

# Open-Domain QA 프로젝트

[Table of Contents](#) ( Ctrl + click )

[Team Introduction](#)

[Members](#)

[Project Overview](#)

[Dataset](#)

[Evaluation](#)

[Project 세부 & 적용 방법론](#)

[Data Split & Augmentation](#)

[Model](#)

[Reader](#)

[Retrieval](#)


[Hyper Parameter Tuning](#)


[Post processing & Ensemble](#)


## Team Introduction


런앤런 (Run & Learn) : 런앤런이라는 팀명은 저희 팀이 가장 처음 세운 2가지 목표를 나타내기 위한 이름으로, 마지막까지 최선을 다해 달려가자는 목표를 나타내는 *Run*, 프로젝트의 성과도 중요하지만 서로의 배움과 성장을 더 중요시하자는 다짐을 나타내는 *Learn* 으로 이루어져 있습니다.

## Members

 강석민 [Github](#)


 김종현 [Github](#)

 김태현 [Github](#)

 오동규 (Me) [Github](#)

 윤채원 [Github](#)

 최재혁 [Github](#)

 허은진 [Github](#)

## Project Overview

### Open Domain Question Answering (ODQA)

: 주어진 지문 없이 사전에 구축되어 있는 지식(Knowledge Resource)을 기반으로 사용자의 질문에 대한 답변을 출력하는 태스크

### Two-Stage Approach

1. Retriever : 질문에 대한 답을 포함하고 있을 것으로 예상되는 문서를 찾아오는 단계. 입력은 사용자의 질문이며, 질문과 유사도가 높은 상위 N개의 문서를 반환.
2. Reader : retriever 단계에서 불러온 문서 내에서 질문에 대한 답을 찾아 사용자에게 전달한다. 질문과 Context로 주어지는 문서가 함께 입력으로 사용되며, 답변을 지문 내에서 추출하거나 직접 생성하는 2가지 방식이 존재한다.

## Dataset

KLUE - MRC (<https://klue-benchmark.com/tasks/72/overview/description>) 데이터셋 일부와 대회에서 주어진 위키피디아 말뭉치를 활용

### KLUE - MRC

데이터 구성: Train 3952개 / Validation 240개

- Train example (context / question / answers)

```
{__index_level_0__: 572,
 answers: {'answer_start': [1028], 'text': ['1890년 5월 18일']},
 context: '중서서원에서 4년간 공부하며 윤치호는 개신교 선교사들의 영향을 받게 되었고, 중서서원 재학 동안 열심히 서양의 문물을 접하며 중국을 세계의 중심으로 보던 조선인들의 중화사상(中華思想)에 입각한 사고방식에서 벗어나게 되었으며, 낙후된 조선과 중국에 대한 강한 비판의식과 낙후된 조선 사회의 현실에 절망, 조선 근대화에 대한 비판적, 부정적인 인식을 갖게 되었다. 상하이에서 3년 반을 보낸 후 청국(淸國) 사회에 대한 그의 소감은 '더러운 물로 가득 채워진 연못'이었다. 반면 일본은 '동양의 한 도원(桃源)'이었다 윤치호에게는 본부인 진주강씨 외에 두 명의 집이 있었던 듯 하다. 그가 상하이에 체류하는 동안 그의 두 번째 집은 다른 남자에게 개가했다. 1886년에는 그의 첫 부인인 진주강씨가 사망했다. 그가 상하이로 망명하고, 그의 아버지 윤용렬은 농주로 유배되었을 무렵이었다. \n\n이후 윤치호는 10여 년간 중국과 미국으로 망명·유학하여 문물을 접하고, 서구 민권사상과 기독교 신앙을 수용했으며, 그는 상해와 미국에서 그의 재능을 높이 평가한 남강리교 선교사들의 지원을 받으며, 마음껏 학업에 정진할 수 있었으나 5년간 미국유학중에는 생활비를 고학으로 충당했다. 영 J. 말렌과 W. B. 보넬 교수의 영향으로 개신교에 귀의를 결심하여 1887년 4월 3일 상하이에서 "예수를 주로 고백하고" 세례를 받고 개신교 신자가 되었다. 그가 개신교 신자가 되게 된 배경에는 4년 여되는 기간 동안의 개신교 연구와 수련이 있는 것으로 보고 있다. \n\n그는 노동을 천시, 경시하는 사농공상의 풍조와 출세욕, 관직열에 빠진 조선의 배관열을 이해할 수 없었다. 유학기간 중 그는 서구의 민권사상과 합리주의, 직업윤리 의식, 민중의 참정권을 수용, 개혁의 필요성을 확신하게 되었다. \n\n내나라 자랑할 일은 하나도 없고, 다만 흉잡질 일만 많으며 일본 한심하며, 일본 일본이 부러워 못견디겠도다. |윤치호|윤치호일기 1888년 12월 29일자\n조선이 지금의 아만적 상태에 머무느니 차라리 문명국의 식민지가 되는 게 낫겠다. |윤치호|윤치호일기 1890년 5월 18일자 } \n\n1890년대 초반 미국 체류시 윤치호는 사회진화론을 최고의 진리로 받아들여 중국인들에 대한 미국 사회의 무시와 억압과 중국인에 대한 인종주의적인 차별 행위까지도 옹호했다. 그러나 합리주의적인 사회를 부정하지는 않았다.,
 document_id: 5279,
 id: mrc-0-000857,
 question: 윤치호가 조선이 문명국의 식민지가 되는 것이 낫다고 한 날은?,
 title: 윤치호,
 difficulty: 3}
```

### 위키피디아 말뭉치

- Data example (text , title)

```
{text: 이 문서는 나라 목록이며, 전 세계 206개 나라의 각 현황과 주권 승인 정보를 개요 형태로 나열하고 있다. 이 목록은 명료화를 위해 두 부분으로 나뉘어 있다. 첫 번째 부분은 바티칸 시국과 팔레스타인을 포함하여 유엔 등 국제 기구에 가입되어 국제적인 승인을 널리 받았다고 여기는 195개 나라를 나열하고 있다. 두 번째 부분은 일부 지역의 주권을 사실상 (데 팩토) 행사하고 있지만, 아직 국제적인 승인을 널리 받지 않았다고 여기는 11개 나라를 나열하고 있다. 두 목록은 모두 가나다 순이다. 일부 국가의 경우 국가로서의 자격에 논쟁의 여부가 있으며, 이 때문에 이러한 목록을 엮는 것은 매우 어렵고 논란이 생길 수 있는 과정이다. 이 목록을 구성하고 있는 국가를 선정하는 기준에 대한 정보는 "포함 기준" 단락을 통해 설명하였다. 나라에 대한 일반적인 정보는 "국가" 문서에서 설명하고 있다.,
corpus_source: 위키피디아,
url: TODO,
domain: None,
title: 나라 목록,
author: None,
html: None,
document_id: 0,
__index_level_0__: 0}
```

## Evaluation

EM score 기준 리더보드 반영.

### 1. Exact match (EM)

- 모델의 예측과 실제 답이 정확하게 일치하는 경우에만 점수가 주어진다.
- 모든 질문이 0점 혹은 1점으로 처리
- 단, 띄어쓰기나 “ . “ 같은 일부 문자를 제외한 후 정답 단어에 대해서만 일치 여부 확인
- 답이 여러가지인 경우, 하나라도 일치하면 정답으로 간주

### 2. F1 score

- Precision (정밀도)과 Recall (재현율)의 조화평균.
- 분류 클래스 간의 데이터 불균형이 존재할 때 적합한 측정지표.

### 3. 최종 리더보드 순위 4위 기록

순위	팀 이름	팀 멤버	EM	F1	제출 횟수
4 (2 )	MRC_1조	E 재혁 종현 채원	70.2800	80.1800	45

## Project 세부 & 적용 방법론

### Data Split & Augmentation

- 데이터 예측 난이도에 따른 Train/Validation Data Split (my own things)

- 1) 전체 데이터를 랜덤하게 5개의 Fold로 분리.
- 2) KorQuAD 데이터셋으로 fine-tuning 되어 있는 MRC 모델을 불러와 1개의 Fold 로만 학습을 진행한다.
- 3) 이후 나머지 4개의 Fold에 대한 예측을 수행하고,
- 4) 해당 과정을 5개의 Fold에 대해 진행하면 5개의 모델이 예측을 수행한 결과가 각 데이터마다 4개씩 만들어진다.
- 5) 4개의 예측 결과에 대한 정답률에 따라 해당 데이터의 예측 난이도를 판단하고, 학습 데이터와 검증 데이터에 난이도 별로 분류된 데이터가 고루 섞일 수 있도록 데이터를 재구성

- Wikipedia 문서 기반 질문 생성 (my own things)

- Pororo 라이브러리의 Question generation , NER 모델 이용.
- 위키피디아 말뭉치의 지문 내에서 개체명을 추출하고 질문 생성 모델에 지문과 함께 입력으로 넣어주면 적합한 질문을 생성해주어 Reader 모델 학습을 위한 데이터로 사용할 수 있다.
- 생성된 데이터의 품질을 판단하기 위해 아래 두가지 조건을 적용해 데이터를 필터링.
  1. Retrieval 모델에 생성된 질문을 넣었을 때, top-5 문서 내에 매칭되는 지문이 포함되어 있다면 적합한 질문을 생성했다 판단.
  2. Data Split 단계에서 사용한 MRC 모델을 통해 Inference를 수행하였을 때 모델이 정답을 잘 찾아내면 질문이 잘 생성된 데이터라 판단.
- 하지만 Augmentation 데이터를 학습에 사용했을 때 오히려 성능이 약간 하락하는 모습을 보였고, 아래의 문제점으로 인해 생성된 데이터가 오히려 모델 학습을 방해하는 요인으로 작용한 것으로 추측된다.
  1. Sparse embedding (BM25) 기반 Retrieval 모델 사용으로 질문과 지문 사이의 유사도를 단순히 매칭되는 토큰을 기반으로만 판단한 점.
  2. fine-tuning 된 모델이 쉽게 맞출 수 있는 데이터만 이용한 점.

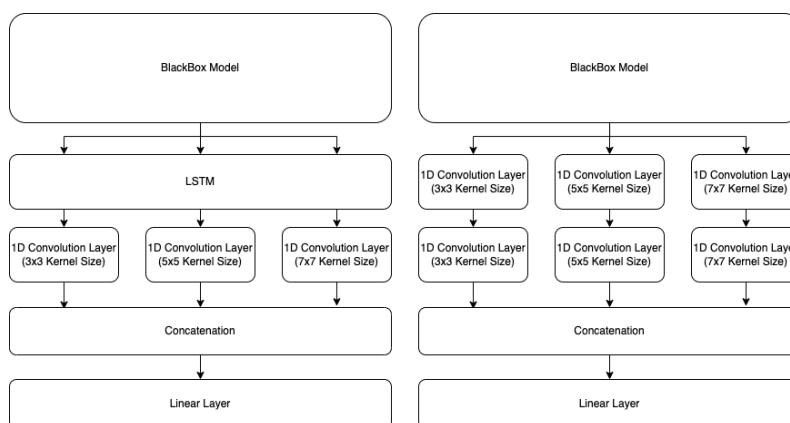
- Distant Supervision (my own things)

- Distant supervision은 QA 모델 학습 시 질문과 연관성이 높지만 실제 정답은 포함하고 있지 않은 지문을 모델 학습에 이용한 Augmentation 방법.
- Retrieval 모델을 이용해 검색한 지문 중 정답을 포함하는 문서 2개와 포함하지 않는 문서 2개를 이용해 학습을 수행.
- 이후 Hyper parameter tuning 단계와 앙상블 단계에서 유의미한 성능 향상을 보여주었음.

## Model

- Reader Model 성능 개선 실험

- Custom Output Layer 추가. (LSTM , Convolution)



- 기본 Linear Layer 대비 일정 수준의 성능 향상과 모델 앙상블에서 효과를 보임.

- Backbone freeze & 커리큘럼 러닝 실험

- Backbone으로 사용한 언어 모델을 Output에 가까운 층부터 순차적으로 학습 진행.
- Data Split 단계에서 구분한 데이터 예측 난이도에 따라 쉬운 데이터부터 어려운 데이터로 순차적으로 학습 진행.

- Span Masking & Multi-task Learning (my own things)

- SpanBERT 논문에서 제시된 방법론을 참조하여 모델의 입력으로 들어가는 질문과 지문 Pair에 Span 단위로 Masking을 적용하고, Masking 된 Token의 예측을 수행

하는 MLM Objective와 질문에 대한 정답을 지문에서 찾아내는 Extraction-based MRC Objective를 함께 학습하도록 구현. (EM score 기준 모델 성능 약 2-3% 가량 개선)

## • Retrieval Model 성능 개선 실험

- 고유명사 기반 Tokenizing
  - BM25 알고리즘 적용 시 Tokenizer를 개선하여 retrieval 성능 개선
  - QA 데이터셋의 특성상 고유명사와 일반명사가 중요한 역할을 차지하고, 많이 등장하는데 기존 klue-bert-base의 서브워드 토큰라이저가 고유명사를 제대로 구분하지 못하는 현상을 발견.
  - Konlpy의 한국어 형태소 분석기를 조사하여 고유명사와 일반명사 구분 성능이 가장 좋았던 Komoran과 동사 정규화 기능이 포함된 Okt를 함께 사용하여 고유명사, 일반명사, 동사만으로 sparse retrieval 알고리즘을 수행하도록 변경.
  - Top-10 문서 검색 정확도 기준으로 82-84%를 기록하던 정확도를 88-90% 가량까지 개선하였음.
- Elasticsearch 이용
  - Elasticsearch 에서 제공되는 Nori 토큰라이저와 BM25 알고리즘 사용.
  - Top-10 문서 검색 정확도 기준 88% 기록.

## Hyper Parameter Tuning

- SigOpt 라이브러리 이용.
- klue-roberta-large 모델과 Distant supervision을 적용한 데이터셋을 이용하여 진행.

## Post processing & Ensemble

- Reader 모델의 출력으로 나온 단어에 조사가 붙어 나오는 경우가 자주 발생.
- Mecab 형태소 분석기를 이용해 끝 단어에 조사가 붙어있는 경우 제거해주었음.
- 학습한 모델들의 예측 결과를 Hard Voting 하여 앙상블.