中文電影對話問答系統資料集
Chinese Movie Dialogue Question Answering Dataset

Shang-Bao Luo
Academia Sinica / Taiwan.
lowzhang@iis.sinica.edu.tw

Cheng-Chung Fan
Academia Sinica / Taiwan.
jjfan@iis.sinica.edu.tw

Kuan-Yu Chen
National Taiwan University of Science and Technology / Taiwan.
kychen@mail.ntust.edu.tw

Yu Tsao
Academia Sinica / Taiwan.
Yu.tsao@citi.sinica.edu.tw

Hsin-Min Wang
Academia Sinica / Taiwan.
whm@iis.sinica.edu.tw

Keh-Yih Su
Academia Sinica / Taiwan.
kysu@iis.sinica.edu.tw

摘要

本論文建構一個中文對話式問答資料集 CMDQA。內容為中文電影資訊獲取的多輪對話場景，總共包含一萬筆對話，共約四萬輪對話。所有問題與背景文檔，皆由網路爬蟲從維基百科彙整而來。每個問題的答案都是其相關文檔內的某個片段。此外，為了模擬真實對話問答的情景，對話中會有代名詞的使用。因此，在 CMDQA 中，問答模型除了需自動地檢索相關文檔外，亦需處理代名詞與歷史資訊的問題。除了對話式多輪問答外，本資料集還可用於評估資訊檢索、機器閱讀理解與問題轉寫等任務的模型成效。除了 CMDQA 以外，本研究提供一個基礎系統並測試其效果。實驗顯示，基礎系統的效能與真人尚有相當大的差異，因此本資料集可對相關研究提供足夠的挑戰性。

Abstract

This paper constructs a Chinese dialogue-based information-seeking question answering dataset CMDQA, which is mainly applied to the scenario of getting Chinese movie related information. It contains 10K QA dialogs (40K turns in total). All questions and background documents are compiled from the Wikipedia via an Internet crawler. The answers to the questions are obtained via extracting the corresponding answer spans within the related text passage. In CMDQA, in addition to searching related documents, pronouns are also added to the question to better mimic the real dialog scenario. This dataset can test the individual performance of the information retrieval, the question answering and the question re-writing modules. This paper also provides a baseline system and shows its performance on this dataset. The experiments elucidate that it still has a big gap to catch the human performance. This dataset thus provides enough challenge for the researcher to conduct related research.

關鍵字：資訊獲取問答系統、對話式問答系統資料集、中文電影問答

Keywords: Information-Seeking Question Answering, Dialogue-based Question Answering Dataset, Chinese Movie QA

1 諸論

近年來，深度學習在資訊獲取問答(Information-seeking QA)的技術突飛猛進。這個任務目標是希望讓機器像人類一樣進行文檔閱讀，並根據使用者給出的問題在文檔中找出對應的答案。這個技術可以讓電腦幫助人類在大量文檔中找到想要的資訊，可以減輕資訊獲取的成本、加速資訊處理的速度以及提升資訊的利用率。此外，對話式的問答系統可進一步讓使用者以漸進式的方式來搜尋答案，使其更具親和性。

D	Document ID : 依照維基百科所蒐集之所有電影相關之文章，並給予相對應的 ID 表。	
Q_1	**凱雷特**寫的電影是什麼？	what is a film written by **Etgar Keret**?
A_1	9.99 美元	$9.99
$Gold\ R_1$	凱雷特（希伯來語：אתגרקרר ת，出生於 1967 年 8 月 20 日）是一位以色列作家，以其短篇小說、平面小說和影視劇本寫作而聞名。9.99 美元是一部 2008 年澳大利亞 …	Etgar Keret (Hebrew: אתגרקרר ת, born August 20, 1967) is an Israeli writer known for his short stories, graphic novels, and scriptwriting for film and television. $9.99 is a 2008 Australian …
Q_2	**這部電影**是哪一年上映的？	what was the release year of **this movie**?
A_2	2008	2008
$Gold\ R_2$	9.99 美元是一部 2008 年澳大利亞定格動作成人動畫片，由塔蒂婭·羅森塔爾（Tatia Rosenthal）撰寫和導演…	$9.99 is a 2008 Australian stop-motion adult animated drama film written and directed by Tatia Rosenthal …

表 1. CMDQA 之題目範例

近年來問答系統已有大量研究與發展(Devlin et al., 2018; Zhu et al., 2021; Chakraborty et al., 2021)，本論文專注在對話式的資訊獲取問答系統。在尋求信息的對話中，系統與使用者彼此會反覆提問，以確定使用者真正想問的問題。對此種對話進行建模具有一定的挑戰性，因為問答模型需要去理解上下文、每輪對話的資訊、代名詞、歷史資訊與主題轉換的狀況，以便找出使用者真正想要的答案。目前在英文上，已建構了 CoQA (Siva et al., 2019)與 QuAC (Choi et al., 2018) 資料集，主要針對多種領域主題、共指、推理與不可回答的問題。

在對話式的問答系統中，常常會以代名詞來指代前輪對話中的專名實體(Named Entity)(Stent and Bangalore, 2010)。對於這種狀況，模型需要能夠解決上下文依賴關係的機制，來正確解釋後續問題的真正含意。在代名詞指涉的問題上，現有研究是透過問題轉寫(Question Rewriting, QR) (Elgohary et al., 2019; Liu et al., 2018; Vakulenko et al., 2021; Tredici et al., 2021)，將含糊資訊變成明確問題，以此來提高問答系統的效能。

此外，文檔搜尋式的問答系統(Yang et al., 2015; Qu et al., 2020; Longpre et al., 2021) 需要先從大量的背景文檔中找到回答問題所需的相關文檔，再根據相關文檔進行作答。由於文檔搜尋通常透過資訊檢索技術來完成，因此資訊檢索的成效會直接影響回答的品質。然而現有的中文問答資料集(Shao et al., 2018; Yiming et al., 2018; He et al., 2018; Zheng et al., 2019; Sun et al., 2019)，都沒有包含文檔搜尋部分，未能模擬真實問答中的情境，因此本篇論文特別針對此問題建構一個中文電影對話式資訊獲取問答資料集(***Chinese Movie Dialogue Question Answering Dataset***；簡稱***CMDQA***)。本資料集將公開供相關研究使用。

本資料集具有以下特點：

- 背景文檔從維基百科電影主題搜集而成，對話式問答系統需要先從大量文檔中篩選出與問題相關的文檔。
- 問答模型不能只使用單輪的文檔與問題來回答問題，還需考慮代名詞與歷史資訊來進行回答。
- 針對每個問題均有提供相關文檔，因此可作為資訊檢索任務的資料集，亦可簡化作為機器閱讀問答任務的資料集。
- 提供各種基礎模型，讓研究者可自由地抽換不同模型，與自行開發的模型相互搭配，並在本資料集上評估效能。

2 任務定義

在對話式多輪問答任務中，每一輪對話中會有一個問題，模型將根據問題、歷史對話內容以及相關文檔來回答這一輪對話的問題。因此，在第t輪對話中，整個問答系統可以表示為：

$$A_t = QA(Q_t, H_t, R_t) \quad (1)$$

$$H_t = [Q_1, A_1, Q_2, A_2, \dots, Q_{t-1}, A_{t-1}] \quad (2)$$

$$R_t = IR(H_t, Q_t) \quad (3)$$

其中A_t為問答系統依照當輪問題Q_t、相關文檔R_t與歷史資訊H_t所找出對應的答案。歷史

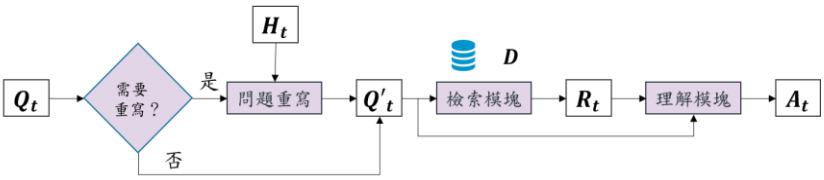


圖 1. CMDQA 之模型概述示意圖

資訊H_t為前$t-1$輪對話中出現過的問題$Q_{1:t-1}$與答案$A_{1:t-1}$。R_t為依照當輪問題Q_t，藉由資訊檢索系統所找出的相關文檔，並依此來回答當輪問題。其中相關的釋例，如表 1 所示。

3 資料蒐集

本論文主要受英文資料集 MovieQA (Tapaswi et al., 2016)啟發，依照英文頁面的維基百科資料來蒐集相關資料，搭配 Google 翻譯取得中文資料。接著由維基百科的標籤資料，透過規則式的模板來產生問題與答案。為了讓品質提升，再經由專名實體檢測與答案對齊等人工整理而得。參照 MovieQA，我們定義了七種標籤：電影、演員、導演、編劇、風格、年份與語言，並依標籤間的關係來建構問題、相關文檔及標準答案的組合。例如：問題提及**導演**，答案回答**電影**，主題就是**導演－電影**，標籤就是電影。針對各個標籤和主題設計的問題數量如表 2 與表 3 所示。

我們依照每一種主題之間的關聯性，將問題生成多輪對話的對話路徑(Dialogue Path)，並且從第二輪對話開始，都會將問題中的關鍵標籤進行代名詞化，讓整個對話情境更加擬真。原來的問題最後被篩選組出二至六輪的對話問題組，相關的資訊如表 4 所示。

4 基線系統概述

本基線系統(Baseline System)依照任務分為三大模塊，分別為檢索模塊、理解模塊與問題重寫模塊，整體示意圖如圖 1 所示。

4.1 檢索模塊

本論文採用 DPR (Karpukhin et al., 2020) 與 BM25 (Trotman et al., 2014)併用的作法(Ma et al., 2021)。資料集會提供每一個問題Q與對應的相關文檔$R \in D$，其中D為維基百科電影主題文檔集合。相關文檔R為問答系統回答問題Q時，檢索模塊認為相關的文檔。

標籤	訓練集	測試集	發展集
電影	12,409	3,675	2,875
演員	3,430	838	495
導演	4,074	1,093	650
編劇	4,880	724	704
風格	20	14	13
年份	101	98	94
語言	63	28	29

表 2. CMDQA 之標籤類別及問題數量

主題	訓練集	測試集	發展集
電影－演員	1,432	773	287
電影－導演	5,720	3,103	1,396
電影－風格	476	217	300
電影－語言	1,351	320	676
演員－電影	2,534	1,793	824
演員－編劇	3,546	732	1,645
演員－年分	7,579	1,597	3,045
編劇－電影	3,012	652	1,542
導演－電影	1,767	550	256

表 3. CMDQA 之主題類別及問題數量

對話輪	訓練集	測試集	發展集
單題數	27,426	9,737	9,971
兩輪	6,390	183	99
三輪	3,038	15	7
四輪	571	0	0
五輪	37	0	0
六輪	2	0	0
合計(多輪)	10,038	198	106

表 4. CMDQA 之多輪問答問題數量

4.2 理解模塊

本資料集回答問題的方式，以從相關文檔中擷取文本答案片段為主，故使用當前標準模型 BERT (Jiao et al., 2019) 為基礎來進行訓練與預測。以第t輪對話為例，輸入X_t的形式如下：

$$X_t = [CLS]q_{t,1}, \dots, q_{t,n}[SEP]r_{t,1}, \dots, r_{t,m}[SEP] \quad (4)$$

其中$R_t = \{r_{t,1}, \cdots, r_{t,m}\}$為檢索模型給予之相關文檔，$Q_t = \{q_{t,1}, \cdots, q_{t,n}\}$為第$t$輪對話所給予的問題。BERT 透過$U$層轉換器(Transformer)，來獲得上下文資訊C：

$$C^0 = XW_{token} + W_{position} + W_{segment} \quad (5)$$

$$C^u = Transformer(C^{u-1}) \; \forall u \in [1, U] \quad (6)$$

其中W_{token}、$W_{position}$與$W_{segment}$為 BERT 所使用到的三種向量。最終透過全連結層FFN_{start}與FFN_{end}來產生答案片段出現在相關文檔中開始P^i_{start}與結束P^i_{end}位置的概率：

$$P^i_{start} = softmax(FFN_{start}(C^U)) \quad (7)$$

$$P^i_{end} = softmax(FFN_{end}(C^U)) \quad (8)$$

最終訓練目標如下：

$$(s, e) = \underset{s \le e}{\operatorname{argmax}} p^s_{start} p^e_{end} \quad (9)$$

即最大化正確的答案開始和結束位置的概率。

4.3 問題重寫模塊

本論文問題重寫採取基於規則的方式（如圖 2 所示）：以當前問題和歷史資訊為基準做分析，來選擇較合理的代名詞替換，並驗證代名詞轉換對於問答模型的影響。

首先將當輪問題Q_t透過句法分析，得出代名詞位置，並預測該代名詞為何種標籤。接著對歷史對話H_t的每一個問題與答案，依照七大標籤來提取關鍵詞後，以具相同標籤的關鍵詞來取代原本的代名詞，最終獲得當輪明確問題Q'_t。當具相同標籤的關鍵詞有多個時，挑選的順序是最近一輪的關鍵詞優先，且答案的關鍵詞優先於問題的關鍵詞。

5 基線模塊評估

5.1 資訊檢索

資訊檢索的實驗結果如表 5 所示。共有四種基礎模型，BM25 模型採用 PyLucene[1]與 Gensim (Rehurek and Sojka, 2011)的版本，表 5 中後者標記為 BM25*。DPR 模型則是採用 Pytorch (Paszke et al., 2017) 的版本[2]。最後 BMD

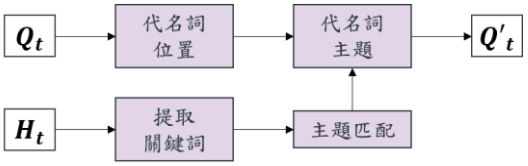


圖 2. 問題重寫之示意圖

(BM25+DPR)則是採用 DPR 與 BM25 模型，透過式(10)進行分數加權後，以最終的文檔s_{BMD}分數來進行篩選。

$$s_{BMD} = s_{DPR} + (1 - \alpha)s_{BM25} \quad (10)$$

訓練 CMDQA 的資訊檢索時，採取段落級(Paragraph-level)的方式來進行背景文檔的切割。呈現數據時，會將段落級的資料恢復成文檔級(Document-level)來進行預測與效能評比。

以召回率(Recall)為評估指標，可以發現在 BMD 的配置下有較好的結果，並且篩選的文檔的數量愈多，其召回率愈高。

5.2 問答模型

問答模型的實驗結果如表 6 所示。我們以 RoBERTa (Liu et al., 2019)為架構，並以中文預訓練模型[3]做為初始。由於簡體中文的預訓練模型使用較多的預訓練文本，我們的實驗皆是將正體中文轉換為簡體中文來進行。在整體訓練中，批次大小為 64、總期(Epoch)次數為 3、學習率為 3e-5。

由表 6 的實驗結果可以發現，在檢索相關文檔R的召回率愈高的情況下，問答模型的效能反而愈低。這主要是由於召回率愈高時，所需輸入問答模型的文字愈長，無關的內容也跟著變多，使得問答模型無法有效地抓取到正確的答案片段。

5.3 問題重寫

問題重寫的實驗結果如表 7 所示。本資料集提供三種不同的方法，說明如下：

- QR[ReA]將本輪問題Q_t的代名詞，直接以上一輪的答案A_{t-1}來進行取代。
- QR[ReQ]將本輪問題Q_t的代名詞，直接以上一輪的問題Q_{t-1}的關鍵詞來進行取

[1] https://cwiki.apache.org/confluence/display/lucene

[2] https://github.com/facebookresearch/DPR

[3] https://huggingface.co/hfl/chinese-roberta-wwm-ext

IR Module	訓練集			測試集			發展集		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
BM25	20.925	51.356	61.510	20.158	48.325	59.469	20.433	49.926	60.158
BM25*	19.492	50.826	62.917	26.740	55.298	65.555	23.278	51.513	63.199
DPR	54.120	75.776	79.435	44.613	72.558	77.963	46.447	71.023	76.206
BMD	68.359	86.955	88.384	62.131	88.022	90.785	63.912	87.223	90.412

表 5. CMDQA 之資訊檢索測試結果

QA Module	訓練集		測試集		發展集	
	EM	F1	EM	F1	EM	F1
Gold P	74.77	87.70	90.78	95.56	91.76	96.49
BM25 @ 1	25.01	44.99	26.72	41.35	25.87	43.49
BM25 @ 5	18.94	37.89	21.51	35.37	21.32	38.08
BM25 @ 10	16.22	34.56	18.42	31.60	18.77	35.61
DPR @ 1	38.36	53.27	31.84	47.43	29.07	48.41
DPR @ 5	24.87	34.91	22.88	37.05	20.72	39.11
DPR @ 10	24.31	38.93	22.52	36.51	20.33	38.59
BMD @ 1	45.61	62.36	40.22	56.56	39.29	58.31
BMD @ 5	29.68	46.71	30.11	44.63	28.68	46.10
BMD @ 10	28.64	44.54	28.35	42.99	26.42	44.05

表 6. CMDQA 之問答模型測試結果

代。此作法的理由為，當輪的問題常針對前一輪的問題進一步追問。

- QR[M]將本輪問題Q_t的代名詞，透過 4.3 節所描述的方法來進行取代。

問題重寫所採用的評估方法為：比較問題重寫後的問題與黃金問題(Gold Q)是否完全相同(Exact Match, EM)，或是採用機器翻譯常用的兩個指標 ROUGE-L 與 BLEU，其中 ROUGE-L 根據召回率衡量重寫後的質量，BLEU 則是根據精確度來進行評量。從表 7 可以看到，使用 EM 評估指標的結果非常不理想。這是因為本論文是將每一組原本的問題，利用不同的句法結構，改寫成相同意思的兩種問句(例如：A 導演曾經參與哪部電影？➔ 哪部電影由這個導演執導？)。因此，我們可以將這個任務視為一種摘要或轉寫，就可以透過機器翻譯常用的兩種指標來進行評估。

5.4 對話問答

多輪對話問答的實驗結果如表 8 所示。我們採用四種實驗設置，說明如下：

- 人工測試數據(Human Performance*)：從訓練、測試與發展集中各抽選 100 題進行人工評測的結果。
- 第一種設置(Gold P + Gold Q)：在完整多輪對話中，每一輪均使用包含答案的文檔(Gold P)與無歧義的問題(Gold Q)讓問答模型來回答。要注意的是，本論文提供的問答模型，只對單輪題目來進行訓練，並沒有針對對話式的架構來特別進行訓練。
- 第二種設置(BMD + Gold Q)：在完整多輪對話中，每一輪均使用無歧義的問題(Gold Q)讓 BMD 檢索模型去尋找相關文檔，最後由問答模型作答。其中@1-10 指的是將檢索回來的相關文檔串接 1 至 10 篇輸入問答模型。
- 第三種設置(BMD + Pron. Q)：在完整多輪對話中，除了第一輪的問題是無歧義的問題外，從第二輪開始，均使用含有代名詞的問題(Pron. Q)。並且每一輪中，BMD 檢索模型均直接使用該輪問題來尋找相關文檔，交給問答模型作答。我們認為此種狀況是最合乎真實情境的狀況。
- 第四種設置(BMD + Rewriting Q) ：與第三種設置相似，但在完整多輪對話中，第二輪起，含有代名詞的問題將透過

QR Module	訓練集			測試集			發展集		
	ROUGE-L	BLEU	EM	ROUGE-L	BLEU	EM	ROUGE-L	BLEU	EM
Original Q	42.683	16.49	0	43.241	14.846	0	42.291	17.289	0.885
QR[ReA]	44.706	18.089	0	45.585	17.834	0.173	45.773	19.974	1.22
QR[ReQ]	42.168	15.389	0.119	42.009	13.186	0	40.72	13.927	0
QR[M]	56.241	40.467	5.622	58.454	41.866	3.286	57.910	43.759	3.540

表 7. CMDQA 之問題重寫測試結果

Dialogue	訓練集			測試集			發展集		
	F1*	F1	EM	F1*	F1	EM	F1*	F1	EM
Human Performance*	88.619	88.246	80.851	89.421	89.990	76.595	86.123	87.026	78.723
Gold P + Gold Q	76.763	76.851	52.212	86.861	86.869	73.737	81.735	82.075	66.038
BMD@1+ Gold Q	45.776	46.019	17.543	43.309	43.434	15.152	49.772	49.686	19.811
BMD@5 + Gold Q	32.209	32.303	7.681	35.28	35.101	9.596	38.356	37.893	10.377
BMD@10 + Gold Q	31.257	31.265	7.362	33.333	33.165	8.586	35.16	35.063	10.377
BMD@1 + Pron. Q	19.676	20.614	0.309	21.898	21.886	0.505	23.288	23.428	0.943
BMD@5 + Pron. Q	15.582	16.228	0.199	18.248	18.182	0.505	17.808	18.082	0.000
BMD@10 + Pron. Q	13.145	13.674	0.004	13.382	13.384	0.505	20.548	20.440	0.000
BMD@1 + Rewriting Q	36.209	37.220	5.529	36.983	37.290	8.586	44.292	44.340	15.094
BMD@5 + Rewriting Q	21.842	22.282	2.740	21.655	21.717	4.545	28.767	28.302	8.491
BMD@10+ Rewriting Q	19.770	20.255	1.624	23.114	23.316	5.556	25.571	25.472	7.547

表 8. CMDQA 之多輪對話問答測試結果

QR[M]的方式將問題重寫後交給 BMD 去查找相關文檔及問答模型去進行作答。

本論文對於對話式多輪問答的評估方式分為三種：F1*、F1 與 EM。F1*為將全部多輪對話的所有題目分別計算 F1 再計算總平均。F1 與 EM 則是先分別計算每個多輪對話的平均 F1 與 EM 後，再按多輪對話的數量去計算平均。

觀察對話式多輪問答的結果(表 8)，可以發現，在第一種設置的狀況下，由於給予正確的文檔與無歧義的問題，於測試集與發展集在 EM 評估指標可分別達到 73.737%與 66.038%。第二種設置探討的是：當需要搜尋相關文檔時，對於問答模型效能的影響。明顯可見的是，在沒有任何微調策略的情況下，問答模型於訓練、測試與發展集上，EM 皆達不到 20%。可見相關文檔搜尋是目前的瓶頸。第三種設置則是進一步探討當問句包含代名詞時，對於整體模型的影響。結果顯示問答模型幾乎已完全無法正確的回答。最後，第四種設置下的結果顯示，透過基礎系統提出的問題重寫方法，可提升約 15%~21%的 F1 分數與 5.22%~ 14.15%的 EM 分數。不過整體效能還是與第二種設置有一定的差距，換言之，本論文提出的資料集，是有一定的困難與挑戰性的。

6 總結

本論文構建一個中文電影對話式資訊獲取問答資料集。其中包含一萬個多輪對話(總計約四萬七千輪)。所有問題與背景文檔，皆由網路爬蟲從維基百科彙整而來。

本論文定義這類問題所需要的框架，依照當輪問題去背景文檔集搜尋相關文檔，並透過對話歷史資訊來改寫需要共指消解的問題，最終交給問答模型來回答當輪問題。

本論文依照上述的框架，提供各種模塊的基礎模型與個別效能。基線實驗結果顯示 CMDQA 資料集有一定的挑戰性與困難性。

未來規劃中，預計將擴充資料集中的問題類型，例如：多文本段與簡答題，讓 CMDQA 資料集更加完整與富有挑戰性。另外，會嘗試持續增加題目，讓測試集與發展集的題目分布更加均勻。

參考文獻

Adam Paszke, Sam Gross, Soumith Chintala and Gregory Chanan. 2017. Pytorch: Tensors and

Dynamic Neural Networks in Python with Strong Gpu Acceleration. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration, 6.

Ahmed Elgohary, Denis Peskov and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context.

Amanda J Stent and Srinivas Bangalore. 2010. Interaction between dialog structure and coreference resolution. In 2010 IEEE Spoken Language Technology Workshop, pages 342-347.

Andrew Trotman, Antti Puurula and Blake Burgess. 2014. Improvements to BM25 and language models examined. In Proceedings of the 2014 Australasian Document Computing Symposium, pages 58-65.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng and Sam Tsai. 2018. DRCD: a Chinese Machine Reading Comprehension Dataset, arXiv:1806.00920.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 778–787, Florence, Italy. Association for Computational Linguistics.

Cui Yiming, Liu Ting, Che Wanxiang, Xiao Li, Chen Zhipeng, Ma Wentao, Wang Shijin and Hu Guoping. 2018. A Span-Extraction Dataset for Chinese Machine Reading Comprehension, arXiv:1810.07366.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context, arXiv:1808.07036.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria and Tat-Seng Chua. 2021. Retrieving and reading : A comprehensive survey on open-domain question answering, arXiv:2101.00774.

Hengrui Liu, Wenge Rong, Libin Shi, Yuanxin Ouyang and Zhang Xiong. 2018. Question rewrite based dialogue response generation. In Proceedings of International Conference on Neural Information Processing, pages 169-180.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805.

Kai Sun, Dian Yu, Dong Yu and Claire Cardie. 2019. Probing prior knowledge needed in challenging chinese machine reading comprehension, aXiv:1904.09679.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4631-4640.

Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng and Adriá de Gispert. 2021. Question Rewriting for Open-Domain Conversational QA: Best Practices and Limitations. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 2974-2978.

Nilesh Chakraborty, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann and Asja Fischer. 2021. Introduction to neural network-based question answering over knowledge graphs. Journal of the Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(3):e1389.

Radim Rehurek and Petr Sojka. 2011. Gensim--python framework for vector space modelling. Journal of NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

Reddy, Siva, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A Conversational Question Answering Challenge, In Proceedings of Association for Computational Linguistics, 7:249–266.

Shayne Longpre, Yi Lu and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. Journal of Transactions of the Association for Computational Linguistics, 9:1389-1406

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pages 355-363.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering, aXiv:2004.4906.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, She Qiaoqiao, Liu Xuan, Wu Tian and Wang Haifeng. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications, arXiv:1711.05073.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding, arXiv:1909.10351.

Xueguang Ma, Kai Sun, Ronak Pradeep and Jimmy Lin. 2021. A replication study of dense passage retriever, arXiv:2104.05740.

Yi Yang, Wen-tau Yih and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 2013-2018.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu and Haifeng Wang. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering, arXiv:2010.08191.

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, aXiv:1907.11692.