上下文語言模型化技術於常見問答檢索之研究

A Study on Contextualized Language Modeling for FAQ Retrieval

曾琬婷　Tseng, Wen-Ting
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
d0409857@gmail.com

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

陳柏琳　Chen, Berlin
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

近年來，深度學習技術有突破性的發展，並在很多自然語言處理的相關應用領域上也有相當亮眼的效能表現，例如 FAQ (Frequently Asked Question)檢索任務。FAQ 檢索無論在電子商務服務或是線上論壇等許多領域都有廣泛的應用；其目的在於依照使用者的查詢(問題)來提供相對應最適合的答案。至今，已有出數種 FAQ 檢索的策略被提出，像是透過比較使用者查詢和標準問句的相似度、使用者查詢與標準問句對應的答案之間相關性，或是將使用者查詢做分類。因此，也有許多新穎的基於上下文的深層類神經網路語言模型被用於以實現上述策略；例如，BERT(Bidirectional Encoder Representations from Transformers)，以及它的延伸像是 K-BERT 或是 Sentence-BERT 等。儘管 BERT 以及它的延伸在 FAQ 檢索任務上已獲得不錯的效果，但是對於需要一般領域知識的 FAQ 任務仍有改進空間。因此，本論文中探討如何透過使用知識圖譜等的額外資訊來強化 BERT 在 FAQ 檢索任務上之效能，並同時比較不同策略和方法的結合在 FAQ 檢索任務之表現。

Abstract

Recent years have witnessed significant progress in the development of deep learning techniques, which also has achieved state-of-the-art performance for a wide variety of natural language processing (NLP) applications like the frequently asked question (FAQ) retrieval task. FAQ retrieval, which manages to provide relevant information in response to frequent questions or concerns, has far-reaching applications such as e-commerce services and online forums, among many other applications. In the common setting of the FAQ retrieval task, a collection of question-answer (Q-A) pairs compiled in advance can be capitalized to retrieve an appropriate answer in response to a user's query that is likely to reoccur frequently. To date, there have many strategies proposed to approach FAQ retrieval, ranging from comparing the similarity between the query and a question, to scoring the relevance between the query and the associated answer of a question, and performing classification on user queries. As such, a variety of contextualized language models have been extended and developed to operationalize the aforementioned strategies, like BERT (Bidirectional Encoder Representations from Transformers), K-BERT and Sentence-BERT. Although BERT and its variants has demonstrated reasonably good results on various FAQ retrieval tasks, they still would fall short for some tasks that may resort to generic knowledge. In view of this, in this paper, we set out to explore the utility of injecting an extra knowledge base into BERT for FAQ retrieval, meanwhile comparing among synergistic effects of different strategies and methods.

關鍵詞：常見問答集檢索，知識圖譜，自然語言處理，資訊檢索，深度學習

Keywords: Frequently Asked Question, Knowledge Graph, Natural Language Processing, Information Retrieval, Deep Learning

一、緒論

隨著網際網路上文本資料或是多媒體資訊的蓬勃發展，FAQ(Frequently Asked Questions)檢索技術的發展已經成為各種應用領域(如電子商務服務、線上論壇等)中極為重要的需求[1],[2]。過去許多網站通常會針對經常被問到的問題與其對應答案經由人工方式整理成問答配對可直接提供給使用者進入網站時瀏覽。但是資料量隨著時間增加，使用者要直接找到需要的答案也越來越困難[3]。時至今日，已經有許多網站提供 FAQ 查詢的服務讓使用者能更快速找到自己的需求，像是 FAQ Finder[4]系統以及 Ask Jeeves[5]服務網頁。

FAQ檢索基本上可視為以自然語言查詢為主的資訊檢索(Information Retrieval)之應用。目前實作於FAQ檢索技術大致可以分為非監督式方法以及監督式方法。前者利用計算使用者查詢和標準問句的相似度方式找到最適合的答案，像是Okapi BM25[6]模型或向量空間模型[7] (Vector Space Model, VSM)。後者主要利用使用者查詢與標準問句的相關答案之間相關性做判斷，BERT (Bidirectional Encoder Representations from Transformers)[8]或K-BERT (Knowledge-enabled Language Representation Model)[9]。其中BERT是能夠理解上下文的語言模型，主要使用Transformers中的注意力機制(Attention Mechanism)[10]讓模型可以學習到上下文的關係。但是BERT較缺乏特定領域知識，因此K-BERT基於知識圖譜並注入至模型中，使得BERT能夠如專家般針對相關知識進行推理。

儘管 BERT 在 FAQ 檢索研究上已經有亮眼的成績，但其效能在需要有一般或特定領域知識的 FAQ 檢索任務上仍然有改進的空間。因此，本論文嘗試比較 BERT 以及它的兩種延伸方法，即 K-BERT 和 Sentence-BERT[11]於兩種不同特性的 FAQ 資料集上的表現。同時，我們也比較使用者查詢和標準問句的相似度、使用者查詢與標準問句的相關答案之間相關性以及將使用者查詢做分類三種 FAQ 檢索策略實作在這兩種不同特性的資料集上之效果。從實驗中發現在含有多個標準問句對應單個答案(類別少)的任務情境適合利用將使用者查詢做分類方式，而單個標準問句對應單個答案(類別多)的任務情境適合利用比較使用者查詢與標準問句的相關答案之間相關性方式。另外，在模型中加入額外知識圖譜的資訊有助於提升 FAQ 檢索的效能。關於詳細方法與實驗討論將於後續章節依序介紹。

二、相關研究

（一）、語言模型

語言模型在自然語言處理的研究中佔有極重要的地位。近年來，預訓練的深層類神經網路語言模型，像是 ELMO (Embeddings from Language Models)[12]、OpenAI GPT (Generative Pre-Training)[13]、BERT 等在各種自然語言處理任務上皆提升整體效能。其中以 BERT 模型效果最為亮眼，它利用 Transformers[10]學習文本中單詞之間上下文關

係的注意力機制。基於此特性應用於現實使用文字的情境中理解到文本的上下文資訊。BERT 模型在多種 FAQ 檢索任務中皆被發現其效能贏過 BiLSTM 和 ELMO 等神經網路的方法，也驗證了 BERT 是一個優質的語言表示模型，能夠達到甚至超越傳統的類神經網路模型。

（二）、知識圖譜

知識圖譜(Knowledge Graph)[14]曾於 2012 年由 Google 所提出，其本質上是基於圖的數據結構主要由節點(實體)和邊(關係)組成三元組(Triplets)。過去知識圖譜經常被應用在金融或醫療專業領域中[9]。近年來，隨著人工智慧的蓬勃發展，知識圖譜又被廣泛得應用在問答系統和聊天機器人中，協助系統更深地理解自然語言並做推理來提升整體問答的效果。目前有許多高品質且大規模的開放式知識圖譜，以英文來說包含 WordNet[15]、Wikidata 和 Yago 等[16]。中文的知識圖譜則包含知網(HowNet)[17]、CnDbpedia[1][9]和 Zhishi.me[2]等。

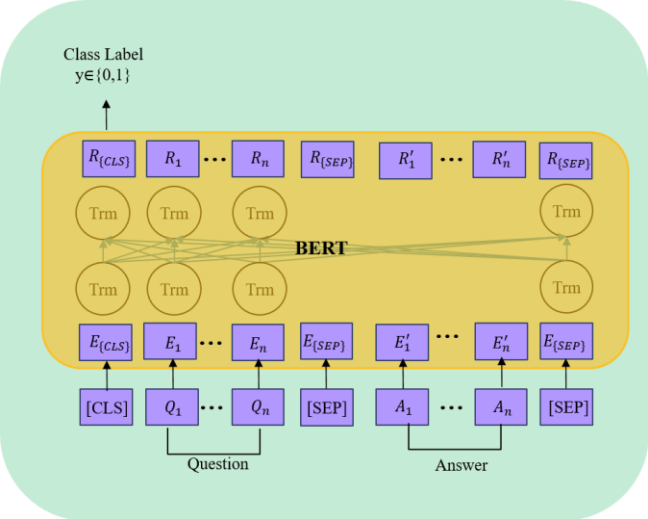
其中 HowNet 為大陸學者於 1998 年所創建，主要為電腦所設計的大型雙語知識庫。提供了設計人工智慧軟體所需的外部知識。HowNet 總共收錄了 50,220 筆漢語詞語，所涵蓋的概念總量達 62,174 筆，目前仍在持續擴充中。詞彙是最小的語言使用單位，但卻不是最小的語意單位。HowNet 使得自然語言理解上更深入地了解詞彙背後豐富的語意。對於一個詞在不同的情境之下可能會有不同的概念。在 HowNet 中 W_C 為一個概念，而 G_C 表示概念所屬於的詞性，DEF 則表示其定義。表一列舉幾個 HowNet 中的例子，其中每個『|』代表一個義原，左邊為英文右邊為中文。

表一、HowNet 定義舉例

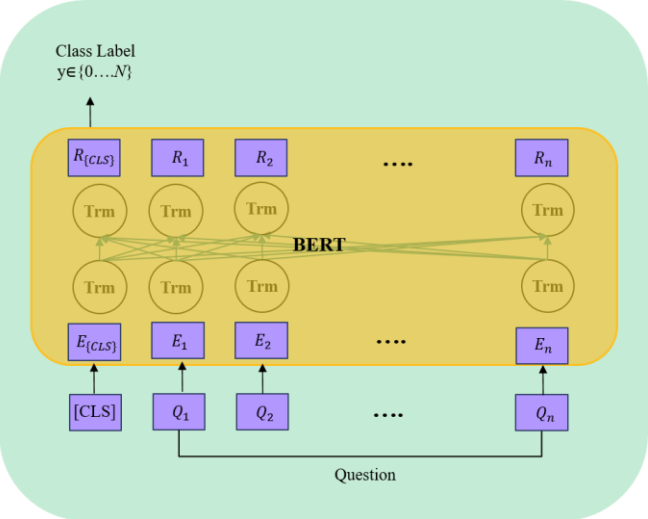
W_C	G_C	DEF
頂點	N	location|位置:belong={angular|角}, modifier={dot|點}
大學生	N	person|人, education|教育, highrank|高等
鮮花	N	FlowerGrass|花草

（三）、FAQ 相關研究

[1] http://kw.fudan.edu.cn/cndbpedia/search/
[2] http://zhishi.me/



圖一、BERT 架構預測問句與答案相關性



圖二、BERT 架構預測問句所屬答案

FAQ 檢索任務主要是利用使用者的查詢來找出一句相對應最適合之答案。而在過去的研究中可透過非監督式學習方式，例如 Okapi BM25 模型或向量空間模型，比較使用者的查詢以及 FAQ 樣本中的標準問句，若兩者的語意相近，則該 FAQ 樣本的答案就可能包含使用者需求的資訊。另外，FAQ 樣本的答案中也可能包含其他額外的資訊。因此，也可以透過監督式學習方式，例如 BERT 或 XLNet[18]，比較使用者查詢與標準問句的相關答案之間相關性。

除此之外，也可以藉由將非監督式學習的排名結果以及監督式學習排名的結果做線性組合並重新排序，以達到同時考慮上述兩種面向的效果。如 Wataru[1]等人使用 Okapi BM25 模型查找使用者查詢和標準問句相似度結合 BERT 模型查找標準問句和答案相關性結果來達到考慮兩個不同面向以提升整體效果，贏過 BiLSTM 和 CNN 等模型。

三、研究方法

本研究主要針對兩個層面做探討。第一個層面，針對不同特性的 FAQ 資料集所適用之策略作探討，分為使用者查詢與標準問句的相關答案之間相關性(Pair)、將使用者查詢做分類(Multiclass)以及比較使用者查詢和標準問句的相似度(Similarity)。第二個層面，我們再針對上述策略透過不同模型方法做實驗測試，使用的模型分為 BERT、K-BERT 以及 Sentence-BERT。

（一）、BERT 神經網路模型

BERT 為雙向 Transformers 的解碼器(Encoder)，主要的特色在於預訓練的方法上使用了 Masked LM (MLM)和 Next Sentence Prediction(NSP)。其中以 MLM 機制更為重要，能隨機性屏蔽掉部分輸入的 Tokens 並預測這些被遮蔽掉的 Tokens，目的在於讓模型依照上下文的資訊學習填補被遮蔽掉的 Tokens。

在實驗中我們使用兩種 Fine-Tuning 的方式來訓練 BERT，分別應用在使用者查詢與標準問句的相關答案之間相關性(Pair)如圖一和將使用者查詢做分類(Multiclass)如圖二，兩種 FAQ 檢索實作策略。前者的輸入分為兩個部分問題和答案中間以[SEP]特殊符號來分隔兩句，問題前面加上[CLS]用於分類，最後輸出{0,1}表示問題和答案是否相關。後者的輸入為單句，以問題作為輸入並在句首加上[CLS]用於分類，最後輸出{0,…,*N*}表示此問句所屬於的答案類別，*N* 表示為答案總共有 *N* 個。

（二）、K-BERT 神經網路模型

BERT 模型透過預訓練方式可以從大型語料庫中獲取通用的語言表示，但是缺少特定領域知識。所以在處理需要領域知識的任務上表現不佳，例如，醫學問答任務或是法律知識問答。在理解領域知識資料時，普通人只能根據其上下文理解單詞，而專家則可以利用相關領域知識進行推斷。為了達到此目的，引入知識圖譜至 BERT 模型中使得模型成為領域專家是一個較佳的方式。

在實驗中我們採用開放式一般領域知識圖譜，例如:HowNet 或是 CnDbpedia，作為知識三元組注入至語句之中。但是過多的知識注入可能會使得語句偏離其真正含義，因而產生知識噪聲（Knowledge Noise）問題。為了克服此問題，K-BERT 架構如圖三引入了軟位置(Soft-position)和可見矩陣(Visible Matrix)來限制不適當知識注入所產生的負面影響。在 K-BERT 中，首先透過知識層(Knowledge Layer)給定輸入句子 $s = \{w_0, w_1, \ldots., w_n\}$和知識圖譜 k，它會將知識圖譜注入到語句中並且轉換成語句樹(Sentence tree) $Tr = \{w_0, w_1, \ldots., w_i\{(r_{i0}, w_{i0}), \ldots., (r_{ik}, w_{ik})\}, \ldots\ldots, w_n\}$，語句樹允許每一個詞最多有兩個分支但其深度只能為一。接著在輸入 K-BERT 前會將資訊轉換成三層

的編碼層(Embedding layer)，包含 Token Embedding、Soft-position Embedding 以及 Segment Embedding。在 K-BERT 中，首先會將語句樹平鋪，平鋪以後的句子是雜亂不易閱讀的。K-BERT 通過 Soft-position 編碼方式來恢復語句樹的信息順序。其中在 Seeing Layer 利用 Masked-Transformer 的概念引入 Visible Matrix，讓詞的詞嵌入只源自於同一句語句樹的枝幹上，不同枝幹的詞之間相互不影響。以矩陣方式表示為:

$$M_{ij}= \begin{cases} 0 & ,\text{visible} \\ -\infty & ,\text{invisible} \end{cases} \quad (1)$$

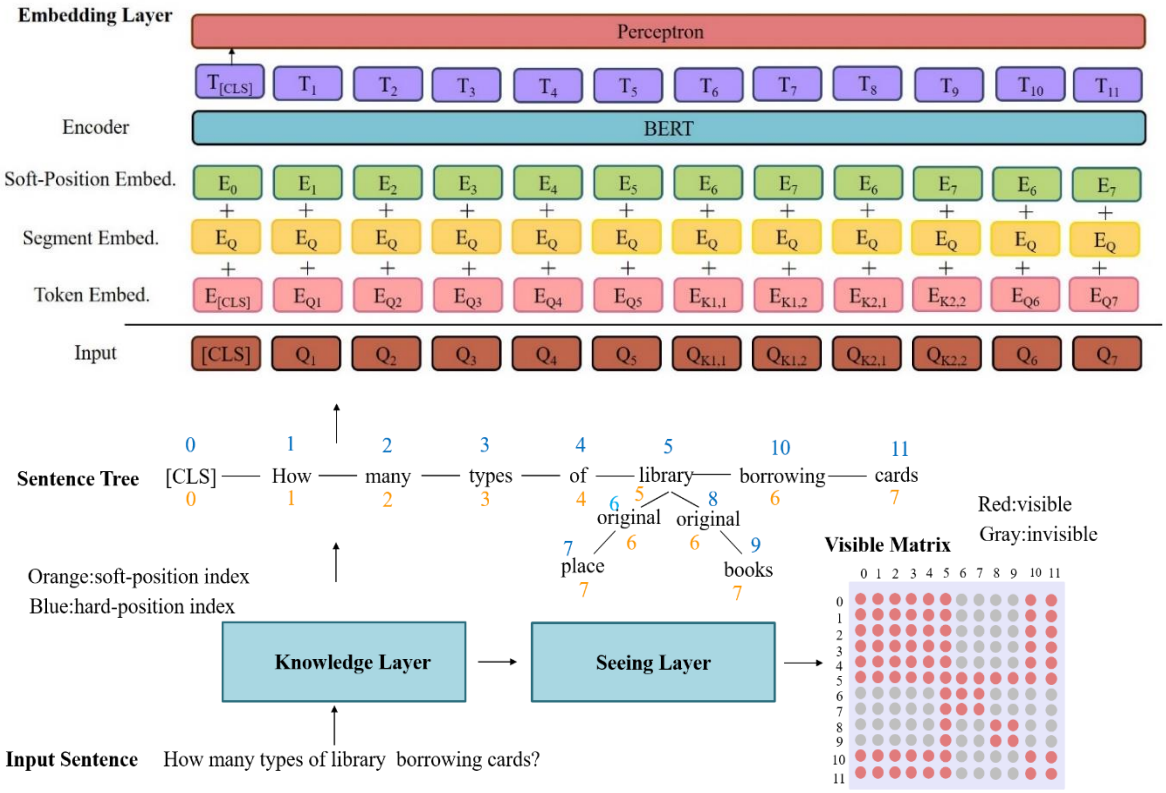
其中 0 表示詞彙之間於語句樹中同一條枝幹上是相互可見的，$-\infty$則表示詞彙之間於語句樹中不同條枝幹上是相互不可見的。將 Visible Matrix 加入至 Self-attention 中表示為:

$$Q^{i+1}, K^{i+1}, V^{i+1}=h^{i}W_{q}, h^{i}W_{k}, h^{i}W_{v} \quad (2)$$

$$S^{i+1} = \text{softmax}(\frac{Q^{i+1}K^{i+1^{T}} + M}{\sqrt{d_k}}) \quad (3)$$

$$h^{i+1}=S^{i+1}V^{i+1} \quad (4)$$

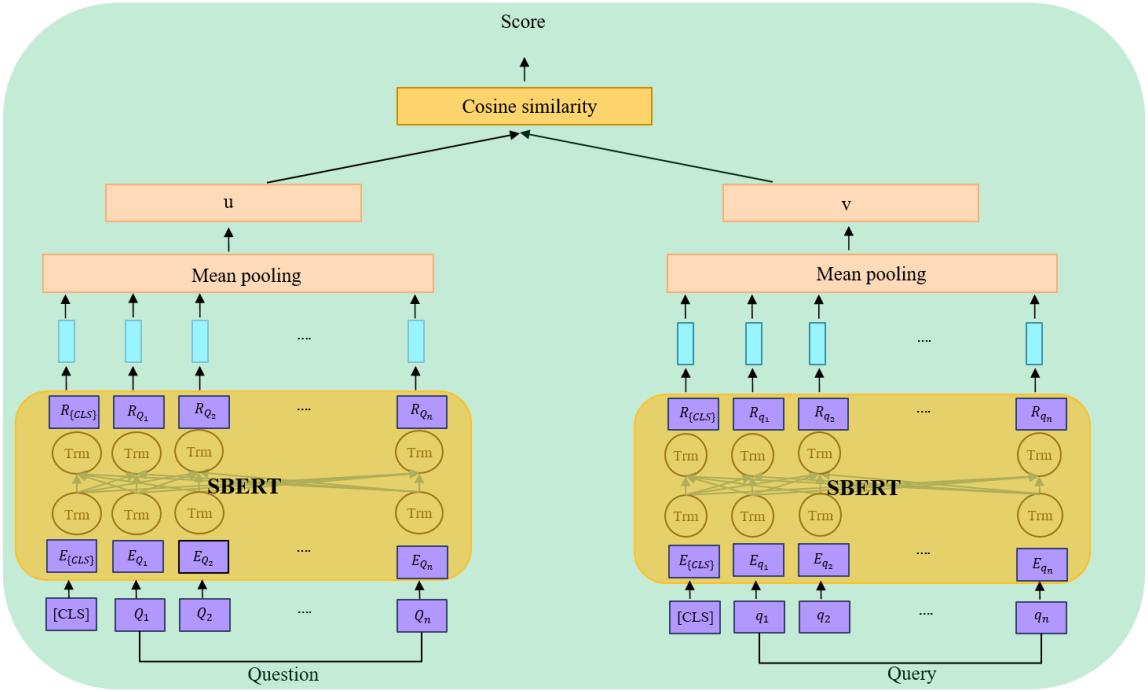
其中式(3)在原始 Self-attention 中多加上遮罩矩陣 M(Mask matrix)，讓語句樹中不同枝幹上的詞彙之間互不影響。



圖三、K-BERT 架構[9]

（三）、Sentence-BERT 神經網路模型

BERT 在語意相似度計算(Semantic Similarity)任務上已經有不錯的表現，但是它必須將兩句語句都輸入至模型中，因而導致大量的計算成本。Sentence-BERT (SBERT)網路結構可以解決 BERT 模型的不足，將兩句不同語句輸入到兩個 BERT 中，其中兩個 BERT 模型的參數是共享的，得到每一句語句的表徵向量。直接利用兩句話的表徵向量計算餘弦相似度，如圖四。



圖四、Sentence-BERT 架構[11]

四、實驗結果與討論

（一）、實驗材料

本研究材料採用兩個公開的中文數據集 TaipeiQA 以及 LawQA[9]。TaipeiQA 是從台北市政府官方網站爬蟲取得的 FAQ 問答數據集，多種問題可能對應至一種答案。LawQA 為法律相關的 FAQ 問答數據集，一句問題只會對應至一句答案。其中將上述兩個數據集整理成三種格式的資料如表二。並將資料切分成訓練集、驗證集和測試集，其中測試集為使用者查詢(Query)。第一種格式資料(Pair)包含多個問題與答案之配對，問題與答案若相關就標記為 1，相反地非相關就標記為 0。第二種格式資料(Multiclass)將每個答案當作一個類別，每個問題皆有相對應的答案。第三種格式資料(Simlarity)包含多個問題與答案之配對以及使用者的查詢問題(Query)當作測試資料。

另外在實驗中我們有加入開放式一般領域知識圖譜的資料，分別使用到 HowNet 總

表二、三種 FAQ 檢索策略資料格式

	Question	Answer	Relevance
Pair	我在網路貸款上借了 1500 百塊不還會怎麼樣？	若無力償還會面臨法院後果的，建議與債權人積極協商，爭取延遲還款。債權人也會起訴你，然後申請執行你的財產的。	1
		我可以幫你	0

	Question	Class
Multiclass	我在網路貸款上借了 1500 百塊不還會怎麼樣？	20

	Question	Query
Similarity	我在網路貸款上借了 1500 百塊不還會怎麼樣？	我貸款借 1500 百元不還會怎樣

表三、資料集統計

資料集	策略	訓練	驗證	測試
TaipeiQA	Pair	147,998	172	2070
	Multiclass	5,821	1,665	1,035
	Similarity	7,485	-	1,035
LawQA	Pair	29,003	3,708	3,631
	Multiclass	14,656	2,657	14,467
	Similarity	14,656	-	1,750

共 52,576 筆以及 CN-Dbpedia 總共 7526,076 筆，並且限制掛載數量最多為 2 深度最多為 1。HowNet 是一個大型的語言知識庫表示每個詞彙背後更深層的語意。CN-DBpedia 是由復旦大學研發並維護的大規模結構化百科知識圖譜資料。百科領域延伸至法律、工商、金融、文娛、科技、軍事、教育和醫療等十多個領域，已經成為業界與學界開放中文知識圖譜的首選。

（二）、評估指標

在評估方法上我們採用 F1 值，主要是為了觀察各個模型分別在三種 FAQ 檢索策略上精確度的一種指標表示為：

$$F1 = \frac{2 * precision * recall}{Precision + recall} \quad (5)$$

F1 值就是精確率 (precision)和召回率(recall)的調和平均值。其中精確率計算所有正確被檢出的結果(TP)占實際上被檢索到的(TP+FP)比例表示為：

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

召回率是計算所有正確被檢出的結果(TP)占所有被應該檢索到的(TP+FN)比例，表示為:

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

（三）、實驗結果

我們的實驗結果顯示於表四中。首先，從三種 FAQ 檢索策略中可以觀察到，TaipeiQA 數據集中多種問題可能對應至一種答案且答案變化性少，所以適合使用 Multiclass 分類的策略直接將每個答案當作一個類別，而使用者查詢透過模型預測相對應最適合的答案。另外，在 LawQA 數據集中一句問題只會對應至一句答案，所以直接使用 Pair 策略計算使用者查詢與標準問句的相關答案之間相關性上效果比較好。

表四、實驗結果

Strategies\Models\Datasets		TaipeiQA	LawQA
Pair	BERT	0.518	0.864[9]
	K-BERT(HowNet)	0.521	**0.873[9]**
	K-BERT(CnDbpedia)	0.507	0.875[9]
Multiclass	BERT	0.697	0.168
	K-BERT(HowNet)	**0.706**	0.178
	K-BERT(CnDbpedia)	0.704	0.139
Similarity	SBERT	0.261	0.827

接著，在模型方法的比較上可以觀察到 BERT 模型中加上 HowNet 知識圖譜後優於原始 BERT 模型。但是在加上 CnDbpedia 知識圖譜後可能會造成過多知識噪音使得準

確率下降。可見加入適當量且高品質的知識圖譜才會增進模型的判斷能力。此外，在 TaipeiQA 數據集中使用者查詢與標準問題問法的差異性較大所以在 Sentence-BERT(SBERT)相似度比對上效果較差。在 LawQA 數據集使用者查詢與標準問題問法較為相似，所以在 Sentence-BERT(SBERT)相似度比對上效果較佳。

最後，我們還嘗試在 LawQA 資料集中的語句樹(Sentence tree)掛載不同數量的知識圖譜三元組(0,1,2,3,4)並比較對實驗結果的影響。在表五中，我們可以觀察到掛載量增加可以提高準確率，但是掛載過多的知識圖譜三元組反而會造成噪音因而造成準確率下降。

表五、掛載不同量知識圖譜之結果

掛載量	準確率
0	0.864
1	0.868
2	**0.873**
3	0.872
4	0.871

五、結論

本研究使用三種基於 BERT 的模型方法，應用到三種自然語言 FAQ 檢索策略上。經實驗驗證 K-BERT 模型中加上知識圖譜效果優於單獨使用 BERT 模型。另外，在 FAQ 檢索任務上資料集屬於多種問題對應至一種答案的情況下適合使用 Multiclass 分類的策略。若資料屬於一句問題只會對應至一句答案，則適合使用 Pair 策略計算使用者查詢與標準問句和相關答案對之間相關性。使用者的查詢與標準問句較為相似，則適合使用 Similarity 策略計算使用者查詢與標準問句相似度並找出對應的答案。從中可以觀察到不同特性的資料集皆有各自合適的方法和策略。在未來，我們希望能夠考慮到資料集中不同面向的資訊；例如，問題與答案相對應的主題，並加入到模型之中與目前的方法做比較。

參考文獻

[1] Wataru Sakata et al., “FAQ retrieval using query-question similarity and BERT-based query-answer relevance,” In *Proceedings of the International ACM SIGIR Conference on*

Research and Development in Information Retrieval, pages 1113–1116, 2019.

[2] Mladen Karan and Jan Šnajder, "Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. Expert Systems with Applications," 91: 418-433, 2018.

[3] Yu-Sheng Lai et al., "Intention Extraction and Semantic Matching for Internet FAQ Retrieval Using Spoken Language Query," In *Proceedings of Research on Computational Linguistics Conference XIII*. 2000.

[4] Kristian Hammond et al., "FAQ finder: a case-based approach to knowledge navigation," In *Proceedings the 11th Conference on Artificial Intelligence for Applications*, IEEE, 1995.

[5] Ask Jeeves, [Online]. Available: http://www.ask.com

[6] Stephen Robertson and Hugo Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, 3(4): 333–389, 2009.

[7] Gerard Salton et al., "A vector space model for automatic indexing," *Communications of the ACM*, 18(11), pages 613–620, 1975.

[8] Jacob Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[9] Weijie Liu et al., "K-BERT: enabling language representation with knowledge graph," In *Proceedings of the AAAI Conference on Artificial Intelligence AAAI*, pages 2901–2908, 2020.

[10] Ashish Vaswani et al., "Attention is all you need," In *Proceedings of Conference on Neural Information Processing Systems*, pages 5998–6008, 2017.

[11] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[12] Matthew Peters et al., "Deep contextualized word representations," In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, 2018.

[13] Alec Radford et al., "Improving language understanding by generative pre-training," 2018.

[14] Shaoxiong Ji et al., "A survey on knowledge graphs: Representation, acquisition and applications," *arXiv preprint arXiv:2002.00388*, 2020.

[15] George A. Miller. "WordNet: a lexical database for English," *Communications of the ACM*, 38(11): 39–41, 1995.

[16] Fabian M. Suchanek et al., "YAGO: a core of semantic knowledge," In *Proceedings of the international conference on World Wide Web*, pages 697–706, 2007.

[17] Zhendong Dong et al., "HowNet and Its Computation of Meaning," In *Proceedings of the International Conference on Computational Linguistics*, pages 53–56, 2010.

[18] Zhilin Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," In *Proceedings of Conference on Neural Information Processing Systems*, pages 5753–5763, 2019.