

제품 스펙 문서 및 숫자 정보처리를 위한 표 질의응답 시스템



저자 1: 권선근

저자 2: 민경언

저자 3: 이상진

지도교수: 권혁철 교수님

목 차

1.	1	
1.1.	1	
1.2.	1	
1.3.	2	
2.	32	
2.1.	62	
2.2.	사용하는 라이브러리 목록	2
2.2.1	Tensorflow	2
2.2.2	CUDA(Compute Unified Device Architectur)	2
2.2.3	pandas	3
2.3.	Transformer	3
2.4.	BERT (Bidirectional Encoder Representations from Transformers)	4
2.5.	TAPAS (Table PArSing)	7
2.6.	KORQUAD 2.0 (The Korean Question Answering Dataset)	8
3.	108	
3.1.	108	
3.1.1.	다양한 형태의 표 데이터 수집오류! 책갈피가 정의되어 있지 않습니다.8	
3.1.2.	130	
3.2.	질의응답 데이터 수집	163
3.2.1	질의응답 데이터셋 형식	13
3.2.2	질의응답 데이터셋 수집 및 관리	13

3.2.3 질문 유형	13
3.3. 모델 학습 및 테스트 2016	
3.3.1. 모델 학습 과정 2016	
3.3.2. 모델 테스트 과정 2116	
3.3.2.1 데이터 평가 값	16
4. 2217	
4.1 설계 변경 내역	17
4.2 테스트 결과 개요	18
4.3 질의 유형별 분석	19
4.3.1. 어휘 포함	16
4.3.2. 어휘 변형	16
4.3.3. 다중 근거	16
4.3.3. 대소 비교	16
4.3.3. 순서	16
5. 오류! 책갈피가 정의되어 있지 않습니다.	
6. 수행 체계	2
6.1 구성원별 역할	
6.2 개발 일정	
7. 산학협력 멘토링 결과 반영 내용	
8. 27	

1. 서론

1.1. 연구 배경

현대 사회에서 많은 정보는 웹이나 데이터베이스 및 문서들에서 찾을 수 있는 표 형식을 사용하여 압축적으로 저장된다. 여기에는 소비자 제품의 기술 사양에서 재무 및 국가 개발 통계, 스포츠 결과 등이 포함된다. 궁금한 점에 대해 답변을 찾으려면 표들을 직접 읽거나 특정 질문에 대한 답변을 제공하는 서비스를 사용해야 한다. 이러한 답변을 자연어를 통해 질의할 수 있는 경우 훨씬 더 유용하며 접근성이 증대된다. 이런 시스템을 질의응답 시스템이라고 하며, 질의응답 시스템은 사용자가 입력한 질의에 정답을 정해진 데이터에서 탐색하여 사용자에게 출력하는 시스템이다.

최근, 이러한 자연어를 전통적인 의미론적 구문 분석(semantic parsing)이 아닌 BERT 아키텍처를 확장하여 테이블 형식의 데이터 구조와 질문을 함께 인코딩하여 답변을 직접 가리킬 수 있는 모델인 TAPAS(Table PArSing) 오픈소스가 등장했다. TAPAS 는 기존의 방식에 비하여 광범위한 도메인의 테이블에 적용할 수 있는 모델로써, 특정 테이블에 대한 질문이 아닌 임의의 질문으로 확장하기 유용하다. 한편, 위키백과 한 페이지 전체에서 표와 리스트가 포함된 지문을 대상으로 질의응답을 수행하는 KorQuAD 라는 웹문서 기계독해를 위한 한국어 질의응답 데이터셋에 BERT 모델을 학습시켜 표에 대한 질의응답을 수행해본 연구 역시 2019 년에 발표되었다. 또한 2020 년에는 오피스와 관련된 질의응답을 위한 테이블 독해에 대한 연구도 실시되었다. 이렇듯 실사용 가능한 자연어 처리 모델 개발은 최근 많은 관심을 받고 있는 분야이다. 따라서 이번 과제에서는 이러한 연구들의 연장선상으로 표 중에서도 제품 스펙 문서에 집중하여 연구한다.

1.2. 기존 문제점

데이터베이스와 같이 정해진 규칙에 맞게 입력되어 선별하고 수치화할 수 있는 데이터는 정형 데이터로 구분된다. 문서, 영상, 그림과 같이 정해진 규칙에 맞게 입력되지 않아 컴퓨터로 의미를 쉽게 파악하기 힘든 데이터는 비정형 데이터로 구분된다. 기존의 TAPAS 는 위키문서로만 구성되어 정형화 된 표에 대해서 높은 성능을 보여주지만 그렇지 않은 표에 대해서는 처리 성능이 다소 떨어진다는 문제점이 있다. 그 중 제품 스펙과 관련된 테이블에서 만들어지는 QA 들에서 처리성능이 매우 떨어지는 점을 착안해 이번 과제에서는 앞서 언급했던 KorQuAD 에 BERT 모델과 TAPAS 를 적용한 것과는 차별화되는, 제품 스펙을

제공하기 위해 표의 숫자 정보처리에 집중한 질의응답 모델을 만들고자 한다.

1.3. 연구 목표

기존 TAPAS 기반 한국어 질의응답 모델들은 제품 스펙 표들에 대한 자연어 질의응답에 비교적 매우 성능이 떨어지는 모습을 보였다. 목표로 하는 대상인 제품 스펙의 경우 같은 품목에 대해서도 많은 제조사들이 다른 스펙 표 레이아웃을 사용하는 경우가 많기 때문에, 이를 모두 학습하여 질의에 응답할 수 있는 모델은 그 활용성이 높다. 또한 제품 스펙 문서 특성상 굉장히 많은 숫자 정보에 대한 질의응답도 잘 처리할 필요가 있다. BERT 모델은 일반적인 자연어의 언어 현상을 모사하는 자연어처리(Natural Language Processing, NLP)모델에서 잘 작동하지만, 바이오 분야와 과학 분야등 사용 단어나 언어의 특성이 다른 특정 분야의 언어모델에 사용하려면 잘 적용되지 않는다. 이번 과제의 목표는 이 두 가지 목적을 달성하는 기존의 BERT 모델이 제품 스펙에 대한 특성을 수집할 수 있도록 언어 데이터들을 수집하고, 언어모델 학습을 추가적으로 진행하여 제품 스펙 문서에 대한 질의응답 모델을 개발하는 것이다. 이번 과제에서는 학습 데이터에서 정답을 얻을 수 있는 정답 위치가 정확히 태깅되어 있어서 정답을 구하기 위해 연산해야 하는 셀의 위치를 예측하도록 학습하는 방법인 감독 학습(supervised learning)을 사용하였다. 아래 그림 1, 그림 2는 TAPAS에서 높은 성능이 나오는 위키피디아의 정형적인 제품 스펙 표와 비정형적인 표의 예시이다.

iMac Pro 사양	
구성 요소 / 프로세서 모델	Intel Xeon 8, 10, 18 코어
모델	2017
출시일	2017년 12월 14일
외관	스페이스 그레이 알루미늄과 유리
디스플레이	27", 5120 x 2880 글로시 유리 커버 16:9, LED 백라이트 P3 지원 IPS 디스플레이 500 니트 밝기 10버트 색상
프로세서	8코어 (W-2145), 10코어 (W-2155), 14코어 (W-2175), 18코어 (W-2195) Intel Xeon 프로세서
메모리	32 GB, 64 GB, 128 GB 2666 MHz DDR4 ECC SDRAM
그래픽	AMD Radeon Pro Vega 56 8GB HBM2 메모리, AMD Radeon Pro Vega 64 16 GB HBM2 메모리
저장공간	1, 2, 4TB PCIe-기반 NVMe SSD
연결	내장 802.11a/b/g/n/ac NBASE-T 이더넷 (1 Gb, 2.5 Gb, 5 Gb, 10 Gb 지원) 블루투스 4.2
카메라	FaceTime HD camera 1080p (1920 x 1080; 2 MP)
주변장치	4x USB 3.0 SDXC 카드슬롯 (UHS-II 지원) 헤드폰/디지털 오디오 출력력 4x Thunderbolt 3 포트 (USB-C type)
무게	9.7 kg ^[4]


크기 및 무게	
높이: 51.6cm 너비: 65.0cm 스탠드 깊이: 20.3cm 무게: 8.92kg ²	
전력 및 사용 환경	
선전 사양: 100V~240V AC 주파수: 50Hz~60Hz(단일) 작동 온도: 10°C~35°C 상대 습도: 5%~90% 미응결 작동 고도: 5,000m까지 테스트	
제한 보증 및 서비스	
iMac에는 90일간의 무상 기술 지원과 1년 제한 보증이 제공됩니다. Mac용 원본 AppleCare+를 구입하면 보증 기간이 AppleCare+ 구입일로부터 3년까지로 연장되며, 12개월당 최대 2건의 우발적인 손상액 대한 보장이 추가로 제공됩니다. 우발적인 손상의 경우, 화면 손상 또는 외부 인물로 인한 손상은 각각 120,000원, 기타 수리 가능한 손상은 각각 370,000원의 본인 부담금이 부과됩니다. 더 알아보기 >	

그림 1(왼쪽), 그림 2(오른쪽) - 위키피디아에서 수집한 iMac Pro의 정형화된 표, 애플

2. 연구 배경 지식

2.1. 개발 환경

- Tensorflow-gpu 1.13.1
- CUDA 10.0
- GPU RTX 3080Ti
- Python 3.6
- 언어모델 : 한국어에 맞게 바꾼 TAPAS

2.2. 사용하는 라이브러리 목록

제품 스펙 문서에 특화된 표 질의응답 시스템의 학습 및 테스트를 위해서 사용한 라이브러리 목록들로는 표 데이터를 추출하기 위해서 파이썬 패키지인 BeautifulSoup, 형태소 분석을 위한 Eunjeon, Tensorflow, CUDA, Pandas 를 사용하였다.

2.2.1. Tensorflow



텐서플로우는 구글이 2015 년에 공개한 머신 러닝 오픈소스 라이브러리이다. 머신 러닝과 딥 러닝을 직관적이고 손쉽게 할 수 있도록 설계되었으며 파이썬을 비롯한 다양한 언어들의 API 를 제공하고 있다. 현재 Tensorflow GPU 사용시 CUDA 언어를 사용하고 있다.

2.2.2. CUDA(Compute Unified Device Architecture)



CUDA 는 NVIDIA 에서 개발한 GPU 개발 툴로써, 그래픽 처리 장치(GPU)에서 수행하는 (병렬 처리) 알고리즘을 C, C++를 비롯한 다양한 언어를 지원하여 작성할 수 있도록 하는 병렬 컴퓨팅 플랫폼 및 API 모델이다. CUDA 를 통해서 많은 양의 연산을 동시에 처리하는 것이 가능하게 하기 때문에 딥러닝, 채굴과 같은 수학적 계산에 많이 쓰인다.

2.2.3. Pandas



Pandas 는 데이터 조작 및 분석을 위해 Python 프로그래밍 언어로 작성된 소프트웨어 라이브러리이다. 특히 숫자 테이블과 시계열을 조작하기 위한 데이터 구조와 연산을 제공한다. 이번 과제에서는 xls 형식의 엑셀 표 데이터 파일을 Pandas Dataframe 으로 변환하여 Pandas 가 제공하는 연산자를 사용하여 xls 파일의 질의응답 데이터가 정확한 것인지 확인한다.

2.3. Transformer

기존의 RNN, CNN 계열의 신경망 구조를 탈피하고자 등장한 Attention 기법이다. transformer 는 RNN 의 순차적인 기계번역 방식을 행렬곱을 사용하여 한번에 처리하게 하였기 때문에 한번의 연산으로도 모든 중요 정보들을 임베딩 할 수 있다. 행렬곱만을 이용했을 때는 기계번역에서 중요한 토큰 위치와 순서값을 제공할 수 없기 때문에 위치

벡터도 함께 입력값으로 넣어주게 된다. 이를 positional encoding 이라고 한다. 즉 기존의 input 또는 output 에 수식을 통해 나온 positional value 값을 아래의 그림처럼 각각 더해주어 최종 input 을 만들어 준다. 학습 모델 중 하나인 Bert 는 아래 그림 4 의 왼쪽 부분인 encorder 만 이용하며, 두가지 대표적인 학습 방법으로 encorder 를 학습시킨 후에 특정 task 의 fine-tuning 을 활용하여 결과물을 얻는 방법으로 사용된다.

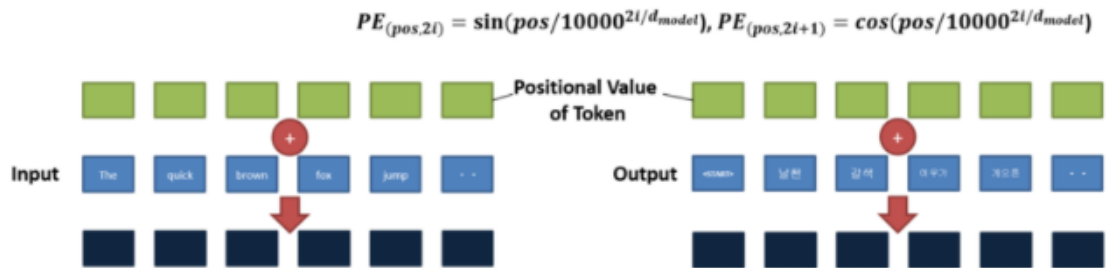


그림 3 - positional value 값을 추가한 최종 input

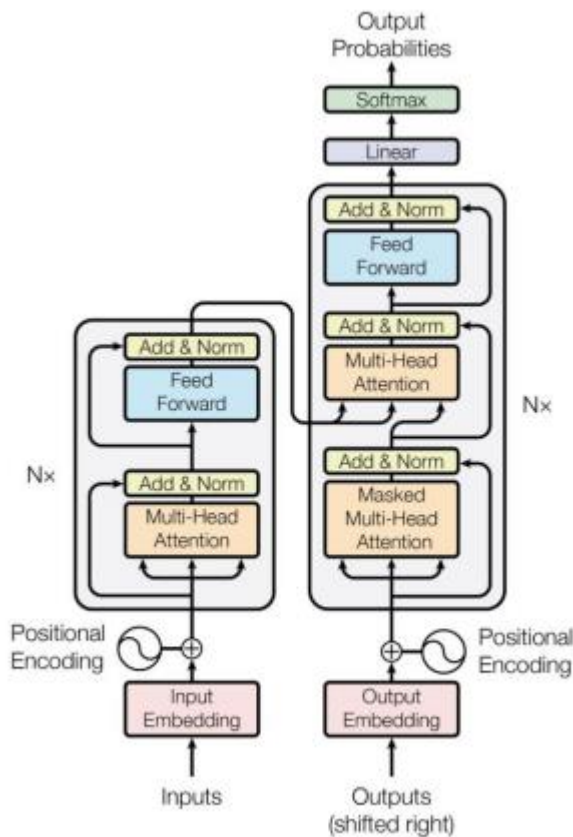


그림 4 - Attention mechanism - Attention Paper

2.4. BERT(Bidirectional Encoder Representations from Transformers)

BERT 는 2018 년 구글에서 공개한 사전학습(pretrained)된 모델로서 자연어 처리 작업에서 최고성능을 받으면서 크게 주목받았다. 이름에서 알 수 있듯이 Transformer 를 이용하여 구현되었다. BERT 가 성능이 높은 이유는 아래 그림 5 처럼 레이블이 없는 방대한 데이터로 학습시켜 이른바 사전학습된 모델(pretrained model) 을 생성하고 특정 목적에 따라 레이블이 있는 데이터들로 추가훈련(fine-tuning)하는 전이학습(transfer learning) 방법을 사용하여 파라미터를 재조정하는 방법을 이용하기 때문이다.

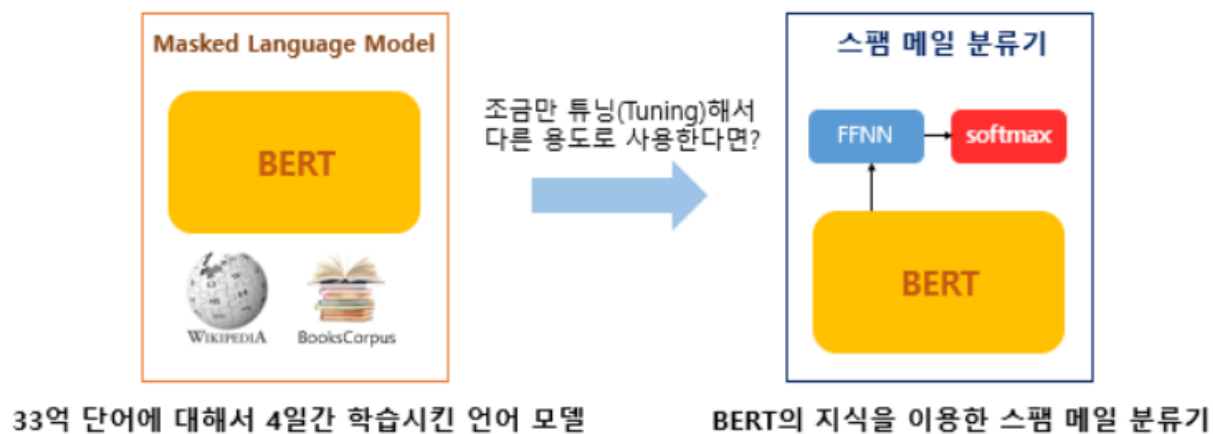


그림 5 - BERT 파인 튜닝

BERT 는 사전훈련 단계에서 전체 단어들중 랜덤으로 15 퍼센트를 뽑아서 마스킹 처리를 한다. 이중 80 퍼센트를 [MASK]로 대체하고 10 퍼센트를 랜덤으로 뽑은 다른 단어로 대체하고 나머지 10 퍼센트는 원본을 유지한다. 이렇게 대체하는 이유는 [MASK]만 사용하면 [MASK] 토큰이 파인 튜닝 단계에서는 나타나지 않으므로 사전 학습 단계와 파인 튜닝 단계에서의 불일치가 발생하는 문제가 있기 때문이다. 이렇게 문장에 뚫린 구멍을 앞뒤 단어 토큰들을 사용해 인공지능망에게 예측시키는 훈련을 통해 마스크드 언어모델(MLM)으로 학습시킨다.

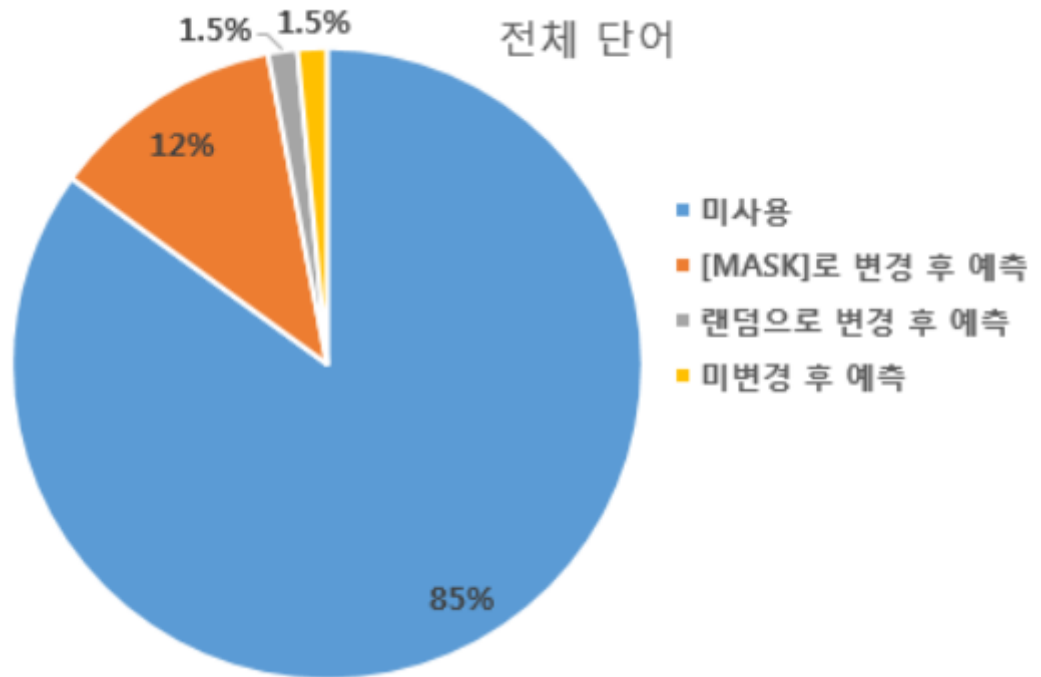


그림 6 - 텍스트 [MASK] 비율

BERT 는 사전훈련에서 두 문장이 이어지는지 아닌지를 판단하는 방법을 통해 훈련된다. 절반은 이어지는 두 문장, 나머지 절반은 이어지지 않는 두 문장을 모델에게 주고 두개가 이어지는지 아닌지를 판단하게 함으로서 학습시킨다. 아래 그림 7 처럼 [CLS] 토큰으로 분류문제의 시작을 알리고 [SEP] 토큰으로 두 문장을 분리한다.

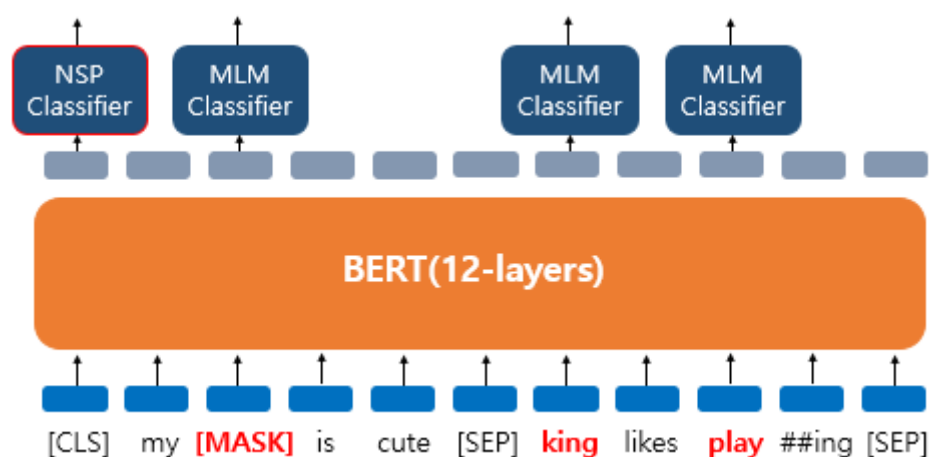


그림 7 - 다음 문장 예측(Next Sentence Prediction, NSP)

아래 그림 8 에서 보듯이 BERT 는 총 세개의 임베딩 층을 가지게 된다. 첫째 Token Embeddings 는 단어 하나하나가 토큰을 이뤄 실질적인 입력을 의미한다. 둘째 Segment Embeddings 는 두 문장을 구별하기 위한 임베딩으로서 그림 8 에서 보듯이 두 문장을 A,B 로 구분한다. 셋째 Position Embeddings 는 위치정보 학습을 위한 층으로서 문장의 단어들을 순서대로 부여한다.

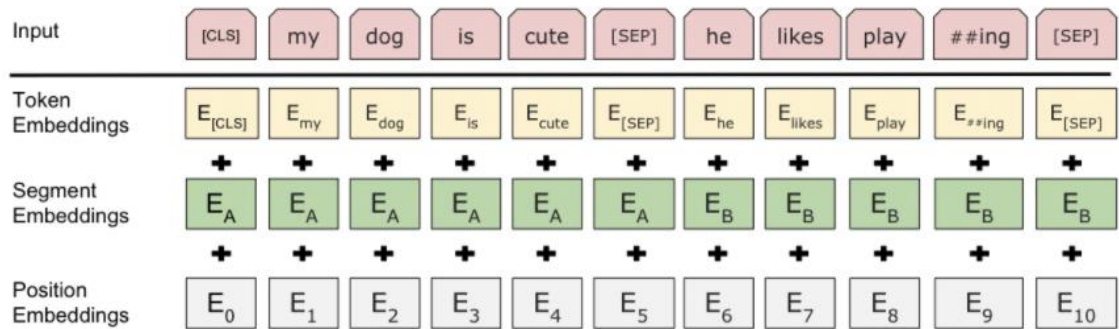


그림 8 - BERT 의 임베딩 층

2.5. TAPAS (Table PARsing)

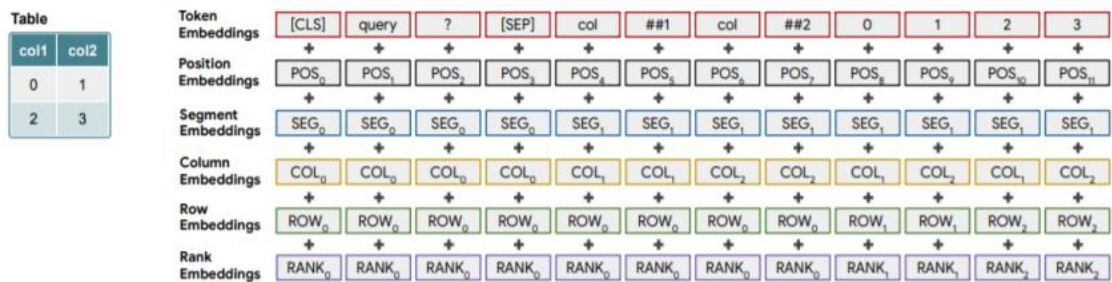


그림 9- TAPAS Word Embeddings

TAPAS 는 이스라엘의 텔아비브 대학의 전산언어학회가 2020 년에 발표한 기술로서 BERT 기술에서 확장하여 표(테이블)에 대한 자연어 질문처리를 수행한다. 위 그림에서는 왼쪽에 표시된 작은 표와 쿼리가 TAPAS 모델에 어떻게 인코딩 되는지 보여준다. 각 셀 토큰에는 열

내에서 행, 열 및 숫자 순위를 나타내는 원소들이 있고, 모델에 쿼리와 테이블을 이해시키기 위해 표현하는데 Token, position, segment embedding 은 기존의 Bert 에서 사용했던 것이고 TAPAS 에서는 table 을 이해시키기 위해 추가적으로 column, row, rank embedding 들을 사용한다. BERT 에서 두문장을 구분하던 [SEP]토큰과 Segment 임베딩은 이제 TAPAS 에서 질문과 테이블을 구분하기 위해 사용된다. column , row 두 embedding 은 표의 행과 열 정보를 나타냄으로서 TAPAS 모델이 표의 형태를 인식하는데 도움을 준다. 마지막 Rank Embeddings 는 숫자로 표현된 셀들의 대소비교를 통해 순서를 적용하여 TOP , WORST 같은 질문들을 처리하게 돕는다. 위 그림 5 에서 보듯이 Token 임베딩에서 col ##1 , col ##2 로 표현하여 열들로 구성됨을 알려주고 그에따라 Rank 임베딩은 COL 1 에서의 rank , COL 2 에서의 rank 를 나타내고 있다. 아래는 학습에 사용한 테이블과 학습한 모델에 대해 질의했을 때 대답을 찾는 과정을 보여준다.

Table				Example questions			
Rank	Name	No. of reigns	Combined days	#	Question	Answer	Example Type
1	Lou Thesz	3	3,749	1	Which wrestler had the most number of reigns?	Ric Flair	Cell selection
2	Ric Flair	8	3,103	2	Average time as champion for top 2 wrestlers?	AVG(3749,3103)=3426	Scalar answer
3	Harley Race	7	1,799	3	How many world champions are there with only one reign?	COUNT(Dory Funk Jr., Gene Kiniski)=2	Ambiguous answer
4	Dory Funk Jr.	1	1,563	4	What is the number of reigns for Harley Race?	7	Ambiguous answer
5	Dan Severn	2	1,559	5	Which of the following wrestlers were ranked in the bottom 3?	{Dory Funk Jr., Dan Severn, Gene Kiniski}	Cell selection
6	Gene Kiniski	1	1,131		Out of these, who had more than one reign?	Dan Severn	Cell selection

그림 10- TAPAS TABLE & Q&A Example

모델이 학습하는 과정에서, 해당 쿼리가 cell selection 문제인지, scalar answer 문제인지, ambiguous answer, aggregation operator 인지 판단하여 단순한 cell selection 이거나 scalar answer 인 경우 쉽게 처리하고, aggregation(집계문제)나 ambiguous 인 경우 softmax 를 적용하여 알맞은 operation 을 찾아낸다. 아래는 간단한 operation 을 찾는 모델 prediction 의 예시이다.

op	$P_s(op)$	compute(op, P_s, T)
NONE	0	-
COUNT	0.1	.9 + .9 + .2 = 2
SUM	0.8	.9×37 + .9×31 + .2×15 = 64.2
AVG	0.1	64.2 ÷ 2 = 32.1

$s_{pred} = .1 \times 2 + .8 \times 64.2 + .1 \times 32.1 = 54.8$

Rank	...	Days	P_s
1	...	37	0.9
2	...	31	0.9
3	...	17	0
4	...	15	0.2
...	0

그림 11 - TAPAS Compute Mode

이번 연구에서는 최종적으로 한국어에 맞게 바꾼 TAPAS 를 기반으로 한 자연어 처리 시스템을 사용한다.

2.6. KorQuAD 2.0 (The Korean Question Answering Dataset)

KorQuAD 는 한국어 기반 기계독해 학습 및 평가 말뭉치로 LG cns 에서 제작되었다. 그 중 KorQuAD 2.0 은 총 100,000+쌍으로 구성된 한국어 질의응답 데이터셋이다. 기존 질의응답 표준 데이터인 KorQuAD 1.0 과는 주어지는 지문이 한두 문단이 아닌 위키백과 한 페이지 전체라는 점, 지문에 표와 리스트도 포함되어 있어 html tag 로 구조화 된 문서에 대한 이해가 필요하다는 점, 답변이 단어 혹은 구의 단위 뿐 아니라 문단, 표, 리스트 전체를 포괄하는 긴 영역이 될 수 있다는 차이점이 있다.

3. 연구 내용

3.1. 표 데이터 수집

3.1.1. 다양한 형태의 표 데이터 수집

이번 과제의 주요 목표 중 하나가 현실에서 접하는 다양한 레이아웃의 표에 대해서 모델을 학습시켜 그 성능을 높이는 것이므로, 한국어로 된 다양한 레이아웃의 표를 얻기 위해서 삼성 전자, ASUS 등의 전자기기 브랜드 공식 홈페이지와 다나와, 행복쇼핑 등의 가격 비교 사이트에 접속하여 제품 스펙을 보여주는 웹브라우저의 html 코드에서 테이블 부분을 잘라내어 저장하는 방법을 사용했다. 아래 그림들은 이번 과제에서 사용된 7 개 레이아웃의 표 중 일부이다.

Color	Dreamy White
Processor	AMD Ryzen™ 5 5500U Processor 2.1GHz(3M Cache, 최대 3.95GHz)
Graphics	AMD Radeon™ Graphics,해당 없음
Display	15.6인치,FHD (1920 x 1080) 16:9,IPS-level 패널,Anti-glare display,LED Backlit,200nits,NTSC: 45%,해당 없음,Screen-to-body ratio: 86 %,해당 없음,해당 없음

그림 12 - ASUS 노트북 스펙 표

UHD 214 cm 스탠드형

디스플레이	화면크기 214 cm	해상도 3,840 x 2,160	
영상	화질엔진 Crystal Processor 4K HLG (Hybrid Log Gamma) 지원	HDR (High Dynamic Range) HDR 디밍 기술 UHD Dimming	HDR10+ 지원 명암비 강화 기술 (Contrast Enhancer) 있음
사운드	스피커 시스템 2CH 사운드출력(RMS) 20W 블루투스 오디오 연결 지원	무빙 사운드 (OTS) 없음 돌비 사운드 있음 멀티 언어/채널 없음	Q-Symphony 있음 Dolby 5.1 Decoder 없음
디자인	디자인 New Bezel-less 스탠드 컬러 TITAN GRAY	전면부 색상 TITAN GRAY	스탠드 타입 SLIM FEET

그림 13 - 삼성전자 tv 스펙 표

제조회사	QCY (제조사 웹사이트 바로가기)	등록년월	2020년 12월
유/무선 구분	무선	용도	음악+통화용
유닛형태	커널형	블루투스	블루투스 v5.0
무선연결거리	10m		
지원			
HFP	○	HSP	○
AVRCP	○	A2DP	○
부가기능			
핸즈프리통화	○	재생/멈춤조절	○
재생시간	4시간(싱글), 2시간(듀얼)	통화시간	3시간
대기시간	1440시간	충전시간	1시간
배터리용량	43mAh	무게	무게:4.5g
충전케이스			
케이스배터리용량	2600mAh		

그림 14 - 다나와 무선이어폰 스펙 표

센서 정보	심박 센서	○	가속도 센서	○
	만보기 센서	-	자이로 센서	○
	기압 센서	○	자외선 센서	-
	3D 터치센서	-	나침반 센서	-
	조도 센서	○	모션 센서	-
스포츠 기능	걷기	○	달리기	○
	자전거 타기	○	일립티컬	○
	로잉머신	○	등산	-
	승마	-	수영	-
	스키	-	골프	-
	요가	-	방수/방진	5ATM
건강 기능	보행 수(만보기)	○	이동거리	○
	칼로리	○	심박측정	○
	혈압체크	-	운동량	○
	체온	-	활동시간	○
	수면정보	○		
부가 기능	진동 알림	-	통화 기능	○
	음악컨트롤	-	스피커	○

그림 15 - 행복쇼핑 스마트워치 스펙 표

3.1.2. 표 데이터 전처리

3.1.1.절에서 언급한 방식으로 수집한 테이블들은 사이트마다 인코딩 방식이 다른 경우도 있고, 테이블을 만드는 html 태그를 다른걸 사용하는 경우도 있었고, 테이블을 꾸미는 CSS 등의 차이가 있기도 했다. 따라서 TAPAS 모델에 학습시키기 전에 전처리를 수행할 필요가 있었다. 전처리의 첫 과정에서는 한글 인코딩을 통일해주었다. euc-kr, utf-8 등이었던 인코딩 방식을 가장 많은 비중을 차지했던 utf-8 로 통일했다. 아래 코드 1 의 코드는 euc-kr 로 인코딩된 파일의 한글 인코딩을 utf-8 로 변경하는 코드이다.

```
for filename in os.listdir(os.getcwd()):
    if filename[-4:] == 'html':
        try:
            with codecs.open(filename, "r", "euc-kr") as sourceFile:
                with codecs.open('./after/' + filename, "w", "utf-8") as
targetname:
                    while True:
                        contents = sourceFile.read(BLOCKSIZE)
                        if not contents:
                            break
                        targetname.write(contents)
```

코드 1 - 인코딩을 utf-8 로 변경

그 다음, 수집한 개별적인 표마다 테이블을 만들때 특정 사이트는 <div>, <dt>나 css 등으로 테이블을 만드는 경우가 있었다. 이것을 TAPAS 에 학습시키게 하기 위해서는 우선 TAPAS 가 인식할 수 있는 테이블 태그인 <table>, <td>, <th>, <tr>등으로 전처리할 필요가 있었다. 따라서 수집한 표의 특성에 맞추어 홈페이지에서 나오는 표의 형태를 최대한 살리는 방향으로 개별적으로 html 태그를 삭제, 변환 등을 하는 프로그램을 작성하였다. 아래의 코드 2 는 파이썬의 정규표현식 모듈인 re 를 이용하여 html 태그를 검색 및 변환(또는 삭제)하는 코드의 일부이다.

```
for filename in os.listdir(os.getcwd()):
    if filename[-4:] == 'html':
        file_after = re.sub('<div class="spec-table">', '<table>',
file_after)
        file_after = re.sub('</div>', '</table>', file_after)

        file_after = re.sub('<dl>', '<tr>', file_after)
```



```

file_after = re.sub('</dl>', '</tr>', file_after)

file_after = re.sub('<dt>', '<td>', file_after)
file_after = re.sub('</dt>', '</td>', file_after)

file_after = re.sub('<dd>', '', file_after)
file_after = re.sub('</dd>', '', file_after)

file_after = re.sub('<ol>', '', file_after)
file_after = re.sub('</ol>', '', file_after)

with open(locate, "w") as file:
    file.write(html.encode('utf8'))
    print(filename+'생성완료')

```

코드 2 - html 태그 전처리

마지막으로는 수집한 데이터에 동일한 css 를 넣은 뒤, 브라우저로 확인해서 수집한 데이터가 서로 레이아웃이 다른, 이번 과제에서 의미가 있는 데이터임을 확인했다. 아래 그림들은 공통적으로 사용된 css 와 서로 다른 사이트에서 수집한 스펙 표이다. 같은 css 를 적용했을 때 두 표가 서로 구별되는 다른 형태임을 알 수 있었다. 또, css 에는 추가적으로 3.2.1. 절에서 후술할 질의 응답 데이터의 구성 요소 중 하나인 정답 셀의 위치 좌표를 넣어주었다.

```

<style>
    table {
        counter-reset: rowNumber -1;
    }

    table tr::before {
        counter-reset: colNumber -1;

        display: table-cell;
        counter-increment: rowNumber;
        content: "";
        padding-right: 0.3em;
        text-align: right;
    }

    table tr td::before {
        display: table-cell;
        counter-increment: colNumber;
        content: "(" counter(rowNumber) ", " counter(colNumber)
    );
        padding-right: 0.3em;
        text-align: right;
    }

```

```

table tr th::before {
  display: table-cell;
  counter-increment: colNumber;
  content: "(" counter(rowNumber) "," counter(colNumber)
)";

  padding-right: 0.3em;
  text-align: right;
}
</style>

```

코드 3 - html 레이아웃 처리

또한, 다양한 형태의 표를 모델이 잘 인식하지 못하는 문제가 발생하여, 질의에 대한 응답을 도출하지 못하는 문제를 해결하기 위하여 위의 re.sub 를 이용한 프로그램을 제작하여 표의 형태를 어느 정도는 비슷한 형태로 바꿔주었다. 아래 그림 16, 그림 17 은 원본 html 표 데이터와 언급한 전처리 작업을 거친 html 파일이 표 데이터로써 최종적으로 사용되는 모습이다.

이미지				(1,1) 제조사/브랜드	(1,2) 레노버 게이밍	(1,3) 레노버 아이디어패드	(1,4) ASUS 전북
	39.6cm / 2.25kg 라이젠5 / RTX3050	39.6cm / 1.64kg 라이젠5 / 광시야각	35.6cm / 1.3kg 라이젠7 / 63Wh	(2,1) 상품명	(2,2) 3 15ACH R5 3050 W10	(2,3) Slim5-15ITL-Premiere i7	(2,4) UM425UAZ-KI028
표준사/브랜	레노버 게이밍	레노버 아이디어패드	ASUS 전북	(3,1) 최저가격	(3,2) 1,048,950원	(3,3) 998,950원	(3,4) 1,097,050원
상세명	3 15ACH R5 3050 (SSD 256GB)	Slim5-15ALC R5 82LN0066KR (SSD 256GB)	UM4250A-KI064 (SSD 512GB)	(4,1) 등록년월	(4,2) 2021.07	(4,3) 2021.01	(4,4) 2021.07
최저가격	930,000원	699,000원	1,221,300원	(5,1) 화면 크기	(5,2) 39.6cm(15.6인치)	(5,3) 39.6cm(15.6인치)	(5,4) 35.56cm(14인치)
등록년월	2021.08	2021.06	2021.07	(6,1) CPU 제조사	(6,2) AMD	(6,3) 인텔	(6,4) AMD
화면 크기	39.6cm(15.6인치)	39.62cm(15.62인치)	35.56cm(14인치)	(7,1) CPU 종류	(7,2) 라이젠5-4세대	(7,3) 코어i7-11세대	(7,4) 라이젠7-4세대
CPU 제조사	AMD	AMD	AMD	(8,1) CPU 코드명	(8,2) 세잔	(8,3) 타이거레이크	(8,4) 루시엔
CPU 종류	라이젠5-4세대	라이젠5-4세대	라이젠7-4세대	(9,1) CPU 넘버	(9,2) 5600H (3.3GHz)	(9,3) i7-1165G7 (2.8GHz)	(9,4) 5700U (1.8GHz)
CPU 코어명	세잔	루시엔	세잔	(10,1) 코어 수	(10,2) 헥사코어	(10,3) 쿼드코어	(10,4) 옥타코어
CPU 넘버	5600H (3.3GHz)	5500U (2.1GHz)	5800H (3.2GHz)	(11,1) 운영체제(OS)	(11,2) 윈도우10	(11,3) 미포함(프리도스)	(11,4) 미포함(프리도스)
코어 수	헥사코어	헥사코어	옥타코어	(12,1) 화면 비율	(12,2) 와이드 16.9	(12,3) 와이드 16.9	(12,4) 와이드 16.9
운영체제 (OS)	미포함(프리도스)	미포함(프리도스)	미포함(프리도스)				
화면 비율	와이드 16:9	와이드 16:9	와이드 16:9				
해상도	1920x1080(FHD)	1920x1080(FHD)	1920x1080(FHD)				
화면 밝기	250nit	300nit	400nit				

그림 16 - 다나와 제품비교 테이블 원본 , 그림 17 - 다나와 제품비교 테이블 전처리 후

3.2. 질의응답 데이터 수집

3.2.1. 질의응답 데이터셋 형식

각 질의응답 데이터는 대상_테이블_html 파일명, 질문유형, 질문, 답변, 대상 카테고리, 답변위치, 작업자, 레이아웃, 비교등의 항목으로 이루어져 있다. 표 1 은 데이터셋의 예시이다.

대상_테이블_html 파일명	질문	답변	답변 위치	...
MacBook Air.html	MacBook Air 는 어느정도로 가볍니?	1.29kg	(15,1)	...
X SFX14-41G MAX, 17 A10SD HDR, TFG7576XG.html	X SFX14-41G MAX, 17 A10SD HDR, TFG7576XG 중에서 등록년월이 가장 나중인 제품의 상품명은?	TFG7576 XG	(1,3)	...

표 1 - 데이터셋 형식 예시 표

여기서 표 1 에 표시하지 않은 작업자, 질문유형, 레이아웃, 비교 항목은 작성을 돕기 위해 추가한 항목이며, 표시된 대상_테이블_html 파일명, 질문, 답변, 답변위치 항목은 학습이나 테스트를 진행하기 전, 데이터가 명확한지 확인하는 파이썬 프로그램에서 pandas 라이브러리를 이용하여 데이터를 불러온다. 파이썬 프로그램에서는 대상_테이블_html 파일명 항목에 해당하는 값으로 불러온 html 파일을 BeautifulSoup 라이브러리로 인코딩 한 뒤, 답변위치에 있는 셀과 답변 항목의 값이 서로 같은지 확인한다. 같다면 이 질문은 정확하게 만들어진 데이터로써 모델에 테스트나 학습용으로 저장된다.

3.2.2. 질의응답 데이터셋 수집 및 관리

질의응답 데이터셋은 조원 세명이 함께 수집하고 엑셀 형식의 파일에 저장하였으며, 접근성과 협업을 위해 구글 스프레드시트를 통해 관리했다. 만든 데이터셋은 테스트되고 그

결과가 이상이 있을 경우 질문을 수정하고, 어떻게 수정하면 보다 퀄리티 높은 데이터셋을 만들 수 있는지에 대한 노하우를 조원끼리 함께 공유했다.

3.2.3. 질문 유형

2장에서 TAPAS에서는 질문을 보고 정답의 유형을 Cell Selection, Compute Mode 등으로 분류한다고 하였다. 아래 표 2에서 보는 것처럼 스펙 표의 특성상 셀들간에 연관성이 없으므로, 앞서 보았던 복수의 셀을 고려해야 할 질문의 경우는 많지 않기 때문에 Cell Selection만 이용했다. 이번 과제에서 만든 Cell Selection을 이용한 질문 유형은 총 5가지로, 어휘 포함, 어휘 변형, 다중 근거, 대소 비교, 순서 유형이다. 이 중 어휘 포함, 어휘 변형 유형은 테이블이 전자기기 한 제품을 물어보는 경우(단일 테이블)의 경우에 대부분이 작성되었다. 아래 표 2와 표 3는 단일테이블의 일부분과 어휘 포함, 어휘 변형 질문의 예시이다. 어휘 포함은 질문에 테이블 내의 셀의 단어가 그대로 포함된 것이고 어휘 변형은 질문에 셀의 단어가 변형되어 들어간 것이다.

저장 용량	<ul style="list-style-type: none"> ● 64GB ● 128GB ● 256GB
방수 및 방진	<ul style="list-style-type: none"> ● IEC 규격 60529 하의 IP68 등급 획득(최대 수심 2m, 최대 30 분)
칩	<ul style="list-style-type: none"> ● A13 Bionic 칩 ● 3 세대 Neural Engine

표 2 - iPhone 11 테이블 부분의 일부분

질문유형	질문
------	----

어휘 포함	iPhone 11 가 지원하는 방수 및 방진등급은 무엇인가요?
어휘 변형	iPhone 11 에는 먼지가 들어오는 걸 막는 기능이 있나요?

표 3 - iPhone 11 의 어휘 포함, 어휘 변형 질문 예시

한편, 스펙 표에는 여러 가지의 비슷한 제품들을 함께 비교하는 유형도 있다. 이번 과제에서는 같은 카테고리의 제품 세가지를 비교하는 테이블(다중 테이블)에 대해서 다중근거, 대소비교, 순서의 질문을 만들었고 적은 개수의 어휘 포함, 어휘 변형 질문도 만들었다. 아래 표 4 와 표 5 는 그 중 다중 근거, 대소 비교, 순서에 해당하는 표의 형태와 질문의 예시이다. 다중 근거는 여러 개의 셀을 살펴보고 정답을 도출해야 하는 질문 유형이고 대소 비교는 '가장 많은/가장 적은'등 여러 셀을 살펴보고 대소를 비교하는 질문 유형이며, 순서는 테이블 내에서 연도나 순서를 묻는 질문 유형이다.

상품명	AN500UHD	쿠카 UC701UHD	U501QLED
제조사/브랜드	이스트라	이스트라	더함 우버
응답속도	9ms	8ms	5ms
지상파 UHD	수신 불가	수신 불가	수신 불가
에너지효율	1 등급	3 등급	1 등급
등록년월	2021.01	2021.03	2020.11

표 4 - AN500UHD, 쿠카 UC701UHD, U501QLED 테이블의 일부

질문유형	질문	답변
다중 근거	AN500UHD, 쿠카 UC701UHD, U501QLED 중에 제조사/브랜드가 더함 우버인 상품명은?	U501QLED

다중 근거	AN500UHD, 쿠카 UC701UHD, U501QLED 은 지상 UHD 가 수신 가능한가?	수신 불가
대소 비교	... 중에서 TV 의 응답속도가 가장 빠른 상품명은?	U501QLED
순서	... 중에서 등록년월 기준으로 제일 나중에 나온 상품명은?	쿠카 UC701UHD

표 5 - AN500UHD, 쿠카 UC701UHD, U501QLED 의 다중 근거, 대소 비교, 순서 질문 예시

그림 7 의 다중근거 첫번째 질문 예시 같은 경우에는 답변은 항상 그림 6 의 테이블에서 제품의 상품명이어야 함을 알 수 있다. 한편, 다중근거 두번째 예시의 경우에는 세 제품의 공통적인 특성을 물어본 경우이므로 답변이 지상파 UHD 의 셀 중 하나여야 한다. 대소 비교, 순서는 다중근거 첫번째 예시와 마찬가지로 제품명이 답으로 와야만 한다.

상품명	Inspire 2	M3 PLUS	Charge 4 SE
사용시간	평균사용:10 일이하	평균사용:7 일 이하	평균사용:7 일 이하
연결	블루투스 NFC	블루투스 GPS	블루투스 GPS NFC

표 6 - Inspire 2, M3 PLUS, Charge 4 SE 테이블의 일부

질문유형	질문	답변
다중 근거	Inspire 2, M3 PLUS, Charge 4 SE 중에서 연결 기능으로 GPS 를 지원하는 제품의 상품명은 무엇인가?	M3 PLUS&&Char ge 4 SE
대소 비교	Inspire 2, M3 PLUS, Charge 4 SE 중 평균사용 이용시간이 가장 짧은 상품명은 무엇인가?	M3 PLUS&&Char ge 4 SE

표 - 7 Inspire 2, M3 PLUS, Charge 4 SE 의 다중 근거, 대소 비교, 순서 질문 예시

추가로, 다중 테이블은 같은 카테고리의 제품들을 비교하므로 두 세가지의 제품들이 공통된 스펙 값을 보유하는 경우가 많았고, 이런 경우에 상품명에 답변으로 오는 질문의 경우 충분히 세 가지 상품명 중 두 가지가 정답이 될 수 있는가능성이 있었다. 이런 경우 표 6 과 7 에서 보듯이 다중 테이블에 대한 질문에 만족하는 정답이 두 개인 경우 답변을 &&를 사용해서 복수 표기하고, 한국어 TAPAS 모델에서 테스트할 때 이들 중 하나만 대답하여도 정답으로 인식하도록 하였다. 수집한 질문 유형 5 가지를 정리하면 아래와 같다.

질문 유형	
어휘 포함 (단일, 다중)	테이블에 있는 어휘가 포함되는 유형(849 개)
어휘 변형 (단일, 다중)	테이블에 있는 어휘가 변형되어 질문에 나타나는 유형(657 개)
다중근거 (다중)	여러 제품 셀을 살펴보고 정답을 도출하는 유형(738 개)
대소비교 (다중)	테이블 내에서 여러 제품 셀을 살펴보고 대소를 비교하는 유형(629 개)
순서 (다중)	테이블 내에서 제품들의 출시년도등의 순서를 묻는 유형(54 개)

표 8 - 질문 유형별 갯수 표

3.3. 모델 학습 및 테스트

3.3.1. 모델 학습 과정

학습 이전의 TAPAS 모델은 KorQuad 2.0 의 데이터에 대해서만 학습이 이뤄진 상태이므로 수집한 전자기기 관련 어휘가 조금이라도 변형되는 어휘 변형의 유형이나 다중 테이블에서 상품명을 답해야 하는 질문의 경우 정답률이 크게 낮아지는 문제가 있어, 과제 목표에 맞게 파인튜닝을 수행하였다. 수집한 3188 개의 전체 데이터셋 중 2927 개를 학습에 사용하였다.

3.3.2. 모델 테스트 과정

학습에 사용되지 않은 261 개의 데이터들은 파인 튜닝된 TAPAS 모델에서 테스트되었다.

3.3.2.1. 데이터 평가 값

언어 모델을 평가할 때는 성능 지표로써 EM 과 F1 스코어를 함께 사용했다. EM(Exact Match)은 모델이 테이블에서 셀의 위치 뿐만 아니라 경계도 정확하게 설정하여 데이터셋의 답변과 정확하게 일치하는 경우 올라가는 수치이다. F1 은 Precision(정밀도)과 Recall(재현율)의 조화평균으로, EM 과 다르게 서로 trade-off 관계인 정밀도와 재현율을 모두 고려해서 정답과 예측치의 겹치는 부분을 고려한 평가 방법이다. 음절 단위의 F1 점수에 대한 수식은 다음과 같다.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$f1 = 2 \times \frac{precision \times recall}{precision + recall}$$

TP 는 실제 정답인 음절을 제대로 예측한 것을 의미하며, TN 은 정답에 없는 음절이 예측된 정답에도 포함되지 않은 것을 의미한다. FP 는 실제 정답에 포함되어 있지 않은 음절이 예측 정답에 포함된 것을 의미하며, FN 은 실제 정답에 포함된 음절이 예측 정답에 포함되지 않은 것을 의미한다.

4. 연구 결과 분석 및 평가

4.1. 설계 변경 내역

4.1.1 표 데이터 카테고리 확대

과제 착수 시점에서는 표 데이터의 수집과 관련해서 카테고리를 하나로 좁힌 후 점차 늘려나가는 방향으로 설정했으며, 표 데이터는 html 파일 뿐만 아니라 이미지나 pdf 형식의 text-based 표까지 수집하려고 했었다. 하지만 오픈소스로 제공하는 OCR(Optical Character Recognition)같은 문자 인식 오픈 소스 라이브러리들이 한글로 된 표를 인식하는데 어려움이 있었기에 html 파일의 표만 모으기로 계획을 변경했다. 또한, 수집하는 표의 카테고리 역시 같은 카테고리의 제품들은 거의 동일한 형태의 표를 제공하는 경우가 많아 한 카테고리에서 여러 가지 형태의 테이블들을 모으기가 어렵다고 판단되어 전자기기라는 넓은 카테고리에서 무선 이어폰, 스마트 워치, 노트북, 태블릿, TV 등 다양한 제품군들의 표를 모으기로 하였다.

4.1.2 질문유형 추가

중간 보고 이후에는 질문 유형이 어휘 포함, 어휘 변형 두 가지 밖에 되지 않는다는 점이 유용성이 낮다고 판단되어 다양한 제품들을 하나의 표에서 비교하는 다중 테이블을 수집하고 질문유형으로 다중 근거, 대소 비교, 순서 세가지를 추가했다.

4.1.3 정답률에 따른 레이아웃 변경

테스트 과정에서 TAPAS 가 정확하게 셀의 위치를 찾아냈지만 셀의 일부분만이 정답인 경우와 전체가 정답인 경우를 판단해서 정확하게 정답을 도출해내지 못하는 경우가 많아 수집한 레이아웃 중 셀의 내용이 너무 많은 표의 경우에는 데이터에서 제외시켰다. 또, 3.1.2의 마지막 부분에서 언급했듯이 같은 <td>셀 안에서 질문과 답변을 유추할만한 키워드가 함께 있는 경우의 정답률이 크게 떨어져 같은 <tr> 내에서 다른 <td>로 분리하는 등의 어느 정도 기존의 레이아웃을 변경하는 등, 다양한 레이아웃의 표에 대한 연구를 진행한다는 기존의 과제목표와 상충되는 변경도 있었다.

4.2. 테스트 결과 개요

433 개의 표 데이터와 3100 여 개의 질의응답 데이터를 수집한 이후, 부산대학교 인공지능연구실의 한국어에 맞게 바꾼 TAPAS 에서 이번 과제에서 수집한 데이터셋과

KorQuAD 를 이용해서 추가적인 학습(파인튜닝)을 진행하였다. 아래의 표 9 는 기존 모델과 파인튜닝을 진행한 모델을 테스트한 성능 비교 표이다.

	단일 테이블		다중 테이블	
성능 지표	F1	EM	F1	EM
한국어에 맞게 바꾼 TAPAS	56.9	49	40.0	18.6
KorQuAD 2.0 와 스펙 표 질의응답 데이터를 함께 파인 튜닝	58.2	49.2	39.0	13.9
스펙 표 질의응답 데이터로 파인 튜닝	81.5	73.6	66.1	48.1
KorQuAD 2.0 로 파인 튜닝 후 스펙 표 질의응답 데이터로 파인 튜닝	81.9	75.2	77.4	64.3

표 9 - 타파스 모델 성능비교 표

수집한 2927 개의 질의응답 데이터에 대한 테스트 결과, F1, EM 값이 가장 높은 모델은 코쿼드 데이터로 파인튜닝한 뒤 스펙질의데이터로 파인튜닝한 모델이었다. 이는 추가로 학습하고자 했던 질문의 유형이 기존 학습된 데이터 및 코쿼드 2.0 의 데이터와 너무 패턴이 달라서 추가 학습시 loss 가 많이 발생했기 때문으로 보인다. 예를 들어, 상품명을 찾는 질문이 대부분인 다중 근거, 대소 비교, 순서의 유형의 경우 질문에서 행과 열을 찾는것 뿐만 아니라, 마지막으로 상품명에 해당하는 행까지 고려해야 하는데, 이는 기존 학습된 데이터에서 찾을 수 없었던 유형이다.

이를 해결하기 위해서는 파인튜닝이 아닌 학습데이터에 아예 현재 데이터를 넣어서 학습을 하거나 현재보다 훨씬 많은 양의 데이터로 파인튜닝을 진행해서 loss 를 줄여야 제대로 된 모델이 만들어질 것으로 보인다. 다음부터 나올 테스트 결과는 가장 EM, F1 값이 높았던 코쿼드와 스펙표 데이터를 따로 학습한 모델로 테스트한 결과이다.

질의응답 데이터셋 실험결과

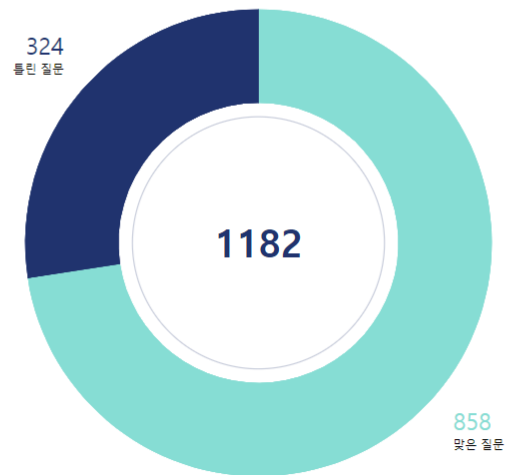


그림 18 - 실험 결과

- 평가한 질의 응답 데이터셋 수: 1182
- 정답을 맞춘 질문의 개수: 858 개
- EM: 72.5, F1: 81.5
- 유형별 정답률
 - 어휘 포함: 434/849 = EM: 85.0, F1: 90.2
 - 어휘 변형: 305/657 = EM: 78.3, F1: 84.2
 - 다중 근거: 191/738 = EM: 77.3, F1: 83.3
 - 대소 비교: 71/629 = EM: 49.4, F1: 69.3
 - 순서: 21/54 = EM: 50.0, F1: 66.8

그림 18 은 최종 테스트 결과이다. 가장 EM, F1 수치가 높은 한국어에 맞게 바꾼 TAPAS 모델은 전체 데이터셋 1182 개에 대해서 858 개가 정확한 정답을 도출해내는데 성공했으며, EM(Exact Match) score 는 72.5, 정확한 답을 찾는다는 실패했지만 정답 셀의

위치 까지 맞춘 F1 Score 는 81.5 이 나왔다. 유형별 정답률은 어휘포함 유형의 경우 85 로 가장 높았으며, 대소 비교의 경우가 49.4 으로 가장 낮았다.

4.3. 유형별 분석

4.3.1. 어휘 포함

- EM: 85.0, F1:90.2

질문	실제 답변	모델 응답	정답 여부
Neo QLED 8K 214 cm 의 디밍기술은 무엇인가?	Ultimate 8K Dimming Pro	Ultimate 8K Dimming Pro	맞음

표의 셀 이름인 디밍기술을 그대로 질의에 삽입하여 질문하는 유형이며, 모든 질문유형 중 가장 높은 EM 수치를 기록하였다.

4.3.2. 어휘 변형

- EM: 78.3, F1:84.2

질문	실제 답변	모델 응답	정답 여부
QLED 214 cm 의 디스플레이 화질이 어떻게 돼?	3,840 x 2,160	3 , 840 x 2 , 160	맞음

위처럼 표에서 해상도라고 나와있는 셀을 질의에 포함하여 해상도를 직접적으로 묻지 않고 디스플레이 화질이 어떻게 되는지 어휘를 변형해서 물어보는 질의이다.

4.3.3. 다중 근거

- EM: 77.3, F1:83.3

질문	실제 답변	모델 응답	정답 여부
----	-------	-------	-------

젠하이저 GTW 270, T5, AcousticANC1 중에 USB 타입 C 단자가 탑재된 제품의 상품명은 무엇인가?	AcousticANC1&& GTW 270	GTW 270	맞음
---	------------------------	---------	----

세 가지 제품의 스펙이 적혀있는 표에서 각 셀에 대한 직접질의를 하여 해당 특성에 해당하는 답변을 도출하는 질의이다. 이 때, 질의의 정답이 여러개일 경우는 &&를 삽입하여 둘 중 하나라도 맞으면 맞은 것으로 모델을 수정하였다.

4.3.4. 대소 비교

● EM: 49.4, F1:69.3

질문	실제 답변	모델 응답	정답 여부
3 15ACH R5 3050, X13 GV301QC-K6026T, TFG7597XG 중에서 가격이 가장 비싼 상품명은?	TFG7597XG	TFG7597XG	맞음

세가지 제품의 스펙이 적혀있는 표에서 각 셀에 대한 직접질의를 하여 해당 특성에 대한 대소를 비교하여 답변을 도출하는 질의이다. 사전에 제작된 모델은 다중제품이 있는 표에 대한 학습이 진행되지 않아 정답율이 상대적으로 낮게 나왔다.

4.3.5. 순서

● EM: 50.0, F1: 66.8

질문	실제 답변	모델 응답	정답 여부
MGN63KH, NT767XCL-KLTE, Aero 13-be0176AU 중 블루투스 버전이 가장 최신인 상품명은?	Aero 13-be0176AU	Aero 13 - be0176AU	맞음

세 가지 제품의 스펙이 적혀있는 표에서 제품들의 출시년도등의 순서를 묻는 유형이다. 사전에 제작된 모델은 다중제품이 있는 표에 대한 학습이 이루어지지 않았고, 모델이 셀값은 찾으나 정답에 해당하는 경계값을 잘 찾지 못하는 경향이 있어 정답률이 낮게 나왔다.

5. 결론 및 향후 연구 방향

연구 결과, 한국어에 맞게 변형한 TAPAS 모델은 KorQuAD 2.0 를 이용하여 스펙 표에 대해서 어느 정도의 변형된 어휘를 사용한 질문에 대한 적절한 답을 찾아내는데 성공했지만, 추가적인 학습 이후에도 이번 과제에서 수행한 전자기기에 특화된 어휘가 조금이라도 변형되면 잘 찾아내지 못하는 경우가 많았다. 이를 위해서는 현재 3,000 개의 질의응답 데이터셋이 아닌 KorQuAD 2.0 에 대응하는 많은 양의 스펙에 특화된 질의응답 데이터셋을 학습하는게 필요해 보인다. 특히, 이번 과제의 주요 주제이기도 했던 제품 스펙 표에서 자주 나오는 단위와 현재 전자기기에서 흔히 사용되는 외래어와 외국어의 혼용등이 성능을 높이는데 크게 작용할 것으로 보인다.

6. 수행 체계

6.1. 구성원별 역할

이름	역할분담
민경언	<ul style="list-style-type: none">● 최종 발표 및 시연 준비● 모델 성능 평가
권선근	<ul style="list-style-type: none">● 시스템 테스트● 모델 성능 평가
이상진	<ul style="list-style-type: none">● 시스템 테스트● 학습용 표 데이터 전처리
공통	<ul style="list-style-type: none">● 전반적인 지식 이해

- 질의 응답쌍 3000 여개 생성
- 학습용 표 데이터 400 여개 수집
- 발표 심사 준비

6.2. 개발 일정

5 월				6 월				7 월				8 월				9 월			
1 주	2 주	3 주	4 주	1 주	2 주	3 주	4 주	1 주	2 주	3 주	4 주	1 주	2 주	3 주	4 주	1 주	2 주	3 주	4 주
TAPAS 및 관련기술 공부																			
착수보고서 준비																			
				표 수집 및 전처리															
								중간 보고서 준비											
											데이터 학습 및 튜닝								

