KorQuAD 2.0: 웹문서 기계독해를 위한 한국어 질의응답 데이터셋

김영민[0], 임승영, 이현정, 박소윤, 김명지
LG CNS, AI빅데이터연구소
{ymk, seungyoung.lim, hyunjeonglee, soyoon.park, kmj0614 }@lgcns.com

KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension

Youngmin Kim[0], Seungyoung Lim, Hyunjeong Lee, Soyoon Park, Myungji Kim
LG CNS, AI/Bigdata Research Center

요 약

KorQuAD 2.0은 총 100,000+ 쌍으로 구성된 한국어 질의응답 데이터셋이다. 기존 질의응답 표준 데이터인 KorQuAD 1.0과의 차이점은 크게 세가지가 있는데 첫 번째는 주어지는 지문이 한두 문단이 아닌 위키백과 한 페이지 전체라는 점이다. 두 번째로 지문에 표와 리스트도 포함되어 있기 때문에 HTML tag로 구조화된 문서에 대한 이해가 필요하다. 마지막으로 답변이 단어 혹은 구의 단위뿐 아니라 문단, 표, 리스트 전체를 포괄하는 긴 영역이 될 수 있다. Baseline 모델로 구글이 오픈소스로 공개한 BERT Multilingual을 활용하여 실험한 결과 F1 스코어 46.0%의 성능을 확인하였다. 이는 사람의 F1 점수 85.7%에 비해 매우 낮은 점수로, 본 데이터가 도전적인 과제임을 알 수 있다. 본 데이터의 공개를 통해 평문에 국한되어 있던 질의응답의 대상을 다양한 길이와 형식을 가진 real world task로 확장하고자 한다.

주제어: KorQuAD, MRC, 딥러닝, 기계독해, 질의응답

1. 서론

기계독해(MRC;Machine Reading Comprehension)는 기계가 주어진 지문과 질문을 이해하여 지문 내에서 답변 영역을 찾아야 하는 자연어처리 과제로 자동 질의응답 기술의 핵심 토대가 되는 기술이다. 기계독해를 위한 한국어 표준 데이터셋으로는 KorQuAD 1.0[1]이 있으며 모델 학습에 이용할 수 있을 뿐만 아니라 여러 모델간의 성능 평가를 위한 객관적 기준이 된다.

KorQuAD를 비롯해 기존에 공개된 한국어 데이터셋은 모두 위키백과나 뉴스 기사의 지문 한 문단과 같이 짧은 평문으로 구성된 지문에서 질의응답을 수행한다. 공개된 데이터셋에 대해 좋은 성능을 내기 위하여 연구되는 기계독해 알고리즘 또한 평문 형태의 지문 입력에 최적화 되어있다. 하지만 실제로 질의응답 기술이 필요한 영역을 살펴보면 정제된 평문에서 기계독해를 수행해야 하는 경우 보다는 웹 문서, 상품 매뉴얼, 이용 약관 문서 등 양식 구조도 다양하며 길이 또한 문단이 아닌 문서 레벨에서 수행해야 하는 경우가 많다.

이처럼 실무에서 필요한 질의응답 태스크와 학계의 연구의 괴리가 있어서 데이터셋에 대해서는 잘 질의응답 할 수 있는 알고리즘도 현실의 문제에서는 난항을 겪고 적용하기 힘든 경우가 많다. 이에 본 논문은 다채로운 구조와 길이를 가진 문서 레벨에서의 기계독해를 위해 구축한 데이터셋 KorQuAD 2.0을 소개하고 특징을 분석한 결과를 제공한다. 기계독해 모델은 KorQuAD 1.0[1]과 달리 자연어로 구성된 질문과 HTML 코드로 이루어진 위키백과 웹 문서를 입력으로 받아 짧은 답변뿐 아니라 매우 긴 문단 단위의 답변, 표로 이루어진 답변, 리스트로 이루어진 답변 영역을 리턴할 수 있어야 한다.

또한 우리는 동시에 데이터 공개 및 리더보드[1]를 운영하여 여러 모델을 동시객관적으로 평가할 수 있도록 한다. KorQuAD 2.0은 정확도(EM, F1 점수) 외에도 실용성을 위해 모델 추론 속도에 대한 지표도 추가하여 답변을 리턴하는데 너무 긴 시간이 걸리는 것을 지양하도록 유도한다. 우리는 KorQuAD 2.0의 공개로 한국어 자연어처리 연구자들이 현실의 문제 해결에 더 가까운 데이터셋을 쉽게 확보하고 객관적인 기준으로 연구 결과의 성능 평가를 하는 데 이바지하고자 한다.

2. 관련 연구

다문단 기계독해는 KorQuAD 1.0[1]과 같은 단일 문단에 대한 기계독해에 비해 읽어야 할 문서의 길이가 길어짐에 따라 추론 속도와 정확도 측면에서 모두 도전적인 과제이다. 영문의 경우 초기 기계독해 데이터로 QASent[2], WikiQA[3]이 있는데 질문의 수가 만개 이하로 적다는 한계가 있었다. 그 이후 크고 정제된 데이터인 SQuAD 1.1[4], 지문과 관련은 있지만 답을 찾을 수 없는 질문을 포함하여 기존의 데이터를 보완한 SQuAD

[1] https://korquad.github.io/

2.0[5] 등이 공개되어 있지만 여전히 단일 문단에서 답을 찾는 과제들이다. 다문단 기계독해 데이터로는 평균 6개의 단락으로부터 답변을 추론해야 하는 TriviaQA[6]와 두 개의 단락에 대해 멀티홉 추론을 해야 하는 HotpotQA[7]이 있다. 그러나 두 데이터는 읽어야 할 문서의 길이가 길어져도 답변 영역은 구 단위를 넘어가지 않을 뿐만 아니라, 표 등이 없는 평문에 한정된다. Natural Questions[8]는 검색엔진에 수집된 쿼리들에 대한 답변을 위키백과 HTML에서 찾아야 하는 데이터셋으로, 다문단 문서 중 문단 혹은 표 전체가 답이 될 수 있다는 점과 웹 구조를 살렸다는 점에서 기계독해의 영역을 구조화된 문서로 확장시켰다는 의의가 있다. 그러나 웹쿼리에서 수집한 데이터인 만큼 정제된 자연어 질문이 아니다. 구조화된 문서에 대한 한국어 데이터로는 TabQA[9]에서 표와 질문을 인공적으로 만들어 사용한 것이 있다. 표와 그림 등이 포함된 교과서에서 질문에 대한 답을 찾아야 하는 TextbookQA[10], 비디오를 보고 텍스트 질문에 답을 해야 하는 TVQA[11] 등과 같이 멀티모달 데이터들도 있다.

KorQuAD 2.0은 구조화된 HTML에 대한 자연어 질문을 수집한 다문단 기계독해 데이터셋이다. 데이터 수집 과정은 KorQuAD 1.0[1]과 유사하나, 기계독해의 대상이 HTML 문서로 확대되었고, 구 단위뿐만 아니라 표의 셀, 표 전체, 리스트 전체, 여러 문단들이 답변이 될 수 있다는 점에서 웹 구조를 활용한 복잡한 자연어 질문들을 포함한다.

3. 데이터 구축

KorQuAD 2.0은 두 가지 방식을 통해 제작되었다. 먼저 크라우드 소싱을 통해 KorQuAD 1.0[1]과 유사한 방식으로 데이터를 수집하였다. 작업자는 사전 테스트를 통해 정상적인 기계독해 질문을 생성하는지 확인된 인원들로, 제공된 문단 안에서 답을 정하고 그에 맞는 질문을 생성하게 된다. 작업자가 태스크에 참여하기 위해서는 사전 테스트를 통과해야 하는데 테스트에서는 다양한 질문 예시를 보여주고 질문이 올바르게 만들어졌는지, 올바르지 않다면 어떤 이유에서 인지 판단하여 태스크의 목적과 방향성에 대해 인지할 수 있도록 하였다. 추가적인 데이터 제작으로 기존 KorQuAD 1.0[1] 데이터 중 일부를 2.0 타입으로 변환함으로써 보다 많은 데이터를 확보하였다.

3.1 문서 수집

우리는 다양한 주제에 대해 구조화된 문서를 모으기 위해 위키백과 문서들을 활용하였다[2]. 위키의 문서 중에서 사람들이 관심 있어하는 문단을 선정하기 위해 3년 동안(2016/06/01 ~ 2019/05/31)의 page view 상위 15만 문서를 선별하였고 더 다양한 문서 도메인을 다루기 위해 임의로 5만개의 문서를 추가하여 HTML 문서 데이터를 수집하였다.

3.2 질문-답변 생성

크라우드 소싱을 통해 질문-답변을 생성할 때 작업자에게는 문서 전체가 아닌 일부 문단만 보여줌으로써 문서 앞쪽 혹은 내용이 쉬운 부분에서만 질문을 만드는 것을 방지하였다. 작업자에게 보여줄 문단을 추출하기 위해 먼저 소제목 단위로 문서를 나누고 유의미한 텍스트 정보가 있는 부분으로 평문(<p>), 표(<table>), 리스트(<ul>, <ol>, <dl>)의 HTML tag 영역을 제시하였다. 추가적으로 HTML 태그를 제외한 순수 텍스트가 90단어 이하로 매우 짧은 문단과 참고문헌과 같이 지문 자체와 관련이 없는 부분들은 제거하였다. 최종적으로 작업자는 소제목과 소제목에 해당하는 문단들 그리고 전체 문서를 볼 수 있는 링크를 제공받아 질문을 만들게 된다.

데이터 제작 작업은 총 4가지로 구성된다. 먼저 답변 길이에 따라 Long 타입과 Short 타입으로 구분할 수 있다. Short 타입은 KorQuAD 1.0[1]과 같이 답변의 영역이 문단 내 단어 혹은 구이고 Long 타입은 질문에 답하기 위한 정보가 최소 한 문단 이상에 걸쳐있는 경우이다. Tag(<p>, <table>, <ul>, <ol>, <dl>)로 쌓인 블록을 하나의 문단으로 보았을 때 Long답변은 최소 한 문단부터 최대로는 위키 문서에서 소제목 아래의 문단 전체까지 가능하다. 또한 답변이 되는 문단의 형태에 따라 각각 평문 혹은 리스트인 경우와 표로 이루어진 경우로 나뉜다. 작업자들은 각 작업에 따라 다른 가이드라인과 테스트를 받게 된다. 이를 통해 Long/Short 타입을 명확히 구분할 수 있도록 하면서 최대한 다양한 어휘와 추론과정을 포함한 질문을 만들도록 권장하였다. 추가적으로 지문의 내용을 그대로 사용하는 것을 방지하기 위해 음절 단위의 4-gram 기준 질문의 50%가 지문과 중복되는 경우 제출하지 못하도록 하였다.

위와 같은 방식으로 질문-답변 데이터를 만든 후 품질 검증을 위하여 검수과정을 거치도록 하였다. 검수자는 과제 성공률 85% 이상인 작업자를 선별하여 선정되었다. 검수 방식은 데이터 한 쌍에 대해 두 명의 검수자가 통과/불통의 판단을 내리고 만약 둘의 판단이 같다면 검수가 종료되지만 다른 경우 추가적으로 다른 검수자의 판단을 추가하여 최종 판정이 되도록 하였다.

3.3 KorQuAD 1.0 데이터 변환

KorQuAD 1.0[1] 데이터를 활용하여 일부를 KorQuAD 2.0으로 변환하였다. 기존 context는 1~2 문단 정도의 평문으로 구성되어 있는데 이를 위키백과 문서 전체에 대한 HTML로 변환하고 정답 인덱스도 그에 맞추어 수정

[2] 크롤링한 날짜는 7월 4일, 7일, 10일로 나누어 진행

하였다. 학습, 검증, 평가 데이터를 포함해 총 21,421개의 질문-답변 쌍이 변환되었고 나머지 변환되지 않은 데이터들은 위키백과 웹 문서가 수정되어 새로 크롤링한 문서에서 기존의 지문을 찾을 수 없는 경우들이다.

4. 데이터 특징 및 분석

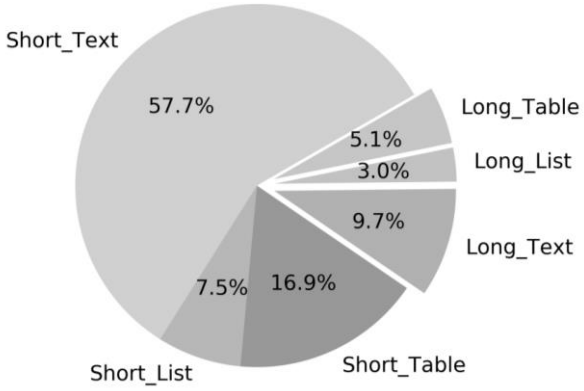
4.1 데이터 구성

질문은 KorQuAD 1.0[1]에서 변환한 데이터를 포함하여 총 102,960개가 있다. 다양한 문서에서 데이터를 만들기 위해 문서당 질문 제작에 제한을 두어 문서 47,957개에 문서당 평균 2.15개의 질문이 매칭된다. 이 중 38,506 문서는 학습 데이터, 4,739 문서는 검증 데이터, 나머지 4,726 문서는 평가 데이터로 나누었다.

[표 1] 문서 및 질문 개수

	학습	검증	평가	합계
문서	38,496	4,736	4,725	47,957
질문	83,486	10,165	9,309	102,960

질문-답변은 먼저 답변 길이에 따라 Short과 Long, 답변의 tag에 따라 평문, 표, 리스트로 구분 가능하다. 전체에서 Long 답변이 차치하는 비율은 17.8%이고 tag 유형에 따라서는 평문이 67.4%로 가장 많은 비율을 차지하고, 표, 리스트가 각각 22.0%, 10.5%를 차지한다.

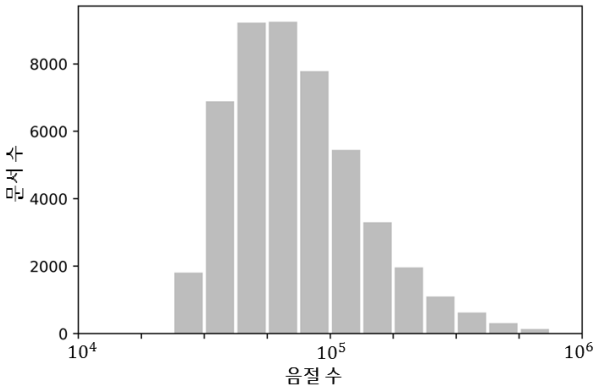


[그림 1] 답변 유형 비율

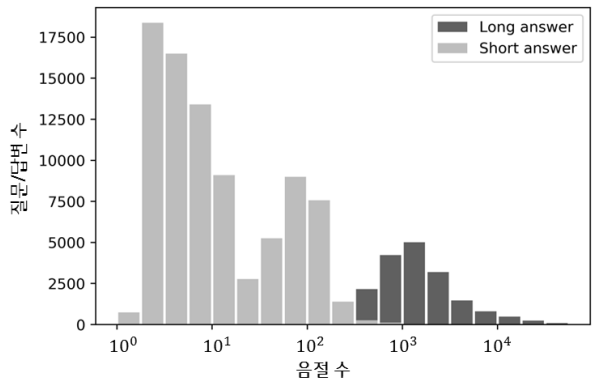
4.2 지문 및 답변 길이

문서의 길이는 위키 페이지의 HTML을 그대로 사용할 경우 평균 90,259음절로 매우 길다. 우리는 최소한의 전처리를 통해 길이를 줄였다. 먼저 HTML의 주석과 <script>를 제거하였고 table 형식과 관련된 정보인 colspan, rowspan을 제외한 모든 attribute를 제거하였다. 그 결과 길이가 평균 19,864음절로 크게 줄어들었다. 배포데이터는 전체 HTML과 전처리된 HTML 두 가지가 포함되어 있다.

답변의 길이분포도 한 글자부터 만 자가 넘는 것까지 다양하다. [그림 3]을 보면 대략 세 부분으로 나뉘는 것을 볼 수 있다. 답변의 길이가 짧은(1~20자) 첫 번째 그룹은 HTML tag 없이 순수 텍스트로 된 Short 답변으로 이루어져 있다. 중간 부분(20~120자)의 그룹은 short 답변이긴 하지만 HTML tag가 포함되어 있어 길어졌다. 긴 부분(120자~)은 long 답변으로 구성되어 있다. 다양한 길이의 지문과 답변이 있기 때문에 길이를 고려한 모델 설계가 필요할 것이다.



[그림 2] 문서의 음절 수당 빈도수



[그림 3] 답변의 음절 수당 빈도수

4.3 질문 유형

질문의 유형을 분석하기 위해서 Short 350개, Long 200개를 검증데이터에서 임의 추출하였고 질문과 근거가 되는 지문과의 관계에 따라 [표 2]과 같이 분류하였다.

먼저 Short의 경우 가장 많은 비율을 차지하는 유형은 구문 변형으로 48.0%를 차지하였고 그 다음으로 표/리스트가 27.7%를 차지하였다. 그 뒤로 어휘 변형(15.4%), 여러 문장 종합적 활용(8.0%)이 있고 출제 오류도 0.9%가 포함되어 있다.

Long의 유형에서는 주어진 소제목을 활용하여 구문 변형 및 유의어 변형 등으로 질문 제작하는 경우가 47%로 가장 많은 비율을 차지했다. 그 다음으로 소제목이 고유 명사여서 변형이 어려운 등의 이유로 그대로 사용하여 질문을 만드는 경우가 38%나 되었다. 소제목을 활용하지 않고 지문 내용을 통해 질문을 제작하여 질문만

으로는 어떤 소제목과 관련이 있는지 알기 어려운 유형은 15%를 차지했다.

[표 2] 질문 유형

Short 답변	
구문 변형 (48.0%)	**Q.** **외국인들을 위해 먹는 샘물이 일시 판매**되었던 년도는 언제일까?
	…1988년 서울 올림픽 무렵 **외국인들을 위하여 일시 판매**를 허용했던 적이 있으나, 다시 판매를 제한하였다. …
어휘 변형 (15.4%)	**Q.** 2009년 시즌 도중 **경질된** 지바 롯데의 감독은?
	…시즌 도중에 바비 밸런타인 감독의 **해임**이 발표되자 일부 팬들은 …
여러 문장 종합적 활용 (8.0%)	**Q.** 'Don't Cha'는 한국 휴대전화 기기 제조사의 휴대전화 CM송으로도 사용되었는데 그 제조사는 어디인가?
	…첫 싱글 **'Don't Cha'**는 영국, 오스트레일리아, 캐나다 등의 나라에서 1위에 …. 또한 **이 노래**는 한국의 휴대전화 기기 제조사 **SKY의 휴대전화 CM송**으로 쓰여, …
표/리스트 (27.7%)	**Q.** 득표율 2위를 한 사람은 어느 정당 소속인가?
	후보 / 정당 / 기반주 / 득표 / 선거인단 제임스 얼 카터 주니어 (롤터 프레데릭 먼데일) / 민주당 / 조지아주 미네소타주 / 40,831,881 / 50.08 / 297 제럴드 로돌프 포드 주니어 (로버트 조셉 돌) / 공화당 / 미시간주 캔자스주 / 39,148,634 / 48.02 / 240 로널드 윌슨 레이건 (로버트 조셉 돌) / 공화당 / 캘리포니아주 캔자스주 / 0 / 0.00 / 1
기타 출제 오류 (0.9%)	**Q.** 꽃가루가 식물에 전이되어 수정을 거쳐 유성 생식에 이를 수 있게 하는 과정을 일으키는 말은? *(지문에서 관련 설명을 찾을 수 없음)*
	… 이것으로 파리를 불러들여 **수분**(꽃가루받이)을 한다고 한다. 꽃덮이조각은 …
Long 답변	
소제목 중복 (38%)	**Q.** 피터슨과 **노먼 그란츠**의 관계는 어떤 과정을 통해 형성되었는가?
	Title. 오스카 피터슨 - #생애 - **#노먼 그란츠**
소제목 변형 (47%)	**Q.** 이경직의 **가족 관계**는 어떻게 이루어져 있는가?
	Title. 이경직 - **#가계**
자체 제작 (15%)	**Q.** 문화재를 보존하기 위해 시행하는 법은 무엇일까?
	Title. 거문오름 용암동굴계 상류동굴군 - #공개제한

5. 실험 및 결과

우리는 공개되어있으면서 여러 태스크들에서 좋은 성능을 내고 있는 BERT multilingual[12] 모델을 사용하여 KorQuAD 2.0 학습 데이터를 학습시키고 검증, 평가 데이터에 대해 성능을 측정하였다. 우리는 HTML tag가 웹 구조에 대한 정보를 담고 있다고 판단하여 속성만을 전처리한 HTML 문서를 입력으로 사용하였다. 이때 BERT 모델의 경우 입력의 길이가 512자로 제한되기 때문에 문서를 128의 stride를 주어 입력하였고, 학습 시 정답 토큰의 길이가 입력 길이 제한보다 긴 경우 제거하였다. 하이퍼 파라메터로는 공개된 모델의 기본값들을 사용하였다. 최종적으로 평가 데이터에 대해 사람의 점수도 측정하여 기계학습 결과와 비교하였다.

5.1 실험 결과

성능 측정으로는 KorQuAD 1.0에서 사용했던 EM, F1 외에 질문 하나당 응답 속도(1-example latency)를 추가하여 총 세 가지 척도를 사용하였다. EM은 답과 예측값이 정확히 일치하는 비율을 나타내고 F1은 음절 기준으로 얼마나 중복되는지를 고려한 점수이다. 이 때 정답 및 예측 텍스트에서 HTML tag와 같이 실제 정보를 가지지 않은 부분은 Python BeautifulSoup 라이브러리를 활용하여 제거하였다. 질문 하나당 응답 속도는 데이터 전처리 시간과 모델의 추론시간을 포함하여 전체 시간을 측정한 후 질문 수로 나누어 하나의 쿼리당 걸리는 평균 시간이다.

실험 결과, BERT 딥러닝 모델은 평가 데이터에서 EM 30.2%, F1 46.0%의 성능을 내었고 한 쿼리당 평균 13,484 밀리초의 응답시간이 소요되었다. 하지만 사람의 점수는 EM 69.8%, F1 85.7%로 기계학습에 비해 두 배 가까이 높은 성능을 보였다. 기계 학습 모델의 답변 길이에 따른 결과를 보면 Short의 경우 54.0%로 Long의 15.0%와 3배 이상 차이가 있다. 유형에 따라서는 표가 가장 낮은 성능을 보이고 평문이 제일 높다.

[표 3] KorQuAD 2.0에 대한 성능 비교

	검증 데이터		평가 데이터	
	EM	F1	EM	F1
Baseline	30.8	46.8	30.2	46.0
Human	-	-	68.8	83.9

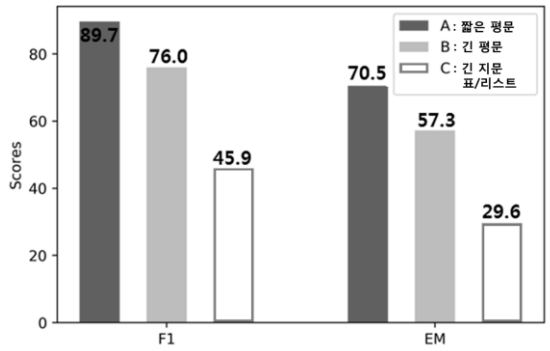
[표 4] 기계학습의 유형에 따른 F1 점수

	평문	표	리스트	합계
Long	15.0	8.9	25.5	15.0
Short	56.9	44.8	51.9	54.0

5.2 데이터 변화에 따른 성능

우리는 기존의 KorQuAD 1.0[1] 데이터로 학습한 모델이 긴 지문과 표 등에 대해서 어느 정도의 성능을 보이는지 확인해보았다. KorQuAD 1.0[1] 데이터로 BERT

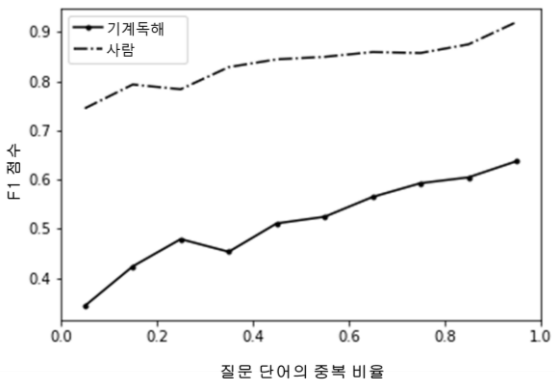
multilingual[12] 모델을 학습하였고 추론 시에는 학습 데이터에 없던 HTML tag를 제거하였다. [그림 4]를 보면 세 종류의 검증데이터가 있는데 A는 짧은 지문을 갖는 기존의 KorQuAD 1.0 데이터이고 B는 A에서 사용한 짧은 지문을 긴 지문으로 변환한 KorQuAD 2.0 데이터이다. A와 B의 F1 성능을 비교해보면 지문을 길게 변환했을 때 성능이 약 13% 정도 하락하는 것을 볼 수 있다. C는 새로 만든 데이터 중 리스트와 테이블만 포함한 데이터인데 A와 비교하여 절반 가까이 성능이 하락한다. 이를 통해 기존 데이터로 학습된 모델은 길고 복잡한 문서에 대해서는 잘 대처하지 못함을 확인할 수 있다.



[그림 4] 데이터의 형태에 따른 성능.

5.3 질문-지문 중복 비율에 따른 성능

지문과 질문의 중복 정도에 따른 성능 측정을 통해 질문의 난이도를 확인했다. 질문을 형태소분석을 통해 조사 및 어미를 제거한 후 답이 포함된 지문과의 중복 비율을 계산하고 그에 따른 F1 점수를 측정하여 기계독해와 사람의 점수를 [그림 5]에 나타내었다. 중복 비율이 낮아질수록 사람과 기계독해 모두 F1 점수도 낮아지는 것을 확인할 수 있다. 그러나 사람의 경우 점수가 가장 높을 때와 낮을 때는 92.0%와 74.4%로 17.8% 차이 나는 반면 기계독해의 경우는 63.7%에서 34.4%로 29.3%가 하락하였다. 중복 비율이 낮을수록 어려운 질문이지만 사람보다 기계독해의 점수가 크게 낮아지는 만큼 중복 비율이 낮더라도 좋은 성능을 내는 모델을 만드는 것이 필요하다.



[그림 5] 질문과 지문의 중복 비율에 따른 F1 점수

6. 결론 및 향후 방향

본 논문에서는 구조를 가진 문서에 대한 질의응답 데이터셋인 KorQuAD 2.0을 소개하고 baseline 모델을 통해 결과를 분석하였다. 또한 데이터 공개 및 리더보드 운영을 통한 공정한 평가로 학계에 기여하고자 한다. 데이터는 위키백과 한 페이지 전체를 대상으로 한다는 점에서 길고 복잡한 양식을 가진 문서를 다루고, 모델의 정확도는 물론 문서가 길어진 만큼 기존에는 고려되지 않았던 추론 속도까지 리더보드에 반영한다는 차이점이 있다. 이로써 평문에 한정되어 있던 기계 독해의 영역을 웹 문서, 약관, 표 등 다양한 양식의 문서로 확장하고, NLP 연구자가 현실적인 문제를 해결하는 데에 필요한 데이터를 확보할 수 있도록 기여한다.

우리는 구축한 KorQuAD 2.0 데이터를 바탕으로 향후 현실에서 마주하게 되는 복잡한 양식의 문서에 대한 질의응답 연구를 지속하고자 한다. 뿐만 아니라 모델 경량화 및 학습 및 추론 속도 개선을 위한 연구를 통해 실사용 가능한 자연어처리 모델 개발에 힘쓰고자 한다.

참고문헌

[1] 임승영; 김명지; 이주열. KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋. *한국정보과학회 학술발표논문집*, 539-541, 2018.

[2] WANG, Mengqiu; SMITH, Noah A.; MITAMURA, Teruko. What is the Jeopardy model? A quasi-synchronous grammar for QA. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. p. 22-32. 2007.

[3] YANG, Yi; YIH, Wen-tau; MEEK, Christopher. Wikiqa: A challenge dataset for open-domain question answering. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. p. 2013-2018. 2015.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[5] RAJPURKAR, Pranav; JIA, Robin; LIANG, Percy. Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv preprint arXiv:1806.03822*, 2018.

[6] Mandar Joshi, Eunsol Choi, Daniel S. Weld, Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

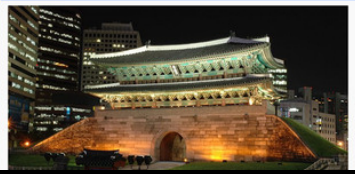
[7] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[8] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle

Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453-466, 2019.
[9] 박소윤, 임승영, 김명지, 이주열. TabQA : 표 양식의 데이터에 대한 질의응답 모델, HCLT pp.263-269, 2018.
[10] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, Hannaneh Hajishirz. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. p. 4999-5007. 2017.
[11] Jie Lei, Licheng Yu, Mohit Bansal, Tamara L. Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

7. 부록

[표 5] KorQuAD 2.0 예시, □는 질문에 대한 답

<table>
<tr><th></th><th>예시</th></tr>
<tr><td>질문</td><td>서울의 GDP는 세계 몇 위야? (Short, 평문)</td></tr>
<tr><td>지문</td><td>
서울특별시
위키백과, 우리 모두의 백과사전.
좌표: 북위 37° 34′ 00″ 동경 126° 58′ 41″
비슷한 이름의 소울에 관해서는 해당 문서를 참조하십시오.
서울특별시(서울特別市)는 대한민국의 수도이자 최대 규모의 도시이다. 백제의 첫 수도인 위례성이었고, 고려 때는 남경(南京)이었으며, 조선의 수도가 된 이후로 현재까지 대한민국 정치·경제·사회·문화의 중심지 역할을 하고 있다. 중앙으로 한강이 흐르고, 북한산, 관악산, 도봉산, 불암산, 인능산, 인왕산, 청계산 등의 여러 산들로 둘러싸인 분지 지형의 도시이다. 동서 간의 거리는 36.78 km, 남북 간의 거리는 30.3 km이며, 넓이는 605.25 km²이다. 면적은 대한민국 전 국토의 0.6%를 차지하지만, 약 977만 명의 인구가 살고 있어 인구밀도가 높다.
시청 소재지는 중구이며, 25개의 자치구로 이루어져 있다. 1986년 아시안 게임, 1988년 하계 올림픽, 2010년 서울 G20 정상회의를 개최한 국제적인 도시이다. 서울의 GDP는 세계 4위이다.[2][3]
서울
특별시
서울特別市

</td></tr>
<tr><td>질문</td><td>서울에 있는 산들에 대해 알려줘 (Long, 평문)</td></tr>
<tr><td>지문</td><td>
지리 [편집]
서울은 북위 37° 34′, 동경 126° 59′의 한반도 중서부에 위치하는 분지 지형의 도시이다. 시의 중심으로 한강이 흐르고, 서울 도심에는 남산, 인왕산(338m)이 있다. 시 주변으로 북한산(837m), 관악산(629m), 도봉산(740m), 수락산(428m), 불암산(510m), 구룡산(306m), 우면산(293m), 아차산, 지양산 등이 서울을 둘러싸며 경기도 및 인천광역시와 자연적 경계를 이루고 있다.[7] 동서 간의 거리는 36.78 km, 남북 간의 거리는 30.3km이며, 면적은 약 605.25 km²이다. 서울의 면적은 대한민국의 0.6%이며 남북한 면적의 0.265%이다. 서울특별시의 최북단은 도봉구 도봉동이고 최남단은 서초구 원지동이며 최동단은 강동구 강일동, 최서단은 강서구 오곡동이다.
산 [편집]
서울의 산 목록 문서를 참고하십시오
서울은 국립공원으로 지정된 북한산을 최고점으로 한 고양·양주구룡과 경기평야가 만나는 지대에 있다. 주위에는 북한산(836m)·도봉산(717m)·인왕산(338m)·관악산(629m) 등 500m 내외의 산과 구릉이 자연성벽과 같이 둘러싸고 있는 분지이다. 광주산맥의 한 줄기인 도봉산은 백운대·인수봉·노적봉의 3개 봉우리가 솟아 있는 북한산과 이어져 있고, 그 산줄기는 다시 남으로 뻗어 북악산(342m)을 솟게 하였다. 그리고 북악산에서 동으로 뻗은 산줄기에 낙산(125m), 서로 뻗은 산줄기에 인왕산이 있다. 인왕산에서 뻗은 산줄기 중 남쪽으로 뻗은 것은 숭례문을 지나 남산(265m)·응봉(175m)과 이어져 있고, 서쪽은 무악재의 안부(鞍部)를 지나 안산(296m)과 이어져 있는데 모두 구릉성 산지이다. 한강 남쪽에는 100m 이하의 구릉지가 펼쳐져 있고, 남쪽에 천연의 요새와 같이 서울의 외곽에 솟아 있는 관악산(629m)외 청계산(618m), 구룡산(306m), 우면산(293m)이 있다. 그 외에 서울 동부에 불암산, 수락산, 망우산, 아차산이 있다.[7]
</td></tr>
<tr><td>질문</td><td>서울의 대학교 집중도는 얼마나 돼? (Short, 표)</td></tr>
<tr><td>지문</td><td>
경제 [편집]
2014년 서울의 지역내총생산(GRDP)은 327조 6020억 원이며, 실질성장률은 2.2%이다.[22]
삼성, LG, 현대자동차, SK, 롯데 등 주요 대기업의 본사가 있다. 대한민국 GDP의 22%를 창출하고 있으며, 금융 기관의 50% 이상이 집중되어 있다.
<table>
<tr><th>2007년[23]</th><th>국내총생산 (10억원)</th><th>사업체 수 (개소)</th><th>은행예금 (10억원)</th><th>내국세 (10억원)</th><th>의료기관 (개소)</th><th>자동차수 (천대)</th><th>대학교 (개소)</th></tr>
<tr><td>전국</td><td>581,516</td><td>3,131,963</td><td>512,419</td><td>82,226</td><td>44,029</td><td>13,949</td><td>180</td></tr>
<tr><td>서울</td><td>127,175</td><td>735,258</td><td>259,355</td><td>35,436</td><td>12,396</td><td>2,691</td><td>42</td></tr>
<tr><td>집중도(%)</td><td>21.87</td><td>23.48</td><td>50.61</td><td>43.1</td><td>28.15</td><td>19.29</td><td>23.33</td></tr>
</table>
</td></tr>
</table>