TAPAS를 이용한 사전학습 언어 모델 기반의 표 질의응답*

조상현[O1], 김민호[2], 권혁철[1]
부산대학교 전기전자컴퓨터공학과[1], 부산가톨릭대학교 소프트웨어학과[2]
delosycho@gmail.com, minho@cup.ac.kr, hckwon@pusan.ac.kr

Table Question Answering based on Pre-trained Language Model using TAPAS

Sanghyun Cho[O1], Minho Kim[2], Hyuk-Chul Kwon[1]
Dept. of Computer Science Pusan National University[1], Dept. of Software Catholic University of Pusan[2]

요 약

표 질의응답은 반-정형화된 표 데이터에서 질문에 대한 답을 찾는 문제이다. 본 연구에서는 한국어 표 질의응답을 위한 표 데이터에 적합한 TAPAS를 이용한 언어모델 사전학습 방법과 표에서 정답이 있는 셀을 예측하고 선택된 셀에서 정확한 정답의 경계를 예측하기 위한 표 질의응답 모형을 제안한다. 표 사전학습을 위해서 약 10만 개의 표 데이터를 활용했으며, 텍스트 데이터에 사전학습된 BERT 모델을 이용하여 TAPAS를 사전학습한 모델이 가장 좋은 성능을 보였다. 기계독해 모델을 적용했을 때 EM 46.8%, F1 63.8%로 텍스트 텍스트에 사전학습된 모델로 파인튜닝한 것과 비교하여 EM 6.7%, F1 12.9% 향상된 것을 보였다. 표 질의응답 모델의 경우 TAPAS를 통해 생성된 임베딩을 이용하여 행과 열의 임베딩을 추출하고 TAPAS 임베딩, 행과 열의 임베딩을 결합하여 기계독해 모델을 적용했을 때 EM 63.6%, F1 76.0%의 성능을 보였다.

주제어: 사전학습 언어모델, 표 질의응답, 기계독해

1. 서론

표 질의응답은 반-정형화된 표 데이터에서 질문에 대한 정답을 찾는 문제이다.

표 질의응답을 위한 영문 데이터셋으로는 WIKISQL[1], WIKITQ[2], SQA[3]가 있다. WIKISQL과 WIKITQ는 학습 라벨로 정답을 제공하며 정답의 연산에 필요한 셀의 위치는 제공하지 않는다. SQA는 정답 셀(Cell)의 위치를 제공하며 이전 질문에서 얻은 답을 다음 질문에서 활용하는 대화형 질의응답 데이터이다. 한국어 표 질의응답을 위한 데이터로는 KorQuAD 2.0이 있다. KorQuAD 2.0은 표, 리스트, 텍스트의 정답과 질문, 그리고 정답을 포함하고 있는 위키피디아 문서를 제공하며 전체 데이터셋에서 약 22%의 데이터가 표에서 정답을 찾을 수 있는 질문과 정답으로 이루어져 있다.

본 논문에서는 한국어 표 질의응답을 위한 TAPAS 모델 기반의 한국어 표 언어모델 사전학습 방법과 KorQuAD 2.0 표 데이터에서의 질의응답을 위한 정답 예측 방법을 제안한다.

* 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2013-0-00131, (엑소브레인-총괄/1세부)휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술개발)

2. 관련 연구

[4]에서는 표의 각 셀과 질문을 인코딩하기 위해서 LSTM[5]을 이용했으며 인코딩된 자질을 셀에 질의와 열의 셀들과의 어텐션 스코어로 가중합한 자질을 추가했다. 추가된 자질은 bi-LSTM에 통과시켜 셀 feature를 보강하고 보강된 feature와 질문 벡터간의 어텐션 스코어를 계산하고 질문과 가장 연관성이 높은 셀을 선택하도록 하였다.

[6]에서는 테이블에서 질문에 대한 답을 하기 위한 어텐션 감독을 이용하는 다중 레이어의 시퀀셜 네트워크인 NEOP(Neural Operator)를 제안했다. NEOP은 여러 개의 SelRUs(Selective Recurrent Units)를 이용했다.

[7]에서는 대화형 테이블 질의응답에서의 성능 개선을 위해서 여러 종류의 지식을 통합한 의미 파싱 기법을 제안했다. 사용된 지식으로는 문법 지식, 전문가 지식, 외부 리소스 지식이 있다. 지식의 자질을 인코딩하고 정답을 얻기 위해서 GRU[8]를 이용했다.

[9]는 사전학습 기반의 약한 감독(Weak Supervision)의 표 질의응답을 위한 모델인 TAPAS(Table Parsing) 모델을 제안했다. TAPAS는 기존의 BERT[10]를 확장한 모델이며 표를 위한 특수한 임베딩을 추가했다. 추가된 임베딩으로는 비교가 가능한 셀들의 크기 순위를 나타내는 랭킹 임베딩, 셀이 속한 행을 나타내는 행 임베딩, 그리고 셀이 속한 열을 나타내는 열 임베딩이 있다. 영문 위키피디아 문서에서 추출한 약 620만 개의 테이블 데이터를 이용하여 모델을 사전학습 시켰으며, 정답 연산을 위

해 필요한 셀에 대한 정보가 제공되지 않는 표 질의응답 데이터의 학습을 위해서 정답을 위해 선택된 셀들의 “카운트, 합산, 평균”의 연산이 적용된 값과 데이터셋의 학습 라벨간의 오차에 대한 Hubber Loss 값을 줄이도록 학습하는 약한 감독 학습 방법을 제안했다. 영문 표 질의응답 데이터셋인 WIKISQL, WIKITQ, SQA에서 모두 높은 성능을 보였다.

본 논문에서는 TAPAS를 한국어의 표 데이터에 맞게 사전학습하고, TAPAS의 임베딩을 활용하여 선택된 셀 내에서 정답 경계를 추출할 수 있는 한국어 표 질의응답 모형을 제안한다.

3. TAPAS 모델 사전학습

TAPAS는 BERT를 확장한 모델이며 기존 BERT의 인코더의 위치 임베딩과 토큰 유형 임베딩 외에 표 데이터에 적합한 랭킹 임베딩과 행, 그리고 열 임베딩을 추가했다. 기존 2차원 형태를 가지는 테이블 데이터는 그림2와 같이 평면화 되어 모델의 입력으로 들어간다. 랭킹 임베딩은 셀 데이터가 부동소수점으로 표현될 수 있는 경우 해당 데이터를 크기에 따라 정렬하고 정렬된 순위에 대한 임베딩을 부여하며, 행과 열 임베딩의 경우 평면화된 표 데이터에 행과 열의 정보를 부여하는 역할을 한다.

TAPAS 모델을 사용하기 위해서 위키피디아에서 추출된 표 데이터를 사전학습에 이용했다. 표 데이터와 함께 학습하기 위한 텍스트 데이터로 위키피디아에서 그림과 같이 해당 표를 설명하는 텍스트 단락을 추출하였다. 표를 설명하는 단락이 존재하지 않는 경우 위키피디아 문서와 표가 존재하는 단락의 제목으로 대체하였다. 한국어 위키피디아 문서에서 약 11만 개의 표를 추출하여 사전학습 데이터를 구축하였다.

표 데이터에 추가적인 임베딩을 실험하기 위해서 각 테이블 셀에 대한 개체명 정보를 추출하여 개체명 임베딩을 추가하였다. 사용된 개체명은 “인물 이름, 시간, 수치, 국가 이름”으로 나누었으며 개체명 인식을 위해서 개체명 사전과 규칙을 사용하였다.

본 연구에서는 표 데이터에 대해서 처음부터 사전학습을 하는 방법 외에 기존에 위키피디아 텍스트에 사전학습된 BERT 모델을 불러와서 새로운 임베딩을 추가하여 학습하는 방법과 텍스트에 사전학습된 모델에서 출력된 표현 값에 새로운 임베딩 값을 더하여 학습하는 실험을 진행했다.


표 설명 텍스트

유튜브는 원도 미디어 비디오(WMV), 오디오 비디오 인터리브(AVI), MPEG, MPEG-4 파트 14(MP4) 포맷들을 업로드할 수 있다.[7]
동영상은 창 모드 또는 전체 화면 모드 가운데 하나를 골라서 볼 수 있다.(2015년부터는 영화관 모드도 사용할 수 있다.) 영상을 다시 불러오지 않고도 두 개의 방식을 바꾸어가며 쓸 수 있다.

표 타이틀 | 유튜브 미디어 인코딩 옵션 비교

itag 값	기본 컨테이너	영상 해상도	영상 인코딩	영상 로파일	영상 비트레이트 (Mbit/초)	소리 인코딩	소리 비트레이트 (kbit/초)
5	FLV	240p	소렌슨 H.263	빈칸	0.25	MP3	64
6	FLV	270p	소렌슨 H.263	빈칸	0.8	MP3	64
13	3GP	빈칸	MPEG-4 Visual	빈칸	0.5	AAC	빈칸
17	3GP	144p	MPEG-4 Visual	심플	0.05	AAC	24
18	MP4	270p/360p	H.264	베이스라인	0.5	AAC	96
22	MP4	720p	H.264	하이	2-2.9	AAC	192
34	FLV	360p	H.264	메인	0.5	AAC	128
35	FLV	480p	H.264	메인	0.8-1	AAC	128
36	3GP	240p	MPEG-4 Visual	심플	0.17	AAC	38
37	MP4	1080p	H.264	하이	3-4.3	AAC	192

표 데이터

그림1. 추출된 표 데이터 예시

3.1 테이블 질의응답 모델

TAPAS 모델에서 출력된 표현을 이용하여 질문에 대한 정답을 얻기 위해서 질의응답 모델에 평면화된 표 데이터인 $C=\{c_0,c_1,\ldots,c_m\}$와 질문 $Q=\{q_0,q_1,\ldots,q_n\}$, 그리고 [SEP],[CLS] 토큰과 함께 $X=\{[CLS],q_0\ldots,q_n,[SEP],c_0,\ldots,c_m\}$의 형태로 입력하게 된다.

그림2는 같이 기존의 [10]에서 기계독해 모델을 학습한 것과 같이 TAPAS 모델에서 출력된 표현에 바로 FFNN(Feed-Forward Neural Network)를 적용한 모델이다. FFNN에서 출력된 값은 정답의 시작에 대한 확률 y^s과 끝의 위치에 대한 확률 y^e이다.

$$h_i = TAPAS(x_i) \in R^{(n+m+2)\times d} \quad (1)$$
$$h^p = FFNN(h) \in R^{(n+m+2)\times 2} \quad (2)$$
$$y^s = softmax(h^{p0}) \in R^{(n+m+2)} \quad (3)$$
$$y^e = softmax(h^{p1}) \in R^{(n+m+2)} \quad (4)$$

그림2. 기계독해 방식의 정답 예측

그림3은 [9]에서 열에 대한 표현 값과 각 셀에 대한 표현 값을 합산하여 표의 셀을 선택하도록 한 것과 같이 행에 대한 표현 값과 열에 대한 표현 값을 합산하여 정답이 있는 행과 열을 예측하도록 한 모델이다. 이 방법은 정답이 있는 셀을 찾는 정확도가 비교적 높지만 그림4와 같이 셀 내에서 세부적인 정답 경계를 추출해야만 하는 문제는 적합하지 않은 문제가 있었다.

$$h^c = h^p c_{ids} \quad (5)$$
$$h^r = h^p r_{ids} \quad (6)$$
$$h^p = FFNN(h) \in R^{(n+m+2)\times 2} \quad (7)$$
$$h^p = FFNN(h) \in R^{(n+m+2)\times 2} \quad (8)$$
$$y_j^c = softmax(FFNN(h_j^c)) \in R^{(C)} \quad (9)$$
$$y_j^r = softmax(FFNN(h_j^r)) \in R^{(R)} \quad (10)$$

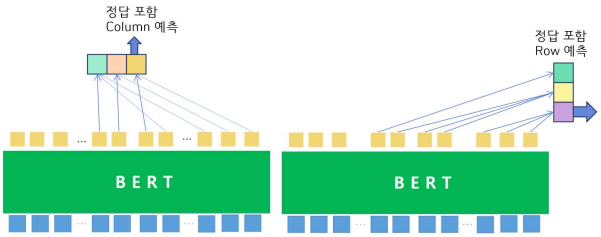

그림3. 행과 열의 임베딩을 이용한 정답 예측

수식의 c_{ids} 와 r_{ids}는 TAPAS에 입력되는 행과 열의 번호를 나타내는 ids를 원핫(one-hot) 형태의 벡터로 변환한 벡터이다.

셀 선택 정확도를 유지하면서 셀 내의 정답 경계를 찾기 위한 모델의 학습을 위해서 TAPAS 모델에서 출력된 표현 값과 합산된 행, 열의 표현 값을 결합(Concatenation)하고 결합된 표현 값에 FFNN을 적용하여 정답의 시작과 끝 위치를 예측하도록 하였다. 길이가 다른 행과 열의 임베딩을 길이를 맞추기 위해 임베딩 벡터에 전치된 c_{ids} 와 r_{ids}를 각각 행렬 곱셈 연산을 수행하였다. 정답을 얻는 방식은 식 와 같다.

$$\bar{h}^c = h^p c_{ids}^T \quad (11)$$

$$\bar{h}^r = h^p r_{ids}^T \quad (12)$$

$$\bar{h}^p = [\bar{h}^c; \bar{h}^r; h^p] \quad (13)$$

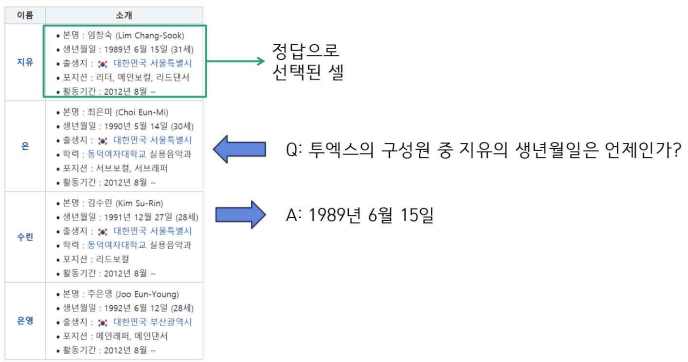

그림4. 셀 내의 경계 예측이 필요한 질의응답 데이터

4. 실험 및 결과

본 연구에서 사용된 KorQuAD 2.0 데이터셋은 83,686개의 학습셋과 10165개의 개발셋으로 이루어져 있으나 표 질의응답 모델의 학습과 평가를 위해 단답형의 정답을 가지는 표에 대한 정답이 태깅되어 있는 13376개의 학습 데이터와 984개의 개발 데이터 추출했다. 추출된 학습 데이터에서 12,000개의 데이터를 모델의 학습 데이터로 사용했고 1376개를 검증 데이터로 사용했으며 추출된 개발 데이터 984개를 평가 데이터로 사용했다.

표1. 추출된 표 데이터 개수

학습 데이터	12,000개
개발 데이터	1,376개
평가 데이터	984개

텍스트 데이터에 사전학습된 BERT는 BERT-base의 설정인 “히든 차원 수: 768, 히든 레이어 수: 12, 어텐션 헤드 수: 12”와 같다.

표2는 TAPAS 모델의 예측 방법 설정에 따른 성능 비교 결과이다.

표2. 예측 방법에 따른 성능 비교 (%, dev)

모델	EM	F1
MRC	46.8	63.8
col&row selection	60.2	74.3
MRC + col&row embedding	62.8	75.2
MRC + NE + col&row embedding	63.6	76.0

텍스트 데이터에서 기계독해를 학습한 것과 같이 TAPAS의 표현을 이용하여 정답과 시작과 끝을 예측하도록 한 기존 MRC 모델의 경우 다른 2가지 방법과 비교하여 낮은 성능이 나타나는 것을 확인했다.

행과 열에 대한 표현의 합산을 통해서 정답이 있는 행과 열을 예측하는 col&row selection 모델의 경우 기계독해 방법에 비해서 더 향상된 성능을 보였지만 셀 내에서 정확한 정답의 경계 추출이 불가능하기 때문에 성능의 하락이 있었다.

마지막으로 행과 열의 임베딩을 TAPAS의 표현과 결합하고 기계독해 방식으로 정답의 시작과 끝을 예측하도록 한 MRC + col&row embedding 모델의 경우 정답이 있는 셀 선택에 대한 높은 성능을 보이면서 셀 내의 정확한 정답 경계에 대한 예측이 가능했기 때문에 가장 좋은 성능을 보였다. NE는 개체명 임베딩을 추가하여 실험한 결과이며 F1 점수에서 0.8% 향상된 것을 보였다.

표3은 사전학습 모델에 따른 성능 비교 결과이다. 비교를 위해서 모델의 예측 방법으로 표2의 기계독해 방법을 이용했다.

표3. TAPAS 모델 설정에 따른 성능 비교 (%, dev)

모델	EM	F1
BERT	40.1	50.9
BERT + 임베딩 추가	50.9	62.1
coldstart TAPAS	42.8	54.1
warmstart TAPAS	46.8	63.8

텍스트에 사전학습된 BERT 모델을 바로 표 데이터에 적용했을 때 가장 낮은 성능을 보였다. coldstart 모델

은 BERT 모델을 처음부터 학습시킨 모델을 의미한다. warmstart 모델은 텍스트에 사전학습된 BERT 모델을 불러와서 TAPAS 모델을 표 데이터에 사전학습 시킨 모델이다.

coldstart 모델의 경우 기존의 텍스트에 사전학습된 BERT 모델에 표를 위한 임베딩을 추가한 모델보다 더 낮은 성능을 보였으며, 표 데이터에 사전학습하지 않고 임베딩만 추가한 모델에 비해서 warmstart TAPAS의 성능 향상폭이 크지 않았다. 이는 [9]에서 620만 개의 표 데이터를 사전학습에 사용한 것에 비해 10만 개의 표 데이터의 개수가 충분하지 않았던 것으로 생각된다.

5. 결론

본 연구에서는 표 질의응답을 위한 한국어 언어모델 사전학습과 표 질의응답을 위한 모델의 예측 방법을 제안하였다. 기존의 텍스트 데이터에 사전학습된 언어모델을 이용하여 TAPAS 모델을 표 데이터에 사전학습하는 것이 가장 성능이 나았음을 보였고, 표 데이터를 사전학습하기 위해서는 더 많은 양의 표 데이터가 필요한 것으로 보인다. 표 질의응답을 위한 예측 모델의 경우, 행과 열에 대한 표현 값을 구하고 TAPAS의 표현과 결합하여 기계독해와 같이 정답의 시작과 끝을 예측하도록 했을 때 가장 좋은 성능을 보였으며, 개체명에 대한 임베딩을 추가했을 때 성능이 향상했음을 보였다.

향후 연구에서는 더 정확한 개체명 태깅 방법과 표 질의응답에 도움이 될 수 있는 개체명을 추가하고 표 질의응답에서 성능을 향상할 수 있는 자질 추가에 대한 연구를 할 계획이다.

참고문헌

[1] Zhong, V., Xiong, C., & Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. arXiv preprint arXiv:1709.00103.

[2] Pasupat, Panupong, and Percy Liang., Compositional semantic parsing on semi-structured tables, arXiv preprint arXiv:1508.00305, 2015.

[3] Iyyer, M., Yih, W. T. and Chang, M. W., Search-based neural structured learning for sequential question answering, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1821-1831, 2017.

[4] 박소윤, 임승영, 김명지, 이주열, TabQA : 표 양식의 데이터에 대한 질의응답 모델, HCLT, pp.263-269, 2018.

[5] Hochreiter, Sepp, and Jürgen Schmidhuber, Long short-term memory, Neural computation 9.8, pp.1735-1780, 1997.

[6] Minseok Cho, Reinald Kim Amplayo, Seung won, Hwang, and Jonghyuck Park, Adversarial tableqa: Attention supervision for question answering on tables, ACML, 2018.

[7] Müller, T., Piccinno, F., Nicosia, M., Shaw, P. and Altun, Y., Answering Conversational Questions on Structured Data without Logical Forms, arXiv preprint arXiv:1908.11787, 2019.

[8] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

[9] 1Herzig, J., Nowak, P. K., Müller, T., Piccinno, F. and Eisenschlos, J. M.. TAPAS: Weakly Supervised Table Parsing via Pre-training, arXiv preprint arXiv:2004.02349, 2020.

[10] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.