Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 6, pp. 502-513, December 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.6.502


# KOreaPAS: TAPAS 기반의 한국어 특화 표 질의응답 모델

오수지 · 고유경 · 이유경 · 강필성†

고려대학교 산업경영공학과

# KOreaPAS: TAPAS based Korean-Specific Table Question Answering Model

Suzie Oh · Yookyung Kho · Yukyung Lee · Pilsung Kang

Department of Industrial & Management Engineering, Korea University

Table Question Answering (QA) aims to answer questions based on semi-structured tables. Unlike text data, tables possess a unique two-dimensional structure, driving the exploration of specialized learning approaches to enhance language models' understanding of tables. However, while Table QA research is advancing rapidly in English, its development in Korean is still in its early stages. To mitigate this gap, we present KOreaPAS, specifically designed for Korean Table QA tasks. KOreaPAS is based on TAPAS's architecture, and its learning process consists of two stages: pre-training and fine-tuning. In the publicly available Korean tabular dataset for pre-training language models, approximately 36.5% instances lack text information related to tables, and it can potentially hinder the models' learning of various correlations between text and the table during pre-training. To address the issue, we introduce a table-text mapping method that retrieves the most relevant text for the table from Wikipedia pages. Further, we propose a multi-granularity fine-tuning strategy that utilizes the three granularities of the table structure for both training and inference. Experimental results robustly confirm the effectiveness of the proposed approaches in enhancing the comprehension abilities of language models towards questions over tables. Specifically, KOreaPAS demonstrated the highest performance among currently published benchmark models in tests conducted on two Korean Table QA datasets, thus establishing a new standard in Korean Table QA tasks.

***Keywords:*** Natural Language Processing, Table Question Answering, Question Answering, TAPAS

## 1. 서 론

질의응답(Question Answering)은 주어진 데이터를 기반으로 사용자의 질문에 대한 정답을 도출하는 과업으로, 사용자가 대규모의 데이터로부터 원하는 질문에 대한 정답을 쉽고 효율적으로 얻을 수 있도록 보조하는 것을 주목적으로 한다(Pandya and Bhatt, 2021; Wang, 2022). 가장 일반적인 형태의 질의응답은 텍스트 데이터를 대상으로 수행되지만, 현실 세계에서 접할 수 있는 문서에는 표, 차트, 이미지 등 다양한 형태의 데이터가 포함되는 만큼 각 데이터의 특성에 맞는 질의응답 모델 또한 활발하게 연구되고 있다(Pasupat and Liang, 2015; Agrawal *et al.*, 2015; Kahou *et al.*, 2017; Kafle *et al.*, 2018; Methani *et al.*, 2019; Masry *et al.*, 2022).

텍스트 이외의 데이터를 다루는 대표적인 질의응답 과업으로는 표 데이터에 대한 질문에 정답을 반환하는 표 질의응답(Table Question Answering)이 있다(Jin *et al.*, 2022). 표는 열과

이 논문은 2023년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2022R1A2C2005455)의 성과물임. 또한, 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00471, 모델링 & 최적화 기반 오류-free 정보인프라 자율제어 기술 개발)

† 연락저자 : 강필성 교수, 02841, 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel: 02-3290-3383, Fax: 02-929-5888,
E-mail: pilsung_kang@korea.ac.kr
2023년 7월 5일 접수; 2023년 8월 12일 수정본 접수; 2023년 9월 1일 게재 확정.

행의 2차원 구조를 가지는 데이터로, 웹페이지, 스프레드시트, PDF, HTML 등의 다양한 문서와 데이터베이스 시스템에서 데이터를 구조화하고 저장하기 위해 널리 사용된다(Dong *et al.*, 2022; Wang *et al.*, 2021). 특히, 표는 복잡한 자료를 구조화하여 사용자가 데이터를 쉽게 읽고 이해하는 데 도움을 주며, 표로 구조화된 데이터를 이용하면 보다 용이하게 데이터를 분석하고, 통계적 특징을 파악할 수 있다. 따라서, 복잡한 수치 데이터나 통계 자료를 주로 다루는 금융, 행정, 과학 등의 분야에서 표 질의응답 모델의 필요성이 더욱 부각되고 있다(Jin *et al.*, 2022; Chen *et al.*, 2021, 2022; Zhao *et al.*, 2022; Zhu *et al.*, 2021).

GPT(Radford *et al.*, 2018)와 BERT(Devlin *et al.*, 2019)를 시작으로 정답 쌍이 존재하지 않는 대규모의 텍스트 데이터(unlabeled large corpus)로 사전 학습된 언어 모델이 다양한 자연어 처리 과업에서 우수한 성능을 나타낼 수 있음이 여러 연구에서 입증되었다(Lewis *et al.*, 2019; Liu *et al.*, 2019; Raffel *et al.*, 2019; He *et al.*, 2020). 하지만 사전 학습 언어 모델이 사전 학습 시 활용한 텍스트 데이터는 식별 가능한 구조가 없는 비정형 데이터인 반면 표 데이터는 2차원의 구조를 가지는 반정형 데이터이므로 현존하는 사전 학습 언어 모델을 표 질의응답 학습에 그대로 적용하는 것은 적합하지 않다(Dong *et al.*, 2022; Liu *et al.*, 2021). 따라서 사전 학습 언어 모델을 이용해 표 내외의 텍스트에 대한 이해 능력은 유지하면서, 표가 가지는 구조적 정보를 추가로 학습하기 위한 다양한 방법이 시도되어 왔다. 대표적인 방법으로는 열과 행을 구분하는 구분자가 추가된 새로운 입력 형태를 활용하는 방법(Glass *et al.*, 2021; Iida *et al.*, 2021; Liu *et al.*, 2021; Yin *et al.*, 2020), 표가 가지는 정보를 나타내기 위한 새로운 임베딩을 추가하는 방법(Deng *et al.*, 2020; Eisenschlos *et al.*, 2021; Herzig *et al.*, 2020; Wang *et al.*, 2021; Yang *et al.*, 2022) 등이 있다.

표 질의응답 과업을 효과적으로 수행하기 위해서는 표 학습에 적합한 입력 형태나 모델 구조를 기반으로 대규모의 표 데이터를 이용한 추가적인 학습이 필수적이다. 그러나 현재까지 공개된 대규모의 표 데이터(Herzig *et al.*, 2020; Liu *et al.*, 2021; Wang *et al.*, 2021; Yin *et al.*, 2020; Yu *et al.*, 2020)는 대부분 영어로 구성되며, 한국어의 경우 한국어에 특화된 표 데이터 및 표 질의응답 데이터의 부재로 인해 표 질의응답 관련 연구가 제한적으로 진행되어 왔다. 하지만 최근 약 120만 개의 한국어 표 데이터(KorWikiTabular)와 약 7만 개의 한국어 표 질의응답 데이터(KorWikiTQ)로 구성된 대규모의 데이터셋(Jun *et al.*, 2022)이 공개되며 한국어에 특화된 표 질의응답 연구가 가능해졌다. 이에 본 연구에서는 표 질의응답 과업을 위한 최초의 대규모 한국어 데이터셋에 대해 학습을 수행한 한국어 특화 표 질의응답 모델인 KOreaPAS를 제안한다. KOreaPAS를 학습하는 과정은 대량의 표 데이터로 사전 학습을 수행하는 과정과 표 질의응답 데이터에 대해 미세 조정(fine-tuning)을 하는 두 가지 단계로 구성된다. 이 과정에서 본 연구는 공개된 한국어 표 데이터셋이 사전 학습 단계에서 가지는 문제점을 보완하는 과정과 표에 존재하는 다양한 granularity를 활용한 새로운 미세 조정 방식에 집중하였다. KOreaPAS를 이용해 두 가지 한국어 표 질의응답 데이터셋에 대해 평가를 진행한 결과 모든 데이터셋에서 우수한 성능을 달성했으며, 이를 통해 본 논문에서 제안한 학습 방식이 사전 학습 언어 모델의 한국어 표 질의응답 능력을 향상시키는 데 크게 기여함을 확인할 수 있다. 본 연구는 최초로 현재까지 공개된 한국어 표 데이터셋 및 한국어 표 질의응답 데이터셋을 이용해 학습을 수행했으며, 향후 진행될 한국어 표 질의응답 연구에서 참조할 수 있는 비교 모델을 제안했다는 점에서 의의가 있다.

본 논문의 구성은 다음과 같다. 먼저 제2장에서는 본 논문에서 제안하는 모델의 기반이 되는 TAPAS(Herzig *et al.*, 2020)를 포함하여 표 데이터에 특화된 사전 학습 연구에 관해 서술한다. 제3장에서는 본 연구에서 모델을 학습하기 위해 사용한 사전 학습 방식과 미세 조정 방식에 대해 자세히 소개하며, 제4장에서는 두 가지 한국어 표 질의응답 데이터셋에 대한 실험 결과를 서술한다. 마지막으로 제5장에서는 본 연구에서 제안한 모델의 의의와 한계를 설명한 후 이를 개선하기 위한 향후 연구 방향을 제시한다.

## 2. 관련 연구

### 2.1 표 특화 사전 학습

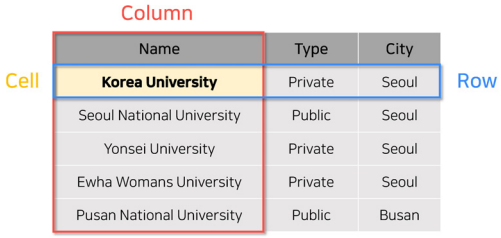


**Figure 1.** An Example of Various Granularities in the Table. The Table Includes Different Granularity Evidence Consisting of Coarse-grained Column and Row and Fine-grained Cell

표에 특화된 사전학습 방법론은 모델의 구조에 따라 크게 두 가지로 구분할 수 있다. 트랜스포머(Vaswani *et al.*, 2017)의 인코더 구조를 기반으로 한 대표적인 모델로는 Herzig *et al.*(2020), Eisenscholos *et al.*(2021), Yang *et al.*(2022), Yin *et al.*(2020), Wang *et al.*(2021), Deng *et al.*(2020)가 있으며, 트랜스포머의 인코더-디코더 구조를 기반으로 한 대표적인 모델로는 Liu *et al.*(2022), Xie *et al.*(2022)가 있다. 트랜스포머의 인코더 구조를 기반으로 한 모델 중 Yin *et al.*(2020), Wang *et al.*(2021), Eisensholos *et al.*(2021), Yang *et al.*(2022) 모두 BERT의 Masked Language Modeling(MLM) 사전 학습 방식을 응용하여 텍스트 또는 표에 속하는 토큰이 전체 문맥상에서 가지는 표현(Contextual Representations)을 학습한다. 이때, Yin *et al.*(2020)

이외의 방법론들은 BERT에서 사용한 WordPiece tokenizer의 토큰 단위 마스킹이 아닌 셀 단위의 마스킹을 사용하는 점에서 차이를 보인다.

표는 일반적인 텍스트와 달리 2차원의 구조를 가지는 만큼 데이터 내에 다양한 granularity가 존재한다. 데이터의 granularity란 데이터가 세분화된 정도를 의미하며, 그 정도에 따라 coarse-grained와 fine-grained로 구분된다. 표 데이터의 경우 <Figure 1>과 같이 가장 세분화된 단위인 셀을 fine-grained로 구분 지을 수 있으며, 셀을 포함하는 열과 행을 coarse-grained로 구분 지을 수 있다. 이처럼 표가 가지는 다양한 granularity의 표현을 학습하기 위해 Yin *et al.*(2020)과 Wang *et al.*(2021)은 MLM 외에 추가적인 사전 학습 과업을 제안하였다. Yin *et al.*(2020)은 셀 수준의 표현을 학습하기 위해 마스킹된 셀을 예측하는 Cell Value Recovery(CVR)와 열 수준의 표현을 학습하기 위해 마스킹된 열의 이름과 데이터 유형을 예측하는 Masked Column Prediction(MCP)을 사전 학습 과정에 추가하였다. 반면에, Wang *et al.*(2021)은 셀 수준의 표현을 학습하기 위해 랜덤하게 마스킹된 셀에 대응되는 원본 문자열을 반환하는 Cell-level Cloze(CLC), 표 수준의 표현을 학습하기 위해 입력으로 주어진 텍스트와 표의 관련 여부를 이진 분류하는 Table context retrieval(TCR)을 추가하였다. 두 연구에선 이러한 추가적인 사전 학습 방식이 표 질의응답, Text-to-SQL, 셀 유형 분류(Cell Type Classification), 표 유형 분류(Table Type Classification) 등 표와 관련된 여러 후속 과업(Down-stream Task)에서 성능 향상에 기여함을 보이며 표가 가지는 다양한 granularity를 학습에 응용하는 것의 중요성을 확인했다.

최근 인코더-디코더 구조를 사용하여 표 질의응답 과업을 Text-to-Text 방식으로 학습하는 모델이 WikiSQL(Zhong *et al.*, 2018), WikiTQ(Pasupat and Liang, 2015), SQA(Iyyer *et al.*, 2017) 등의 대표적인 표 질의응답 벤치마크 데이터셋에서 좋은 성능을 기록하고 있다(Liu *et al.*, 2021; Xie *et al.*, 2022). Liu *et al.*(2022)은 사전 학습된 BART(Lewis *et al.*, 2019)를 기반으로 표와 SQL 쿼리를 입력으로 받고 정답을 생성하는 방식으로 사전 학습을 진행한 후, 질문 텍스트와 표를 입력으로 받고 정답을 생성하는 방식으로 미세 조정을 수행하였다. 또한, Xie *et al.*(2022)은 T5(Raffel *et al.*, 2020)를 이용해 총 21개의 자연어 처리 과업을 Text-to-Text 문제로 푸는 모델을 제안하였다. 이러한 방법론은 표에 특화된 구조나 사전 학습 방식을 사용하지 않으므로 학습이 용이하며, 입력과 출력의 구성을 변경하여 여러 과업에 쉽게 적용할 수 있다는 장점이 있다. 하지만 Liu *et al.*(2022)의 경우 SQL 쿼리-정답 쌍으로 구성된 추가적인 데이터를 필요로 하며, Xie *et al.*(2022)은 사전 학습을 진행한 모델에 비해 낮은 성능을 보인다는 한계를 가진다.

본 연구는 추가적인 데이터 구축 과정이 필요 없으면서도 영어 표 질의응답 벤치마크 데이터셋(Pasupat and Liang, 2015; Iyyer *et al.*, 2017)에 대해 좋은 성능을 보인 TAPAS 구조를 학습에 활용하였다. 또한, 표가 가지는 다양한 granularity를 사전 학습에 반영한 연구(Wang *et al.*, 2021; Yin *et al.*, 2020)를 참고하여 TAPAS를 기반으로 표 내에 존재하는 다양한 granularity를 학습과 추론에 활용한 미세 조정 방식을 제안하였다.

### 2.2 TAPAS

TAPAS는 대규모의 표 데이터에 대해 사전 학습과 미세 조정을 수행한 대표적인 표 질의응답 모델이다. TAPAS는 BERT 구조를 기반으로 하며, BERT의 기존 입력 임베딩에 표가 가지는 특수한 정보를 학습하기 위한 추가적인 임베딩을 도입하였다. 새롭게 추가된 임베딩에는 특정 토큰이 속한 행과 열의 인덱스를 나타내는 행 임베딩, 열 임베딩과 동일한 열에 속하는 셀들 간의 상대적 순위 정보를 나타내는 순위 임베딩이 해당한다. TAPAS는 위키피디아로부터 약 620만 개의 표와 약 2,130만 개의 텍스트를 수집하여 대규모의 표-텍스트 데이터 쌍을 구축하였으며, 이를 이용해 표 데이터에 특화된 사전 학습을 진행하였다. TAPAS의 사전 학습 과정은 BERT와 유사하나, i) 표와 텍스트의 결합 표현(joint representation)을 학습할 수 있도록 텍스트와 선형화된 표를 함께 입력으로 구성하는 점과 ii) MLM 학습 시 텍스트 부분에 대해서는 단어 단위로, 표 부분에 대해서는 셀 단위로 마스킹을 수행한다는 점에서 미세한 차이를 보인다.

미세 조정 단계에서는 정답을 도출하는 과정에서 집계 연산의 필요 유무에 따라 문제의 유형을 세 가지로 분류하고, 유형별로 손실 함수(loss function)를 상이하게 구성한다. 이 중 집계 연산이 필요하지 않고 표에 존재하는 특정 셀이 바로 정답이 되는 문제는 셀 선택(Cell Selection) 유형으로 분류된다. 현재까지 공개된 한국어 표 질의응답 데이터셋의 경우, 대부분 정답이 표 내의 특정 셀에 대응되는 문제로 구성되어 있으므로 본 논문은 미세 조정 단계에서 셀 선택 유형의 문제에 한하여 학습과 성능 평가를 수행하였다.

본 연구에서 새롭게 제안한 미세 조정 방식은 TAPAS의 미세 조정 방식과 유사하지만 두 가지 측면에서 차이를 보인다. 첫 번째 차이점은 TAPAS의 경우 셀 선택 유형에 해당하는 문제를 학습하기 위한 손실 함수가 정답 셀을 선택하기 위한 손실 함수와 정답 셀을 포함한 열을 선택하기 위한 손실 함수로 구성되는 반면 본 연구에선 정답 셀을 포함한 행을 선택하기 위한 손실 함수도 학습 과정에 추가하여 표가 가지는 모든 granularity를 학습에 활용한다는 점이다. 두 번째 차이점은 추론 과정에 있다. TAPAS는 표가 가지는 두 가지의 granularity를 학습에 활용하는 것과 별개로 추론 시에는 셀에 대한 점수만을 사용한다. 그러나 본 연구에선 각기 다른 granularity별로 산출된 점수를 모두 추론 과정에 활용하였다. 이에 대한 상세한 설명은 각각 3.2장과 3.3장에 포함된다.

### 2.3 한국어 표 질의응답

본 연구 이전까지 진행된 한국어 표 질의응답 연구는 한국

어 표 질의응답 데이터셋의 부재로 인해 연구 수행 기관에서 자체적으로 제작한 데이터셋을 사용하거나(Park *et al.*, 2018), 위키피디아 문서를 기반으로 제작된 질의응답 데이터셋에서 정답이 표 내부에 존재하는 데이터를 선택적으로 추출하여 학습과 평가를 진행하였다(Cho *et al.*, 2020). 이와 같은 연구들은 한국어 표 질의응답 과업을 위해 공개된 대규모의 데이터가 없는 상황에서 직접 데이터를 구축하고, 이를 바탕으로 한국어 표 질의응답 과업에 특화된 모델을 제안하였다는 점에서 중요한 의의가 있다. 그러나 두 연구 모두 모델의 학습 및 평가에 사용한 데이터를 공개하지 않아 연구 간 정량적 성능 비교가 어렵다는 한계를 가진다. 이후, Jun *et al.*(2022)과 AI Hub(https://www.aihub.or.kr/)를 통해 대규모의 한국어 표 질의응답 데이터셋이 공개되며 한국어 표 질의응답 연구에서도 공정한 비교가 가능해졌다. 그럼에도 불구하고 두 데이터셋을 모두 활용한 한국어 표 질의응답 연구는 현재까지 진행된 바 없으며, 해당 데이터셋으로 향후 진행될 연구를 위한 대표적인 연구가 필요한 실정이다. 이에 따라 본 연구에서는 KOreaPAS를 활용해 최초로 두 가지 데이터셋에 대한 학습 및 성능 평가를 진행하였다

표 질의응답 과업의 입력은 질문과 표로 구성되므로 사전 학습 단계에서 표와 텍스트의 결합 표현을 학습하는 것이 중요하다. 그러나 공개된 한국어 표 데이터셋의 약 36.5%에 해당하는 데이터에서 표 관련 텍스트 정보가 누락된 것이 확인되었다. 따라서 본 논문에선 정보 검색(Information Retrieval) 기술을 활용하여 사전 학습 데이터를 보강하는 방법을 제안한다. 해당 방법론은 한국어 표 데이터셋이 사전 학습 단계에서 가지는 문제점을 보완하며, 추후 새로운 한국어 표 데이터셋 구축 시에도 활용 가능하다는 점에서 중요한 의의를 가진다. 사전 학습 데이터 보강 절차에 대한 상세한 설명은 3.1.2장에 포함된다.

## 3. 제안 방법론

본 연구에서는 TAPAS 구조를 기반으로 한국어 표에 특화된 언어 모델인 KOreaPAS를 제안한다. KOreaPAS의 전체 학습 프레임워크는 <Figure 2>와 같이 표-텍스트 매핑(Table-Text Mapping) 과정을 통해 보강된 사전 학습 데이터를 이용한 사전 학습 단계와 표의 다양한 granularity를 고려한 미세 조정 단계로 구성된다. 특히, 사전 학습 단계에서 Kim *et al.*(2021)과 Lee *et al.*(2021)의 연구를 참조하여, 대규모의 영어 데이터로 사전 학습된 언어 모델을 한국어 데이터로 추가 학습하는 언어 간 전이 학습(Cross-lingual Transfer Learning) 방식을 적용하였다. 이를 위해 KOreaPAS의 인코더를 Huggingface에 공개된 TAPAS의 인코더 파라미터로 초기화하여 사전 학습을 수행하였다(https://huggingface.co/google/tapas-base). 해당 모델이 사전 학습에 활용한 영어 표 데이터는 약 620만 개로, 본 연구에서 사전 학습에 활용한 한국어 표 데이터 개수가 약 120만 개임을 고려하면 언어 간 전이 학습을 수행하기 적절한 모델로 판단된다.

### 3.1 사전 학습

(1) 사전 학습 데이터 구성

표에 특화된 사전 학습을 수행하기 위해선 대규모의 표 데이터가 필요하다. 따라서, 본 연구는 사전 학습에 사용할 데이터를 구축하기 위해 KorWikiTabular(Jun *et al.*, 2022)와 KorQuAD 2.0(Kim *et al.*, 2020)을 활용하였다. KorWikiTabular는 위키피디아 문서를 기반으로 제작된 최초의 대규모 한국어 표 데이터셋으로 위키피디아 페이지 내에 존재하는 Infobox 및 Table 정보를 추출하여 표 데이터를 구축하였으며, 표 관련 텍스트, 표의 제목 및 캡션 등의 정보를 함께 제공한다. 추가적으로, 본 연구는 더욱 다

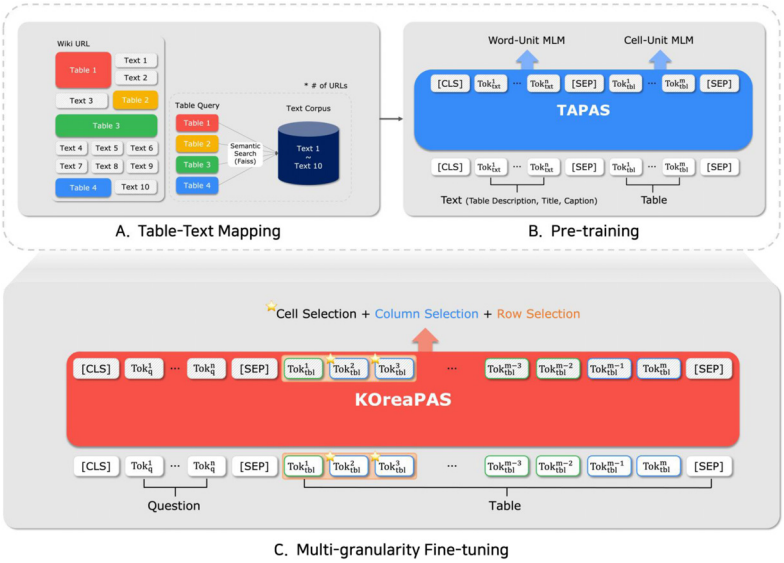


**Figure 2.** Overall Training Framework

**Table 1.** An Example of KorWikiTabular Dataset. In the Table, it can be Observed that the 'Description' field is Empty

| Title | 이쿠지역 | | |
|---|---|---|---|
| URL | https://ko.wikipedia.org/w/index.php?title=이쿠지역 | | |
| Description | None | | |
| Caption | 이쿠지역_인접역 | | |
| Table | 아이노카제토야마 철도선 | 아이노카제토야마 철도선 | 아이노카제토야마 철도선 |
| | 구로베도야마 방면 | - | 니시뉴젠도마리 방면 |

양한 위키피디아 문서로부터 표 데이터를 확장하기 위해 KorQuAD 2.0을 활용하였다. KorQuAD 2.0은 대표적인 한국어 질의응답 데이터셋으로, 질문에 대한 답을 찾기 위한 context가 표, 리스트 등의 데이터를 포함한 하나의 위키피디아 문서로 구성되어 사전 학습을 위한 추가적인 표 데이터를 추출하는 것이 가능하다. 최종적으로 KorQuAD 2.0에 포함된 위키피디아 문서 URL 42,352개 중 KorWikiTabular에 등장한 URL을 제외한 10,744개의 URL으로부터 9,850개의 표 데이터를 추가 수집하여 사전 학습 데이터를 구성하였다.

(2) 사전 학습 데이터 보강 절차

대표적인 표 사전학습 연구들은(Herzig *et al.*, 2020; Jun *et al.*, 2022; Yin *et al.*, 2020) 표와 텍스트의 결합 표현을 학습하기 위해 표와 관련된 텍스트를 함께 입력으로 구성하여 사전 학습을 진행한다. 그러나 KorWikiTabular의 약 36.5%에 해당하는 데이터에서 <Table 1>과 같이 표와 관련된 텍스트를 나타내는 'Description' 항목이 비어 있는 것이 확인되었다.

이 경우, 표와 관련된 유일한 텍스트인 표의 제목 및 캡션만을 모델의 입력으로 활용하게 되므로 표와 텍스트의 결합 표현을 충분히 학습하기 어렵다는 한계를 가진다. KorQuAD 2.0에서 추가로 수집한 데이터 또한 표와 관련된 텍스트 정보가 부재하므로 동일한 한계를 지닌다. 따라서 본 연구에서는 <Figure 3>과 같이 표와 관련된 텍스트 정보를 매핑(mapping)하는 정보 검색 모듈을 설계하여 사전 학습 데이터를 보강하였다.

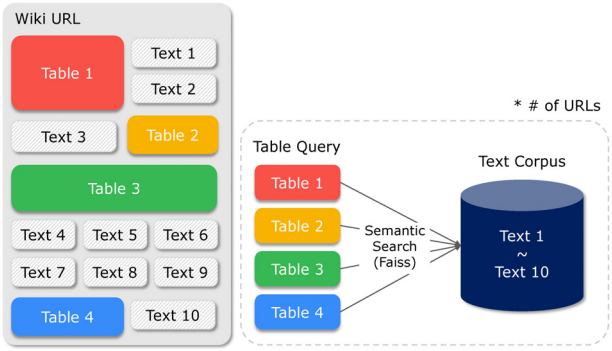


**Figure 3.** Table-Text Mapping Process

사전 학습 데이터를 보강하는 작업은 특정 표가 위치한 위키피디아 문서 내에서 표와 가장 관련 있는 텍스트를 찾아 기존 데이터에서 비어 있는 'Description' 항목을 보충하는 것을 목표로 한다. 표와 가장 관련 있는 텍스트를 반환하는 과정은 2단계로 구분지을 수 있으며, i) 표를 포함한 위키피디아 문서로부터 검색에 활용할 텍스트 코퍼스(corpus)를 구축하는 단계와 ii) 구축된 텍스트 코퍼스에서 표와 가장 유사한 텍스트를 찾는 단계로 구분된다. 첫 번째 단계에서는 HTML 형식의 위키피디아 문서로부터 한 문단의 시작과 끝을 나타내는  및  태그로 감싸진 부분의 텍스트를 추출하여 표와 관련된 텍스트 코퍼스를 구축한다. 이때, 표 데이터를 포함한 위키피디아 문서를 수집하기 위해 KorWikiTabular와 KorQuAD 2.0에 포함된 URL 정보를 활용하였다. 두 번째 단계에서는 FAISS 라이브러리(Johnson *et al.*, 2017)를 이용해 텍스트 코퍼스에서 특정 표와 의미적으로 가장 유사한 텍스트를 선별한다. 특정 표에 대응되는 여러 텍스트의 표현(representation) 중 표의 표현과 가장 높은 코사인 유사도를 보이는 상위 1개의 텍스트가 해당 표와 관련된 텍스트로 매핑되며, 표 및 텍스트의 표현을 구하기 위해서는 BERT를 기반으로 대규모의 한국어 데이터에 대해 사전 학습한 KLUE-BERT 모델을 사용하였다(https://huggingface.co/klue/bert-base). 이와 같은 2단계의 사전 학습 데이터 보강 절차를 통해 KorWikiTabular의 누락된 'Description' 항목을 적절한 텍스트로 채울 수 있으며, 추가적으로 수집한 KorQuAD 2.0 표 데이터에도 관련된 텍스트를 매핑할 수 있다.

(3) 사전 학습

KoreaPAS는 TAPAS의 모델 구조 및 사전 학습 전략을 활용하여 한국어 표 데이터에 특화된 모델을 구축하였다. 사전 학습 단계에서 모델의 입력은 식 (1)과 같이 구성된다.

$$\mathrm{x} = [\mathrm{CLS}]\mathrm{Tok}\ {}^{1}_{\mathrm{txt}}, \cdots, \mathrm{Tok}\ {}^{\mathrm{n}}_{\mathrm{txt}}[\mathrm{SEP}]\mathrm{Tok}\ {}^{1}_{\mathrm{tbl}}, \cdots, \mathit{Tok}\ {}^{m}_{tbl}[\mathit{SEP}] \tag{1}$$

식 (1)에서 $\mathrm{Tok}^{\mathrm{i}}_{\mathrm{txt}}$ 는 텍스트 시퀀스의 i번째 토큰을 나타내며, [SEP] 토큰을 기준으로 텍스트 시퀀스와 구분되는 $\mathrm{Tok}^{\mathrm{j}}_{\mathrm{tbl}}$ 는 표 시퀀스의 j번째 토큰을 의미한다. 텍스트 시퀀스는 한국어 표 데이터에 존재하는 'Caption' 정보와 'Description' 정보를 이어 붙여 구성하였으며, Caption 정보가 따로 존재하지 않는 KorQuAD 2.0의 경우 'Description' 정보로만 구성하였다. 표 시퀀스의 경우 표를 셀 단위로 분할한 후 차례대로 이어 붙

이는 방식으로 평면화하여 구성한다. 사전 학습은 Herzig *et al.* (2020)에서 진행한 방식과 동일하게 전체 시퀀스에서 15%에 해당하는 토큰을 [MASK] 토큰으로 대체하여 MLM을 수행하였으며, 이때 텍스트 시퀀스는 단어 단위로, 표 시퀀스는 셀 단위로 마스킹을 진행하였다. 마스킹 이후의 입력 시퀀스를 $x^{mask}$라 할 때, KOreaPAS는 식 (2)와 같이 BERT에서 제안한 MLM 손실 함수를 따라 학습이 진행된다.

$$L_{MLM} = \frac{1}{n}\sum_{i=1}^{n} -\log P([\mathrm{MASK}_i] = y_i \mid x^{mask}) \quad (2)$$

이때, $[\mathrm{MASK}_i]$와 $y_i$는 각각 입력 시퀀스의 i번째 마스크 토큰과 정답 토큰을 의미하며, n은 $x^{mask}$에서 마스킹된 전체 토큰의 개수를 의미한다.

### 3.2 Multi-Granularity Fine-Tuning

KOreaPAS의 미세 조정 단계에서 모델의 입력은 식 (3)과 같이 구성된다. 사전 학습 단계의 입력과 동일하게 질문을 나타내는 텍스트 시퀀스와 표 시퀀스가 차례대로 입력되며, [SEP] 토큰을 기준으로 구분된다. 모델의 입력 $x'$를 사전 학습이 완료된 KOreaPAS와 선형 계층(Linear Layer)에 차례대로 통과시켜 token logit $t \in R^{d_{input_len} \times hidden_dim}$를 얻게 되며, 식 (4)를 따른다.

$$x' = [\mathrm{CLS}]\mathrm{Tok}_q^1, \cdots, \mathrm{Tok}_q^n [\mathrm{SEP}] \mathrm{Tok}_{tbl}^1, \cdots, \mathrm{Tok}_{tbl}^m [\mathrm{SEP}] \quad (3)$$

$$t = \mathrm{Linear}(\mathrm{KOreaPAS}(x')) \quad (4)$$

본 논문에서 제안하는 미세 조정 방식은 표에 존재하는 세 가지 granularity를 모두 학습에 활용한다. 이에 따라 손실 함수는 총 세 가지의 개별 손실 함수로 구성되며(식 (9)), 정답에 해당하는 셀을 선택하기 위한 셀 선택 손실 함수 (식 (6)), 정답 셀을 포함한 열을 선택하기 위한 열 선택 손실 함수(식 (7)), 정답 셀을 포함한 행을 선택하기 위한 행 선택 손실 함수 (식 (8))가 이에 해당한다. 셀 선택 손실 함수(식 (6))는 특정 셀 c가 정답으로 선택될 확률 $p_{cell}^{(c)}$와 셀 단위의 정답 레이블 간의 교차 엔트로피(Cross-entropy)를 통해 계산되며, 열 선택 손실 함수(식 (7))는 특정 열 co가 정답 셀을 포함한 열로 선택될 확률 $p_{col}^{(co)}$와 열 단위의 정답 레이블, 행 선택 손실 함수(식 (8))는 특정 행 ro가 정답 셀을 포함한 행으로 선택될 확률 $p_{row}^{(ro)}$와 행 단위의 정답 레이블 간의 교차 엔트로피를 통해 계산된다. 최종적으로 세 가지 손실 함수를 모두 합한 전체 손실 함수(식 (9))를 통해 모델은 정답에 해당하는 셀과 정답 셀을 포함한 열 및 행에 대한 선택 확률이 커지도록 학습하게 된다. 아래의 수식에서 CE는 교차 엔트로피 손실 함수, 1은 지시 함수를 나타내며, $cell^*$, $col^*$, $row^*$ 는 각각 정답 셀, 정답 셀을 포함한 열 및 행을 나타낸다.

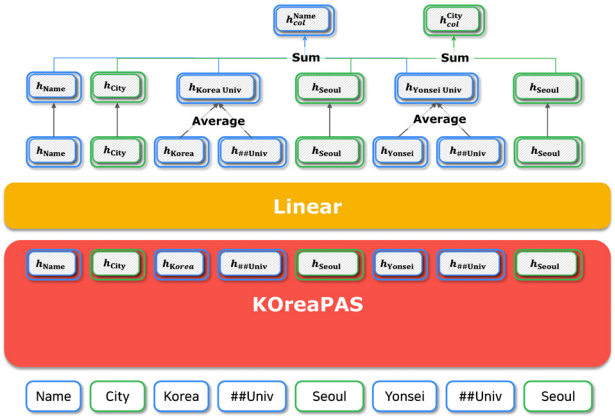


**Figure 4.** The process of calculating column logits. The final column logits can be passed through the softmax function to determine the probability of each column containing the answer

$$L_{cells} = \frac{1}{|\mathrm{Cells}|}\sum_{c \in \mathrm{Cells}} \mathrm{CE}(p_{cell}^{(c)}, 1_{c=cell^*}) \quad (6)$$

$$L_{columns} = \frac{1}{|\mathrm{Columns}|}\sum_{co \in \mathrm{Columns}} \mathrm{CE}(p_{col}^{(co)}, 1_{co=col^*}) \quad (7)$$

$$L_{rows} = \frac{1}{|\mathrm{Rows}|}\sum_{ro \in \mathrm{Rows}} \mathrm{CE}(p_{row}^{(ro)}, 1_{ro=row^*}) \quad (8)$$

$$L_{total} = L_{cells} + L_{columns} + L_{rows} \quad (9)$$

각 granularity별 선택 확률을 계산하기 위해서는 granularity별 logit이 필요하며, 이는 token logit을 이용해 계산된다. <Figure 4>와 같이 선형화된 표는 모델의 입력으로 들어가기 전 토큰화 과정을 거치므로 하나의 셀에 속하는 문자열이 여러 개의 토큰으로 나누어질 수 있다. 따라서, 각 셀에 대한 logit이자 $p_{cell}^{(c)}$은 각 셀에 속하는 토큰들의 token logit을 평균하여 구하며, 이때 각 셀은 모수가 $p_{cell}^{(c)}$인 베르누이 분포를 따르는 독립적인 베르누이 변수로 표현된다. 앞서 구한 셀 별 logit을 동일한 열에 속하는 셀끼리 합하여 각 열의 logit이 계산되며, 전체 열의 logit을 소프트맥스(softmax) 함수에 통과시켜 $p_{col}^{(co)}$을 구한다. $p_{row}^{(ro)}$를 구하는 과정도 이와 유사하게 행 별로 셀의 logit을 합산하여 각 행의 logit을 구한 후, 전체 행의 logit을 소프트맥스 함수에 통과시켜 구한다.

### 3.3 추론

미세 조정까지 완료된 KOreaPAS를 이용해 추론을 진행할 때, 특정 셀 c의 점수 $Score_c$는 식 (10)과 같이 셀 c가 정답으로 선택될 확률 $p_{cell}^{c}$, 셀 c가 포함된 행 가 정답 셀을 포함한 행으로 선택될 확률 $p_{row}^{ro}$, 셀 c가 포함된 열 co가 정답 셀을 포함한 열로 선택될 확률 $p_{col}^{co}$을 모두 합하여 산출한다. 이때, 하나의 셀에 대한 확

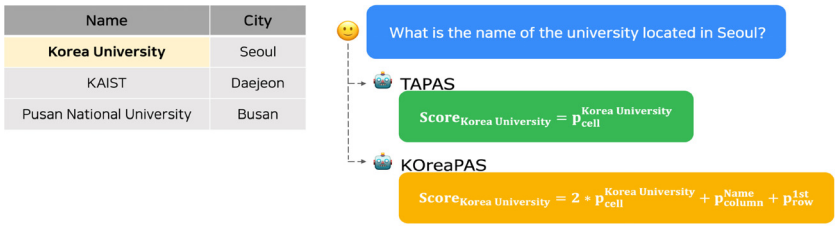


**Figure 5.** An Example of the Inference Process

률 $p_{cell}^{c}$을 셀이 속한 행과 열의 확률로 간주하여 합산 시 $p_{cell}^{c}$에 2의 가중치를 주었다. 최종적으로 표에 존재하는 모든 셀 중 $Score_c$ 값이 가장 높은 셀이 질문에 대한 정답으로 반환된다.

$$Score_c = 2*p_{cell}^{c} + p_{row}^{ro} + p_{col}^{co} \tag{10}$$

## 4. 실험 및 결과

### 4.1 실험 설계

(1) 데이터셋

본 연구에서 제안한 한국어 표 질의응답 모델을 학습하기 위해서 각 학습 단계별로 두 가지 데이터셋을 활용하였다. 사전 학습 단계에서는 최초의 대규모 한국어 표 데이터셋인 KorWikiTabular와 한국어 질의응답 데이터셋인 KorQuAD 2.0으로부터 추출한 표 데이터를 병합해 약 120만 개의 표 데이터를 활용했으며, 미세 조정 단계에서는 KorWikiTQ(Jun *et al.*, 2022)와 행정문서 대상 기계독해 데이터를 활용하였다. KorWikiTQ는 위키피디아 문서를 기반으로 제작된 한국어 표 질의응답 데이터셋이며, 행정문서 대상 기계독해 데이터는 공공데이터포털, 공공기관 보유 행정문서를 기반으로 구축된 한국어 질의응답 데이터셋이다. 행정문서 대상 기계독해 데이터는 정답 경계 추출형, 다지선다형, 절차형 등 다양한 질문 유형의 데이터를 포함하지만, 본 연구에선 표에 정답이 존재하는 표 정답 추출형 데이터만을 학습 및 평가에 사용하였다.

표 데이터를 정제된 리스트 형태로 제공하는 KorWikiTQ와 달리 행정문서 대상 기계독해 데이터는 HTML 언어 형태로 제공하며, 셀 내에서 일부분만이 정답이 되는 문제도 포함하고 있다. 그러나, 본 연구에서는 하나의 셀이 정답에 대응되는 셀 선택 유형의 문제에만 집중하고 있으므로 셀 내에서 특정 부분만이 정답이 되는 문제는 데이터에서 제거한 후 학습과 평가를 수행하였다. 두 가지 데이터셋 모두 검증 데이터셋이 따로 주어지지 않기 때문에 전체 학습 데이터 중 10%를 랜덤하게 추출하여 검증 데이터셋으로 활용하였으며, 최종적으로 학습, 검증, 평가에 사용한 데이터 개수는 <Table 2>와 같다.

**Table 2.** Dataset Statistics

| Dataset | Train | Dev | Test |
|---|---|---|---|
| KorWikiTQ | 52,386 | 5,821 | 11,771 |
| MRC dataset for administrative documents | 90,468 | 10,051 | 12,521 |

(2) 베이스라인 모델

본 연구에서 제안한 모델과의 비교 모델로는 Ko-TABERT(Jun *et al.*, 2022)를 사용하였다. Ko-TABERT는 Jun *et al.*(2022)에서 대규모의 한국어 표 데이터셋과 함께 제안한 모델로, MLM을 통해 사전 학습을 진행한 후, 입력에서 정답의 시작과 끝을 예측하는 스팬 추출(span extraction) 방식의 미세 조정을 수행하였다. 본 논문에서 활용한 데이터셋과 동일한 데이터셋으로 사전 학습과 미세 조정을 수행하였다는 점에서 비교에 적합한 모델로 판단되나, 모델의 체크포인트가 외부에 공개되지 않아 KorWikiTQ에 한해서만 비교를 수행하였다.

(3) 평가 지표

표 질의응답 모델을 정량적으로 평가하기 위해 질의응답 분야의 대표적인 평가 지표인 Exact Match(EM)와 F1 score(F1)를 사용하였다. EM은 전체 데이터 중 모델을 통해 선택된 셀의 값과 실제 정답 셀의 값이 정확히 일치하는 데이터의 비율을 뜻한다. 전체 데이터의 개수가 N, 질문 $q_i$와 표 $t_i$로 구성된 데이터 $D_i$에 대한 예측 값과 정답이 각각 $Predictions_i$, $Answers_i$인 경우 EM은 식 (11)과 같이 계산된다.

$$EM = \frac{1}{N}\sum_{i=1}^{N} 1_{Predictions_i = Answers_i} \tag{11}$$

F1은 모델을 통해 선택된 셀의 값과 실제 정답 셀의 값 간에 음절 단위로 겹치는 부분의 비율을 평가에 반영한 점수이다(식 (12)~식 (14)). 오답일 경우 점수를 아예 부여하지 않는 EM과 달리 F1의 경우 정답과 겹치는 정도에 따라 점수가 부여될 가능성이 있으므로 EM보다 완화된 평가 척도로 여겨진다.

$$Precision_i = \frac{|Predictions_i \cap Answers_i|}{|Predictions_i|} \tag{12}$$

$$Recall_i = \frac{|Predictions_i \cap Answers_i|}{|Answers_i|} \tag{13}$$

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\left(2\times\frac{Precision_i \times Recall_i}{Precision_i + Recall_i}\right) \tag{14}$$

(4) 구현 상세

본 연구에서는 각 학습 단계별로 다른 종류의 GPU를 사용했으며, 사전 학습 단계에서는 NVIDIA A100 8대, 미세 조정 단계에서는 NVIDIA A6000 4대를 사용하였다. 사전 학습 시엔 배치 사이즈를 256으로 설정하여 약 1,500,000번의 Step까지 학습을 진행했으며, 검증 데이터셋 기준 유의미한 손실 값 변화를 보이지 않는 Step에서 학습을 종료하였다. 미세 조정 단계에서는 배치 사이즈를 64로 하여 20 에폭까지 학습을 진행하였으며, 매 에폭마다 검증 데이터셋의 정량적인 성능을 측정하여 가장 성능이 좋은 에폭의 모델을 최종 평가에 사용하였다. 옵티마이저는 자연어 처리 과업에 널리 활용되는 AdamW를 활용하였으며, 학습률의 경우 사전 학습 단계에서는 5e-5, 미세 조정 단계에서는 1e-5를 이용하였다.

### 4.2 실험 결과

(1) 주요 실험 결과

본 연구에서 제안한 KOreaPAS의 정량적인 성능은 <Table 3>과 같다.

**Table 3.** Main Results

| Dataset | Model | EM | F1 |
|---|---|---|---|
| KorWikiTQ | Ko-TaBERT(Jun *et al.*, 2022) | 87.20 | 91.20 |
| | KOreaPAS (Ours) | **94.73** | **96.01** |
| | w/o Multi-Granularity Fine-Tuning | 94.02 | 95.56 |
| | w/o Korean Table Pre-training | 38.66 | 53.62 |
| MRC dataset for administrative documents | KOreaPAS (Ours) | **94.67** | **96.11** |
| | w/o Multi-Granularity Fine-Tuning | 93.57 | 95.40 |
| | w/o Korean Table Pre-training | 46.48 | 58.55 |

KorWikiTQ의 경우 Ko-TaBERT에 비해 EM 기준 7.53% 높은 성능을 달성하였으며, 두 가지 데이터셋에서 모두 본 논문에서 수행한 사전 학습 및 미세 조정 과정이 성능 향상을 보였다. 특히 대규모의 한국어 표 데이터로 진행된 사전 학습의 경우 각 데이터셋에 대해 EM 기준 55.36%, 47.09%의 큰 성능 향상을 보인 것을 확인할 수 있다. 이러한 큰 성능 향상은 대규모의 한국어 표 데이터를 이용한 사전 학습이 일반적인 텍스트와 달리 표가 가지는 구조적인 특징을 언어 모델이 학습하는 데 도움이 되었음을 시사한다. 또한, 표가 가지는 다양한 granularity를 학습과 추론 과정에 반영한 새로운 미세 조정 방식도 각 데이터셋에서 EM 기준 0.71%, 1.1%의 유의미한 성능 향상을 보였다. 이는 하나의 granularity를 학습할 때보다 여러 granularity를 동시에 학습할 때 모델이 더 다양한 정보를 학습할 수 있으며, 여러 granularity의 점수를 모두 활용하여 추론을 진행할 경우 개별 granularity가 잘못 예측한 부분을 상호 보완할 수 있기 때문이다.

질문에 나와 있는 중요 키워드가 표에서 한번도 등장하지 않는 경우, 모델이 질문과 표의 관계를 파악하는 데 어려움이 존재한다. KorWikiTQ와 행정문서 대상 기계독해 데이터 모두 표의 제목에 대한 정보를 포함하고 있으며, 질문에 등장하는 중요 키워드가 표의 제목에 등장할 확률이 높다는 데이터 분석 결과에 따라 모델의 입력에 표의 제목 정보를 함께 구성하여 학습 및 평가를 진행하였다.

**Table 4.** Comparison of Model Performance Based on Input Format

| Dataset | Input Format | EM | F1 |
|---|---|---|---|
| KorWikiTQ | Question + Table | 94.73 | 96.01 |
| | Question + Table Title + Table | **94.81** | **96.18** |
| MRC dataset for administrative documents | Question + Table | **94.67** | **96.11** |
| | Question + Table Title + Table | 94.59 | 96.06 |

결과적으로 KorWikiTQ에서는 약간의 성능 향상을 보였지만, 행정문서 대상 기계독해 데이터에 대해선 성능 하락을 보이며 데이터셋마다 차이가 있음을 확인하였다. 새로운 입력 구성에 대한 모델의 추론 결과를 분석한 결과, 대부분의 질문이 정답 셀을 포함한 열과 행을 찾는데 필요한 핵심 단서를 포함하고 있기 때문에 질문과 표의 관계를 완벽히 파악하지 못해도 정답을 도출할 수 있었던 것으로 보인다. 또한, 표의 제목에 나타난 정보가 질문에서도 대부분 비슷하게 등장하는 점도 표의 제목 정보 추가 여부가 유의미한 성능 차이로 이어지지 않은 원인으로 판단된다.

(2) 한국어 표 사전 학습 효과

대규모 한국어 표 데이터를 이용한 사전 학습의 효과를 확인하기 위해 KorWikiTQ 평가 데이터셋에 대한 MLM accuracy를 확인하였다. MLM accuracy는 전체 마스킹 된 토큰 중 모델이 올바르게 복원한 토큰의 비율을 나타내는 평가 지표로, MLM accuracy가 높다는 것은 모델이 입력으로 주어진 문장을 잘 이해하고 있음을 나타낸다. 모델의 입력은 표의 제목에 해당하는 텍스트 시퀀스와 표 시퀀스로 구성되며, 마스킹을 수행하는 위치에 따라 4가지로 구분하여 성능을 확인하였다. <Table 5>에서 all은 입력 전체, title은 표의 제목, header는 표의 헤더에 속하는 셀, cell은 표에서 헤더를 제외한 셀에 한해 마스킹을 수행했음을 의미한다. 실험을 위한 비교 모델로는 대표적인 한국어 사전 학습 언어 모델인 KLUE-BERT 모델을 사용하였다(https://huggingface.co/klue/bert-base).

**Table 5.** MLM accuracy on KorWikiTQ Test Dataset

| | all | title | header | cell |
|---|---|---|---|---|
| klue/bert-base | 18.39 | 36.21 | 22.68 | 21.40 |
| KOreaPAS | **50.31** | 36.25 | **66.51** | **50.66** |

결과적으로 모든 경우에서 비교 모델보다 KOreaPAS가 우수한 성능을 보였다. 특히, 일반적인 텍스트의 형태를 가지는 표의 제목에 마스킹한 경우에는 근소한 성능 차이를 보이지만, header, cell과 같이 표에 속하는 토큰에 마스킹한 경우에는 큰 성능 차이를 보이는 것이 확인되었다. 이러한 결과를 통해 현존하는 한국어 사전 학습 언어 모델이 텍스트에 대한 이해 능력은 갖추었지만, 표의 구조적인 정보를 파악하는 능력은 부족하며, 따라서 언어 모델의 표 이해에 있어 표 데이터에 특화된 사전 학습은 필수적임을 알 수 있다.

(3) Multi-granularity Fine-tuning 효과

본 연구에서는 표가 가지는 세 가지 granularity를 학습 단계와 추론 단계에 모두 활용하는 미세 조정 방식을 제안하였다. 이에 따라 다양한 표의 granularity를 활용하는 것이 실제로 성능 향상에 도움이 되는지 단계별로 실험을 통해 확인하였다.

학습 단계에서의 효과를 확인하기 위해 TAPAS에서 제안한 미세 조정 방식으로 학습을 수행한 모델과 KOreaPAS의 성능을 비교하였다. 이때, 학습 단계에 국한된 비교가 가능하도록 KOreaPAS의 경우에도 TAPAS와 동일하게 셀의 점수만을 활용하여 추론을 진행하였다. 또한 두 모델 간의 성능 차이가 통계적으로 유의미한지 검정하기 위해 McNemar's test를 수행하였다. 그 결과 <Figure 6>과 같이 KOreaPAS가 TAPAS에 비해 우수한 성능을 보였으며, 모든 데이터셋에서 낮은 p-value를 기록하여 본 논문에서 제안한 미세 조정 방식이 학습 단계에서 유의미한 성능 향상에 기여함을 확인하였다.

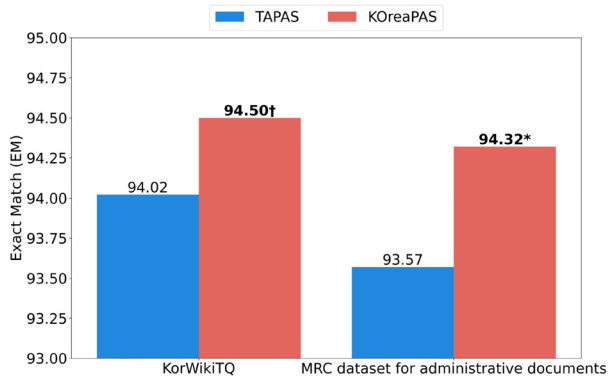


**Figure 6.** Test EM Results with Different Fine-tuning Methods. An (*) Indicates a Statistical Significance $p < 0.005$ and an (†) Indicates A Significance at $p < 0.05$ for McNemar's Test

추론 단계에서의 효과를 확인하기 위해 추론 시 한 가지 혹은 두 가지 granularity의 점수만을 활용한 결과와 세 가지 granularity의 점수를 모두 활용한 결과를 비교하였다. 결과적으로 모든 데이터셋에서 세 가지 granularity의 점수를 활용하였을 때가 가장 높은 성능을 보였으며, McNemar's test 수행 결과 대부분의 경우에서 매우 낮은 p-value를 기록하였다. 이를 통해 추론 단계에서 세 가지 granularity의 점수를 모두 활용하는 것이 성능 향상에 유의미한 영향을 끼침을 확인할 수 있다.

**Table 6.** Test EM Results with Different Combinations of Granularities Used for Inference. The Numbers within Parentheses Indicate the p-values via McNemar's Test Against the Original KOreaPAS Utilizing all Three Granularities for Inference

| # of Granularities | Granularities used for inference | Dataset | |
|---|---|---|---|
| | | KorWikiTQ | MRC dataset for administrative documents |
| 1 | Cell | 94.503 (0.006655) | 94.322 (2.08E-06) |
| 2 | Cell + Row | 94.503 (0.000971) | 94.537 (0.004059) |
| | Cell + Column | 94.707 (0.621873) | 94.473 (0.001135) |
| | Row + Column | 94.198 (1.25E-06) | 94.545 (0.140772) |
| 3 | Cell + Row + Column | **94.733** | **94.673** |

(4) 사례 연구

KorWikiTQ 평가 데이터셋에 대한 추론 결과를 분석한 결과 257개의 데이터에 대해 본 논문에서 제안한 미세 조정 방식으로 학습한 모델의 경우 정답을 예측했지만, TAPAS에서 제안한 미세 조정 방식으로 학습한 모델의 경우 오답을 예측한 것이 확인되었다. <Table 7>의 예제와 같이 '벤 애플랙이 2016년에 후보에 오른 부문 알려줘.'라는 질문이 주어졌을 때, TAPAS의 미세 조정 방식으로 학습한 모델의 경우 정답을 포함한 행을 선택하는 데 성공했지만, 정답을 포함한 열을 선택하지 못하여 오답을 예측한 것을 확인할 수 있다.

이는 추론 시 셀에 대한 점수만을 반영하는 TAPAS의 특성상 정답 셀에 대한 점수가 낮은 경우 그대로 오답으로 이어질 가능성이 높기 때문이다. 그러나 본 논문에서 제안한 추론 방식의 경우 정답 셀에 대한 점수가 낮더라도 정답 열에 대한 점수가 이를 보완할 수 있기 때문에 정답을 예측할 수 있었던 것으로 해석할 수 있다.

**Table 7.** An example of Inference Results Based on Different Fine-tuning Methods

| Question | 벤 애플랙이 2016년에 후보에 오른 부문 알려줘. | | | |
|---|---|---|---|---|
| Answer | 초이스 무비: 드라마/SF 판타지 배우상 | | | |
| Prediction (TAPAS) | 배트맨 대 슈퍼맨: 저스티스의 시작 | | | |
| Prediction (KOreaPAS) | 초이스 무비: 드라마/SF 판타지 배우상 | | | |
| Table | 연도 | 작품 | 부문 | 결과 |
| | 1999 | 아마겟돈 | 초이스 무비: 남자배우상 | 후보 |
| | … | … | … | … |
| | 2016 | 배트맨 대 슈퍼맨: 저스티스의 시작 | 초이스 무비: 드라마/SF 판타지 배우상 | 후보 |

(5) 오답 분석

KorWikiTQ 평가 데이터셋에서 KOreaPAS가 오답을 예측한 620개의 데이터 중 오답 셀의 추론 점수가 가장 높은 상위 50개의 데이터에 대해 오답 유형을 분석하였다. 이 중 19개에 해당하는 데이터가 데이터의 라벨링(labeling)이 잘못되어 실제로는 모델의 예측이 정답이거나, 모델의 예측 또한 정답이 될 수 있는 경우로 확인되며 오히려 모델의 좋은 예측 성능을 뒷받침하는 결과를 보였다.

19개를 제외한 31개의 데이터에서 자주 등장하는 오답 유형은 두 가지로 구분 지을 수 있다. i) 첫 번째 유형은 질문에서 올바른 행 또는 열을 예측하는 데 필요한 2개의 단서 중 1개의 단서만을 고려하여 정답을 도출한 경우이다. <Table 8>과 같이 '2018년 동계 올림픽 스노보드 여자 빅에어 예선 1차에서 가장 높은 점수를 기록한 선수는 누구입니까?'라는 질문에선 '1차'와 '가장 높은 점수'가 정답을 찾기 위해 고려해야 하는 핵심 단서가 된다.

그러나 KOreaPAS는 '가장 높은 점수'만 고려했기에 '1차'가 아닌 전체 순위가 가장 높은 '아나 가서'를 정답으로 예측하였다. ii) 두 번째 유형은 질문에 등장한 핵심 단서가 표에는 핵심 단서의 이음동의어로 등장한 경우이다. 예를 들어, '북아일랜드 정당 신 페인의 첫 번째 대표는 누구였어?'라는 질문이 주어졌을 때, 정답을 포함한 열은 '초대 대표'이지만 모델은 '현 대표'에 속하는 대표의 이름을 예측하였다. 이는 '첫 번째'와 '초대'가 동일한 의미를 가진다는 사실을 모델이 파악하지 못했기 때문이다. 두 번째 유형에 해당하는 문제의 경우 언어 모델이 표를 이해하는 능력보다는 언어를 이해하는 능력과 직결되는 것으로, 성능이 더 좋은 언어 모델을 사용하여 학습을 수행하면 개선 가능할 것으로 보인다.

## 5. 결론

본 연구는 대규모의 한국어 표 데이터를 이용해 사전 학습 및 미세 조정을 수행한 한국어 표 질의응답 모델 KOreaPAS를 제안하였다. 본 논문에서는 공개된 한국어 표 데이터셋이 사전 학습 시 가지는 한계를 보강하는 과정에 집중했으며, 보강된 데이터를 이용한 사전 학습이 언어 모델의 한국어 표 이해에 큰 도움이 됨을 실험적으로 확인했다. 더불어, 표의 다양한 granularity를 학습과 추론에 반영하는 신규 미세 조정 방식은 본 연구가 활용하고 있는 TAPAS의 미세 조정 방식에 비해 정량적, 정성적으로 모두 유의미한 결과를 보였다. 결과적으로, 본 연구에서 제안한 KOreaPAS는 두 가지 한국어 표 질의응답 데이터셋에 대해 현재까지 공개된 성능 중 가장 우수한 성능을 보였으며, 이를 통해 한국어 표 질의응답 과업에서 KOreaPAS 의 유효성을 입증하였다. 또한, 해당 연구가 최초로 두 가지 데이터셋에 대해 동일한 모델로 성능 평가를 수행하였다는 점을 고려할 때, KOreaPAS가 추후 한국어 표 질의응답 과업에서 중요한 비교 기준이 될 것으로 기대된다.

그러나 본 연구는 표 질의응답 데이터셋의 다양한 문제 유형 중 셀 선택 유형에 주력하였으므로, 셀 내의 일부만이 정답인 문제나 합, 차, 평균 등의 집계 연산이 필요한 문제에는 한계를 보인다. 이러한 한계를 극복하기 위한 향후 연구 방향은 두 가지로 제시될 수 있다. 첫 번째 연구 방향은 본 연구의 학습 방식을 확장하여 셀 내에서 일정 부분만이 정답인 문제나

**Table 8.** An Example of Error Case of KOreaPAS

| Question | 2018년 동계 올림픽 스노보드 여자 빅에어 예선 1차에서<br>가장 높은 점수를 기록한 선수는 누구입니까? | | | |
|---|---|---|---|---|
| Answer | 로리 블루앵 | | | |
| Prediction | 아나 가서 | | | |
| Table | 순위 | 이름 | 1차 | 2차 |
| | 1 | 아나 가서 | 88.25 | 98.00 |
| | … | … | … | … |
| | 4 | 로리 블루앵 | 90.25 | 92.25 |

집계 연산이 필요한 문제 등 보다 더 다양한 질문 유형에 대응 가능한 모델을 수립하는 것이다. 다음으로, 두 번째 연구 방향은 앞서 개선된 모델을 학습시키기 위한 추가적인 한국어 표 질의응답 데이터셋을 구축하는 것이다. 현재까지 공개된 한국어 표 질의응답 데이터셋의 경우 대부분의 문제가 집계 연산이 필요하지 않은 문제로 구성되어 있다. 반면, 영어의 경우 금융 분야에서의 높은 수요에 따라 다양한 집계 연산에 특화된 표 질의응답 데이터셋 구축이 최근까지 활발히 진행되고 있다 (Zhu *et al.*, 2021; Chen *et al.*, 2021, 2022). 따라서 향후에는 실제 산업에서의 수요에 맞춰 다양한 유형의 한국어 표 질의응답 데이터셋을 구축하고, 이를 활용해 KOreaPAS기반의 고도화된 모델을 개발하는 연구가 이루어지기를 기대한다.

## 참고문헌


Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. (2015), VQA: Visual Question Answering, *International Journal of Computer Vision*, **123**, 4-31.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R.N., Beane, M.I., Huang, T.', Routledge, B. R., and Wang, W. Y. (2021), FinQA: A Dataset of Numerical Reasoning over Financial Data, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3697-3711.

Chen, Z., LI, S., Smiley, C., Ma, Z., Shah, S., and Wang, W. Y. (2022), ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6279-6292.

Cho, S., Kim, M., and Kwon, H. (2020), Table Question Answering based on Pre-trained Language Model using TAPAS, *Proceedings of the 32th Annual Conference on Human and Cognitive Language Technology*, 87-90.

Deng, X., Sun, H., Lees, A., Wu, Y., and Yu, C. (2020), TURL: Table Understanding through Representation Learning, *Proceedings of the VLDB Endowment*, **14**(3), 307-319.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.

Dong, H., Cheng, Z., He, X., Zhou, M., Zhou, A., Zhou, F., Liu, A., Han, S., and Zhang, D. (2022), Table Pre-training: A Survey on Model Architectures, Pretraining Objectives, and Downstream Tasks, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 5426-5435.

Eisenschlos, J. M., Gor, M., Müller, T., and Cohen, W. W. (2021), MATE: Multi-view Attention for Table Transformer Efficiency, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7606-7619.

Glass, M. R., Canim, M., Gliozzo, A., Chemmengath, S. A., Chakravarti, R., Sil, A., Pan, F., Bharadwaj, S., and Fauceglia, N. R. (2021), Capturing Row and Column Semantics in Transformer Based Question Answering over Tables, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1212-1224.

He, P., Liu, X., Gao, J., and Chen, W. (2020), DeBERTa: Decoding- enhanced BERT with Disentangled Attention, *Proceedings of the 9th International Conference on Learning Representations* (ICLR 2021).

Herzig, J., Nowak, P. K., Müller, T., Piccinno, F., and Eisenschlos, J. M. (2020), TaPas: Weakly Supervised Table Parsing via Pre-training, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4320-4333.

Iida, H., Thai, D.N., Manjunatha, V., and Iyyer, M. (2021), TABBIE: Pretrained Representations of Tabular Data, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3446-3456.

Iyyer, M., Yih, W., and Chang, M. (2017), Search-based Neural Structured Learning for Sequential Question Answering, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1821-1831.

Jin, N., Siebert, J., Li, D., and Chen, Q. (2022), A Survey on Table Question Answering: Recent Advances, *China Conference on Knowledge Graph and Semantic Computing*, 174-186.

Johnson, J., Douze, M., and Jégou, H. (2017), Billion-Scale Similarity Search with GPUs, *IEEE Transactions on Big Data*, **7**, 535-547.

Jun, C., Choi, J., Sim, M., Kim, H., Jang, H., and Min, K. (2022), Korean-Specific Dataset for Table Question Answering, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6114-6120.

Kafle, K., Cohen, S. D., Price, B. L., and Kanan, C. (2018), DVQA: Understanding Data Visualizations via Question Answering, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5648-5656.

Kahou, S. E., Atkinson, A., Michalski, V., Kádár, Á., Trischler, A., and Bengio, Y. (2017), FigureQA: An Annotated Figure Dataset for Visual Reasoning, *Workshop Track Proceedings of the 6th International Conference on Learning Representations* (ICLR 2018).

Kim, J. and Kang, P. (2021), K-Wav2vec 2.0: Automatic Speech Recognition based on Joint Decoding of Graphemes and Syllables, *Proceedings of the Annual Conference of the International Speech Communication Association*, 4945-4949.

Kim, Y., Lim, S., Lee, H., Park, S., and Kim, M. (2020), KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension, *Journal of KIISE*, **47**, 577-586.

Lee, C., Yang, K., Whang, T., Park, C., Matteson, A., and Lim, H. (2021), Exploring the Data Efficiency of Cross-Lingual Post-Training in Pretrained Language Models, *Applied Sciences*, **11**(5).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019), BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871-7880.

Liu, Q., Chen, B., Guo, J., Lin, Z., and Lou, J. (2021), TAPEX: Table Pre-training via Learning a Neural SQL Executor, *Proceedings of the 10th International Conference on Learning Representations* (ICLR 2022).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019), RoBERTa: A Robustly Optimized BERT Pretraining Approach, *CoRR*, abs/1907.11692.

Masry, A., Do, X., Tan, J. Q., Joty, S. R., and Hoque, E. (2022), ChartQA: A Benchmark for Question Answering about Charts with Visual and

Logical Reasoning, *Findings of the Association for Computational Linguistics: ACL 2022*, 2263-2279.

Methani, N., Ganguly, P., Khapra, M. M., and Kumar, P. (2019), PlotQA: Reasoning over Scientific Plots, *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1516-1525.

Pandya, H. A. and Bhatt, B. S. (2021), Question Answering Survey: Directions, Challenges, Datasets, *Evaluation Matrices*, ArXiv, abs/2112.03572.

Park, S., Lim, S., Kim, M., and Lee, J. (2018), TabQA: Question Answering Model for Table Data, *Proceedings of the 30th Annual Conference on Human and Cognitive Language Technology*, 263-269.

Pasupat, P. and Liang, P. (2015), Compositional Semantic Parsing on Semi-Structured Tables, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1470-1480.

Radford, A. and Narasimhan, K. (2018), Improving Language Understanding by Generative Pre-Training.

Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019), Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, **21**(140), 1-67.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017), Attention is All you Need, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 5998-6008.

Wang, Z. (2022), Modern Question Answering Datasets and Benchmarks: A Survey, ArXiv, abs/2206.15030.

Wang, Z., Dong, H., Jia, R., Li, J., Fu, Z., Han, S., and Zhang, D. (2021), TUTA: Tree-based Transformers for Generally Structured Table Pre-training, *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1780-1790.

Xie, T., Wu, C., Shi, P., Zhong, R., Scholak, T., Yasunaga, M., Wu, C., Zhong, M., Yin, P., Wang, S.I., Zhong, V., Wang, B., Li, C., Boyle, C., Ni, A., Yao, Z., Radev, D. R., Xiong, C., Kong, L., Zhang, R., Smith, N. A., Zettlemoyer, L., and Yu, T. (2022), UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 602-631.

Yang, J., Gupta, A., Upadhyay, S., He, L., Goel, R., and Paul, S. (2022), TableFormer: Robust Transformer Modeling for Table-Text Encoding, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 528-537.

Yin, P., Neubig, G., Yih, W., and Riedel, S. (2020), TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8413-8426.

Yu, T., Wu, C., Lin, X. V., Wang, B., Tan, Y. C., Yang, X., Radev, D. R., Socher, R., and Xiong, C. (2020), GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing, *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.

Zhao, Y., Li, Y., Li, C., and Zhang, R. (2022), MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 6588-6600.

Zhong, V., Xiong, C., and Socher, R. (2018), Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning, *CoRR*, abs/1709.00103.

Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., and Chua, T. (2021), TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 3277-3287.

## 저자소개

**오수지:** 숙명여자대학교 IT공학과에서 2021년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학과 석사과정에 재학 중이다. 연구 분야는 자연어 처리 및 표 질의응답이다.

**고유경:** 고려대학교 미디어학부와 통계학과에서 2021년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학과 석사과정에 재학 중이며, 연구 분야는 프롬프트 기반 학습을 비롯한 자연어처리이다.

**이유경:** 한국외국어대학교 산업경영공학과에서 2019년 학사학위를 취득하고 고려대학교 산업경영공학과 석박사 통합 과정으로 재학 중이다. 연구분야는 자연어 처리, 대화 시스템, 거대 언어모델, 텍스트 생성이다.

**강필성:** 서울대학교 산업공학과에서 2003년 학사, 2010년 박사학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수로 근무하였으며, 현재는 고려대학교 산업경영공학부 정교수로 재직 중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.