

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. 서론

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

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)

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