표 기계독해 언어 모형의 의미 검증을 위한 테스트 데이터셋

유재민* · 조상현 · 권혁철

부산대학교

Test Dataset for validating the meaning of Table Machine Reading Language Model

Jae-Min YU* · Sanghyun Cho· Hyuk-Chul Kwon

Pusan University

E-mail : lian6605@gmail.com / delosycho@gmail.com / hckwon@pusan.ac.kr

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 언어 모형 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.

ABSTRACT

In table Machine comprehension, the knowledge required for language models or the structural form of tables changes depending on the domain, showing a greater performance degradation compared to text data. In this paper, we propose a pre-learning data construction method and an adversarial learning method through meaningful tabular data selection for constructing a pre-learning table language model robust to these domain changes in table machine reading. In order to detect tabular data sed for decoration of web documents without structural information from the extracted table data, a rule through heuristic was defined to identify head data and select table data was applied. An adversarial learning method between tabular data and infobax data with knowledge information about entities was applied. When the data was refined compared to when it was trained with the existing unrefined data, F1 3.45 and EM 4.14 increased in the KorQuAD table data, and F1 19.38, EM 4.22 compared to when the data was not refined in the Spec table QA data showed increased performance.

키워드

기계독해, 표 기계독해, 사전학습 언어모형, 의미 데이터 선별

Ⅰ. 서 론

상품의 정보, 기업의 재무 회계, 정부의 정책 등 다양한 정보를 효율적으로 알리는데 사용되는 표는 구조를 가진 정형 데이터로 구조적인 해석하여 평문보다 정확하고 신속한 정보 전달을 활용된다.

이렇게 구조를 가지는 자료형인 표는 평문 기반의 지문 및 질문을 이해하고 답변을 도출하는 기존의 기계독해(Machine Reading Comprehension,

* speaker

MRC) 기법으로 해석하기에는 그 한계가 존재한다[6]. 따라서, 표 데이터를 기계독해 기술을 통해 해석하기 위한 새로운 기계독해 기법인 표 기계독해(Table MRC)가 요규되며 이를 위한 기법[6] 및 한국어 기반 학습 데이터셋(Dataset)[5]의 보강이 이루어져왔다. 그 중, 한국어 질의응답 쌍으로 구성된 KorQuAD2.0 데이터셋[5]은 기존 KorQuAD1.0 데이터셋으로 학습된 기계독해 모형이 할 수 없었던 표 및 리스트와 같은 구조를 가지는 데이터에 대해서도 질의에 대한 정답을 도출할 수 있게 하는 유의미한 결과를 보여주었다.

하지만, 위키피디아 문서(Wikipedia Article) 기반으로 구축된 KorQuAD2.0 데이터셋의 특성으로 인해 공개된 웹 및 DB 등에서 확인 가능한 국가 및 기업에 관련된 통계, 특정 제품의 성분 정보 등 일반 상식에 관련된 표 데이터 해석에서는 비교적 준수한 성능을 보이고 있으나 실제 중요한 업무들에 활용되는 문서의 해독 및 이해를 위해 활용되기에는 한계가 존재한다.

따라서, 본 논문은 정부에서 발표한 2018 행정업무운영 편람의 기준에 따라 공무서 데이터 80건을 수집하여 기계 독해 모델의 표 데이터 학습 및 해석 능력 검증을 위한 1,231건의 표 데이터 질의응답 쌍 데이터셋을 구축하였으며 실험을 통해 해당 데이터셋을 활용하여 학습된 기계독해 모델의 성능이 유의미하게 향상됨을 보인다.

본 논문의 구성은 다음과 같다. 2장에서는 기계독해 모델 학습을 위한 기존의 언어 모형에 대해 서술하며 3장에서는 제안하는 테스트 데이터셋 구축 기법과 질의응답 쌍에 대한 내용을 다룬다. 4장에서는 기존 모델과 제안하는 데이터셋을 활용하여 학습한 모델의 실험 결과를 보이며 5장에서는 결론을 서술하며 논문을 마친다.

II. 관련 연구

기계독해를 위한 데이터셋은 영어의 SQuAD가 있으며 한국어 데이터셋은 KorQuAD1.0&2.0, 그리고 NIA Dataset 등이 있다. 표 기계독해를 위한 한국어 데이터셋은 TabQA가 있다.

영어로 된 대표적인 표 기계독해의 경우 TableQA[12]가 있다. 모델을 통해 표 형식의 데이터에서 자연어 질의를 처리하고 응답하는데 특화되어 있다.

SQuAD 영어 위키 문서와 질문-답변 쌍으로 10만건의 데이터셋으로 질의응답에서 활용되고 있다. ETRI 공개 말뭉치로 SQuAD 한국어 데이터셋이 존재한다.

LG CNS에서 공개한 KorQuAD1.0은 SQuAD 구축기준에 맞추어 한국어 위키 문서를 통해 구축한 70,000+개의 질의응답 쌍으로 구성된 데이터셋으로 한국어를 바탕으로 구성된 질의응답 데이터셋이다. 이를 기준으로 KorQuAd2.0이 구축되었으며, 1.0과 달리 위키피디아 문서에서 페이지 전체를 대상으로 추출한 HTML Tag를 활용하여 질의응답 쌍을 구축하였다. KorQuAD2.0은 이전의 1.0 데이터셋과 달리 리스트와 표를 포함하여 구축되어 있다. KorQuAD2.0에서 구축한 표 데이터의 경우 HTML 태그에서 표와 관련된 태그를 살려 질의응답을 구성하였다. 기존의 1.0과 달리 위키백과 페이지 전체를 지문으로 보며, 지문의 길이가 긴 경우 단락을 나누어 질의응답쌍을 구성하였다. 답변 역시 한 두 문장이 아닌, 문단, 표, 리스트를 포괄한다.

표 1 기계독해를 위한 데이터셋

언어	데이터셋 이름	규모(개)	
영어	SQuAD 1.0&2.0	107,702	151,054
한국어	KorQuAd 1.0&2.0	66,181	102,960
한국어	Nia Dataset	약 450,000	
한국어 표	TabQA	약 100,000	
한국어 논문 초록	KorSciQA	2,490	

표 기계독해와 관련된 연구로 조상현[8]의 표와 관련된 기계독해 질의응답을 다루고 있다. 한국어 기반의 BERT와 TAPAS를 이용한 표 질의응답 모형을 제안하였다.

본 연구에서는 조상현[8]의 모델을 활용하여 표 기계독해 모델에서 공문서와 관련된 테스트 데이터셋을 구축하여 특정 분야에서의 성능을 실험하였다.

III. 테스트 데이터셋

3.1 데이터셋 생성 방식

데이터셋 생성 방식은 데이터 획득, 데이터 정제, 라벨링, 생성의 과정을 거친다.

데이터 획득이란 '원시 데이터'를 확보하는 과정으로 정부에서 배포하는 공문서와 부산대학교 강의계획서 PDF를 활용하였다. 데이터 정제 과정에서는 PDF를 HTML 파일로 변환한 후 KorQuAd의 HTML 형식에 맞게 Table %Tag를 제외한 모든 Tag를 제거하였다. 이후 정답 태깅을 위한 라벨링 과정을 거친 후 질의응답 쌍을 생성하였다.

데이터셋의 신뢰성을 위한 검수 과정은 세 명의 작업자와 한 명의 검토자가 교차 검증을 통하여 질의유형과 정답, 그리고 태깅을 검토하였다.

Office_data는 평문과 표가 함께 나타나는 공문서 pdf를 웹 문서 독해를 위한 HTML_TAG로 변환한 후 기계독해를 위한 질의응답 쌍을 구축하였다. 데이터셋은 데이터 전처리, 정답 태깅, 등을 모두 포함한 데이터셋으로 의의를 지닌다.

3.2 데이터셋 질문 유형

KorQuAD2.0의 경우 표와 리스트가 전체 질의응답 쌍의 27.7%를 차지하지만 이와 관련된 질의응답 유형에 대한 분류는 없기에 기존 1.0의 질문 유형을 기준으로 벤치마킹하였다.

사전학습 언어 모형의 기본 데이터셋인 KorQuAD 2.0의 질문 유형에 맞게 여섯 가지로 분류한 후 1,221개의 질의응답 쌍을 구성하였다.

표 2 Office_data 질문 유형 분류

유형1	어휘포함	테이블에 있는 어휘가 질문에 포함되는 유형
유형2	어휘변형	테이블에 있는 어휘가 변형되어 질문에 나타남
유형3	다중근거	여러 개의 셀을 살펴보고 정답 도출
유형4	대소비교	여러 개의 셀을 비교하여 숫자 관련 지식 개념이 있어야 정답 도출
유형5	순서비교	테이블 내에서 순서와 관련된 시간적 지식 개념이 있어야 정답 도출
유형6	Cell_Count	`Cell을 계산하여 정답 처리

질문 유형에 따른 분포는 다음 표와 같다. 어휘를 포함하거나 변형하는 유형이 다수를 차지하며, 숫자를 비교하거나 계산해야 하는 유형이 그 후를 차지하였다. 전후 순서의 개념을 이해해야 하는 유형이 가장 적게 분포하였다.

표 3 Office_data 구축량 및 유형별 분포

	유형 1	유형 2	유형 3	유형 4	유형 5	유형 6
개수	420	225	125	30	36	22
비율(%)	48.49	25.98	14.43	3.4	4.1	2.5

Ⅳ. 실험 및 결과

Office_데이터셋의 성능 측정은 KorQuAD 데이터셋에서 학습한 BERT기반 TAPAS 기계독해 모델을 이용하여 Office_data 기계독해 평가셋에서의 일반화 성능을 평가하였다. 실험에 사용한 질의응답은 총 859건으로 구축한 데이터셋만을 대상으로 한 실험과 KorQuAD에 추가하여 따로 실험을 진행하였다.

실험모델의 경우 [8]의 모델을 차용하였으며 이는 텍스트에 사전학습된 언어 모델을 기반으로 테이블에 적합한 TAPAS 언어 모델을 학습한 것이다. BERT-base 설정인 은닉 차원 : 768, 은닉 계층 수 : 12, 어텐션 헤드 : 12를 사용하였다.

기계독해 성능 지표로 Exact Match와 음절단위 F-1 Score를 사용하였다. F1-Score는 실제 정답과 출력 응답 간의 부분 일치를 나타내며, EM-Score는 모델이 출력한 응답과 실제 정답이 완전히 일치하는 경우를 나타낸다.

실험 결과 KorQuAD를 기반으로 한 기본 모델과 Office_data의 학습 결과 그리고 두 가지를 합한 학습 결과는 다음과 같다. 총 859건의 질의응답 쌍을 대상으로 하였다.

표 4 실험 결과

EM/F1	BERT		Ko_Electra	
	EM(%)	F1(%)	EM(%)	F1(%)
KorQuAD	57.33	75.87	**73.12**	81.94
Office_data	32.81	51.79	39.43	59.06
KorQuAD+Office_data	**66.19**	**76.23**	72.87	**81.95**

Office_data만을 실험한 결과는 기존의 BaseLine 모델인 KorQuAD기반의 데이터셋보다 성능이 좋지 않았는데 이는 데이터 질의응답의 부족에 기인한다고 판단되었다. KorQuAD모델에 Office_data를 추가한 결과는 EM-Score가 8.86% 향상되었으며, F1-Score는 0.36% 향상되었다. 특이한 점은 한국어의 경우 EM-Score보다 음절 단위의 F1-Score를 구하지만, BERT에서 표 기계독해 테스트셋 실험 결과 F1-Score가 아닌 EM-Score에서 큰 상승폭을 보인 점인데 이는 Cell 단위로 정답을 추출하는 TAPAS 언어모형에서 원인을 찾을 수 있다고 생각한다.

표 5 유형별 실험 결과(Office_data)

	BERT		Ko_Electra	
	EM(%)	F1(%)	EM(%)	F1(%)
유형1	24.27	46.53	41.41	64.11
유형2	22.65	38.61	42.20	58.76
유형3	11.99	34.35	36.47	56.23
유형4	16.61	31.48	46.51	57.38
유형5	19.39	29.53	33.24	42.31
유형6	18.09	39.10	45.24	57.48

유형별 실험 결과는 유형별 데이터셋의 개수에 비례하여 성능의 차이가 드러났다. 다른 유형에 비하여 유형5 '순서비교'가 성능이 좋지 않았는데 이는 순서와 관련된 시간적 개념이 부족한 것으로 유추한다.

Ⅴ. 결 론

본 논문에서는 공문서와 관련된 질의응답을 위해 Office_데이터셋을 정의하고 수집하여 구축하였다. KorQuAD데이터셋에 Office_데이터셋을 추가하여 BERT와 TAPAS모델에 학습시킨 결과 표를 대상으로 하는 기계독해의 성능 향상이 이루어짐을 실험을 통해 확인할 수 있었다. 이번 연구에서는 적은 양의 데이터셋을 구축하였고 질문 유형의 비율이 제한적인 단점이 있었다. 따라서 향후 연구에서는 데이터를 추가로 수집하여 구축하고, 다양한 분야의 공문서에서 나올만한 질의응답 쌍을 유형별로 분류하여 표 기계 독해의 성능을 향상시킬 수 있도록 연구를 진행할 예정이다.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT : Pretraining of Deep Bidirectional T ransformers for Language Understanding”, NAACL 2019.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. “Attention Is all you need,” Advances in neural information processing systems, pp. 5998-6008, 2017.

[3] Pranav Rajpurkar, Robin Jia, Percy Liang. “Know What You don’t Know: Unanswerable Questions for SQuAD”. ACL2018

[4] 임승영, 김영지, 이주열. “KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋”. 한국정보과학회 학술발표논문집, pp. 539-541, 2018.

[5] 김영민, 임승영, 이현정. “KorQuAD 2.0: 웹문서 기계독해를 위한 한국어 질의응답 데이터셋”. 제 31회 한글 및 한국어 정보처리 학술대회 논문집, 2019.

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

[7] 함영균, 정용빈, 정희석, 황혜경, 최기선. “KorSciQA: 한국어 논문의 기계독해 데이터셋”. 제 31회 한글 및 한국어 정보처리 학술대회 논문집(2019년).

[8] 조상현, 김민호, 권혁철. “TAPAS를 이용한 사전학습 언어 모델 기반의 표 질의응답”. 제 32회 한글 및 한국어 정보처리 학술대회 논문집(2020년).