도메인 변화에 강건한 사전학습 표 언어모형

조상현○, 최제훈, 권혁철
부산대학교 정보융합공학과
delosycho@gmail.com, hoon941209@pusan.ac.kr, hckwon@pusan.ac.kr

Domain-agnostic Pre-trained Language Model for Tabular Data

Sanghyun Cho○, Choi Jae-Hoon, Hyuk-Chul Kwon
Dept. of Information Convergence Engineering

요 약

표 기계독해에서는 도메인에 따라 언어모형에 필요한 지식이나 표의 구조적인 형태가 변화하면서 텍스트 데이터에 비해서 더 큰 성능 하락을 보인다. 본 논문에서는 표 기계독해에서 이러한 도메인의 변화에 강건한 사전학습 표 언어모형 구축을 위한 의미있는 표 데이터 선별을 통한 사전학습 데이터 구축 방법과 적대적인 학습 방법을 제안한다. 추출한 표 데이터에서 구조적인 정보가 없이 웹 문서의 장식을 위해 사용되는 표 데이터 검출을 위해 Heuristic을 통한 규칙을 정의하여 HEAD 데이터를 식별하고 표 데이터를 선별하는 방법을 적용했으며, 구조적인 정보를 가지는 일반적인 표 데이터와 엔티티에 대한 지식 정보를 가지는 인포박스 데이터간의 적대적 학습 방법을 적용했다. 기존의 정제되지 않는 데이터로 학습했을 때와 비교하여 데이터를 정제하였을 때, KorQuAD 표 데이터에서 f1 3.45, EM 4.14가 증가하였으며, Spec 표 질의응답 데이터에서 정제하지 않았을 때와 비교하여 f1 19.38, EM 4.22가 증가한 성능을 보였다.

주제어: 표 기계독해, 사전학습 언어모형, 의미 데이터 선별

1. 서론

기계독해 QA는 입력된 문서나 단락 내에서 주어진 질문에 대한 정답을 찾아내는 문제이다. 최근 데이터의 양의 급격히 증대되고, 단순한 텍스트로 이루어진 데이터뿐만 아니라 표나 리스트 등의 다양한 데이터로부터 필요한 정보를 적시에 사용하는 QA 시스템의 성능이 중요해졌다.

표가 다루고 있는 도메인이나 표의 형식이 변하게 되면 기존에 학습되었던 성능이 크게 감소하며, 이를 방지하기 위해서 정형화된 표 데이터 외에도 다양한 형태와 도메인을 가지는 표 데이터에 대한 사전학습이 필요하다. 기존의 연구에서 사용된 정형화된 위키피디아 표 데이터와는 달리 나무위키의 표 데이터는 다양한 형태를 가지기 때문에 사전학습에 사용하기 위해 데이터 전처리 과정이 필요하다. 먼저, 나무위키의 주제별로 의미 있는 표와 의미 없는 표로 분류한다. 다음으로 각 주제에서 나눠지는 단락별로 존재하는 의미 있는 표와 관련된 텍스트를 그룹화하여 실제 사용될 사전학습 데이터를 만든다.

본 논문에서는 정형화되지 않은 데이터에서 의미 있는 정보를 추출하고 분류하는 방법을 통해 구축한 사전학습 데이터를 사용해 향상된 표 언어모형을 구축하는 방법과 일반적인 표 데이터와 인포박스 데이터간의 적대적 학습을 통한 강건한 표 언어모형 구축 방법을 제안한다.

2. 관련 연구

TABERT[1]는 자연어와 구조화된 텍스트에 대해서 마스킹 된 행의 이름과 타입을 예측하는 MCP(Masked Column Prediction)와 마스킹된 셀의 값을 예측하는 CVR(Cell Value Recovery)의 방법으로 사전학습한 언어모형이다. TAPAS[2]는 기존의 BERT 기반 모형이 구조화된 표 데이터를 잘 처리할 수 있도록 기존의 임베딩 외에 행, 열, 랭킹 임베딩을 추가하여 표 데이터에 대해 사전학습한 언어모형이다.

위키피디아의 표 데이터와 함께 입력된 텍스트 데이터가 표의 내용에 수반되는지 혹은 반하는지에 따라서 ENTAILED 혹은 REFUTED 라벨을 부착한 TABFACT[3] 형식의 데이터를 반자동으로 구축하고, 마스크 언어모형 외에 해당 태스크를 사전학습에 추가하는 표 언어모형 사전학습 방법[4]으로 표 질의응답에서 향상된 성능을 보였다.

이전 표 전처리 관련[5]의 연구에서는 일반적인 HTML 문서의 표에서 정보를 추출하기 위해서 다양한 표를 필터링(Filtering) 기준에 따라 명백한 Decorative Tables를 분류하고 Meaningful Tables에 7가지 Heuristic을 적절한 가중치와 함께 선형 보간해 표에서 추상화 수준인 HEAD를 식별하는 기법을 제안하였다.

기계독해 모형이 추가의 파인튜닝 없이 새로운 도메인에 적응할 수 있도록 Domain Generailzation을 통해 적대적 학습을 적용한 방법[6]은 다양한 도메인의 기계독해 데이터셋에서 향상된 성능을 보였다.

3. 표 언어모형

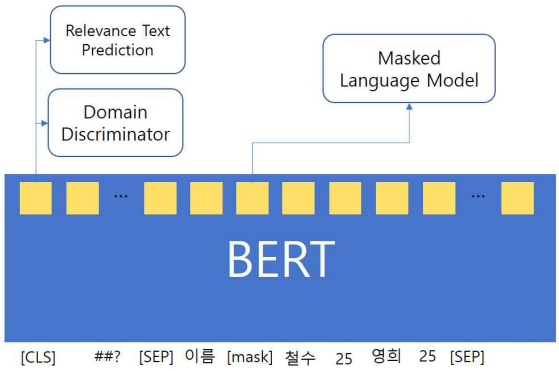

그림1. 언어모형 사전학습 구조

표 데이터에서는 2차원의 반-구조화된 정보를 인코딩하기 위해 기존의 BERT 모형의 토큰, 토큰 유형, 위치 임베딩 외에 표의 행과 열에 대한 임베딩을 사용한다. 본 논문에서는 BERT의 임베딩 레이어에서 행과 열의 임베딩을 추가하는 대신, BERT에서 출력된 벡터를 행과 열을 기준으로 합산하여 행과 열의 임베딩만을 이용하였으며, 각 셀의 순서 정보에 관한 인코딩하기 위해서 GRU를 이용하였다.

최종적으로 출력된 벡터를 이용하여 마스킹된 토큰예측하는 마스킹 언어모형과 [CLS] 토큰에서 출력된 벡터를 이용하여 함께 입력된 텍스트와 표가 서로 관련되어 있는지를 예측하는 태스크를 이용하여 표 언어모형을 사전학습 하였다.

3.1 의미 표 선별

기존의 연구에서 사용되었던 위키피디아 표 데이터의 경우 정형화된 형태를 가지고 있다. 다양한 형태나 도메인의 표 데이터를 처리하기 위해서는 정형화된 표 데이터 외에 다양한 사전학습 데이터가 필요하다. 정형화되지 않은 표 데이터를 사전학습 데이터로 사용할 때, 표의 구조적인 정보를 이용하는 대신 웹페이지에서 장식을 위한 목적으로 사용되는 표들이 존재한다. 본 논문에서는 이러한 데이터로 인해서 생기는 성능저하를 방지하기 위해서 이전 연구[5]에서 사용되었던 규칙 방법을 이용하여 의미 없는 표 데이터를 제거하고 나무위키에서 약 50만 개의 표 데이터를 사전학습 데이터로 이용하였다. 의미 없는 표 검출에 사용된 규칙은 다음과 같다.

연도구분				
기원전 422년	←	기원전 **421년**	→	기원전 420년
세기구분				
기원전 6세기	←	기원전 **5세기**	→	기원전 4세기
밀레니엄구분				
기원전	←	제1천년기		

그림 2. 의미 없는 표 데이터 예시

1. 표의 행 개수의 합이 2개 이하라면 의미 없는 표로 판단한다.
2. 표의 행이 가지는 셀의 개수 중 가장 작은 개수에 맞춰 각 행을 분해한다. 이때 그 개수가 2개 이하라면 의미 없는 표로 판단한다.
3. 이전 연구[5]에서 정의된 Heuristic을 나무위키 데이터에 맞게 적용 후 이진 행렬을 생성한다.

3.1) 표가 2개의 배경색으로 나뉘면 위쪽 행과 왼쪽 열이 HEAD가 될 수 있다.

3.2) 표 내에서 텍스트 속성으로 두 개의 영역으로 나뉜다면 위쪽 행 혹은 왼쪽 열이 HEAD가 될 수 있다.

3.3) 행이나 열을 구성하고 있는 셀들의 콘텐츠 타입 혹은 속성이 동일하게 적용되었다면 행이나 열의 끝이 HEAD가 될 수 있다.

3.4) 행이나 열을 구성하고 있는 셀들의 패턴이 동일하게 적용되었다면 행이나 열의 끝이 HEAD가 될 수 있다.

3.5) 가장 위의 행 또는 가장 왼쪽의 열이 병합된 셀일 때, HEAD가 될 수 있다.

3.6) 표에서 가장 왼쪽 위에 있는 셀이 비어있다면 그 셀을 포함하는 행과 열은 HEAD가 될 수 있다.

3.7) 위 과정들에서 생성된 이진 행렬을 결합해 셀이 HEAD가 될 수 있는지 판단한다. 결과에서 HEAD가 존재하지 않는다면 의미 없는 표로 판단한다.

4. 표의 전체 셀 중 비어있는 셀의 비율이 30% 이상이면 의미 없는 표로 판단한다.

아래의 수식은 선형보간법을 이용해 Heuristic 행렬을 결합하는 방법과 결합한 행렬 및 중간값을 이용해 최종 행렬을 만드는 방법을 나타낸다.

$$S = \sum_{i=1}^{6} \lambda_i H_i,$$

$$where\, 0 \le \lambda_i \le 1 \text{and} \sum_{i=1}^{6} \lambda_i = 1,$$

$$MID = \frac{\max(S_{ij}) + \min(S_{ij})}{2}$$

$$R_{ij} \begin{cases} 1 \text{ if } S \ge Mid \\ 0 \; otherwise, \end{cases}$$

여기서 H_i는 각 Heuristic 별로 생성된 행렬, S는 모든 Heuristic 행렬의 합, λ_i는 행렬별 가중치이다. Mid는 S의 최댓값과 최솟값의 합에 대한 중간값, R는 최종 행렬로써 $S_{ij} \ge Mid$라면 R_{ij}에 '1' 을 부여한다.

3.2 의미 텍스트 선별

기존의 연구에서는 표와 함께 입력할 텍스트를 추출하기 위해서 같은 문단 내에서 표를 설명하는 텍스트를 사용했으며, 표를 설명하는 텍스트가 없는 경우 단순하게 동일하게 등장한 토큰의 빈도수가 높은 텍스트 단락이나 문서의 제목 등을 사용했다. 본 논문에서는 사전학습에

서 표와 함께 입력되는 불필요하거나 관련이 없는 텍스트 데이터를 최소화하기 위해서 위키피디아 데이터로 먼저 사전학습한 모형을 통해서 입력된 텍스트와 표 데이터가 연관이 있는지 없는지에 대한 예측값을 구하고 해당 예측값이 가장 높게 나오는 텍스트 단락을 사전학습을 위한 텍스트로 사용하였다.

위키피디아 데이터로 사전학습하는 데이터는 같은 단락 내에서 표를 설명하고 있는 텍스트들이 존재하는 표 데이터를 추출하고 학습을 진행하였다.

3.3 적대적 학습

위키피디아와 나무위키에 존재하는 인포박스 데이터는 주로 엔티티에 대한 지식 정보들을 구조적인 정보로 저장하고 있다. 일반적인 표 데이터에서는 문맥적인 정보 외에 지식 정보를 학습하기 어렵다. 본 논문에서는 사전학습 과정에서 인포박스 데이터와 일반적인 표 데이터에 대해서 적대적 학습을 적용하였다.

[CLS] 토큰의 출력 벡터에 입력된 시퀀스가 인포박스 데이터인지 일반적인 표 데이터인지를 구분하는 판별기를 적용하였다. 이전 연구[6]를 따라서, 적대적 학습에 대한 손실 값으로 판별기의 예측값과 정규 분포 간의 쿨백-라이블러 발산 값으로 설정하였다. 이에 대한 수식은 다음과 같다.

$$L_D = -\sum_{i=1}^{N}\log P(l_i^{tab}|h_i^{tab}) - \sum_{i=1}^{N}\log P(l_i^{info}|h_i^{info})$$

$$L_{adv} = KL(u(l) \parallel \log P(l_i^{tab}|h_i^{tab})) + KL(u(l) \parallel \log P(l_i^{info}|h_i^{info}))$$

여기서 $l = \{l^{tab}, l^{info}\}$는 입력된 표가 일반적인 표인지 인포박스인지에 대한 라벨을 나타내며, $h = \{h^{tab}, h^{info}\}$는 판별기의 예측값을 나타낸다. 판별기는 L_D를 최소화하도록 학습되며, BERT 모형의 최종 손실값은 마스크 언어모형에 대한 손실 값과 적대적 학습의 손실 값을 더한, $L = L_{lm} + L_{adv}$이다.

4. 실험 및 결과

본 연구에서는 텍스트 데이터에 사전학습된 RoBERTa[8] base 모델[1]을 불러와서 표 데이터의 사전학습에 적용하였다. 학습 및 평가를 위한 데이터셋으로 KorQuAD 2.0 데이터셋에서 표에 정답이 태깅된 데이터를 추출하였으며, 다른 도메인의 표 질의응답 데이터의 실험을 위해서 제품 리뷰 사이트에서 추출한 표 데이터를 통해 제품의 스펙 데이터에 대한 질의응답셋을 구축한 스펙 표 질의응답 데이터셋을 이용하였다. KorQuAD 2.0의 데이터는 학습 및 개발 데이터셋 13,376개, 평가 데이터로 984개를 사용하였으며, 스펙 표 질의응답 데이터셋은 학습 및 개발 데이터셋 2,776개, 평가 데이터로 279개를 사용하였다.

사전학습의 경우, 약 50만 개의 나무위키 표 데이터 및 인포박스 데이터를 사용하였으며 학습 횟수는 2회, 학습률은 0.00001를 적용하였다.

파인튜닝의 경우, KorQuAD Table과 Spec 표 질의응답 데이터 모두 학습 횟수는 4회, 학습률은 0.00005를 적용하였다.

표 1. 사전학습 모형에 따른 KorQuAD 표 데이터에서의 기계독해 성능 비교

사전학습	파인튜닝 QA 데이터	KorQuAD Table	
		f1	EM
나무위키	KorQuAD	79.33	69.22
나무위키 정제	KorQuAD	**82.78**	**73.36**
적대적 학습	KorQuAD	80.41	70.21

표 2. 사전학습 모형에 따른 Spec 표 데이터에서의 기계독해 성능 비교

사전학습	파인튜닝 QA 데이터	Spec Table	
		f1	EM
나무위키	KorQuAD	37.13	23.10
	Spec	54.92	24.89
	KorQuAD->Spec	42.11	24.47
나무위키 정제	KorQuAD	43.08	33.55
	Spec	56.88	26.58
	KorQuAD->Spec	60.49	28.69
적대적 학습	KorQuAD	49.75	**35.90**
	Spec	57.93	25.73
	KorQuAD->Spec	**66.97**	31.64

표 1은 사전학습 방법에 따른 KorQuAD Table에서의 성능 비교를 나타낸다. 여기서 나무위키는 별다른 표나 텍스트 데이터의 정제없이 나무위키에서 추출한 표 데이터 전체를 사전학습에 사용하여 학습한 언어모형을 통해 튜닝한 모형을 나타낸다. 나무위키 정제는 의미없는 표 데이터를 걸러내는 의미 표 선별과 표와 가장 관련이 높은 텍스트를 선별하는 의미 텍스트 선별을 통해 구축한 사전학습 데이터를 이용한 모형을 의미한다. 적대적 학습은 선별된 데이터와 함께 인포박스 데이터와 일반적인

1) https://huggingface.co/klue/roberta-base

표 데이터에 적대적인 학습을 적용하여 사전학습한 언어 모형을 통해 파인튜닝한 모형을 의미한다.

표 2는 사전학습 방법에 따른 Spec 표 질의응답 데이터에서의 사전학습 방법 및 파인튜닝 데이터에 따른 성능 비교를 나타낸다. 여기서 “KorQuAD->Spec”은 사전학습된 언어모형을 KorQuAD 표 데이터로 먼저 파인튜닝 한 뒤, Spec 표 질의응답 데이터에 파인튜닝을 한 모형을 나타낸다.

표에서 구조적인 의미를 담고 있지 않고 시각적 효과를 위해서 사용된 장식 표 데이터들을 제거하고, 관련 텍스트 예측 태스크를 통해서 표 데이터와 관련성이 높은 텍스트들을 선별하여 구축한 사전학습 데이터를 통해 학습하였을 때 KorQuAD와 Spec 표 데이터에서 모두 향상된 성능을 보였다.

일반적인 표 데이터와 인포박스에 대하여 적대적인 학습 방법을 적용했을 때, KorQuAD 표 데이터에서의 성능은 떨어졌지만, Spec 표 데이터에서는 향상된 성능을 보였다.

5. 결론

본 연구에서는 도메인 변화에 강건한 표 질의응답 구축을 위한 사전학습 데이터 정제 방법 및 적대적 학습 방법을 제안하였다. 기존의 정제되지 않는 데이터로 학습했을 때와 비교하여 데이터를 정제하였을 때, KorQuAD 표 데이터에서 f1 3.45, EM 4.14가 증가하였으며, Spec 데이터에서 f1 19.38, EM 4.22가 증가하였다. 정제된 사전학습 데이터에 적대적 학습 방법을 추가했을 때, KorQuAD 표 데이터에서는 성능이 하락하였지만, Spec 표 데이터에서는 f1 6.48, EM 2.95가 증가한 성능을 보였다.

향후 연구에서는 더 다양한 도메인에 대한 표 질의응답 학습 및 평가 데이터를 구축하고 Spec 표 데이터 외에도 다양한 여러 도메인의 표 데이터에서도 강건한 기계독해 모형의 구축 및 평가에 대한 실험을 진행할 예정이다.

참고문헌

[1] YIN, Pengcheng, et al. TaBERT: Pretraining for joint understanding of textual and tabular data. arXiv preprint arXiv:2005.08314, 2020.

[2] HERZIG, Jonathan, et al. TaPas: Weakly supervised table parsing via pre-training. arXiv preprint arXiv:2004.02349, 2020.

[3] CHEN, Wenhu, et al. Tabfact: A large-scale dataset for table-based fact verification. arXiv preprint arXiv:1909.02164, 2019.

[4] EISENSCHLOS, Julian Martin; KRICHENE, Syrine; MÜLLER, Thomas. Understanding tables with intermediate pre-training. arXiv preprint arXiv:2010.00571, 2020.

[5] JUNG, Sung-Won; KWON, Hyuk-Chul. A scalable hybrid approach for extracting head components from web tables. IEEE Transactions on Knowledge and Data Engineering, 2005, 18.2: 174-187.

[6] LEE, Seanie; KIM, Donggyu; PARK, Jangwon. Domain-agnostic question-answering with adversarial training. MRQA@EMNLP, 2019.

[7] Devlin, Jacob et al, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019.

[8] LIU, Yinhan, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[9] 조상현, 김민호, 권혁철, "TAPAS를 이용한 사전학습 언어 모델 기반의 표 질의응답", 제32회 한글 및 한국어 정보처리 학술발표 논문집, pp.87-90, 2020.

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

[11] Tzeng, Eric, et al. "Adversarial discriminative domain adaptation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.