多段提取模型之技術報告

簡介

多段提取模型之目的為給定文本與問題，從文本提取多個文本段(Text Span)之後，並將其進行串接視為答案回答。基於多段提取的資料集 DuReader 與 DROP 進行展開研究，由於本任務的實際應用為中文語系的問答模型，牽涉到命名實體辨別、多階段預測、遷移學習、BIO 標記與集合學習等相關技術需求。最終於 FGC 資料集上達到一定之成效。

一、介紹

該技術報告介紹共牽涉到以下技術：命名實體擴充、詞性規則修正、遷移學習、BIO 標記任務。基礎模型則是使用 BERT 與 Efrat et al. (2019)提出之 Tag-based MSPE 模型，進行主要建構，除了額外特徵的引入與各模型的訓練細節之外，主要是過濾各式模型所提取的各種文本段進行串接或擴充來進行規則上的修正。鑒於多段提取模型的輸出要求，通常都是以詞組為單位的，基於字元為主的 BERT 相關系列架構，都面臨文本段提取不夠精確；並且由於該多段提取模型的答題領域過於發散，所以該技術報告過分迭代大量的規則，讓輸出之答案至少要以名詞為最高原則，而非以有實質意義的句子為輸出點。

二、多段提取模型

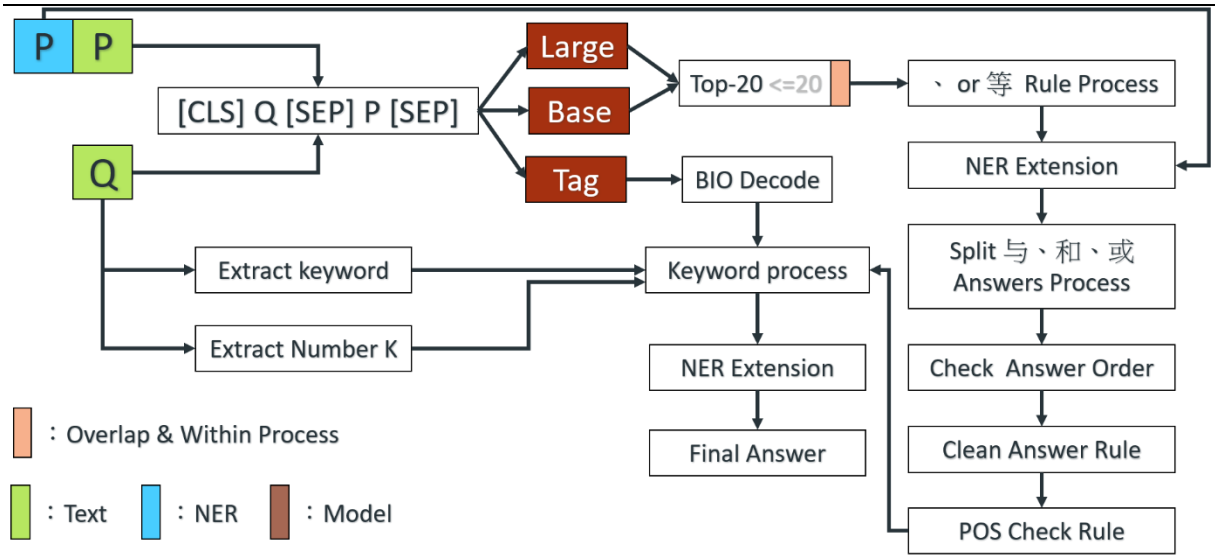
二之一、系統架構

整個系統架構如圖(一)、所示，主要共分為兩大部分；第一部分為基礎模型相關，負責提取文本中符合長度規則之答案，又分為 BERT(Large)、BERT(Base)與 Tag-based MSPE 模型；第二部分為規則庫過濾，包含符號處理、命名實體辨別擴充處理、答案交疊與重疊處理、詞性合法性處理、關鍵詞提取處理與特定數量回答提取處理。

二之二、基礎模型相關

二之二之一、基礎模型訓練

基礎模型都是以 BERT 為主的方法，其中 BERT(Large)與 BERT(Base)採用微調策略在本團隊蒐集之中文語料上，進行遮掩語言模型的上游任務，中文語料包含：中文維基、警廣語 CAN 新聞等資料；訓練該兩模型，只透過訓練於 DRCD 資料集來進行後續預測之需求。再者 Tag-based MSPE，則是在本團隊提供之中文化 DROP 與 DuReader 資料集上，先進行 BIO 標記的訓練後，在遷移學習在本團隊提供的 FGC 資料集的訓練集上。



圖(一)、系統架構之示意圖

二之二之二、基礎模型預測

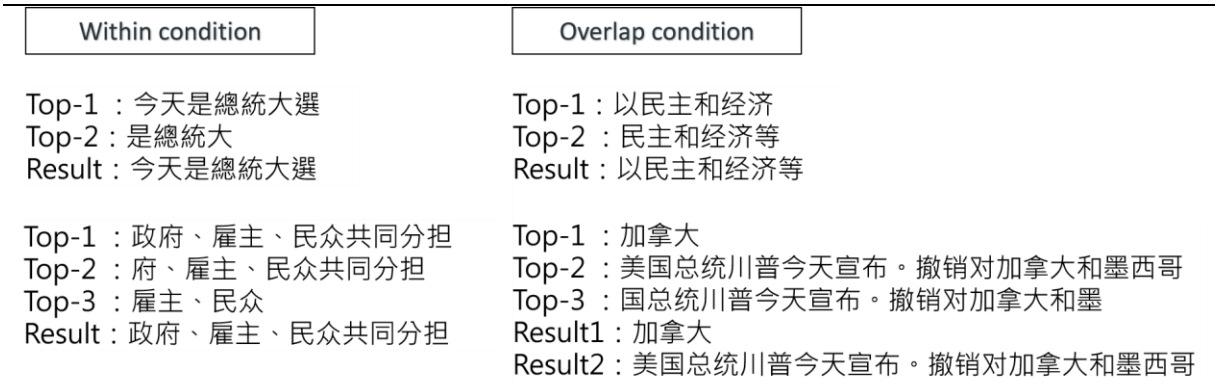
針對 BERT(Large)與 BERT(Base)原先的 DRCD 訓練上，是採用單段提取之任務，為了應付多段提取之需求，採用 Top-k 的方式，去產生多文本段之需求。

實際情況下，BERT(Base)的預測狀況多為短詞組，反之 BERT(Large)為長詞組預測。鑑於這兩者的差別，使用 Base 的模型時，會進行額外設定答案的最高長度為 20；Large 模型則是除了長度限制外，還會透過命名實體辨別去做額外的答案處理，得到更多的候選文字段。針對 Tag-Based MSPE 模型則是在 BIO 解碼(BIO Decoding)上，進行 Beam Search 提取多個可能答案，並供與後續模塊進行處理。

二之三、規則庫過濾

本研究生在針對該 FGC 資料庫進行研究時，面臨以下狀況：答案格式不統一、問題曖昧不清、答案正確要求嚴格、題目領域沒有集中性、訓練資料不足、額外訓練資料庫品質不佳、常規機器學習訓練經驗法則不適用與英文與系上最佳之模型於中文應用上沒有基礎之成效等狀況。尤其是答案格式的不統一、導致在搜尋多個文本段落時，只能以回答至少**具有基本意義之詞組為首要目的**。其餘的狀況說明則闡述該任務之挑戰與困難性。

本技術報告會基於上段所陳述之形況進行以下規則之說明：符號處理、命名實體辨別擴充、答案交疊與包含處理、詞性合法性處理、關鍵詞提取處理與特定數量回答提取處理。



圖(二)、交疊與包含示意圖

二之三之一、符號處理

本方法有針對、，。！？『』《》【】〔〕；等常見之符號，進行分段、取代與消除之處理。由於 FGC 資料庫的文章通常高於 512 個字，除了讓文章可以盡量保持原有的資訊外，也發現在不考慮符號的情況下，有更好的成效。

除此之外，基於回答要以**具有基本意義之詞組為首要目的**，故當輸出的答案包含到句號或是特定結尾符號的時候，將其視為包含多個句子以及完整性問題，會直接捨去。

二之三之二、命名實體辨別擴充(NER Extension)

本部分則是要說明，由於基礎模型的侷限性，是以字元為最小單位的情況，所以輸出的片段有機率輸出不完整的詞組，尤其是命名實體的部分，所以當輸出答案有包含到該命名實體辨別的部分文字，則會直接把該命名實體辨別擴充出來，並且視為絕對正確答案之一。

二之三之三、答案交疊與包含處理(Overlap and Within Process)

關於交疊與包含的示意圖如圖(二)、所示，在各種測試下，最終以聯集的方式去合併答案，其餘後續也會在進行命名實體辨別擴充的後處理。

二之三之四、詞性合法性處理(Clean Answer Rule and POS Check Rule)

此部分合法性的處理，是針對各別答案的輸出進行過濾。依表(一)、所示，並呈現依照結巴給語的 POS 結果和處理之形況。

關於捨棄：將此單詞答案，直接整組丟掉，不進行處理。

關於刪除：將此多詞輸出下，擁有該詞性的詞組刪除，其餘答案不動。

關於保留：判別為該詞性的，不做任何其餘處理。

關於保留[特殊處理]：由於答案的回答不統一性，因此有針對特定情況做保留處理。

多段提取模型之技術報告，羅上堡。

詞姓名稱(英文)	單詞輸出	多詞輸出	舉利用詞
形容詞(a)	捨棄	刪除	大，好，新
形語素(ag)	捨棄	刪除	奇，私，秀
副形詞(ad)	捨棄	刪除	完全，突然，直接
名形詞(an)	保留	保留	安全，困難，矛盾
區別詞(b)	捨棄	保留	主要，副，總
連結詞(c)	捨棄	保留[特殊處理]	和，而，但
副詞(d)	捨棄	保留[特殊處理]	不，也，就
副語素(dg)	捨棄	捨棄	俱，輒
能願動詞(df*)	捨棄	保留	不要
嘆詞(e)	捨棄	捨棄	嗯，哎，咦
外語(eng*)	保留[特殊處理]	保留[特殊處理]	English
趨向動詞(f)	捨棄	捨棄	上，中，後
語素(g)	捨棄	捨棄	浠，僭，涔
前接成分(h)	捨棄	捨棄	非，超低
成語(i)	捨棄	捨棄	一口氣，大吃一驚，九曲迴腸
簡稱略語(j)	捨棄	刪除	法，人大，漢
後接成分(k)	捨棄	保留	們，者，型，式
慣用語(l)	保留	刪除	發言人，是不是，沒想到
數詞(m)	捨棄	保留	年，一，月，多
干支(mg*)	捨棄	保留	巳，寅
指示代詞(mq*)	捨棄	保留	這件，這場，一方面
名詞(n)	保留	保留	人，時，國家
名語素(ng)	保留	保留	子，身，師，眾
人名(nr)	保留	保留	連，王，楊
完整人名(nrfg*)	保留	保留	李自成，張居正，康熙
外國名詞(nrt)	保留	保留	二人，闖王，崇禎
地名(ns)	保留	保留	臺灣，美國，日本
機構團體(nt)	保留	保留	國務院，外交部
其他專名(nz)	保留	保留	百科，和平，英語
擬聲詞(o)	捨棄	捨棄	哈哈，砰，嗚，嘿嘿
介詞(p)	捨棄	刪除	在，為，對
量詞(q)	捨棄	保留[特別處理]	道，個，家
代詞(r)	捨棄	捨棄	他，我，這
茲(rg*)	捨棄	捨棄	茲
多數代詞(rr*)	捨棄	捨棄	其他人，妳們，偺們
這位(rz*)	捨棄	捨棄	這位

方位名詞(s)	捨棄	捨棄	心中, 國內, 身上
時間詞(t)	捨棄	捨棄	當, 現在, 當時
時語素(tg)	捨棄	捨棄	現, 晚, 春
助詞(u)	捨棄	保留[特殊處理]	等, 之, 來說
結構助詞: 得(ud*)	捨棄	刪除	得
結構助詞: 的(uj*)	捨棄	刪除	的
結構助詞: 地(uv*)	捨棄	刪除	地
動態助詞: 過(ug*)	捨棄	刪除	過
動態助詞: 了(ul*)	捨棄	捨棄	了
動態助詞: 著(uz*)	捨棄	刪除	著
動詞(v)	捨棄	刪除	是, 有, 說
動語素(vg)	捨棄	保留	喝, 言, 怒
副動詞(vd)	捨棄	保留	持續, 狡辯, 逆勢
內動詞(vi*)	捨棄	保留	等同於, 徜徉於, 沉溺於, 沉緬於
名動詞(vn)	保留	保留	發展, 工作, 研究
完成動詞(vq*)	捨棄	保留	去過, 去淨, 唸過, 捱過
標點符號(w)	捨棄	捨棄	， 。
非語素字(x)	捨棄	捨棄	嗎, 姆, 灞
語氣助詞(y)	捨棄	捨棄	呢, 吧, 嗎
狀態詞(z)	捨棄	捨棄	涓, 優良, 最佳
副狀態詞(zg*)	捨棄	保留	很, 此, 較

表(一)、各詞性與答案輸出之處理表

二之三之四、關鍵詞提取處理與數量回答處理(Keyword Process and Extract number K)

此部分是一起完成的，針對問題的文本上，進行關鍵詞提取以及透過特定字眼去尋找有沒有特定數量回答之需求。關鍵字的處理則是針對以下詞組進行旁邊的名詞尋找；像是，**哪裡、那裡、那些、這些、多少、幾個、幾處或幾朵等字眼**。特定數量的搜尋則是只有從二至七的搜尋。在該系統架構上，如果沒有特定標明數量，每個獨立模型都統一以搜尋 20 個候選答案為主要原則。

二之三、其餘細節

由於該模型的演化進程過於複雜，本部分說明尚未說明的部分或狀況。

1.全部的訓練資料都為簡體中文。

2.正確答案的符號也都進行清除處理。

3.模型之間的答案分數是採取開始跟結尾的機率進行相乘的。

4.答案會依照內容最剛開始出現的排序進行輸出排序。

Data[# of Q]	Number of Question			Passage Length		
	Train	Test	Dev	Max	Min	Mean
FGC	98	29	33	1,276	227	618
DuReader	500,00	1,866	4,601	-	-	-
DRCD	26,932	3,485	3,524	-	-	435.8

表(二)、資料集資訊表

三、資料集與評估方式

三之一、資料集介紹

本模型共使用 DuReader、DRCD 與 FGC 資料集進行訓練。關於每個資料集相關資訊如表(二)、所示。由於 DROP 的訓練成效一直不佳，不提供 DROP 資料集的相關資訊。其中 DuReader 是將原本提供的訓練集隨機採樣 10%的數據量，視為驗證集來使用。

三之二、評估方式

本技術報告採用 Micro EM 與 EM 的評估方式，來評估模型成效。其中 EM 是該模型回答的答案要與正確答案完成一致，才能得分；Micro EM 則是該模型回答的答案答對正確答案的其中幾個答案，並以模型回答與正確答案的最高長度視為分母，進行評估。

四、系統效能

由於該模型有經過各大環節之改善，主要呈現歷代演進的成效表，如表(三)、所示。

MSPE	NGC		
	Train83	Dev31	Test26
Original	32.06%/12.05%	27.12%/0%	18.08%/0%
+ Pre-processing	32.86%/12.05%	29.81%/0%	18.08%/0%
+ Clean answer 1	32.86%/12.05%	29.81%/0%	24.49%/7.69%
+ Clean answer 2	34.31%/13.25%	31.40%/0%	26.64%/7.69%
+ Split answer rule	34.55%/13.25%	31.40%/0%	26.64%/7.69%
+ POS Check rule	34.64%/13.25%	31.40%/0%	27.12%/7.69%
+ Tag-Based & keyword [v12_b7]	36.13%/15.66%	28.18%/0%	26.48%/7.69%
+ Fix_Rule_v18_b13	50.55%/37.35%	37.10%/12.90%	37.14%/19.23%
+ Fix_Rule_v18_b14	50.55%/37.35%	38.06%/16.12%	46.55%/34.62%
+ Fix_Rule_v18_b15	51.87%/40.96%	39.16%/19.35%	50.27%/42.31%
+ Fix_Rule_v18_b16	62.87%/55.42%	56.25%/32.26%	60.26%/50.00%
+ Fix_Rule_v18_b17	67.76%/61.45%	59.58%/35.48%	71.98%/65.38%
+ Fix_Rule_v18_b18	73.20%/66.27%	71.60%/58.06%	83.47%/77.42%
+ Fix_Rule_v18_b19	76.41%/69.88%	71.60%/69.23%	82.66%/77.42%
+ Fix_Rule_v18_b20	78.17%/73.49%	71.6%/69.23%	82.66%/77.42%
Final	79.13%/74.70%	71.6%/69.23%	85.15%/77.42%

五、觀察與總結

關於各大模型的演進有以下主要改善與觀察總結：

1. 將詞性標註的規則分析清楚，除了命名實體辨別的擴充外，詞性上也進行擴充。
2. 總共訓練 6 組 Base 模型、4 組 Large 模型與 3 個 Tag-based 模型，並且要選用在原本資料集沒有特別好的模型，這是因為 Domain Mismatch 以及資料集領域過於發散之問題。
3. 引入 NER 的 Embedding 給與 BERT 模型，有辨別為 NER 的詞，拆為字元給 1 的 embedding 給予該字元，反之給 0 的 embedding，並且在訓練時遮罩 0 的梯度進行訓練。
4. Tag-based 的 Beam Search 需要額外的規則去進行限制，由於 DuReader 的資料集下，訓練出來的模型會盲目的固定一個長度就進行一個標籤輸出。儘管在 Beam Search 找到更多的答案，也需要額外的詞性規則去做大量的限制。
5. 針對特定數量提取的規則限制，實質上沒有實質利益的改善，除了 DuReader 幾乎沒有這種問法之外，就是該 FGC 的資料集，對於特定數量的題目少之又少，只能寫特定的規則去進行改善。
6. 針對問題關鍵字，去進行很簡單的詞性篩選與檢查，是有非常好的改善空間的，呈現於表(三)、Fix_Rule_v18_b15~b17 之間。
7. 由於資料集的答案標準不一，本研究對於正確答案的處理，除了去除符號之外，也允許順序不同與形容詞的出現與否進行特別的撰寫評估規則。
8. 針對親屬題型的題目，本研究採取查表法，結構化的題目，目前是無法處理的。
9. 針對國家、地名與朝代的題目，除了詞性規則的幫助之外，也檢查非常大量的合法性與資料可能性進行了規則撰寫。
10. 針對罕見名詞，採取加入關鍵詞的方式，透過命名實體辨別的方式去解決。
11. 針對作品名稱、特定格式之回答要求，是以不過濾符號的文本進行直接篩選，再透過特定名稱之上下文的詞向量去決定哪部分的特定名稱視為答案。
12. 以 BERT 為基礎模型的侷限性，就是很難要求框出正確的名詞，需要做很大量的後續規則動作進行修正。
13. 以增加特定特徵給與基礎模型，對於單體模型表現是會成長的，但是幾乎都差別於一個實體或是名詞上的完整性。