●○ 한국어 텍스트 과제

도서자료 기계독해

●○ 개요: 기계독해 데이터셋이란?

딥러닝 기반 기계 독해(MRC: Machine Reading Comprehension) 기술 개발에 활용할 수 있는 학습 데이터셋으로 ㈜유클리드 소프트에서 구축했으며, 100만 건의 한국어 질문과 대답으로 구성되어 있다. (구축건수 진행상황에 따라 변동 가능성 있음)

기계 독해 기술은 사람이 맥락을 이해하고 논리적으로 답을 찾는 것처럼 질의에 대한 답을 찾는 기술이며, AI 챗봇 상담, 비정형 텍스트 형태의 전문지식에 대한 질의응답 등 키워드가 아닌 자연어 형태의 질의를 받아 지식 정보를 제공하는 서비스에 활용될 수 있다.

기계 독해의 사례 대해서는 아래를 참고할 수 있다.

Context: 인도정부는 영주권 취득을 위한 외국인 투자 하한을 ① 18개월 내에 1억 루피(150만 달러), 혹은 ② 36개월 내에 2억5000만 루피(360만 달러)로 책정하였다. 해당 투자가는 매 회계연도 기준 최소 20명의 인도인들을 고용해야 한다. PRS를 통해 외국인 투자자는 최대 20년까지(기본 10년+추가 10년 연장) 횟수제한 없이 인도 입출국이 가능하며, 정부의 이와 같은 결정은 "Make in India"정책에 따른 자국 내 투자 활성화 및 일자리 창출을 도모에 기반한다. 인도 신정부는 Make in India 정책의 일환으로 자동차, 섬유, 전자, 항공, IT, 건설 등 25개 중점사업을 지정하고 해당 분야 투자가들에게 부문별 인센티브를 제공하고 있다.

[일반적인 질의응답]
Q. 외국인 투자자는 36개월 내에 얼마를 투자해야 인도 영주권을 취득할 수 있어?
A. 20억5000만 루피

[지문에 정답이 없는 질문의 학습]
Q. 외국인 투자자가 인도를 설치하려면 36개월 동안 얼마를 투자해야 해?
(Plausible) A. 20억 50만 루피

그림1 | 기계 독해의 사례

●○ 데이터셋의 구성

본 데이터셋은 대표적인 기계독해 학습용 데이터셋인 SQuAD와 같은 지문(Context)-질문(Question)-답변(Answer)으로 이루어진 데이터셋 형태로, 질문의 답변 여부에 따라 정답이 있는 데이터셋 70만 건과 정답이 없는 데이터셋 30만 건으로 구성되어 있다. 정답이 없는 데이터셋은 질문에 대한 답이 지문에 없을 경우 기계 독해 모델이 정답이 없음을 판단할 수 있도록 학습하기 위해 사용된다. 정답이 있는 데이터셋과 동일하게 지문-질문-답변의 형태이지만, 답변은 질문의 의도에 대응하는 정답처럼 보이지만 의미적으로 틀린 대답(plausible answer)을 채택하고 있다.

데이터 종류	포함 내용	제공 방식
지문-질문-답변 데이터셋	정답이 있는 데이터셋 (70만 건)	JSON 파일
	정답이 없는 데이터셋 (30만 건)	

●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

항목		설명	예시
version	버전	SQuAD 버전 정보	v2.0
time	생성일시	질문답변 데이터가 생성된 일시	202010131500
title	표제	지문의 출처인 자료의 제목	주요국 투자유치인센티브 제도 조사
agency	발행처	자료를 발행한 기관	KOTRA
year	발행연도	도서자료의 발행연도	20090226
content_id	컨텐트 번호	도서자료의 ID	CNTS-00049179205
KDC	주제분류	KDC 중분류	320
paragraphs	지문	도서자료에서 추출한 지문과 작업자가 생성한 질문-답변으로 이루어진 데이터쌍	-
context	지문	질문을 생성할 지문 텍스트	-
qas	질문답변 셋	지문에 대한 질문답변 셋	-
question	질문	질문 텍스트	인도에서 외국인 투자유치를 지원하는 기관이 어디야

항목		설명	예시		
id	질문 번호	질문의 아이디	KOTRA0001-1		
is_impossible	정답 여부	정답이 있는 질문과 정답이 없는 질문을 boolean으로 구분	false: 정답이 있는 질문 true: 정답이 없는 질문		
answers	답변	-			
text	답변 텍스트	-	산업정책진흥국		
answer_start	답변 위치	지문에서의 답변의 오프셋	85		

●○ 데이터 예시

이 데이터 셋 예시는 정답이 있는 데이터와 정답이 없는 데이터 모두에 해당된다.


```
{
  "version": "v2.0",
  "data": [
    {
      "title": "주요국 투자유치인센티브 제도 조사",
      "agency": "KOTRA",
      "time": "202010131500",
      "year": "201809--",
      "content_id": "abs000000",
      "kdc": "320",
      "paragraphs": [
        {
          "context": "(국가 투자유치 전담기관 활성화) 인도의 외국인 투자유치 담당 부처는 산업정책진흥국(DIPP; Department of Industrial Policy and Promotion)이며, Invest India라는 산하 투자유치 전담기관을 통해, 잠재적 투자자들에게 분야별, 지역별 정보를 제공하고 사전 투자 결정에서 사후 관리까지 전체적인 투자유치 과정을 지원",
          "qas": [
            {
              "question": "인도에서 외국인 투자유치를 지원하는 기관이 어디야?",
              "id": "KOTRA0001-1",
              "is_impossible": false,
              "answers": [
                {
                  "text": "Invest India",
                  "answer_start": 122
```

```
                    }
                ]
            },
            {
                "question": "인도에서 내국인 투자유치 과정을 지원하는 정부부처가 어디야",
                "id": "KOTRA0001-2",
                "is_impossible": true,
                "answers": [
                    {
                        "text": "산업정책진흥국",
                        "answer_start": 85
                    }, …]
            }, …]
        },…]
    },…]
}
```


※ 한 지문(“content”)에 대해 질문답변 데이터(“qas”)가 여러 개일 수 있다.

●○ 데이터 구축 과정

‘국립중앙도서관’에서 저작권에 문제가 없는 60,506건의 자료를 공급받아 텍스트 추출이 가능한 도서자료 22,129건을 기준으로 다음과 같이 정제 및 가공작업을 진행함

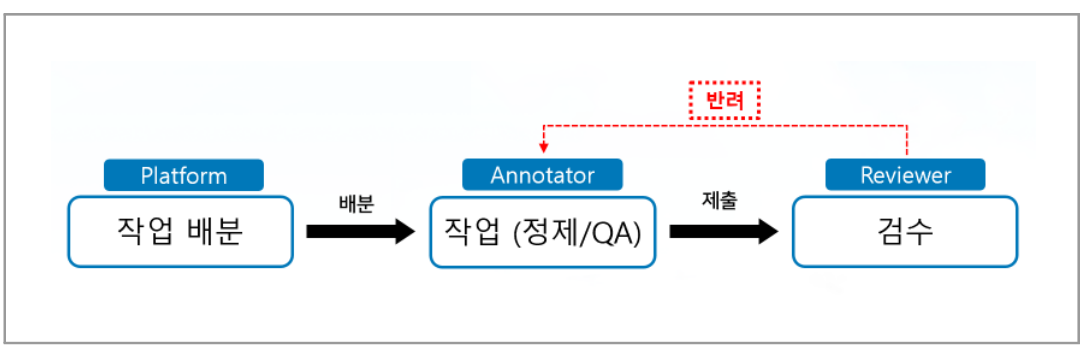

그림2 | 데이터 구축 프로세스

- 정제작업
  - 도서내용 중 중요내용을 담은 300~500글자의 지문을 선택하고 표, 각주, 그림 등을 제거하는 작업을 진행함

- 정제가 어려운 형식이거나, 내용이 어려워 질문, 답변을 할 수 없는 형식에 대해서는 작업을 제외함(논문, 국한문 혼용, 영문자료, 각종 행정보고서 등)
- 목표량 : 정제된 지문 25만개

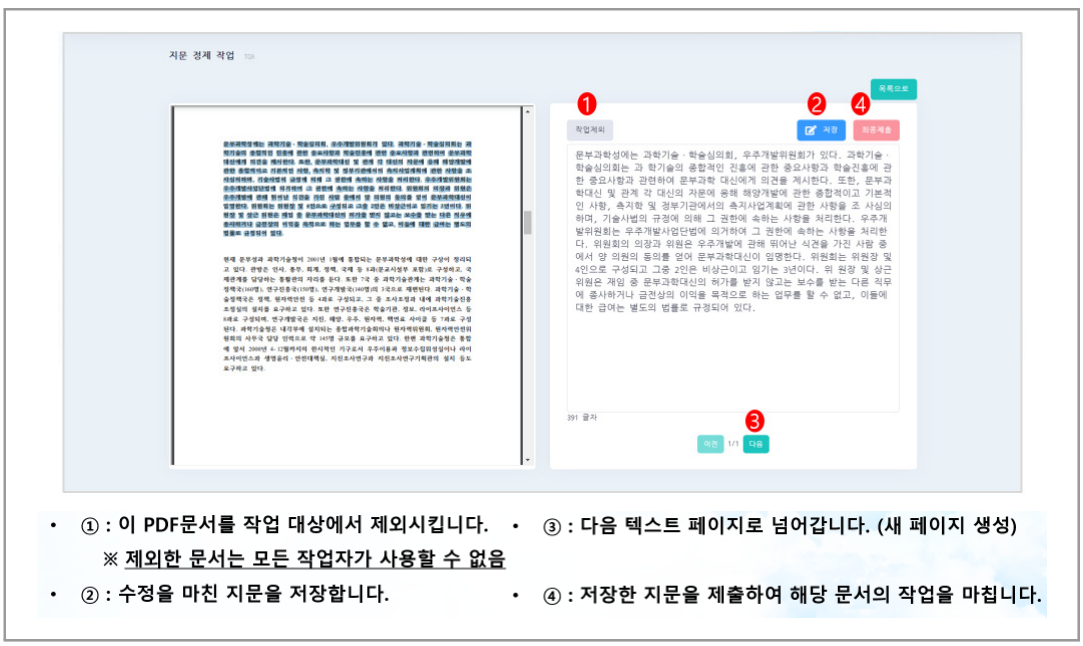

그림3 | 지문정제 작업도구 화면 및 설명

- 가공작업
  - 정제된 지문 안에서 적정한 답변을 선택하고, 답변에 호응하는 질문을 3개, 답변과 호응하지 않는 질문을 1개 작성하도록 함
  - 답변에 호응하는 질문은 1.지문내 어휘사용 2.동의어,유사어 사용 3.일상용어 사용, 이렇게 세 가지 유형으로 작성함
  - 질문, 답변 작성이 어려운 지문에 한하여 작업제외함
  - 목표량 : '지문-질문-답변'으로 이루어진 데이터 셋 100만개

그림4 | 가공작업도구 화면 및 설명

●○ 검수와 품질 확보

- 품질을 위한 교육
  - 정제와 가공 모두 난이도가 있는 작업이라고 여겨져 주 2~3회 온오프라인 교육을 통해 크라우드소싱 인력을 교육함
  - 검수과정을 통해서도 반려사유를 알려줌으로써 교육이 이루어짐
- **검수체계**
  - 정제 검수
    - 정제된 지문에서 삭제해야 할 부분이 있는지, 띄어쓰기 및 맞춤법이 틀린 부분이 있는지, 중복된 지문이 있는지 체크 후 수정
  - 가공 검수
    - 작성된 질문과 답변이 작성 기준에 맞는지 체크 후 반려 및 수정 제출
  - 샘플링된 데이터 셋 검토
    - 구축된 데이터를 초기부터 일정 기간마다 샘플링하여 주관기관으로 전송하여 구축 가이드와 학습에 유효한 데이터인지 검토받음
    - 검토사항을 데이터 구축에 적용함

●○ 데이터 구축 담당자

수행기관: ㈜포티투마루
(전화: 02-6952-9201, 이메일: bd@42maru.ai)

참여기관: ㈜유클리드소프트
(전화: 042-365-6589, 이메일: tqa@euclidsoft.co.kr)