표 기반 질의응답 데이터와 문맥 기반 질의응답 데이터 기계 독해 모델 성능 비교 및 한계 분석

실용언어 분석 프로그래밍
영어영문학과
2020115848
강준영

< 목 차 >

I. 프로젝트 진행 배경 및 필요성

1. 기계독해 모델이란

기계 독해 모델(Machine Reading Comprehension, MRC)은 자연어 처리 분야의 연구 주제이다. 사람이 텍스트를 읽고 질문의 답변을 추론하듯 인공지능이 스스로 질문을 분석하고 이해하며 주어진 문서에서 최적의 정답을 찾아내는 기술이다.

MRC의 대표적인 활용 분야는 질의응답 시스템이다. MRC 기반의 질의응답 시스템은 검색된 문서 내에서 실시간으로 정답을 찾아 제공하는 방식이며 기존의 지식기반 시스템의 단점을 보완해주는 역할로 딥러닝 기반 언어처리 기술에서 두각을 나타내고 있다.

MRC 기반 질의응답 시스템은 (1) 언어분석(NLU), (2) 문서 검색(Search), (3) 기계 독해(MRC)의 3단계로 이루어지며 대표적인 모델로는 양방향으로 문맥을 이해하는 'BERT(Bidirectional Encoder Representations from Transformer)'와 'GPT(Generative Pre-trained Transformer)'가 있다.

2. 기계독해 모델의 발전과 현재 연구 동향

기계독해는 검색 엔진, 챗봇, 교육 등 다양한 분야에서 사용되고 있다. 초기 기계독해 시스템은 규칙기반 모델로 제한된 데이터셋과 특정 도메인에 국한되었으나, 빅데이터와 딥러닝의 발전과 더불어 기계독해 분야도 함께 성장하며 관련 연구 논문의 수도 눈에 띄게 증가했다. 최근 연구 동향은 지식기반 기계독해와 답변이 불가능한 질문 처리 방법 등 실질적인 측면에서의 문제들에 대한 논의가 이루어지고 있다.

지식기반 기계독해는 문맥만으로 답할 수 없는 경우 상식이 들어가는 인간의 독해 능력과 차이점이 있으며 이를 MRC에 도입하려는 연구 과제이다.

또한, 기계독해에서는 주어진 문맥 속에 정답이 존재한다는 암묵적인 가정이 있지만, 실제 상황에서는 이런 가정이 항상 일치하지는 않는다. 그렇기에 MRC 모델은 무엇을 모르는지 알아야하며. 답변이 불가능할 경우, 불가능한 것으로 표기해야 하며, 모델 자신의 예측 결과에 대한 검증이 필요하다.

대화형 기계독해(CMRC)에서는 이전 대화 기록은 질의응답을 위한 중요한 문맥의 역할을 하지만 MRC는 각 질의응답 쌍이 독립적이다. 이를 해결하기 위해 대화 기록이 응답 출력을 위한 입력으로 제공되는 CMRC 모델이 소개되었으나, 여전히 공통참조(Coreference)를 모델이 처리하는 방식에 있어 해결할 과제가 남아있다.

이처럼 기계독해의 현재 연구 동향은 인간의 독해 능력과 비슷한 수준을 목표로 하며 다양한 데이터 유형과 상황에 대한 문제 해결 능력을 위주로 논의되고 있다.

3. 데이터에 따른 기계독해 모델 성능 비교 필요성

기계독해를 위한 데이터셋은 다양한 유형과 복잡성을 지니고 있다. 질의응답을 위한 외부 지식의 필요성, 답변 불가 질문의 가능성 등 다양한 문제 유형에서 모델의 성능이 어떻게 다른지 비교하는 것은 특정 분야에서의 모델의 적합성을 판단하는 데에 필요하다.

또한, 모델과 데이터셋의 관계를 분석해 모델의 약점을 파악하고 개선하는 데에 중요한 역할을 할 수 있으며 이를 기반으로 데이터셋 특성에 따른 모델의 최적화가 가능하다.

특히, 본 프로젝트의 비교 대상인 표 기반 데이터와 문맥 기반 데이터는 각각의 응용 분

야가 다르고 각각의 질의응답 처리는 구조화된 데이터를 처리해 특정 값을 추출하는 방식과 비정형 데이터에서 의미를 이해하고, 추론하는 방식으로 서로 상이하다. 두 가지 결과 비교를 통해 특정 분야에서의 모델의 효율성을 평가할 수 있으며 모델의 강점과 약점을 파악해 개선 방향을 제시한다.

II. 프로젝트 실험 과정 및 결과

1. 프로젝트 개요 및 과정

본 프로젝트의 비교 대상 데이터는 표 기반 질의응답 데이터[1]와 문맥 기반 질의응답 데이터[2]이며 모두 한국어 기반 데이터이다. 전자의 경우 한국학술정보㈜, 국가통계포털 등의 데이터로 구축되었고, 후자는 KoWIKI 본문을 기반으로 구축된 데이터이다.

모델은 한국어 질의응답 작업에 특화된 KoELECTRA 기반 사전학습 모델(arogyaGurkha/koelectra-base-discriminator-finetuned-squad_kor_v1)을 사용했으며, 성능평가 지표로 Accuracy(정확도), Precision(정밀도), Recall(재현율), F1 Score를 계산하였고 직관적인 비교를 위해 시각화를 진행하였다.

두 데이터의 형식이 달라 데이터 파싱을 진행해 동일한 형식으로 전처리하였다. 전처리 데이터의 구조는 'Document Title', 'Context', 'Question', 'Answer'를 Key로 가진다. 다음 표는 데이터 구성이다. 기존의 표 정보 데이터는 질문에 대한 정답 답변 위치 항목이 있었지만, 전처리 데이터에선 제외하였다.

Key	표기반 질의응답	문맥기반 질의응답	Type(표 / 문맥)
Document Title	문서 제목	제목	string / string
Context	표 제목 및 내용	문맥 정보	array / string
Question	질의	질의	string / string
Answer	응답	응답	string / string

▲ 전처리 데이터 구성

	id	Document Title	Context	Question	Answer
0	0	성안동 일원 분류식화 하수관로 정비사업 기본 및 실시설계용역 사업수행능력 평가서 제출안내	성안동 일원 분류식화 하수관로 정비사업 기본 및 실시설계용역 사업수행능력 세부평가기...	성안동 일원 분류식화 하수관로 정비사업의 용역업자 사업수행능력 평가시 참여기술인 ...	50
1	1	성안동 일원 분류식화 하수관로 정비사업 기본 및 실시설계용역 사업수행능력 평가서 제출안내	성안동 일원 분류식화 하수관로 정비사업 기본 및 실시설계용역 사업수행능력 세부평가기...	성안동 일원 분류식화 하수관로 정비사업의 용역업자 사업수행능력 평가시 신용도에 대...	10
2	2	성안동 일원 분류식화 하수관로 정비사업 기본 및 실시설계용역 사업수행능력 평가서 제출안내	성안동 일원 분류식화 하수관로 정비사업 기본 및 실시설계용역 사업수행능력 세부평가기...	유사용역 수행실적 항목의 평가방법은 어떻게 되는가	-건수 : 6점 -금액 : 6점 -전차용역 : 1점 -용역수행성과 ...

▲ 표 기반 질의응답 데이터 (전처리)

1) https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71565
2) https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=106

	id	Document Title	Context	Question	Answer
0	0	다테_기미코	재팬 오픈에서 4회 우승하였으며, 통산 단식 200승 이상을 거두었다. 1994년 ...	다테 기미코가 최초로 은퇴 선언을 한게 언제지	1996년 9월 24일
1	1	Ave;new	ave;new(아베;뉴, アベニュ?)는 도쿄 치요다구에 본 거처를 둔 일본의 음악 ...	ave;new 본거지 어디야	도쿄 치요다구
2	2	사카이_다다요시_(1714년)	사카이 다다요시(일본어: 酒井忠休, 1714년 9월 24일 ~ 1787년 6월 3일...	사카이 다다요시의 아버지가 누구지	사카이 나오타카(酒井直隆)
3	3	일반성면	일반성면은 동부 5개 면의 교통, 문화, 교육, 상업의 중심지로서 일찍부터 상업이 ...	일반성면의 면적이 얼마야	19.41 km²
4	4	금나라	지방은 전국을 19개 로(路)로 나누고, 그 아래에 부(府)나 주(州)를 두고, 다...	금나라를 세운 사람이 누구야	태조 아구다

▲ 문맥 기반 질의응답 데이터 (전처리)

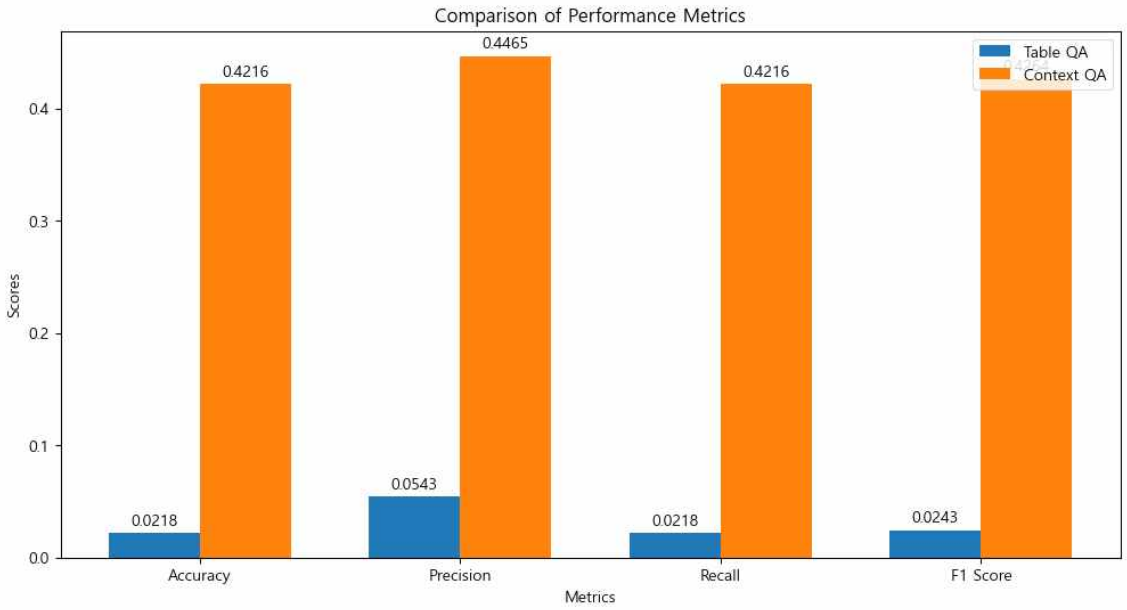
2. 실험 결과 및 해석

2.1 실험 결과

데이터 / 성능 지표	표 기반 질의응답	문맥 기반 질의응답
Accuracy	0.0218	0.4216
Precision	0.0543	0.4465
Recall	0.0218	0.4216
F1 Score	0.0243	0.4264

▲ 모델 예측 성능 지수 표

전체적으로 문맥 기반 질의응답 데이터에 대한 모델의 예측 성능이 높게 나타났고, 표 기반 질의응답 데이터에 대해서는 굉장히 낮은 성능을 보여주었다.



▲ 모델 예측 성능 비교 그래프

2.2 **실험 결과 해석 및 연구 의미**

예측에 사용된 MRC 모델 koelectra-base-discriminator-finetuned-squad_kor_v1는 표 기반 질의응답보다 문맥 기반 질의응답 처리 작업에 더 적합하다고 볼 수 있다. 해당 모델의 사전학습 데이터 셋은 표와 같이 구조화된 형식의 질의응답이 아닌 일반 자연어 텍스트의 질의응답 데이터 셋으로 fine-tuning했다는 것을 알 수 있다.

또한, 각각의 성능 지표가 약 8~10배 가량 차이가 난다는 점은 모델의 사전학습 방식(모듈)과 데이터 셋의 유형이 모델이 새로운 데이터를 처리하는 방식과 능력에 상당한 영향을 끼친다고 해석이 가능하다.

나아가 해당 실험 결과는 발전되고 특화된 모델 개발을 위한 다양한 유형의 데이터 셋 구축의 중요성을 시사한다고 볼 수 있다.

III 프로젝트 결론

1. **한계점 및 향후 연구에 대한 제언**

본 실험 결과는 기본적으로 모델 학습 방식과 데이터셋의 차이에 따른 성능 차이를 나타내고 있다. 하지만, 데이터 전처리 과정에서 표 기반 질의응답 데이터의 정답 답변 위치가 삭제되어 성능이 더 낮아졌을 가능성이 있다는 점, 표 기반 데이터셋의 context 항목의 array 형식이 문맥 기반 데이터셋과 동일한 string 형식이 아닌 점, 다양한 모델 학습 성능 결과 비교가 나타나지 않고 하나의 모델만이 사용되어 일반화하기는 어렵다는 점, 두 가지 유형의 데이터 셋에 동일한 평가 기준을 사용한 점 등의 한계가 있다.

본 프로젝트에서 나타난 한계점을 바탕으로 각각의 데이터셋 유형에 따라 특화된 모델들 간의 비교 분석하는 연구가 필요하다.

IV. 참고문헌

1. Liu S, Zhang X, Zhang S, Wang H, Zhang W. Neural Machine Reading Comprehension: Methods and Trends. Applied Sciences. 2019; 9(18):3698. https://doi.org/10.3390/app9183698

2. Zeng C, Li S, Li Q, Hu J, Hu J. A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets. Applied Sciences. 2020; 10(21):7640. https://doi.org/10.3390/app10217640

3. KT 지니랩스 https://genielabs.ai/tech/detail?domain=nlp&contentsSeq=25