內的樣本數量。根據實驗結果的觀察，資料集中都會有一個組別特別大，囊括了超過三分之一的文章類型，剩餘的文章分布則是數個具有一定樣本數的小組別。

標題分群的部分， DRCD 的離群值樣本數為 657 個，大組別的樣本數 1,097 個，小組別則是有 19 組，樣本數量分佈在 10～193 個 。CMRC 的離群值樣本數為 810 ，大組別樣本數為 2,035，其餘共有 14 個小組別，樣本數分別分佈在 9～116 。

在文章關鍵字分群的部分，分成只取一個和兩個關鍵字作為分群基準，以避免過多文字產生過多雜訊而造成分群效果不佳的情況。在取一個關鍵字分群的部分，DRCD 的離群值樣本數為 768 個，大組別的樣本數 5,105 個，小組別則是有 75 組，樣本數量分佈在 10～436 個 。 CMRC 的離群值樣本數為 678 ，大組別樣本數為 2,117，其餘共有 26 個小組別，樣本數分別分佈在 10～43 。取兩個關鍵字分群的部分，DRCD 的離群值樣本數為 346 個，大組別的樣本數 8,483 個，小組別則是有 7 組，樣本數量分佈在 10～40 個 。CMRC 的離群值樣本數為 591 ，大組別樣本數為 2,453，其餘共有 12 個小組別，樣本數分別分佈在 10～29 。

4.5 實驗結果

本篇實驗微調了 BERT 、 BERT-wwm 及 MacBERT 作為基線進行比較，如表 5 的前三

列，主要實驗為加入分群資訊與集成模型的兩個實驗。

加上分群資訊的部分，可參考表 5 中的 Add clustering info.的欄位。兩個 Bert-based 的模型加上標題資訊進行模型的重新訓練，並與原本的微調結果作比較。在微調表現較佳的 MacBERT 的實驗部分，差距主要在 EM 上，F1-Score 的變化不大。在加入標題資訊 (+ Title Label) 的方面， DRCD 資料集的 EM 微幅下降了 0.201 個百分點；在 CMRC 資料集的部分，EM 提高了 0.435 個百分點。在加入標題的實驗中，CMRC 的 EM 獲得提升，代表文章標題與內文關鍵字的資訊分類確實對模型有一定的幫助；在 DRCD 的表現卻差強人意，其原因可能是作為最初作為分群的標題資訊量不足所造成。DRCD 與 CMRC 在標題資訊上最大的差異是在於標題與文章的形式；DRCD 是多篇文章共用一個較大的標題；CMRC 則是一個文章對應一個標題不同，可能因此輸入了不足以代表文章內容的資訊，使得引入的雜訊影響原本的判斷，進而造成模型表現下降。

根據前述原因，使用了全文的斷詞並取出關鍵字 (+ Keyword Label) 進行分群，此設定可以確保用於分群的資訊與文章內容是相關且具有一定重要程度的。實驗結果的部分，DRCD 在加入一個關鍵字的 EM 下降了 0.334 個百分點，但是加入兩個關鍵字時，提升了 0.286 個百分點；CMRC 加入一個關鍵字有 0.662 個百分點的提升，加入兩個則沒有變化。兩個資料集的實驗結果各有不同。DRCD 加入一個關鍵字的表現略差於兩個關鍵字的原因，可能是只使用一個關鍵字不足以代表整篇文章，分類資訊反而使表現下降。CMRC 則是在用一個關鍵字時有最佳的效果。BERT-wwm 的表現變化的趨勢與 MacBERT 相似。詳細數據可見表 5。另外，由於兩個資料集在不同數量 keyword 的分群數量表現變化差異較大，故此處不針對加入 keyword 數量造成的影響做探討。可參考表 4。

集成學習部分的結果，可參考表 5 的 ensemble 欄位。此處分成兩個實驗進行，結果將與兩個集成前的模型做比較，主要觀察是否比原先的模型有更佳的表現。首先是僅經過微調的預訓練模型集成學習，在 DRCD 資料集的實驗上，相較於原先就表現較佳的 MacBERT 微調結果，EM 提高了 0.063 個百分點，F1-Score 也有 0.011 個百分點的微幅進步。在 CMRC 實驗的部分，進步表現較為顯著，EM 提高了 1.305 個百分點，F1-Score 也有 0.843 個百分點的進步。另外，第二個部分是從前述的微調與加入分群資訊結果中，分別挑出最佳的兩個結果進行集成學習。DRCD 實驗結果中，選擇的是微調的 MacBERT 與加入兩個關鍵字的 MacBERT 模型，其 EM 結果相較於集成前的分數有 0.601 和 0.315 個百分點的進步，F1-Score 也有微幅提升；CMRC 的實驗中，選擇的是加入標題的 MacBERT 與加入一個關鍵字的 MacBERT 模型，其 EM 分別有 1.647 和 1.46 個百分點的進步，F1-Score 也有平均 0.0065 的提升。集成模型作為一個相對較簡單快速方法，也在此處讓模型預測進步方面達到了不錯的效果。

5 結論

本篇研究使用兩種基於 BERT 的語言模型：BERT-wwm 以及 MacBERT，分別在繁體中文語簡體中文的兩個機器閱讀理解任務上進行微調、加入分群資訊重新訓練與模型集成等實驗。加入分群資訊實驗的部分，兩個資料集的分群資訊皆讓模型學到與文章類型相關的潛在回答模式，使預測結果有所提升；在集成模型的部分，歸功於兩種預訓練語言模型的基礎能力，其合併預測結果的方式也得到了不錯的結果。

由於本篇實驗中的分群結果並不算理想，非常多文章被歸在同一個較大的聚類而無法顯示其差異性，可能因此浪費許多可以運用的資訊，也讓分群之間的結果較難做比較。未來，將會針對分群的部分做改進；另外，提升此篇單輪問答模型的表現也可以拓展到多輪問答上使用，故未來也將會針對加入歷史對話以運用在多輪的閱讀理解任務上進行進一步的研究。

參考文獻

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical*

Methods in Natural Language Processing, 2383–2392.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Inter- national Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019a. Pre-training with whole word masking for chinese bert. arXiv preprint arXiv:1906.08101.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. arXiv preprint arXiv:2004.13922.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, Weiming Zhang. 2019. Neural Machine Reading Comprehension: Methods and Trends. *J. Applied Sciences*, 2019, 9(18): 3698

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar and Jan Kleindienst. 2016. Text Understanding with the Attention Sum Reader Network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 908–918.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen and Ruslan Salakhutdinov. 2017. Gated-Attention Readers for Text Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1832–1846.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang and Ming Zhou. 2017. Gated Self-Matching Networks for Reading Comprehension and Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 189–198.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu adn Guoping Hu. 2017. Attention-over-Attention Neural Networks for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 593–602.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi and Hannaneh Hajishirzi. 2016. Bidirectional Attention Flow for Machine Comprehension. In International Conference on Learning Representations. arXiv preprint arXiv: 1611.0160.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever, I. 2018. Improving language understanding by generative pre-training. *Technical report*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo- gies, Volume 1 (Long and Short Papers)*, 4171–4186.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. 2019. XLNET: Generalized autoregres- sive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma and Radu Soricut. 2020. ALBERT: A Lite BERT for Self- supervised Learning of Language Representations. In I*nternational Conference on Learning Representations.*

Kevin Clark, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representation.*