생성형 언어모델을 이용한 테이블 질의응답 평가

민경구°, 최주영, 심묘섭, 정해민, 박민준, 최정규
LG AI연구원
{kyungkoo.min, jooyoung.choi, myoseop.sim, minjun.park, stanleyjk.choi}@lgresearch.ai

Evaluating Table QA with Generative Language Models

Kyungkoo Min°, Jooyoung Choi, Myoseop Sim, Haemin Jung, Minjun Park, Jungkyu Choi
LG AI Research

요 약

문서에서 테이블은 중요한 정보들을 축약하여 모아 놓은 정보 집합체라고 할 수 있다. 이러한 테이블을 대상으로 질의응답하는 테이블 질의응답 기술이 연구되고 있으며, 이 중 언어모델을 이용한 연구가 좋은 결과를 보이고 있다. 본 연구에서는 최근 주목받고 있는 생성형 언어모델 기술을 테이블 질의응답에 적용하여 언어모델과 프롬프트의 변경에 따른 결과를 살펴보고, 단답형 정답과 생성형 결과의 특성에 적합한 평가방법으로 측정해 보았다. 자체 개발한 EXAONE 1.7B 모델의 경우 KorWiki 데이터셋에 대해 적용하여 EM 92.49, F1 94.81의 결과를 얻었으며, 이를 통해 작은 크기의 모델을 파인튜닝하여 GPT-4와 같은 초거대 모델보다 좋은 성능을 보일 수 있음을 확인하였다.

주제어: Table QA, Question Answering, GPT, EXAONE

1. 서론

문서의 중요 정보를 이해하기 쉽게 표현한 테이블은 재무보고서, 과학 논문 등 다양한 실제 환경에서 널리 사용되고 있다[1]. 최근에는 테이블에 포함된 정보를 활용하기 위해 일반 문서에 적용되던 질의 응답 기술을 테이블 대상으로 적용시킨 테이블 질의응답(Table QA)이 연구되고 있다[2, 3, 4, 5, 6].

테이블 질의응답은 사용자의 질문에 대해 테이블에서 정확한 답변을 제공하는 기술을 말하는 것으로 테이블은 텍스트 또는 데이터 베이스의 형태로 구성되어 있으며, 테이블에 포함된 값은 주로 숫자나 짧은 텍스트로 구성되어 있으며 그 값 자체로는 이해하기 어렵고 테이블의 행/열의 정보가 함께 포함되어 있어야 의미를 해석할 수 있다는 특징을 가진다. 따라서 테이블 데이터 질의응답을 위해서는 이러한 테이블 데이터의 특성을 잘 이해하는 모델의 설계와 평가방법이 필요하다.

기존의 테이블 질의응답 연구 중에는 검색을 이용하거나[3, 5], BERT와 같은 트랜스포머 디코더 구조에 테이블에 대한 토큰을 추가하여 변경하는 방식의 사전 학습된 언어모델(pre-trained language model)을 이용하는 방법이 사용되어 왔으며[4, 7, 8], BERT의 임베딩 구조 변경 없이 테이블 데이터의 포맷 변경을 통해 테이블 구조를 반영하여 학습하는 방법도 좋은 성능을 보이고 있다[9].

최근 주목받고 있는 초거대 언어모델[10, 11, 12]은 zero-shot과 few-shot에서 좋은 성능을 보이고 있으며 프롬프트를 이용해 다양한 분야에 적용이 가능해서 주목받고 있다. 하지만 생성형 결과의 다양성과 프롬프트에 따라 다른 결과가 생성되는 불안정성 때문에[13] 테이블 질의응답 분야에 적용할 경우에 오히려 불리한 결과가 나올 수 있다.

본 연구에서는 테이블 질의응답에서 서로 다른 생성형 언어모델을 이용해(GPT-3.5/ GPT-4/ EXAONE 1.7B) 프롬프트의 변화에 따른 결과와 zero-shot/ few shot/ fine-tuning에 따른 성능 변화 결과를 살펴보았다. 또한 단답형 질의응답 데이터셋에 적합한 성능측정 방법인 SM(Semantic Matching)을 제안하고 결과를 측정하였다.

2. 관련연구

2.1 테이블 질의응답

테이블 질의응답에 대한 연구는 크게 두 가지로 분류할 수 있다[1]. 첫번째 방법은 테이블을 데이터베이스로 간주하고 테이블에 대한 자연어질의를 데이터베이스에 대한 질의가 가능한 SQL로 번역하는 방식이다[6, 14]. 이 방법은 실제 테이블이 관계형데이터베이스로 구축되어 있어야 하고 데이터베이스 스키마정보도 알 수 있어야 한다. 이러한 연구에는 여러 테이블 간의 관계나 연산자에 따라 생성되는 결과 SQL의 난이도를 구분해 벤치마크 데이터셋[13]을 구성하고 T5, BERT 등의 언어모델을 적용한 연구들이 있다 [15, 16].

두번째 방법은 테이블 정보를 일종의 텍스트로 간주하고 문서에서 정답을 찾듯이 정보검색과 기계독해 방법을 적용하는 것이다[3, 5]. 텍스트 문서 대상의 Open-domain QA와 같이 정보검색 모델을 통해 정답을 포함한 테이블을 추출하고(Retriever), 테이블로부터 정답을 선택한다(Reader). 정답추출은 사전학습을 통한 weakly supervised 테이블 파싱을 사용하는 TAPAS[4]와 같은 추출형 방식과 FiD[17] 기반의 생성형 모델

을 사용하는 UniK-QA[18] 등의 모델이 있다.

3. 테이블 질의응답 특성

테이블 구조의 특성상 테이블 질의응답의 정답은 문장이 아닌 단어나 구의 형태로 짧게 구성되어 있는 경우가 많다[9, 19]. 본 연구에서 사용하는 생성형 언어모델은 답을 완전한 문장으로 구성하는 경향이 있는데, 생성형 모델이 출력한 답은 테이블 질의응답에서 많이 사용하는 Exact Matching(EM)이나, F1으로 평가하기에 적합하지 않다.

예를 들어, "나익진은 몇 대 체신부 차관을 지냈나요?"라는 질문에 대해 정답인 '7대' 라고 답하지 않고 '7대 입니다', '정답은 7대입니다', '나익진은 7대 체신부 차관을 지냈습니다'와 같은 답을 출력했다면 의미상 정답인 값을 출력했음에도 EM/F1 측정에서는 각각 0.0/28.6, 0.0/20.0, 0.0/10.5으로 답을 맞추지 못한 것으로 측정되는 결과를 얻게 된다 [표 1].

표 1. 모델 예측값에 따른 성능 변화

질문	나익진은 몇 대 체신부 차관을 지냈나요?		
정답	모델예측값	EM	F1
7대	7대 입니다	0.0	28.6
7대	정답은 7대입니다	0.0	20.0
7대	나익진은 7대 체신부 차관을 지냈습니다	0.0	10.5

따라서 테이블 질의응답의 성능을 평가할 때 좋은 평가를 받기 위해서는 언어모델의 결과를 정답과 유사한 형태로 만들어 줄 필요가 있다. 또한 예측값이 정답과 의미상 일치하는지를 반영할 수 있는 평가방법이 필요하다.

4. 데이터 구성 및 실험

먼저 생성형 언어모델의 프롬프트가 예측값의 패턴에 미치는 영향을 알아보기 위해 질문의 프롬프트를 변경하면서 성능 변화를 알아보았다. 다음으로 다양한 생성형 언어모델별 성능 차이를 살펴보고, 마지막으로, 정답과 의미상 일치하는 결과를 반영하기 위한 평가방법을 확인해 보았다.

4.1 테이블 질의응답 데이터 및 언어모델

본 연구에서는 깃헙에 공개된 KorWikiTQ[1] 데이터셋[9, 19]과 KorQuad2Table 데이터셋[9]을 이용하였다[표 2].

KorWikiTQ 데이터셋은 테이블에 대한 설명과 실제 테이블 내용, 테이블 내용에 대한 질문과 답이 포함되어 있으며 질문 작성 규칙에 따라 1–5의 질문유형을 갖는 한국어 위키백과 문서에서 추출한 문서셋이다. KorQuad2Table은 KorQuAD 2.0[2]에 포함된 테이블 관련 질문답변 데이터셋 중 단답형의 정답이 표에 존재하는 데이터셋 약 1만 개를 추출한 데이터 셋이다.

성능 비교를 위해 위키에서 추출한 140만 개의 텍스트 테이블 쌍을 사용해 MLM으로 사전학습(pre-train)한 BERT기반 모델인 KoTaBERT[19]와 OpenAI의 'gpt-3.5-turbo', 'gpt-4' API[3]를 사용했고, 자체 개발한 EXAONE 1.7B 모델을 파인튜닝하였다.

표 2. 테이블 질의응답 데이터

구분	Train	Dev	Total
KorWikiTQ	58,221	11,771	69,992
KorQuad2Table	3,850	427	4,277

4.2 프롬프트 비교 실험

동일한 질문에 대해 프롬프트를 변경해가면서 결과를 측정하였다. 표 3은 KorQuad2Table 데이터셋에서 100개의 질문을 임의로 추출하여 GPT-3.5 모델을 대상으로 few-shot 테스트한 결과이다.

표 3. GPT-3.5 모델에서 프롬프트에 따른 성능 변화

번호	프롬프트	EM	F1
1	질문에 해당하는 답을 context에서 찾아서 답하세요	39.58	59.96
2	질문에 해당하는 답을 context에서 찾아서 짧게 답하세요.	63.54	79.09
3	질문에 해당하는 답을 context에서 찾아서 짧게 답하세요. 찾은 답을 문장으로 구성하지 말고 그대로 답하세요	69.47	79.83
4	질문에 해당하는 답을 context에서 찾아서 아주 짧게 답하세요. 찾은 답을 문장으로 구성하지 말고 그대로 답하세요	75.79	86.38

프롬프트에 별다른 제한을 두지 않을 경우 생성형 언어모델은 예측값을 문장의 형태로 구성하는 경향이 있다. 프롬프트를 자세히 작성해 결과값의 형태를 구체화시킬수록 정답과 가까운 '단어'나 '구'의 형태를 출력해 EM이 높아지는 것을 확인할 수 있었다.

4.3 생성형 언어모델 비교 실험

위키문서의 테이블 데이터로 학습한 모델인 KoTaBERT, 공개된 생성형 언어모델인 GPT-3.5, GPT-4.0과 함께 자체 개발한 EXAONE 1.7B 모델을 비교 실험하였다. 표 4와 표 5는 각각 KorWikiTQ, KorQuad2Table 데이터 셋을 사용한 결과이다. EXAONE 1.7B 모델은 각 데이터셋의 학습셋으로 파인튜닝하였다. EM과 F1은 KorQuad 데이터셋 평가를 위한 평가 스크립트를 일부 변경하여 사용했다.

KorWikiTQ 데이터셋에 대해 GPT-4모델의 경우 세 개의 질문/정답 쌍을 보여준 few-shot 모델이 기존의 사전학습 모델인

[1]https://github.com/LG-NLP/KorWikiTableQuestions
[2]https://korquad.github.io/
[3]https://platform.openai.com/docs/guides/gpt

표 4. 모델별 성능 비교(KorWikiTQ 데이터셋)

Model	KoTaBERT		GPT-3.5 Zero-shot		GPT-3.5 few-shot		GPT-4 few-shot		EXAONE 1.7B	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
level 1	89.6	93.1	47.15	69.30	66.06	77.99	87.55	95.29	96.82	97.14
level 2	89.1	92.3	48.06	71.26	73.58	86.44	87.24	95.46	94.42	96.29
level 3	86.1	89.4	52.20	70.87	61.78	76.43	83.43	92.92	91.67	94.20
level 4	81.7	85.5	45.12	68.71	64.10	76.96	89.14	95.49	93.26	95.37
level 5	67.8	70.7	42.62	59.20	51.98	65.44	84.17	88.20	82.44	87.57
Avg.	87.2	91.2	47.11	68.56	64.94	78.60	86.61	94.01	92.49	94.81

표 5. 모델별 성능 비교 (KorQuad2Table 데이터셋)

Model	KoTaBERT		GPT-3.5 zero-shot		GPT-3.5 few-shot		GPT-4 few-shot		EXAONE 1.7B	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
	69.1	78.4	56.60	76.55	70.12	81.90	84.07	92.29	67.99	78.65

KoTaBERT와 유사한 성능을 보여주었으며 EXAONE 1.7B 모델은 GPT-4에 비해 작은 크기의 모델이지만 파인튜닝을 통해 가장 좋은 성능을 보여줄 수 있었다. KorQuad2Table 데이터셋에 대해서는 KorWikiTQ 데이터셋과 달리 GPT-4 few-shot 모델이 가장 좋은 성능을 보여주었다. EXAONE 1.7B 모델의 성능이 떨어진 이유는 KorWikiTQ와 비교해 학습 데이터가 1/15 수준으로 적어 EXAONE 모델을 파인튜닝하기에 부족했던 것이 원인으로 분석된다.

4.4 의미적 일치를 반영한 평가

3장에서 언급한 것처럼 테이블 질의응답에서 생성형 언어모델을 사용해서 답을 예측할 경우에는 의미적으로는 일치하는 답을 찾았지만 평가방법에 의해 오답으로 분류되는 경우가 있다. 정답과 형태가 일치하지 않더라도 정답과 동일한 의미를 갖는 답이라면 이를 정답으로 측정하는 평가 방법이 필요하다.

기존의 EM(Exact Matching) 평가 방법은 예측값이 정답과 정확히 일치하는 경우만을 정답으로 계산한다. 우리가 제안하는 SM (Semantic Matching)은 기존의 EM 평가 조건을 완화하여 '예측값이 정답 문자열을 모두 포함하는 경우'를 정답으로 계산하도록 변경한 평가 방법이다. 앞서의 실험 결과를 SM을 이용해 평가하고 표 6에 비교하였다.

EM 측정에서는 오답으로 구분되었지만 SM에서는 정답으로 올바로 분류된 예측값(정답을 포함하고 있는 예측값)의 유형을 알아보고, 실제는 오답이지만 SM에 의해 잘못 정답으로 분류된 값(False positive)이 얼마나 포함되어 있는지를 확인해 보았다

표 6. EM과 SM 결과 비교

	KorWikiTQ		KorQuad2Table	
	EM	SM	EM	SM
KoTaBERT	87.2	87.2	69.1	69.1
GPT-3.5 0-shot	47.11	69.70	59.60	75.19
GPT-3.5 few-shot	64.94	71.59	70.12	82.03
GPT-4	86.61	89.31	84.07	91.15
EXAONE 1.7B	92.49	92.76	67.99	71.26

표 7. SM 유형 분석

유형	개수	%	Description
E1	1,269	84.21	질문의 내용을 답에 반복
E2	42	2.79	정답에 단위를 추가
E3	52	3.46	정답에 추가정보를 제공
E4	137	9.10	정답임을 표시하는 설명 추가
E5	7	0.47	다른 내용을 포함
Total	1,507	100	

[표 7]. 실험을 위해 KorWikiTQ 데이터셋에 대한 GPT-3.5 모델 결과 중 SM에 의해 추가로 정답으로 분류된 3,014개의 결과 중 50%를 샘플링하여 분석하였다. 모델의 예측값이 정답과 정확히 일치하지는 않지만 정답을 포함하는 경우를 총 다섯 개의

표 8. SM평가에 의해 정답으로 추가된 예측 값 유형 분석 예

type	question	ground truth	prediction
E1	엘로디 융이 2014년에 출연한 스틸의 원제는 뭔가요?	still	스틸의 원제는 still 입니다
	해릉왕의 후궁 중에서 신비의 이름은 뭐야?	소씨 蕭氏	신비의 이름은 소씨 蕭氏 이다
E2	원주시 귀래면의 인구는 총 몇 명이에요?	2194	2194명
	도네강 유역 주젠 지호의 면적은 얼마인가요?	118	118 km²
E3	콘스탄스 우가 2011년에 출연한 드라마는 뭔가요?	토치우드	토치우드 torchwood
	41기 국수전 도전 4국의 대국은 언제입니까?	11월 24일	1997년 11월 24일
	이성재가 최초로 받은 상은 뭔가요?	신인상	1998년 kbs 연기대상 신인상
E4	유니버시아드 축구 여성부 경기에서 메달 순위 3위 한 나라가 어디예요?	중국	중국이에요
	2019년 3월 K리그 이달의 선수는 누구인가요?	세징야	답변 세징야
	비니 펠드스틴이 제일 처음 출연한 영화는 무엇입니까?	팬 걸	답 팬 걸
E5	라디오 오늘 같은 밤에서 슬리피가 2016년 08월 22일 월요일부터 08월 24일 수요일까지 진행한 코너는 무엇입니까?	쇼 미 더 라임	커너 쇼 미 더 라임
	FC 서울 코칭스태프 6대 감독은 누구인가요?	이장수	고재욱 조영증 박병주 조광래 이장수 귀네슈

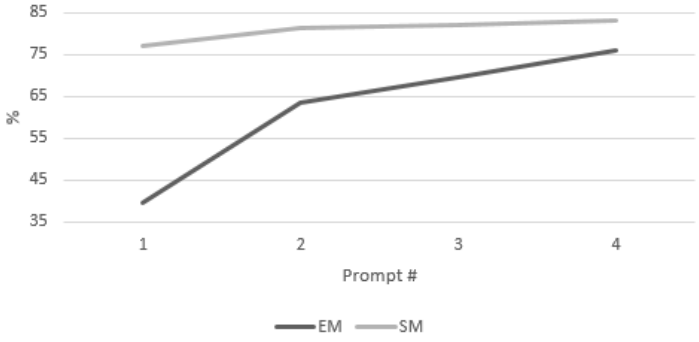

그림 1. 프롬프트 변경에 따른 EM, SM 변화

유형(E1–E5)으로 구분하였다. 표 8에서 각 유형의 예제를 확인할 수 있다. 가장 많은 유형(E1)은 답을 출력하면서 '질문의 내용을 반복하여 답을 명확히 하는 경우'이다. 이 때 반복되는 내용은 답을 구체화하거나 추가정보를 제공하지는 않고, 문장 형식으로 완성시켜 준다. 그 다음으로 많은 유형(E4)은 답 앞에 정답을 표시하는 '답변','정답' 등의 단어를 출력한 후 답을 표시하는 형식이다. 다음으로 E2 유형은 실제 정답에는 포함되지 않은 '단위'를 붙여 답을 구체화하는 것으로 주로 숫자형 답에 적용된다. E3 유형은 정답을 구체화하는 정보가 추가된 경우이다. E5는 정답이 아니지만 SM평가 방법에 의해 정답으로 잘못 분류된 경우에 해당한다.

E1-E4 유형의 결과는 정답과 정확히 일치하지는 않지만 정답으로 간주해도 무방하다. 전체 1,507개의 결과 중 SM평가에 의해 잘못 분류된 경우인 E5는 전체의 0.46%에 불과해 SM 평가방법이 단순하지만 생성형 결과의 품질을 측정하는 방법으로 적절함을 보여주었다.

앞서 4.2절에서는 예측값이 정답과 가깝게 출력되도록 프롬프트를 변경하면서 EM 변화를 살펴봤었다. 이 실험에 SM을 적용해보면 그림 1처럼 EM의 변화에 비해 SM의 변화가 작은 것을 알 수 있다. 즉, EM이 낮은 프롬프트에서도 의미상 정답이라고 할 수 있는 예측값이 출력되었으며, 프롬프트의 변화가 출력값의 형태는 변화시켰을지라도 의미상 정답의 출력에는 큰 영향이 없었다고 할 수 있다.

5. 결론

본 연구에서는 생성형 언어모델을 이용해 테이블 질의응답 성능을 살펴보고 생성형 언어모델의 질의 프롬프트에 따라 성능이 크게 변할 수 있음을 알아보았다. GPT-4 모델의 경우 few-shot임에도 불구하고 사전학습된 BERT 기반 모델과 비슷한 성능을 보였다. 자체개발한 EXAONE 1.7B 모델은 파인튜닝을 통해 GPT-4 보다 우수한 성능을 보였으나 학습데이터가 부족한 데이터셋에 대해서는 성능의 큰 차이점을 보이지 못했다. 정답 문장을 생성하려 하는 생성형 언어모델의 특성상 짧고 간결한 테이블 질의응답 데이터의 정답과 잘 맞지 않기 때문에 의미상 일치하는 정답을 출력했음에도 정답이 아닌 것으로 측정되는 경우가 많다. 그렇기 때문에 이를 보완할 수 있는 Semantic Matching 방법을 제안하였다. GPT-3, GPT-4와 같은 초거대 언어모델에 적절한 프롬프트가 주어진다면 SM 평가방법으로 성능을 측정하는 것이 적절함을 알 수 있었다. 향후에는 큰 사이즈의 EXAONE 모델을 적용해 few-shot 결

과를 확인하고, 패턴만을 고려하는 현재의 SM 평가를 개선할 계획이다

참고문헌

[1] N. Jin, J. Siebert, D. Li, and Q. Chen, "A survey on table question answering: recent advances," *China Conference on Knowledge Graph and Semantic Computing*, pp. 174–186, 2022.

[2] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua, "Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance," *arXiv preprint arXiv:2105.07624*, 2021.

[3] P. Dasigi, M. Gardner, S. Murty, L. Zettlemoyer, and E. Hovy, "Iterative search for weakly supervised semantic parsing," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2669–2680, 2019.

[4] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. M. Eisenschlos, "Tapas: Weakly supervised table parsing via pre-training," *arXiv preprint arXiv:2004.02349*, 2020.

[5] W. Chen, M.-W. Chang, E. Schlinger, W. Wang, and W. W. Cohen, "Open question answering over tables and text," *arXiv preprint arXiv:2010.10439*, 2020.

[6] W. Hwang, J. Yim, S. Park, and M. Seo, "A comprehensive exploration on wikisql with table-aware word contextualization," *arXiv preprint arXiv:1902.01069*, 2019.

[7] P. Yin, G. Neubig, W.-t. Yih, and S. Riedel, "Tabert: Pretraining for joint understanding of textual and tabular data," *arXiv preprint arXiv:2005.08314*, 2020.

[8] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang, "Tabfact: A large-scale dataset for table-based fact verification," *arXiv preprint arXiv:1909.02164*, 2019.

[9] 심묘섭, 전창욱, 최주영, 김현, 장한솔, and 민경구, "표 질의응답을 위한 언어 모델 학습 및 데이터 구축," *33회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 335–339, 2021.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[11] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[12] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

[13] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman *et al.*, "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task," *arXiv preprint arXiv:1809.08887*, 2018.

[14] Q. Lyu, K. Chakrabarti, S. Hathi, S. Kundu, J. Zhang, and Z. Chen, "Hybrid ranking network for text-to-sql," *arXiv preprint arXiv:2008.04759*, 2020.

[15] J. Li, B. Hui, R. Cheng, B. Qin, C. Ma, N. Huo, F. Huang, W. Du, L. Si, and Y. Li, "Graphix-t5: Mixing pre-trained transformers with graph-aware layers for text-to-sql parsing," *arXiv preprint arXiv:2301.07507*, 2023.

[16] T. Scholak, N. Schucher, and D. Bahdanau, "Picard: Parsing incrementally for constrained auto-regressive decoding from language models," *arXiv preprint arXiv:2109.05093*, 2021.

[17] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 874–880, 2021.

[18] B. Oguz, X. Chen, V. Karpukhin, S. Peshterliev, D. Okhonko, M. Schlichtkrull, S. Gupta, Y. Mehdad, and S. Yih, "Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering," *arXiv preprint arXiv:2012.14610*, 2020.

[19] C. Jun, J. Choi, M. Sim, H. Kim, H. Jang, and K. Min, "Korean-specific dataset for table question answering," *arXiv preprint arXiv:2201.06223*, 2022.