ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 50, No. 12, pp. 1102-1110, 2023. 12
https://doi.org/10.5626/JOK.2023.50.12.1102

다양한 도메인 변화에 강건한 한국어 표 기계독해

(Robust Korean Table Machine Reading Comprehension across Various Domains)

조 상 현 † 김 혜 린 † 권 혁 철 ††
(Sanghyun Cho) (Hye-Lynn Kim) (Hyuk-chul Kwon)

요 약 표 데이터는 일반적인 텍스트 데이터와 다르게 구조적인 특장점으로 정보를 압축해 표현할 수 있다. 이는 표가 다양한 도메인에서 활용되는 것으로 이어지며, 기계독해 영역에서의 표 기계독해 능력이 차지하는 비중은 점점 커지고 있다. 하지만 도메인마다 표의 구조와 요구되는 지식이 달라 언어 모델을 단일 도메인으로 학습했을 때 다른 도메인에서의 모델의 평가 성능이 하락해 일반화 성능이 낮게 나타날 가능성이 크다. 이를 극복하기 위해서는 다양한 도메인의 데이터셋 구축이 우선이 되어야 하며, 단순 사전학습한 모델이 아닌 다양한 기법을 적용하는 것이 중요하다. 본 연구에서는 도메인 일반화 성능을 높이기 위해 도메인 간 불변하는 언어적 특성(Invariant-feature)을 학습하는 언어 모델을 설계한다. 각 도메인별 평가 데이터셋에서의 성능을 높이기 위해서 적대적 학습을 이용하는 방법과 표 데이터에 특화된 임베딩 레이어와 트랜스포머 레이어를 추가하는 모델의 구조를 변형하는 방법을 적용하였다. 적대적 학습을 적용했을 때는 표와 관련된 특화된 임베딩을 추가하지 않는 구조의 모델에서 성능이 향상되는 것을 확인했으며, 표에 특화된 트랜스포머 레이어를 추가하고 추가된 레이어가 표에 특화된 임베딩을 추가로 입력받도록 했을 때, 모든 도메인의 데이터에서 가장 향상된 성능을 보였다.

키워드: 기계독해, 표 질의응답, 합성 데이터 생성, 도메인 적응, 도메인 일반화

Abstract Unlike regular text data, tabular data has structural features that allow it to represent compressed information. This has led to their use in a variety of domains, and machine reading comprehension of tables has become an increasingly important aspect of Machine Reading Comprehension(MRC). However, the structure of tables and the knowledge required for each domain are different, and when a language model is trained for a single domain, the evaluation performance of the model in other domains is likely to be reduced, resulting in poor generalization performance. To overcome this, it is important to build datasets of various domains and apply various techniques rather than simply pre-trained models. In this study, we design a language model that learns cross-domain invariant linguistic features to improve domain generalization performance. We applied adversarial training to improve performance on evaluation datasets in each domain and modify the structure of the model by adding an embedding layer and a transformer layer specialized for tabular data. When

· 이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음

† 학생회원 : 부산대학교 정보융합공학과 학생
delosycho@gmail.com
helen6339@naver.com
†† 종신회원 : 부산대학교 정보컴퓨터공학부 교수(Pusan Nat'l Univ.)
hckwon@pusan.ac.kr
(Corresponding author임)

논문접수 : 2023년 8월 9일
(Received 9 August 2023)
논문수정 : 2023년 9월 5일
(Revised 5 September 2023)
심사완료 : 2023년 9월 6일
(Accepted 6 September 2023)

정보과학회논문지 제50권 제12호(2023. 12)

applying adversarial learning, we found that the model with a structure that does not add table-specific embeddings improves performance. On the other hand, while adding a table-specific transformer layer and having the added layer receive additional table-specific embeddings as input, shows the best performance on data from all domains.

Keywords: machine reading comprehension, table question answering, domain adaptation, domain generalization

1. 서 론

기계독해(Machine Reading Comprehension, MRC)는 자연어처리 태스크 중 하나로서 주어진 본문을 기반으로 질문에 대한 답을 찾는다. 일반적으로 기계독해를 수행할 수 있는 신경망 모델은 BERT[1]같은 인코더 기반 모델인데, 이러한 모델은 사람이 구축한 기계독해 데이터셋을 학습하여 동작하게 된다. 이전의 기계독해는 텍스트같은 비정형 데이터를 중점적으로 동작할 수 있도록 하였지만, 최근에는 데이터를 효율적으로 표현하기 위해 밀도 있고 압축된 형태의 표 데이터 구조를 자주 사용하게 되었다. 자연스럽게 표 데이터가 문서에서 차지하는 비중이 커지게 되었고, 이는 표 데이터에 대한 표 기계독해(Table Machine Reading Comprehension, Table MRC) 능력이 필수적으로 요구되도록 하였다.

앞서 언급했듯이, 표 데이터는 다양한 문서에서 자주 등장한다. 도메인마다 등장하는 표 데이터는 다양한 형태 및 크기를 가질 뿐만 아니라, 일반적으로 언급되지 않는 전문적인 용어(법령, 공식, 단위, 수치 등의 용어)를 사용하고 있다. 표 1, 2, 3은 법령, 공문서, 제품 스펙에서 나타나는 표 데이터의 예시이며, 이를 통해 세 가지 도메인의 표 데이터 간 차이를 확인할 수 있다. 이 차이는 곧 도메인마다 표를 해석하는 방식 및 필요로 하는 지식이 달라짐을 의미하고, 단순 한 가지 도메인에서 학습된 표 기계독해는 다른 도메인에 대해서 성능이 하락할 가능성을 가진다[2]. 그러므로 표 기계독해 모델을 학습하기 위해서는 다양한 도메인의 표 기계독해 데이터셋을 구축할 필요성이 있다.

그러나 모든 도메인에 대해서 데이터셋을 구축하는 것은 분명히 한계가 존재하기 때문에, 제한된 도메인의 학습 데이터셋으로 신경망 모델이 표 기계독해를 수행하도록 해야한다. 이때 모델이 데이터셋을 학습하여 태스크에 대해 수행 가능한 상태가 되더라도 학습한 도메인 외의 테스트 데이터셋을 만나게 된다면, 학습 데이터와 테스트 데이터 간의 특징, 타겟값의 분포가 달라지는 도메인 변화(Domain shift)가 일어나기 때문에 모델의 성능 저하를 피할 수 없다.[3] 따라서 기계독해를 수행하는 신경망 모델에게는 다른 도메인에도 대응할 수 있는 능력이 분명히 필요하다.

모델이 학습 데이터셋과 테스트 데이터셋 간의 도메인 변화를 완화할 수 있는 방식에는 대표적으로 두 가지가 있다. 첫 번째는 도메인 적응(Domain Adaptation)이다. 도메인 적응은 학습 데이터셋의 도메인인 소스 도메인(source domain)의 정보를 우리가 테스트하고자 하는 도메인인 타겟 도메인(target domain)으로 옮기는 기법을 말한다. 이 기법은 모델의 decision boundary를 타겟 도메인으로 이동시켜 적응시키기 때문에 오히려 기

표 1 법령 표 데이터 예시

Table 1 Example of law table data

제5조(위원회 구성)	
① 위원회는 위원장 1명과 … 중에서 호선한다.	
② 당연직 위원은 기획예산담당관 … 포함할 수 있다.	
③ 위촉직 위원은 「양성평등기본법」 제21조에 따라 다음 각 호 중 군수가 위촉한다.	의령군의회 의원
	2. 학계 전문가 및 언론인
	3. 관광 및 문화예술에 관한 풍부한 경험과 식견을 갖춘 사람

표 2 공문서 표 데이터 예시

Table 2 Example of office table data

Strategy	Detailed	Key takeaways
Special Videos Special Facilities and Production Advancements	Custom facilities	Specialised unloading facilities for large equipment
	Build advanced equipment	Next-generation video data, motion capture, and versatile robot control
	One-Stop Service	Production process

표 3 제품 스펙 표 데이터 예시

Table 3 Example of spec table data

Model Code	EO-IA500BBEGKR	
Sound	Impedance	32 ohm
	Frequency Response	20 khz
	Sensitivity	93.2 dB
Connectivity	3.5mm Headphone Jack	Yes
General	Connectivity	3 Buttons(Play/Pause, Volume Up/Down)
	Mic	Yes
	Packages	Eartops (small, medium, large)

존의 소스 도메인에서의 성능을 기대하기 어려운 단점을 가지고 있다. 두 번째는 도메인 일반화(Domian Generalization)이다. 도메인 일반화는 앞선 도메인 적응처럼 소스 도메인을 타겟 도메인에 적응시키는 것이 아닌 모든 도메인에 대해 적용할 수 있도록 일반화하는 기법을 일컫는다. 주로 학습 단계에서 각 도메인의 불변하는 특징(invariant feature)을 추출하여 학습에 이용하는 알고리즘을 주로 사용한다[4]. 이를 통해 모델은 특정 도메인에 적응하거나 데이터셋의 형태에 과적합(over-fitting)을 피해서 다양한 도메인에서 좋은 성능을 낼 수 있게 한다.

우리가 실제로 기계독해를 위한 신경망 모델을 사용하는 데에 있어서, 학습할 때 전혀 보지 못한 도메인을 만나는 경우는 빈번하게 존재하기 때문에 도메인 일반화가 도메인 적응보다 더 현실적이고 실용적으로 문제를 해결할 수 있는 방법이다. 따라서 본 연구에서는 표 기계독해 모델의 도메인 일반화 능력을 GAN[5]의 discriminator의 개념을 차용하여 도메인의 불변하는 특징의 학습을 통해서 구현해보고자 한다.

또한, 이전의 연구[13-16]를 통하여 모델 구조나 임베딩의 추가로 표 기계독해에 대한 성능 향상이나 강건성을 심화시킬 수 있었다. 따라서 언급한 특정한 기법의 적용에 추가로, 표 기계독해를 위한 새로운 모델 구조와 특화된 임베딩을 제안하여 실험하고 모델 성능의 변화와 차이에 대해서 관찰한다.

2. 관련 연구

2.1 한국어 표 기계독해 데이터셋

표 기계독해 연구는 한국어 기계독해 벤치마크 데이터셋인 KorQuAD 2.0[6]의 공개를 시작으로 활성화되었다. KorQuAD 1.0과 다르게 텍스트 뿐만 아니라 웹 문서에 존재하는 표나 리스트에 태깅된 정답을 추가하였다. 이후 위키피디아 문서를 기반하는 한국어 표 기계독해 데이터셋인 KorWikiTabularQuestion[7]과 AIHUB에서 행정 데이터 표 기계독해 데이터셋을 공개하면서 연구에 더욱 박차를 가했다.

물론 앞서 언급한 것처럼 다양한 한국어 표 기계독해 데이터셋이 공개되어왔지만, 영어 표 기계독해 데이터셋에 비하면 여전히 부족한 양이다. 영어 표 기계독해 데이터셋에는 FinQA(금융)[8], TabMCQ(과학 시험)[9], TAT-QA(금융)[10], AIT-QA[11]와 같이 특정 도메인에 특화된 다양한 표 기계독해 데이터셋이 공개되며 모델이 다양한 도메인을 학습할 수 있는 기반이 된다. 한국어 표 기계독해는 영어 데이터셋에 비해 상대적으로 데이터셋 양이 부족하기 때문에, 학습하지 않은 도메인에 대응할 수 있는 도메인 일반화 능력을 더욱 필요로 한다.

2.2 표 기계독해 모델

이전에도 국내외에서 표 데이터를 활용하기 위해서 다양한 관점으로 연구가 진행되어왔다.

Tapas[14]는 사전학습 기반의 약한 감독(Weak Supervision) 방식 표 질의응답 모델을 제안했다. Tapas는 기존 BERT를 확장한 모델로서, 표를 위한 몇 가지 특수 임베딩을 추가했다. 추가된 임베딩은 총 3가지인데, 이들은 표의 셀 간의 비교를 통해 크기 순위를 나타내는 랭킹 임베딩, 셀이 속한 행을 의미하는 행 임베딩, 셀이 속한 열을 나타내는 열 임베딩이 있다. Tapas는 영문 위키피디아 문서에서 추출한 약 620만 개의 테이블 데이터를 이용하여 사전학습 되었다. 또한, 정답 도출을 위해 연산 과정이 필요한데 필요한 셀에 대한 정보가 제공되지 않은 경우, 표 질의응답 데이터의 학습을 위해 선택된 셀들의 연산값(count, summation, average)과 데이터셋의 학습 레이블간의 오차에 대한 Hubber Loss 값을 줄이도록 학습하는 약한 감독 학습 방법도 사용하였다. 이를 통해 Tapas는 영문 표 질의응답 데이터셋인 WIKISQL, WikiTableQustions, SQA에서 모두 높은 성능을 확인하였다.

또 다른 연구로는 표 기계독해 모델의 강건성에 대해 집중한 Tableformer[16]가 있다. 이전엔 제안된 모델들[14,15]은 같은 표 데이터에 대해서 행이나 열의 순서를 바꾸는 변형을 가하면 성능이 떨어지거나, 주어진 표 데이터와 질문에 대해서 구조적인 편향(structural bias)를 제대로 학습하지 못해 성능이 떨어지는 약점을 보였다. 이를 극복하기 위해 Tableformer에서는 두 가지 새로운 방법론을 제안했다. 첫 번째는 단순한 위치 임베딩(Position Embedding)이 아닌 셀에 따른 위치 임베딩(Per Cell Position Embedding)을 제안하여 행과 열의 변화에 더욱 강건해지도록 했다. 두 번째는 셀 하나가 절대적으로 어디에 위치하는지(Absolute Order)가 아닌 셀끼리 상대적으로 어디에 위치하는지(Relative Order)를 학습하도록 하여 구조적인 편향을 제대로 학습하게 했다.

위와 같은 다양한 연구들에서 표 데이터에 대한 질의응답을 가능하게, 그리고 더욱 강건하게 할 방법에 관해서 연구했다. 본 논문에서는 언급된 연구들을 이어받아 표 데이터를 더욱 효과적으로 해석하는 것과 더불어 다양한 형태와 도메인의 표 데이터에 대해 강건하게 동작할 수 있는 모델 구조와 방법론을 제안한다.

2.3 도메인 일반화

현재까지 도메인 변화 시에 모델의 성능 하락을 방지하기 위해서 다양한 연구가 진행되었다. 텍스트 기계독

해에서의 도메인 일반화 모델 구축 연구[12]에서 저자는 언어 모델이 도메인 간 변하지 않는 언어적 특성(Domain-Invariant Feature Representation)을 학습하게 된다면, 모델이 도메인 변화에서 자유로울 수 있음을 언급했다. 이를 위해 언어 모델을 구축한 뒤, 모델의 최종 출력층에 숨겨진 표현(hidden represetnation)을 통해 도메인을 구별하는 Discriminator를 구현했다. 이를 통해 언어 모델이 기계독해를 수행할 때, 예측 정답과 실제 정답 간의 오차는 줄여서 정답을 잘 예측하도록 하고 Discriminator가 예측하는 도메인과 실제 도메인 간의 오차는 유지하려는 작업을 통해 교란시킨다. 이 작업을 통해 Discriminator는 데이터셋의 도메인을 정확하게 구별하지 않게 되며, 도메인에 구애받지 않고 도메인 간 변하지 않는 특성을 학습해 6개의 학습하지 않은 도메인에서 좋은 성능을 보였다.

본 연구에서도 모델이 도메인 간 변하지 않는 특성을 학습하는 기법을 차용하여 6개의 도메인에서 언어 모델이 일반화된 능력을 보여줄 수 있는지 확인해보려 한다.

3. 표 기계독해 시스템

그림 1은 본 논문에서 제안하는 표 기계독해 모델의 구조를 나타낸다. 우리는 입력되는 표 데이터의 도메인 변화에 잘 적응하는 모델을 구현하기 위해서 모델 구조 그리고 학습 방법의 두 가지 주요 관점에서 비교 분석을 수행하였다. 모델 구조의 경우, 모델에 표에 특화된 새로운 임베딩을 추가하는 방법과 표에 특화된 인코딩 레이어를 추가하는 두 가지 실험을 진행하였다. 표에 특화된 인코딩 레이어는 BERT의 출력 벡터를 이용해서 행과 열의 표현 벡터를 생성하여 정답 예측에 활용했던 이전 연구[13]를 베이스라인 모델로 설정하고 모델의 추가적인 변형을 하였다. 표-특화 임베딩 실험에서는 Tapas[14]에서 표의 구조적인 정보를 포착하기 위해 추가한 특수 임베딩을 추가하고 도메인 일반화에 어떤 영향을 주는지 비교하는 실험을 진행하였다. 학습 방법에서는 이전 연구[12]에서 기계독해의 도메인 일반화를 위해서 적용했던 방법을 따라서 적대적 학습을 표 기계독해에 적용하였다.

3.1 표 특화 임베딩

이전 연구[13]에서는 사전학습된 언어모형이 표의 구조적인 정보를 포착하게 하도록 다양한 추가 임베딩을 이용해왔다. 그림 2는 표 특화 임베딩의 예시를 나타낸다. Tapas에서는 표의 행, 열 위치를 나타내는 임베딩과 순위 임베딩을 추가하였다. 순위 임베딩은 입력 표에서 순위를 매길 수 있는 열에 한해서 행별로 순위를 매기고 해당 순위를 위치 임베딩과 같이 추가 임베딩으로 사용한 것이다. [13]에서는 인명, 나라 이름, 숫자, 날짜, 시간 금액 등의 입력 토큰의 엔티티를 태깅하고 해당 엔티티 정보를 추가 임베딩으로 사용하였다. 하지만 엔티티 임베딩의 경우, 단일 도메인의 데이터로 학습 및 평가를 했을 때는 성능의 향상이 존재했지만 다양한 도

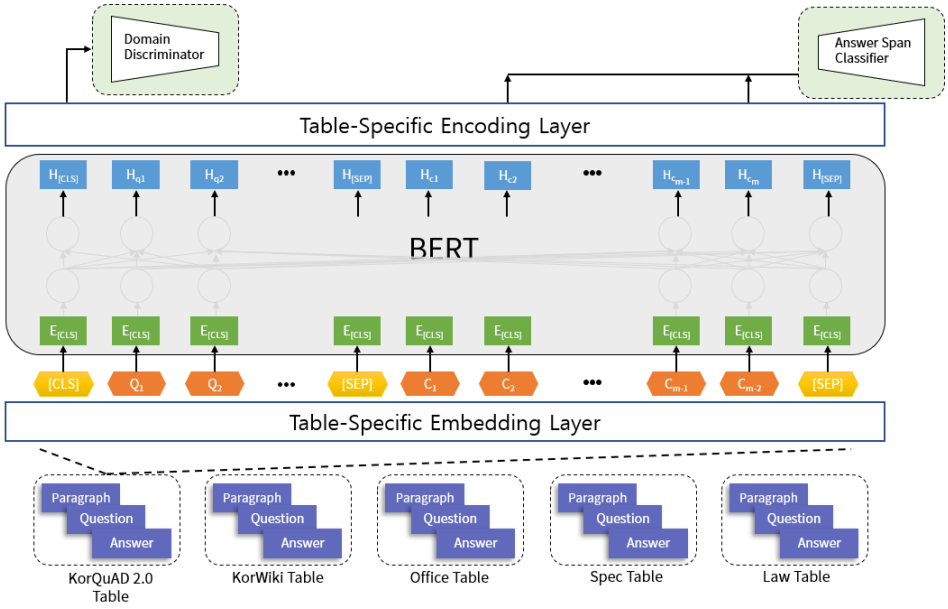

그림 1 제안 표 기계독해 모델 전체적 구조

Fig. 1 The entire structure of the proposed MRC model

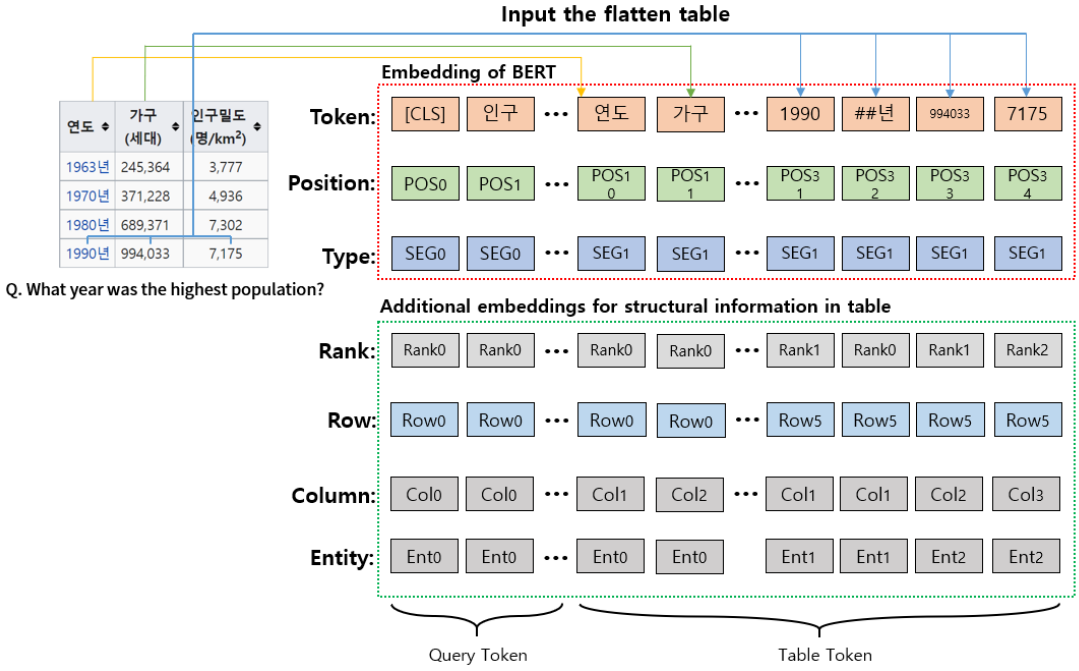

그림 2 표 가계독해 특화 임베딩 예시

Fig. 2 Example of specialized embedding of Table MRC

메인의 데이터를 동시에 학습하는 경우 오히려 성능이 하락하는 것을 보였다. 이에 본 연구에서는 추가 임베딩 실험에서 Tapas에서 적용했던 행, 열, 순위 임베딩을 적용하였다.

3.2 표 특화 인코딩 레이어

그림 3은 본 연구의 베이스라인 표 기계독해 모델[13]의 구조를 나타낸다. 이전 연구[13]에서는 임베딩 레이어에서 표의 구조적인 정보를 인코딩하기 위한 행과 열의 추가적인 임베딩을 이용하였다. 베이스라인 모델에서는 임베딩 레이어에서 행, 열 임베딩을 추가하는 방법을 사용하지 않고 언어모형에서 출력된 표현 벡터를 행과 열의 단위로 합산하여 행과 열의 표현 벡터를 만들었다. 이 방법을 이용하게 되면 기존의 방법과 비교하여 두 가지 장점을 가지게 된다. 첫 번째는 TableFormer[16]에서 언급했던 기존 임베딩 방식에서 행과 열의 순서 변형에 취약한 문제를 해결할 수 있다. 두 번째는 다양한 도메인의 표 데이터를 입력으로 사용하는 경우, 법령, 제품 정보, 행정 용어 등과 같이 특정 도메인의 전문적인 지식을 필요로 하는 용어들이 많이 등장하게 된다. 자연어 텍스트를 이용한 언어모형의 사전학습 과정에서는 이러한 도메인의 지식을 담고 있거나 배경으로 하는 문서나 텍스트가 다수 포함될 수 있지만, 표와 같은 반-구조화된 데이터에서 다루고 있는 정보의 종류나 도메인은 제한되어있는 경우가 많다. 사전학습된 언어모형의 임베딩 레이어에 새로운 임베딩을 임의로 추가하게 되면 텍스트 데이터에 사전학습된 언어모형의 가중치를 그대로 가져와서 적용하더라도 추가된 임베딩으로

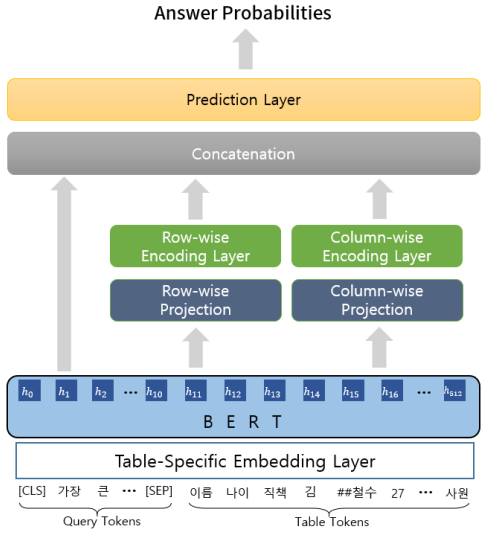

그림 3 베이스라인 표 기계독해 모델 구조

Fig. 3 Structure of the baseline model

인해서 기존의 모형이 변형될 수 있다. 이에 두 번째 장점으로 해당 베이스라인 모델에서는 임베딩 레이어에서 표의 추가적인 임베딩을 사용하지 않기 때문에 이러한 문제를 해결할 수 있다.

본 연구에서는 도메인 일반화에 영향을 주는 기계독해 모델의 영향을 다양하게 비교하기 위해서 베이스라인 모델의 구조에 추가 표 특화 레이어를 추가하였다. 그림 4는 레이어를 추가한 모델의 전체 구조를 나타낸다. 추가한 표 특화 트랜스포머 레이어에는 MATE[15]에서 제안했던 어텐션 마스크 방법을 적용했는데, 같은 행, 열에 존재하거나 질문과 관련된 토큰일 경우에만 어텐션을 적용하도록 하였다.

3.3 적대적 학습

적대적 학습(Adversarial Training)의 근간은 적대적 공격(Adversarial Attack)으로부터 시작된다. 적대적 공격은 학습된 모델이 가지는 허점을 이용하여 모델이 잘못된 예측을 하도록 유도하는 것이다. 이를 이용하여 모델의 학습 단계에서 오히려 취약한 부분을 공략하여 극복할 수 있도록 하는 기술을 적대적 학습이라고 한다.

적대적 학습을 활용하는 대표적인 기술은 GAN(Generative Adversarial Network)이다. 이는 실제에 가까운 이미지나 사람이 쓴 것과 같은 가짜 데이터를 생성하는 모델이다. GAN은 서로 다른 두 개의 네트워크인 Generator(생성기)와 Discriminator(판별기)를 적대적으로 학습시키며 실제와 더욱 비슷한 데이터를 생성하려고 한다.

이를 기반으로 본 연구에서는 기계독해 모델의 첫 번째 숨겨진 표현(hidden representation)을 통해 도메인을 구별하는 Discriminator를 구현한다. 그리고 여기에 Discriminator를 교란시키기 위해 모든 도메인의 확률이 동일한 분포인 균등 분포(uniform distribution)를 생성하여 이와 Discriminator의 결과가 가까워지도록 적대적으로 학습시킨다. 이를 통해 Discriminator는 특정 도메인에 대한 구별보다, 도메인 간 변하지 않는 언어적 특성을 학습하여 다양한 도메인에 대한 강건성을 가질 수 있게 된다.

4. 표 질의응답 데이터셋

사용하는 데이터셋은 법령 표 데이터셋, 공문서 표 데이터셋, 제품 스펙 표 데이터셋[17], KorWikiTabular Questions, KorQuAD 2.0, 행정분야 표 데이터셋이다.

법령 표 데이터셋은 '국가법령정보센터'에 존재하는 안전 법령 문서를 기반으로 표 데이터를 수집했다. 법령 데이터의 경우 표 형태보다 긴 줄글 형태로 표현된 경우가 빈번하게 발생하여 분리한 뒤 표 형태로 재구성하여 평가 데이터셋을 생성하였기 때문에 단순히 기존 데이터만을 이용하는 것보다 풍성하게 데이터셋을 구축되어있다. 대부분 복잡한 법률 용어를 사용하며, 데이터의 특징을 살려 범위나 수치 관련 정보에 강화된 질문이 생성되었다. 데이터셋 형식은 [질문|답변|사용한 문서|문서 내 sheet|답변 셀 위치]로 구성되어있다. 총 2,005개의 데이터를 가지고 있다.

공문서 표 데이터셋은 '공공데이터포털'외 5개의 도메인을 기반으로 표 데이터를 수집했다. 공문서 도메인에서는 매우 다양한 형태의 표가 존재하기 때문에 의미가 부족한 테이블 수집은 지양하고, 적절한 크기의 데이터 선별을 위해 4행 4열 미만의 표 데이터는 제외했다. 이를 바탕으로 복잡한 조건의 질문 유형이 다수 구축되어 있다. 데이터셋 형식은 [사용한 문서|질문|답변|답변 셀 행 위치|답변 셀 열 위치]로 구성되어있다. 총 10,500개의 데이터를 가지고 있다.

제품 스펙 표 데이터셋의 경우, '다나와'를 기반으로 표 데이터를 수집하였다. 제품 스펙 데이터의 표 형태는 대부분 비슷하지만, 제품 카테고리마다 사용되는 용어가 전문적이고 특수하다. 따라서 복잡한 특수 용어를 이용하여 비교하는 질문을 중심으로 생성했다. 데이터셋 형식은 [카테고리|사용한 문서 이름|질문 유형|질문|답변|답변 셀 위치]로 구성되어있다. 총 3,055개의 데이터를 가지고 있다.

KorWikiTableQuestions, KorWikiTQ.는 한국어 위키피디아를 기반으로 반구조화된 데이터를 수집한 질문-답변 데이터셋이다. 위키피디아에는 방대한 양의 표 데이터가 존재하고, 다양한 형태와 크기를 가지기 때문에 최소 8행 5열의 표를 수집하였다. 크라우드 작업자들이 표에 대한 질문을 직접 작성하였기 때문에, 사전에 정의된 템플릿이 아닌 다양한 구조를 보여주어 높은 언어적 범위를 보여주었다. 총 22,033개의 데이터를 가지고 있다.

KorQuAD 2.0은 한국어 Wikipedia article 전체에서 답을 찾아야하는 기계독해 데이터셋이다. 비정형 텍스트 데이터와 표 데이터에 대한 기계독해 데이터셋 모두 가지고 있다. 학습 데이터 83,485개, 평가 데이터 10,165개로 구성되어있다.

행정분야 표 데이터셋[1]은 비정형 데이터인 텍스트 데이터를 이용하여 표와 일반 텍스트 데이터에 대한 다양한 형식을 가진 질의응답 데이터셋이다. '공공데이터포털'과 공공기관 보유 행정문서를 기반으로 데이터를 추출하였다. 데이터 구성은 [표-질문-답변]으로 구성되어있다. 이중 원천 데이터 65,534건을 이용하여 총 131,068개의 질의응답 데이터를 가지고 있다. 본 연구에서는 도메

1) https://aihub.or.kr/

표 4 도메인 일반화 실험 결과 (EM/F1)
Table 4 Test result of domain generalization

Model \ Dataset	Law dataset	Spec dataset	office dataset	KorWikitq	Korquad 2.0 (Table)	Administrative dataset
Vanilla	17.33/36.63	28.00/50.90	27.20/35.34	35.57/40.33	60.00/69.73	25.69/29.42
Projection	36.66/72.56	64.66/76.90	72.01/81.82	91.05/94.95	72.99/79.98	86.25/90.39
TAPAS	36.66/73.11	73.33/82.75	71.72/81.93	90.62/94.72	72.86/80.20	84.82/89.40
Tableformer	36.66/72.22	65.00/76.56	72.51/82.01	91.62/95.44	70.82/78.35	85.18/89.73
TST Layer	36.33/72.70	64.33/77.74	71.86/81.04	91.76/95.60	72.15/79.55	86.18/90.57
TST Layer +Embedding	37.66/**74.05**	**77.00/85.30**	72.15/81.27	**92.77/96.07**	**73.36/81.12**	**86.32/90.95**
Vanilla-adv	12.00/33.91	33.33/56.42	23.20/32.56	33.57/38.45	60.32/71.16	24.33/27.77
Projection-adv	**38.33**/72.75	68.00/79.41	73.08/82.60	91.33/95.34	72.18/79.46	83.89/88.87
TAPAS-adv	35.33/71.29	68.33/79.29	71.43/80.95	90.55/94.81	72.41/78.88	82.82/87.15
Tableformer-adv	37.66/72.12	65.33/77.94	70.79/80.71	91.69/95.41	72.74/79.64	85.18/90.05
TST Layer-adv	37.33/73.41	64.66/77.49	**73.58/82.68**	91.91/95.55	72.73/80.04	85.46/90.18
TST Layer-adv +Embedding	34.66/71.89	62.66/76.08	66.28/77.03	90.19/94.93	68.66/77.72	74.65/83.57

인 일반화 성능을 위해 미지의 도메인(unseen domain)의 역할로써 평가된다. 평가 데이터셋은 총 6,979개이다.

5. 실 험

5.1 실험 환경 및 구성

본 연구에서 학습에 사용한 데이터셋은 총 5가지이고. 평가에 사용한 데이터셋은 총 6가지이다.

법령 표 데이터셋은 학습에 1546개, 평가에 300개 사용하였다. 제품 스펙 표 데이터셋은 학습에 2014개, 평가에 300개 사용하였다. 공문서 표 데이터셋은 학습에 6826개, 평가에 1397개 사용하였다. KorWikiTQ는 학습에 10000개, 평가에 11214개 사용하였다. KorQuAD 2.0 표 데이터셋은 학습에 12655개, 평가에 1618개 사용하였다. AIHUB 공문서 표 데이터셋은 학습은 하지 않고 평가에만 6979개 사용하였다.

학습할 때 적용한 배치(batch) 함수 방식은 총 2가지를 실험했다. 첫 번째는 데이터셋 크기에 상관없이 일관된 순서의 도메인으로 배치를 구성하였다. 하지만 이러한 데이터 배치 방식은 함께 학습되는 여러 데이터 개수의 불균형을 고려하지 않은 방법이기 때문에 일반화 성능에서 안 좋은 성능을 보였다. 이에 본 연구에서는 데이터셋의 크기에 따라 한 배치당 들어갈 도메인의 순서를 직접 설정하여 한 에폭(epoch)을 학습했을 때 모든 도메인이 같은 한 에폭을 학습할 수 있도록 하는 배치 방법을 적용하였다. 이 방식을 통해 모델은 조절된 양을 일정하게 학습하며 더 좋은 결과를 도출할 수 있었다.

5.2 실험 결과

표 4는 학습 방법 및 모델 구조를 달리하여 학습한 모델별 평가 결과를 나타낸다. 표 1에서 '-adv'라고 표시된 모델은 3.3에서 설명했던 도메인 일반화를 위한 적대적 학습 방법을 적용한 모델들의 결과를 나타낸다. Vanilla는 기존의 BERT 모델을 그대로 적용한 결과이다. Projection은 본 논문에서 베이스라인 모델로 사용하고 있는 이전 연구의 모델[13]을 나타낸다. Tableformer와 Tapas는 Projection 모델을 베이스 모델로 하여 필요한 임베딩이나 어텐션 바이어스를 추가하였다. TST Layer는 그림 4에서 나타내고 있는 표 특화 트랜스포머 레이어(Table-specific Transformer Layer)를 추가한 모델을 나타낸다. TST Layer는 표에 특화된 추가 임베딩을 추가하지 않고 추가 레이어만 적용한 모델이며, TST Layer+Embedding은 추가 레이어의 입력으로 표

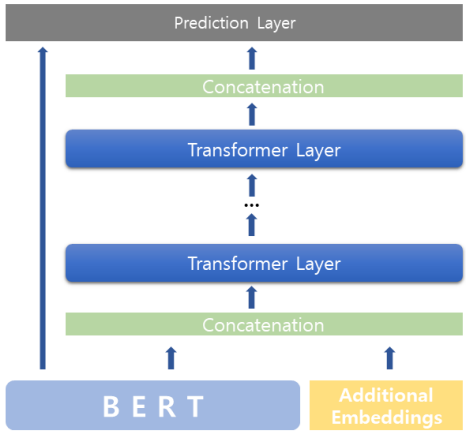

그림 4 레이어 추가 모델 전체 구조 (Add Layer)
Fig. 4 The entire structure of layer adding model

표 5 단일 데이터셋(KorQuAD 2.0) 학습 실험 결과 (EM/F1)
Table 5 Test result of training single dataset

Model \ Dataset	Law dataset	Spec dataset	office dataset	KorWikitq	Korquad 2.0 (Table)	Administrative dataset
Vanilla	8.00/17.71	4.00/18.73	18.40/27.11	28.20/33.32	67.11/76.52	22.61/26.63
Projection	18.00/43.35	16.00/35.53	64.56/75.85	83.17/89.05	78.36/85.50	80.88/87.93
TAPAS	20.00/45.92	13.33/29.68	65.78/76.79	82.39/88.35	78.04/85.90	81.31/87.80
Tableformer	**20.66/48.38**	**19.00/38.70**	65.42/76.23	83.03/88.94	78.30/85.18	81.67/88.52
TST Layer	16.66/46.18	14.00/31.62	65.78/76.64	82.89/88.72	79.52/85.86	81.60/88.03
TST Layer+Embedding	18.33/44.20	13.33/30.70	**67.85/78.68**	**90.05/94.02**	**81.73/87.63**	**85.11/90.43**

에 특화된 임베딩을 추가한 모델을 나타낸다.

적대적 학습 방법을 적용했을 때는 모델의 구조에 따라서 성능이 전체적으로 오르거나 혹은 하락하는 것을 보였다. 적대적 학습을 통해 성능이 향상된 모델은 Tableformer, Projection이며, 성능이 하락한 모델은 Vanilla, Tapas, Add layer이다.

표 5에서는 학습 데이터로 범용적인 도메인을 다루고 있는 KorQuAD 2.0 데이터만을 이용했을 때, 학습되지 않은 다른 도메인의 데이터에서 일반화가 잘 되는지 평가하였다. 학습한 데이터셋인 KorQuAD 2.0에서는 TST Layer 모델이 가장 높은 성능을 보였으며, KorQuAD 2.0과 비교적 유사한 표 형태나 도메인을 다루고 있는 데이터셋인 공문서, KorWikiTQ, 행정 데이터셋에서도 가장 높은 성능을 보였다. 특히 KorQuAD 2.0과 같은 데이터인 위키피디아 데이터를 사용하는 KorWikiTQ에서는 다른 데이터셋들과 비교하여 더 큰 성능 차이를 보여주었다. Tableformer 모델의 경우, 학습한 도메인인 KorQuAD 2.0에서는 높은 성능을 보이지 않았지만 베이스라인 모델인 Projection 그리고 Tapas와 비교했을 때 학습되지 않은 데이터셋에서 높은 성능을 보였다. 특히 표의 형태가 다양하게 나타날 수 있는 법령 데이터셋과 제품 스펙 데이터셋에서 가장 높은 성능을 보였다.

6. 결론 및 토의

본 논문에서는 다양한 도메인의 표 데이터를 이용하여 질의응답을 할 때, 학습 방법과 다양한 표를 위한 모델 구조를 비교하는 실험을 통해서 도메인 일반화에 적합한 표 기계독해 모델을 제안하였다. 기존 기계독해 연구에서 사용되었던 적대적 학습 기법을 표 기계독해에 적용했을 때는 모델의 구조에 따라서 일부 모델의 결과에서는 전체적으로 성능이 향상되는 것을 보였으나, 일부 모델에서는 성능이 하락한 것을 보였다. 성능이 하락한 모델들의 공통적인 특징으로는 표를 위한 특수 임베딩으로 행, 열, 순위, 개체명 임베딩을 추가하였는데, 이 점이 적대적 학습에서 성능 하락에 영향을 준 것으로 생각된다. 모델의 구조 변형에서는 적대적 학습을 하지 않고 표에 특화된 트랜스포머 레이어를 추가하고 해당 레이어에 표를 위한 특수 임베딩을 추가한 모델이 가장 높은 성능을 보였다. 학습되지 않은 도메인의 데이터에 관한 일반화 실험에서는 TST Layer 모델이 많은 데이터셋에서 높은 성능을 보였지만, Tableformer 모델에서는 다양한 형태의 표가 등장하는 데이터셋에서 높은 성능을 보였다. 향후 연구에서는 이러한 특성을 반영하여 TST 레이어와 Tableformer의 구조를 적절하게 조합할 수 있는 모델의 연구가 필요할 것으로 보인다. 또한, 향후 연구에서는 본 연구에서 실험했던 결과를 기반으로 적대적 학습에 좀 더 최적화된 모델을 이용하여 도메인 일반화를 위한 향상된 적대적 학습 방법에 관한 연구를 진행할 예정이다.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics. 2019.

[2] Joon-Ho Lim, Hyun-ki Kim. Evaluating of Korean Machine Reading Comprehension Generalization Performance via Cross-, Blind and Open-Domain QA Dataset Assessment. *Journal of KIISE*, 48(3), 275-283, 2021.

[3] Krueger, David, et al. "Out-of-distribution generalization via risk extrapolation (rex)." International Conference on Machine Learning. PMLR, 2021.

[4] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th*

International Conference on International Conference on Machine Learning - Volume 28 (ICML'13). JMLR.org, I-10-I-18, 2013.
[5] Goodfellow, Ian, et al. "Generative adversarial networks." Communications of the ACM 63.11 : 139-144, 2020.
[6] Kim, Youngmin et al. "KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension." (2020).
[7] Pasupat, Panupong and Percy Liang. "Compositional Semantic Parsing on Semi-Structured Tables." Annual Meeting of the Association for Computational Linguistics (2015).
[8] Chen, Zhiyu et al. "FinQA: A Dataset of Numerical Reasoning over Financial Data." ArXiv abs/2109.00122, 2021.
[9] Jauhar, Sujay Kumar et al. "TabMCQ: A Dataset of General Knowledge Tables and Multiple-choice Questions." ArXiv abs/1602.03960, 2016.
[10] Zhu, Fengbin et al. "TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance." ArXiv abs/2105.07624, 2021.
[11] Katsis, Yannis et al. "AIT-QA: Question Answering Dataset over Complex Tables in the Airline Industry." ArXiv abs/2106.12944, 2022
[12] Lee, Seanie et al. "Domain-agnostic Question-Answering with Adversarial Training." Conference on Empirical Methods in Natural Language Processing (2019).
[13] Sanghyun Cho, Minho Kim, Hyuk-chul Kwon, "Table Question Answering based on Pre-trained Language Model using TAPAS", *Annual Conference on Human and Language Technology*, pp. 87-90, 2020.
[14] Rao, Susie Xi et al. "TableParser: Automatic Table Parsing with Weak Supervision from Spreadsheets." ArXiv abs/2201.01654 (2022): n. pag.
[15] Eisenschlos, Julian Martin et al. "MATE: Multi-view Attention for Table Transformer Efficiency." Conference on Empirical Methods in Natural Language Processing (2021).
[16] Yang, Jingfeng et al. "TableFormer: Robust Transformer Modeling for Table-Text Encoding." WAnnual Meeting of the Association for Computational Linguistics (2022).
[17] Hye-Lynn Kim, Sanghyun Cho, JoongMin Sin, and Hyuk-Chul Kwon, "Evaluation of Generalized Performance of Korean Machine Reading According to the Evaluation Dataset for each Domain," *Proc. of the KIISE Korea Software Congress*, pp. 365-367, 2022.

조 상 현

2019년 부산대학교 정보컴퓨터공학 학사
2021년 부산대학교 정보컴퓨터공학 석사
2021년~부산대학교 정보융합공학 박사 재학. 관심분야 : 자연어처리, 기계독해, 인공지능, 머신러닝

김 혜 린

2022년 부산대학교 도시공학 학사
2022년~부산대학교 정보융합공학 석사 재학. 관심분야 : 자연어처리, 언어모델, 기계독해, 인공지능, 머신러닝

권 혁 철

1982년 서울대학교 공과대학 컴퓨터공학 학사. 1984년 서울대학교 공과대학 컴퓨터공학 석사. 1987년 서울대학교 공과대학 컴퓨터공학 박사. 1988년~현재 부산대학교 정보컴퓨터공학 교수. 관심분야 : 자연어처리, 인공지능, 머신러닝