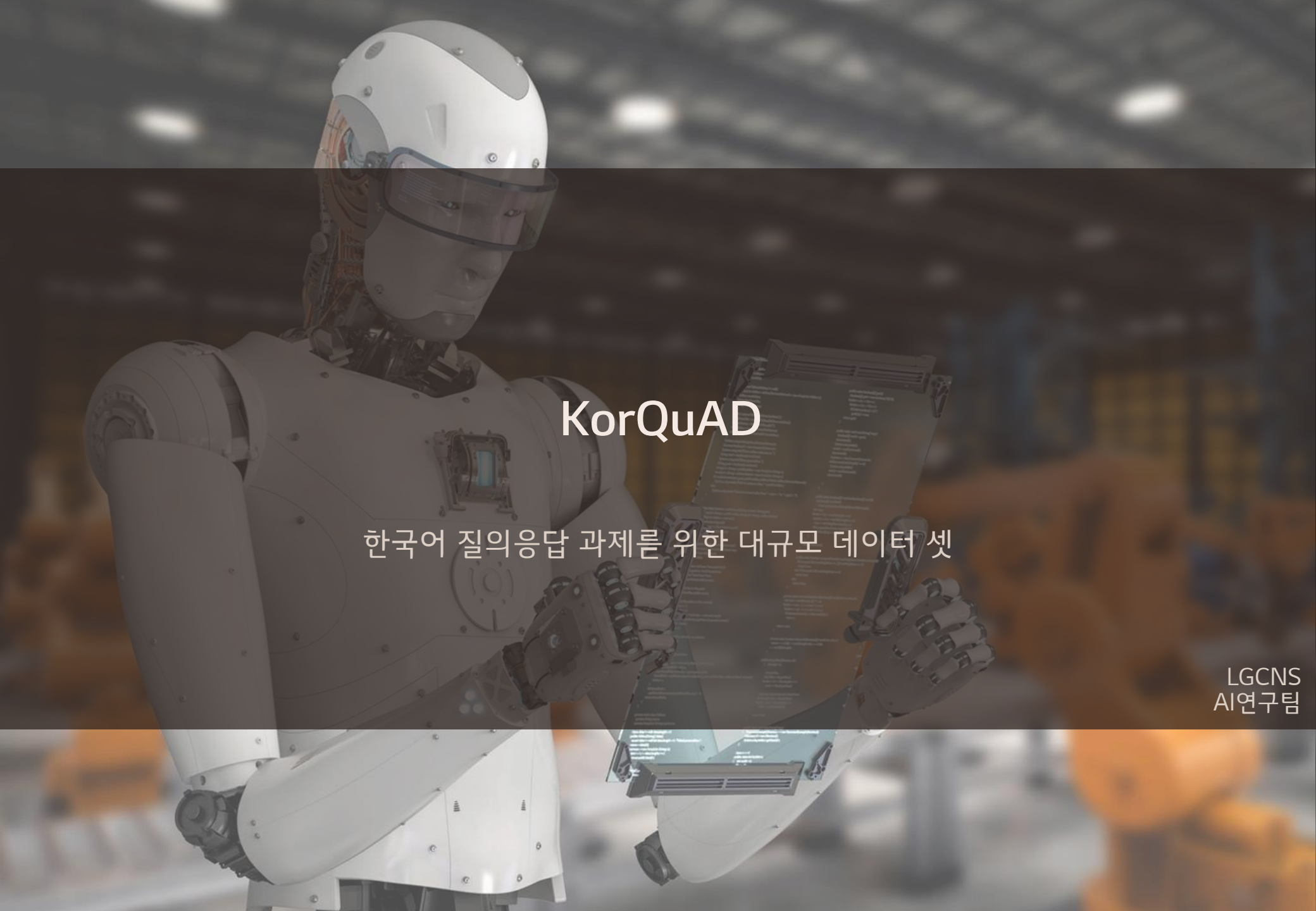


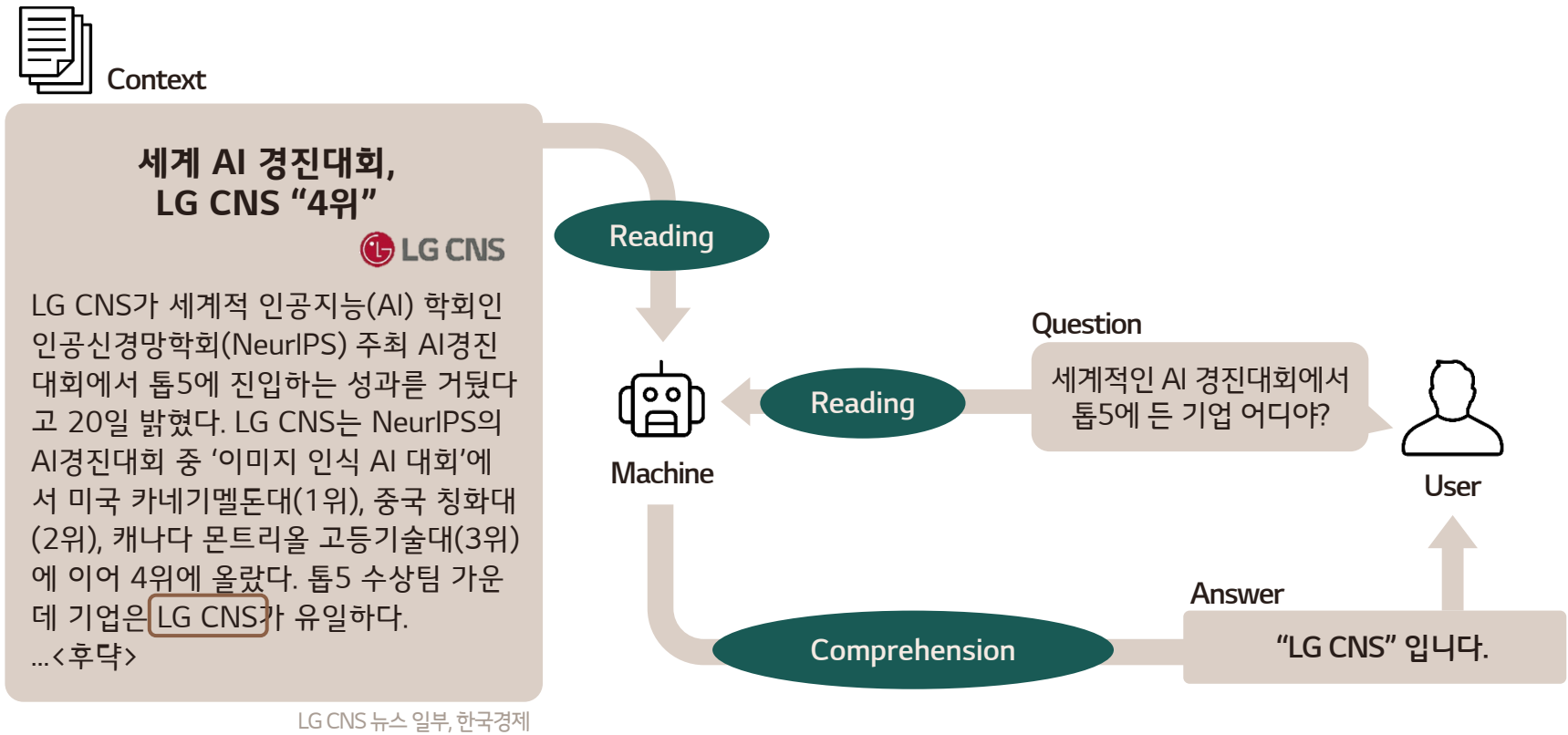
KorQuAD

한국어 질의응답 과제를 위한 대규모 데이터 셋

LGCNS
AI연구팀

1. Machine Reading Comprehension 이란?

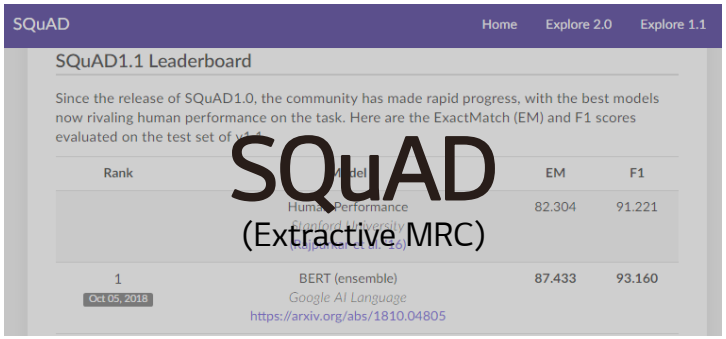
기계 독해 기반의 Question Answering

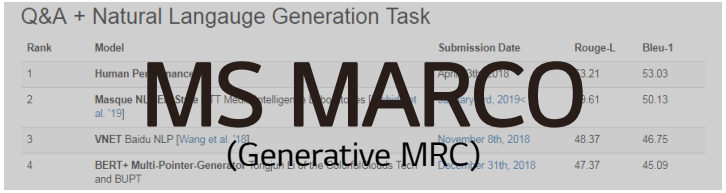


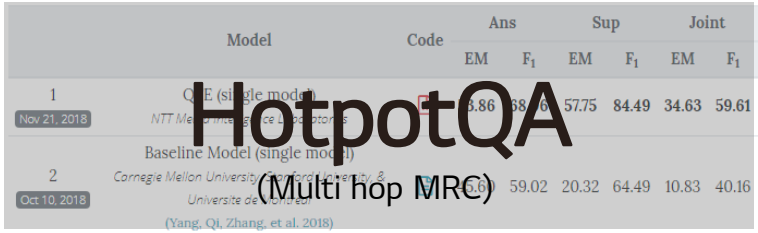
2. KorQuAD

한국어 Machine Reading Comprehension을 위한 데이터 셋

현존하는 다양한 영문 데이터







표준 데이터

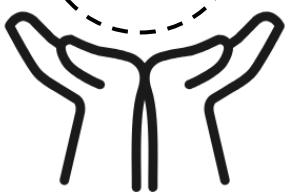
- 연구자들이 쉽게 데이터 확보
- 논문 저술 등 연구 환동에 환받하게 환용

리더보드

- 연구 성과 공유의 장
- 객관적인 기준으로 알고리즘 성능 평가

VS

한국어 데이터 셋 부재



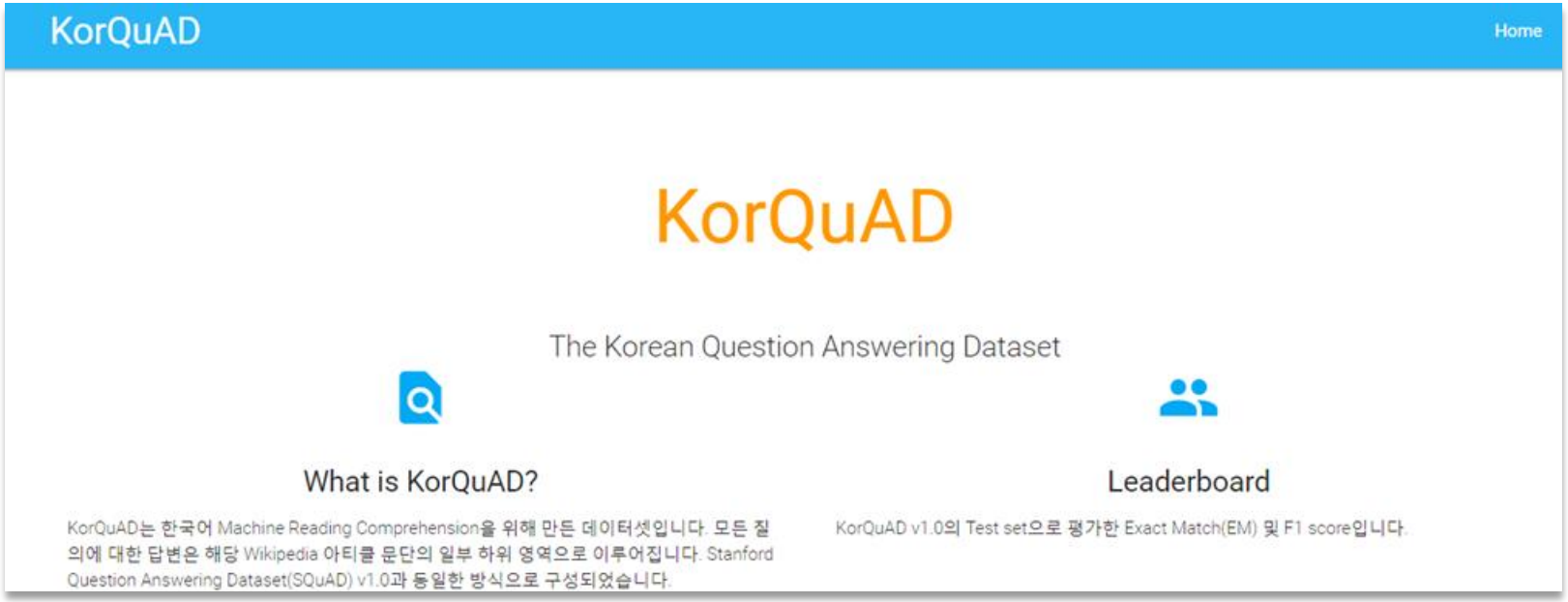
LG CNS

2. KorQuAD

한국어 위키백과를 대상으로 대규모 extractive MRC 데이터 구축

규모: TRAIN 60,407 / DEV 5,774 / TEST 3,898

주소: https://korquad.github.io/



SQuAD v1.0
데이터 생성 방식은
벤치마크하여 **표준성** 확보

Extractive MRC 연구용
다량의 학습데이터를
쉽게 확보

객관적인 기준을 가진
연구 결과 **성능 공유의 장**
마련

LG CNS

2. KorQuAD

KorQuAD Main Page 소개

Getting Started

KorQuAD는 한국어 Machine Reading Comprehension을 위해 만든 dataset입니다. 모든 질의에 대한 답변은 해당 Wikipedia 아티클 문단의 일부 하위 영역으로 이루어집니다. Stanford Question Answering Dataset(SQuAD) v1.0과 동일한 방식으로 구성되었습니다. 전체 데이터는 1,560 개의 Wikipedia article에 대해 10,645 건의 문단과 66,181 개의 질의응답 쌍으로, Training set 60,407 개, Dev set 5,774 개의 질의응답쌍으로 구분하였습니다.

Download Link

- TRAINING SET (37MB)
- DEV SET (3.9MB)

모델을 평가하기 위한 공식적인 evaluation script와 입력 샘플 prediction 파일을 제공합니다. 평가를 실행하려면 python evaluate-korquad_v1.0.py [path_to_dev-v1.0] [path_to_predictions] 를 입력하세요.

공식 평가 스크립트 제공

- EVALUATION SCRIPT
- SAMPLE PREDICTION FILE

Dev set에 대해 만족하는 모델을 만들었다면 공식 점수를 얻고 leaderboard에 올리기 위해 모델을 제출하세요. 테스트 결과의 무결성을 위하여 Test set은 공개되지 않습니다. 대신 모델을 제출하여 Test set에서 실행할 수 있도록 해야 합니다. 다음은 모델의 공식적인 평가를 위한 과정 안내 튜토리얼입니다.

CodaLab 튜토리얼 제공

- SUBMISSION TUTORIAL

Submission Tutorial (Coda Lab)

KorQuAD Submission Guide

? Keyboard Shortcuts

View | View source

이 튜토리얼은 KorQuad 1.0에 대한 공식적인 평가(evaluation)를 위해 모델 및 결과를 제출하는 과정에 대한 설명입니다. 모델이 공식적으로 평가되면 점수가 leaderboard에 추가됩니다.

시작하기에 앞서, CodaLab 계정을 만들고 CodaLab 튜토리얼을 숙지하십시오. 먼저 dev set에 대한 평가 방법을 자세히 설명한 다음, test set에 대한 평가 방법을 살펴보겠습니다.

Dev set 평가

Dev set에 대한 평가를 하기 위해 prediction 파일을 CodaLab에 업로드합니다. 업로드 과정을 자세히 설명하기 전에 예제와 같이 형식이 지정된 json 파일에 대한 prediction 파일이 있는지 확인하십시오.

uuid[0:8]	name	summary	data_size	state	description
0xcd4c5e	KorQuAD_v1.0_dev.json	[uploaded]	3.8m	ready	
0xbe7d3c	dev-sample-v1.0.json	[uploaded]	264k	ready	
0xbcb437	run-make	! make dev-evaluate-in1 -- dev-sample-v1.0.json{0xbe}	4.1k	failed	

uuid[0:8]	name	summary	data_size	state	description
0x8731ef	dev-evaluate-v2.0-in1	= elmo-pred.json{0x2d}	486k	ready	

CodaLab을 통해 모형 평가 후 TEST Set에서의 Score은 리더보드에 업로드

Leaderboard

Rank	Reg. Date	Model	EM	F1
-	2018.10.17	**Human Performance**	80.17	91.20

(Test dataset은 홈페이지에 공개하지 않습니다)

2. KorQuAD_ data collection

▶ 대상 문서 수집

Document Crawling

위키 백과

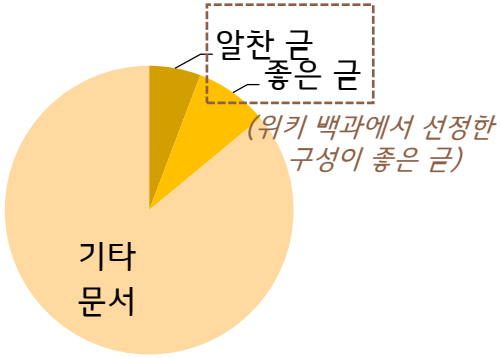
[**알찬 글**]& [**좋은 글**] 목록으로부터

각각 100건, 143건의 문서 우선 수집

→ 양질의 문서 우선 확보

추가적으로 위키백과 랜덤 탐색을 통해

총 **1,637**건의 문서 수집



▶ 질문/답변 생성

Extract Passages

수집한 문서는 **문단 단위**로 정제

이미지/표/URL 제거

▶ 2차 답변 태깅

Passage Curation

300자 미만의 짧은 문단 제거

수식이 포함된 문단 제거

2. KorQuAD_ data collection

▶ 대상 문서 수집 | ▶ 질문/답변 생성 | ▶ 2차 답변 태깅

작업 환경 예시

《해리 포터》(Harry Potter)는 1997년부터 2007년까지 연재된 영국의 작가 J.K. 롤링의 판타지 소설 시리즈다. 이모네 집 계단 밑 벽장에서 생활하던 11살 소년 해리 포터가 호그와트 마법학교에 가면서 겪게 되는 판타지 이야기를 그리고 있다. 1997년 6월 첫 번째 책인 《해리 포터와 마법사의 돌》이 출판되었으며, 2007년 7월 마지막 책인 《해리 포터와 죽음의 성물》이 출판되었다. 해리포터 시리즈가 큰 성공을 거두면서 전 세계적으로 인기를 얻었으며, 영화를 비롯한 비디오 게임 및 다양한 상품들이 제작되었다.

질문을 입력하세요:

답변 영역을 드래그하세요:

띄어쓰기 단위도 정답 영역 선택 후
조사 등은 삭제해 최소 영역 선택

작업자 구성

- 크라우드소싱을 통해 QA 70,000+쌍 생성
- 일정 등급 이상의 작업자만 참여 가능

작업 방식

- 한 사람은 하나의 문단에 대해 2-3개 질문 생성
- 하나의 문단은 3명에게 할당함
  → 한 문단 당 총 6~9개의 질문 생성
  → 질문 어휘의 다양성을 유도

환경 구성

- 작업 환경 구성 시 Copy& Paste를 방지
- 자신의 단어도 질문을 생성할 것을 강력하게 명시
- 질문 예시에 대한 상세 가이드라인 제시

LG CNS

2. KorQuAD_ data collection

▶ 대상 문서 수집 | ▶ 질문/답변 생성 | ▶ 2차 답변 태깅

KorQuAD 데이터 셋 통계

	TRAIN	DEV	TEST	TOTAL
문서	1,420	140	77	1,637
질문	60,407	5,774	3,898	70,079

140개 문서에 대해
각각 2개의 질의를 랜덤 추출

2차 답변 태깅 대상

목적

- Human performance 측정

방법

- 2차 작업자는 문단 & 질문을 보고 답변 영역 선택

결과

- TEST 데이터 EM 80.17% / F1 91.20%
- (참고) SQuAD v1.1 human performance

Model	EM	F1
Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221

LG CNS

2. KorQuAD_ evaluation metric

METRIC

EM:
실제 정답과 예측치가 정확하게 일치하는 비율.

F1:
실제 정답과 예측치의 겹치는 부분을 고려한 점수로, EM보다 완화된 평가 척도.
어절 단위의 F1이 표준인 영문과 달리 한국어에서 어절 단위로 F1을 구할 경우 어절 내 **다양한 형태소 활용** 등으로 인해 점수가 다소 낮게 측정됨. 따라서 보다 적합한 **음절 단위의 F1**을 도입.

▽ KorQuAD F1 계산 예시

런던 대화재(Great Fire of London)는 ... 조기에 진화되지 않아, 5일간 87채의 교회, 1만 3천 채의 집이 붙탔다.

Q: 런던 대화재 당시 수많은 가구의 피해는 며칠 동안에 일어났는가?

Ground Truth : 5일간 (*영문: for 5 days*)
Predicted Answer : 5일 (*영문: 5 days*)

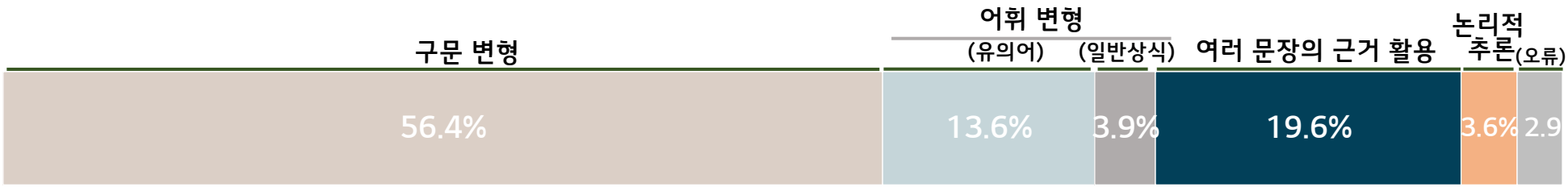
어절 단위 F1	음절 단위 F1	SQuAD (영문)
0%	80%	80%

- 띄어쓰기 (어절) 단위 F1 측정 → 형태소의 조합으로 생성된 다양한 활용형은 0점으로 채점
- 형태소 단위의 F1 측정 → 동일된 형태소 분석기 등의 부재로, 표준으로는 부적합하다는 판단
- **글자 (음절) 단위의 F1** 측정 → 위의 단점을 보완 & 표준으로 채택함

2. KorQuAD_ dataset 특성

* Dev set140개 문서에 대해
각각 2개의 질의를 랜덤 추출하여 분석한 결과임

KorQuAD 질문 유형



구문 변형 (56.4%)	Q: 제 1회 문학동네 작가상을 수상한 작품으로, 96년 발표된 장편소설은?
	... 작품활동을 시작하였고 이듬해 96년 장편《나는 나를 파괴할 권리가 있다》로 제1회 문학동네 작가상을 수상하였다.
어휘 변형 (유의어) (13.6%)	Q: ... 서덜랜드가 무엇을 발전시킨 것을 시작으로 연구가 **시작**되었는가?
	서덜랜드가 see-through HMD를 발전시킨 것을 **시초**로 연구... 증강현실은 ...
어휘 변형 (일반상식) (3.9%)	Q: **해외**에서 활동하는 Kayip, Superdrive와 결성한 프로젝트 그룹의 이름은?
	영국에서 활동하고 있는 Kayip, **베를린**에서 ... Superdrive와 ... '모텟'을 결성

여러 문장 근거 종합적 활용 (19.6%)	Q: 클레멘스가 명예의 전당에 입성하지 못한 이유는?
	... 이 투수들이 **클레멘스**를 제외하고 모두 명예의 전당에 올랐기 때문이다. 클레멘스만이 ... 받았다. 그는 경기력 향상 약물 사용에 연루되어 있기 때문에 **입성 여부가 불확실**...
논리적 추론 요구 (3.6%)	Q. 대한민국 제17대 대통령 선거 당시 후보로 등록했으나 예비경선의 경선 후보로 뽑히지 못한 사람 중 법무부와 관련 있는 사람은?
	정동영 전 열린우리당 의장, ..., **천정배 전 법무부 장관**, ... 등이 후보로 등록하였고... 예비경선으로 **정동영, 손학규, 이해찬, 한명숙, 유시민 후보가 경선 후보로 결정되었다**.
기타 출제 오류 (2.9%)	Q. 티베트 고원에서 발원하는 강은?
	... 강들이 티베트 고원에서 발원하는데, **창장, 황허, (...), 살윈 강** 등이 포함된다.

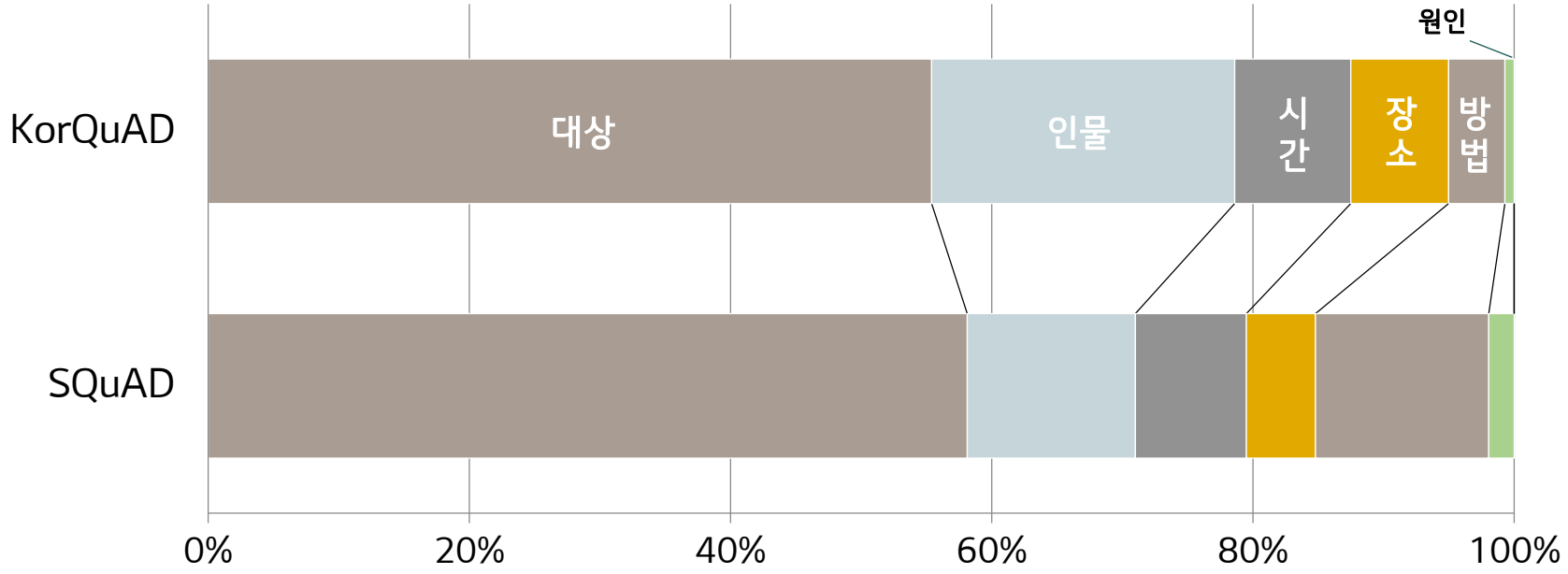
LG CNS

2. KorQuAD_ dataset 특성

* Dev set140개 문서에 대해
각각 2개의 질의를 랜덤 추출하여 분석한 결과임

KorQuAD 답변 유형

대상	인물	시간	장소	방법	원인
55.4%	23.2%	8.9%	7.5%	4.3%	0.7%



⇒ 영문 표준 데이터와 특성이 유사함을 확인

LG CNS