ODQA Solution 발표

쿼터백 조

김다영 * 김다인 * 박성호 * 박재형 * 서동건 * 정민지 * 최석민

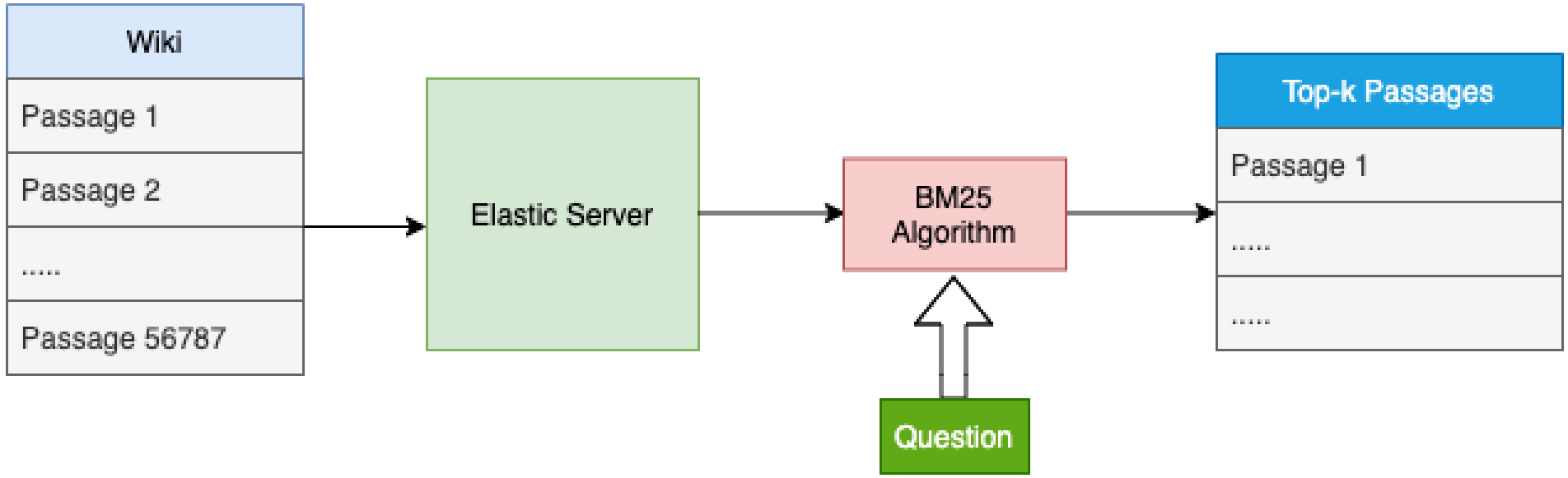
목차

Preprocessing

```python
def preprocess(text):
    text = re.sub(r"\n", " ", text)
    text = re.sub(r"\\n", " ", text)
    text = re.sub(r"\s+", " ", text)
    text = re.sub(r"#", " ", text)
    text = re.sub(
        r"[^a-zA-Z0-9가-힣ㄱ-ㅎㅏ-ㅣぁ-ゔァ-ヴー々〆〤一-鰀<>()\s\.\?!》《≪≫\'<> 〈〉 :‘’%,『」 「」 < > · \"-“”∧ ]",
        "",
        text,
    )
    return text
```

- 토론 게시판 김범찬 캠퍼님의 게시글을 바탕으로 정답에 들어가는 특수문자를 파악하여 특수문자 전처리 진행
- Retrieval과 Reader 모두 예측 성능이 향상되었음

Retrieval



Retrieval - Elastic Search

Elastic Search Settings

```
settings": {
 "analysis": {
   "filter": {
     "my_shingle": {
       "type": "shingle"
     },
     "my_stemmer": {
       "type": "stemmer"
     }
   },
   "analyzer": {
     "nori_analyzer": {
       "type": "custom",
       "tokenizer": "nori_tokenizer",
       "decompound_mode": "mixed",
       "filter": ["my_shingle", "my_stemmer"]
     }
   },
```

Elastic Search Filter	Top k = 1	Top k = 10	Top k = 20
BM25 + nori	0.686	0.890	0.923
Shingle	0.709	0.897	0.929
Stemmer	0.686	0.891	0.923
Shingle + Stemmer	**0.710**	**0.898**	**0.929**

Retrieval - Elastic Search

"제 2차 세계 대전에 참전하여 사망한 자식은?"

[('제2차', 'QUANTITY'), (' ', 'O'), ('세계 대전', 'EVENT'), ('에', 'O'), (' ', 'O'), ('참전하여', 'O'), (' ', 'O'), ('사망한', 'O'), (' ', 'O'), ('자식', 'CIVILIZATION'), ('은?', 'O')]

```
query = {
        "query": {
            "bool": {
                "must":[
                    {"match": {"document_text": question_text}}
                ],
                "should":[
                    {"match": {"document_text": ' '.join([i[0] for i in tagged if i[1]!='O'])}}
                ],
            }
        }
    }
```

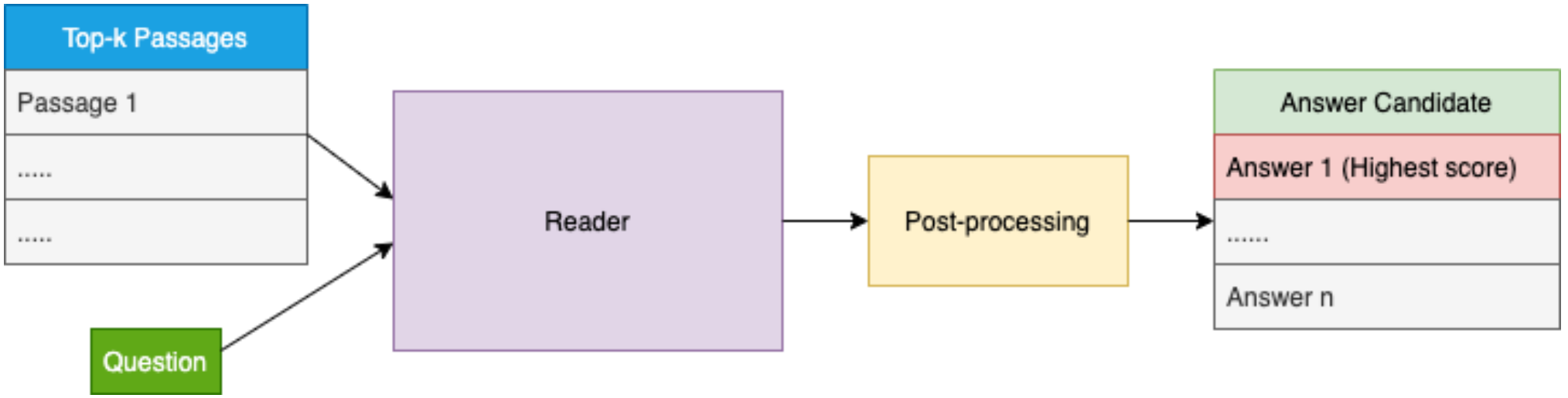
Retrieval - Elastic Search

Retrieval 최종 성능

topK	1	5	10	15
Accuracy	0.7347	0.8839	0.9160	0.9325

Reader

1. Data augmentation



MRC dataset size가 작아 Reader 모델이 biased 될 위험성이 존재

주어진 wiki documents를 이용하여 Data를 늘려보자!

Reader - Data augmentation

(1) Negative Sampling

How to mitigate training bias?

1. Train negative examples

훈련할 때 잘못된 예시를 보여줘야 retriever 이 negative 한 내용들은 먼 곳에 배치할 수 있음

⇒ Negative sample 도 완전히 다른 negative 와 비슷한 negative 에 대한 차이 고려 필요함

(다음 슬라이드에 계속)

2. Add no answer bias

입력 시퀀스의 길이가 N 일시, 시퀀스의 길이 외 1개의 토큰이 더 있다고 생각하기

⇒ 훈련 모델의 마지막 레이어 weight 에 훈련 가능한 bias 를 하나더 추가

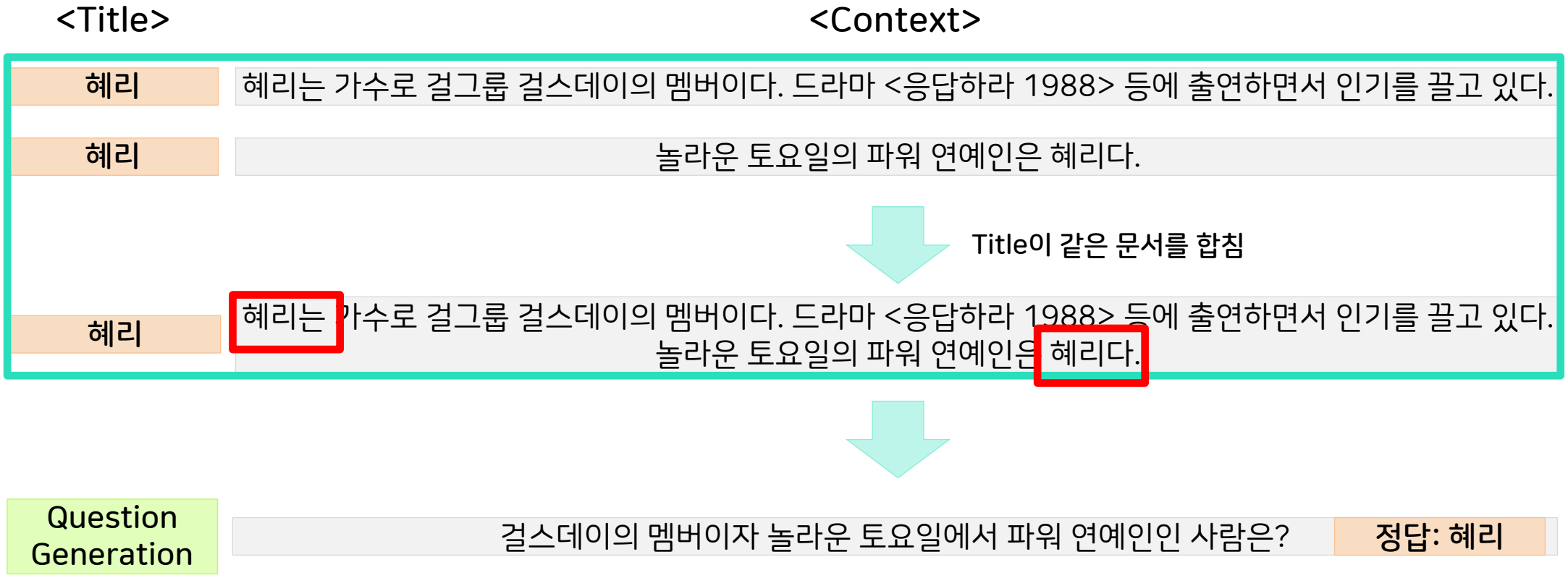
⇒ Softmax 로 answer prediction 을 최종적으로 수행할 때, start end 확률이 해당 bias 위치에 있는 경우가 가장 확률이 높으면 이는 "대답 할 수 없다" 라고 취급

출처 : MRC 8강 Reducing Training Bias

- retrieval를 이용해 선택한 **Negative passage**를 No answer로 라벨링 하여 사용
- Model의 Robustness를 향상시킬 수 있음.

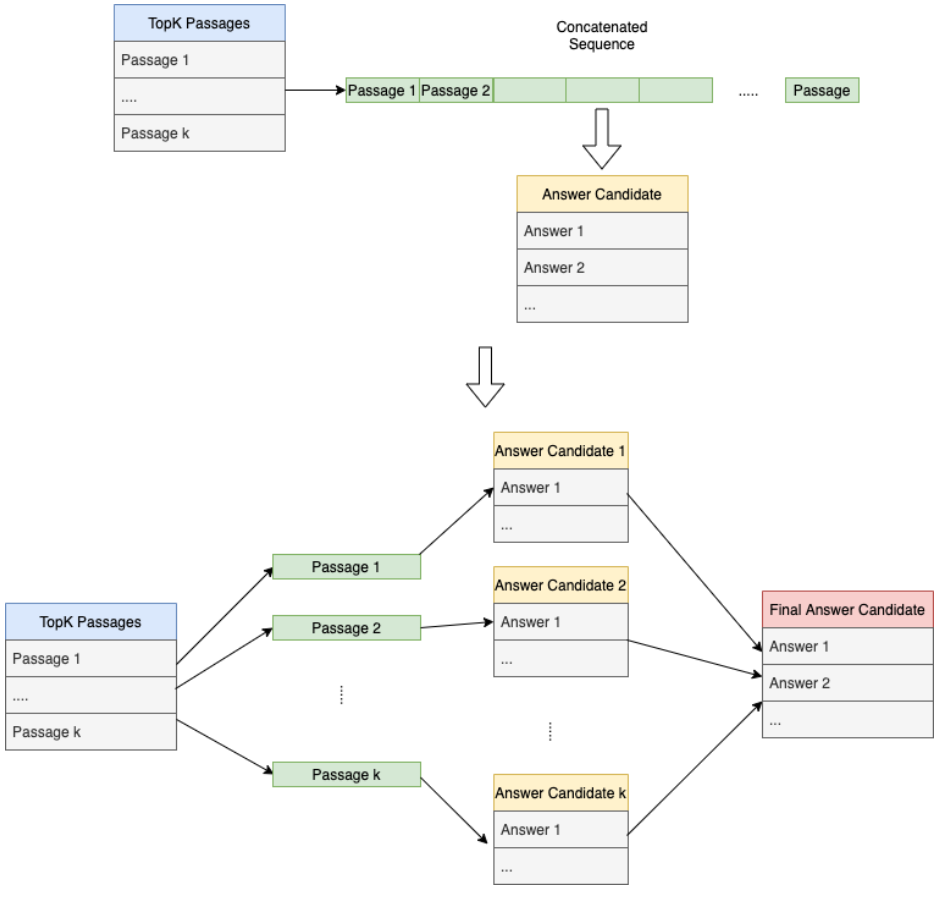
Reader - Data augmentation

(2) Question Generation



Reader

2. Post-Processing

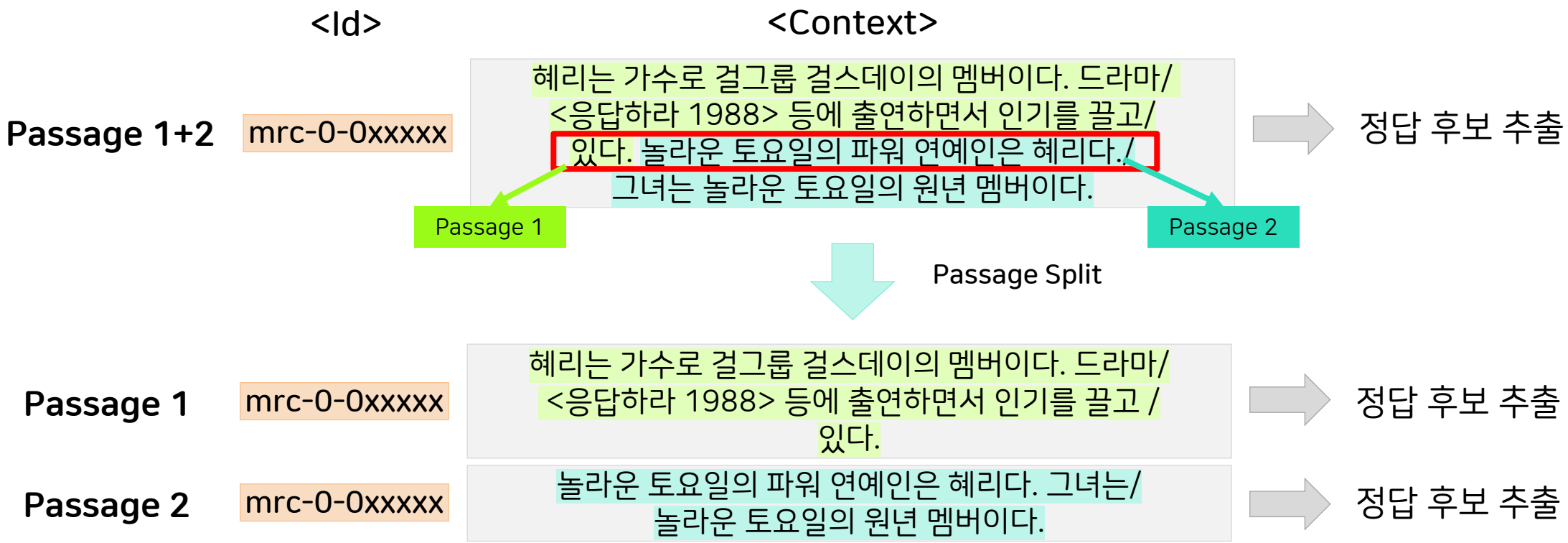


(1) 기존 방식

(2) Top-k passage split

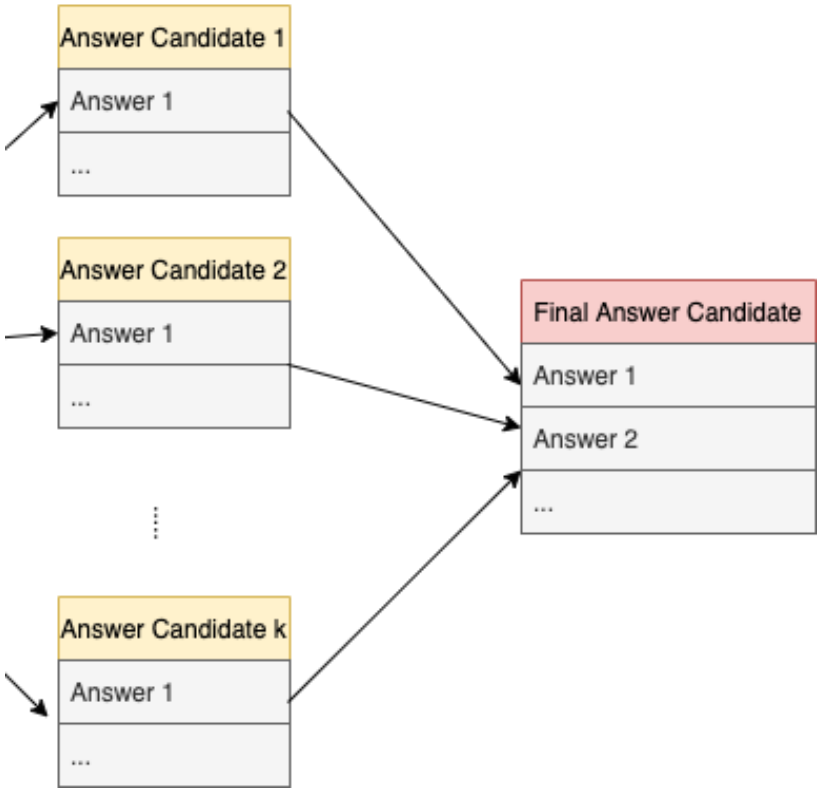
Reader

(1) Top-k Passages Split



Reader

(2) Answer Score 계산 방식 변경



<기존 방식>

- Answer 후보의 start_logit + end_logit

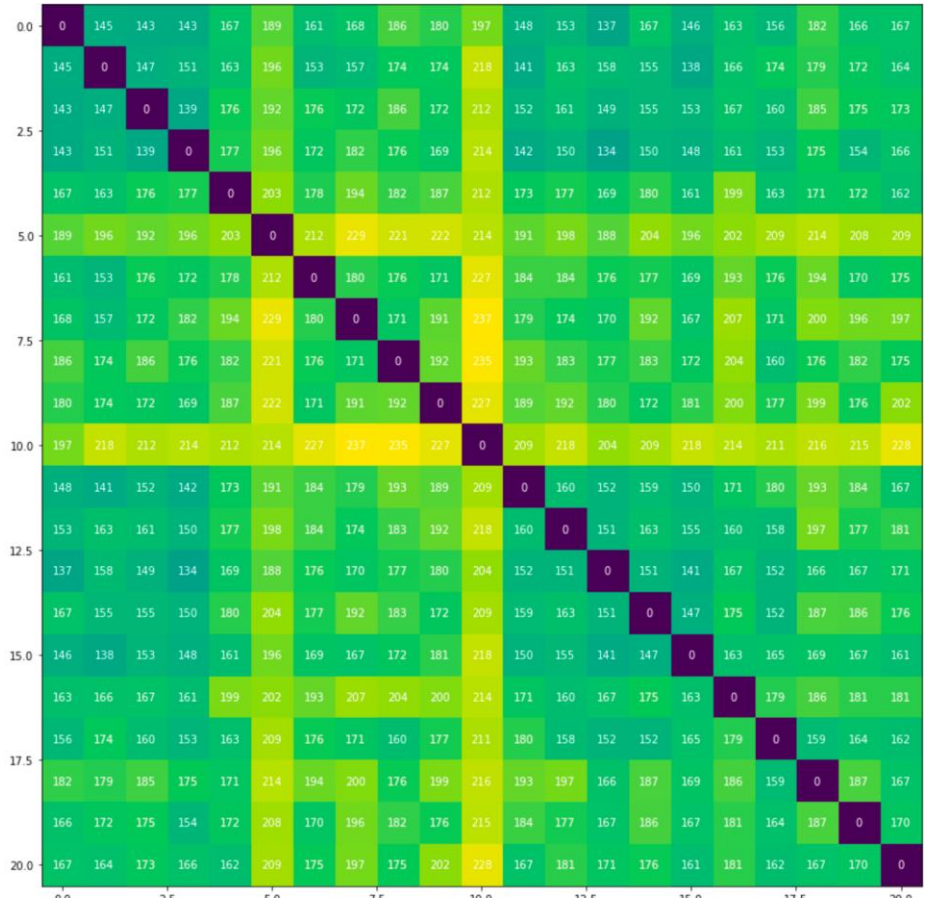
<문제점>

- 각 Passage마다 출력되는 Logit값의 분포가 다르므로 더해서 대소관계를 비교하는 것이 정확할까?

<해결 방법>

- 확률을 이용하여 정렬
  (Start Probability * End Probability)

Ensemble



<Soft voting>

- score가 들어있는 n_best_predictions를 이용하는 방법.
- 모든 제출의 n_best_predictions가 보존되어 있지 않아 더 많은 파일을 활용하는 hard vote가 유리하다고 판단

<Hard voting>

- 20개 가량의 predictions.json 파일을 앙상블
- 포함관계, 양 끝 조사 포함 여부 등을 고려하여 동률일 경우에 대응
- 이 때 예측 답안이 다른 파일들과 100개 이상 다른 것만을 선정

쿼터백 소개 및 Gather Town 😂

쿼터백 소개 및 Gather Town 😂

재형

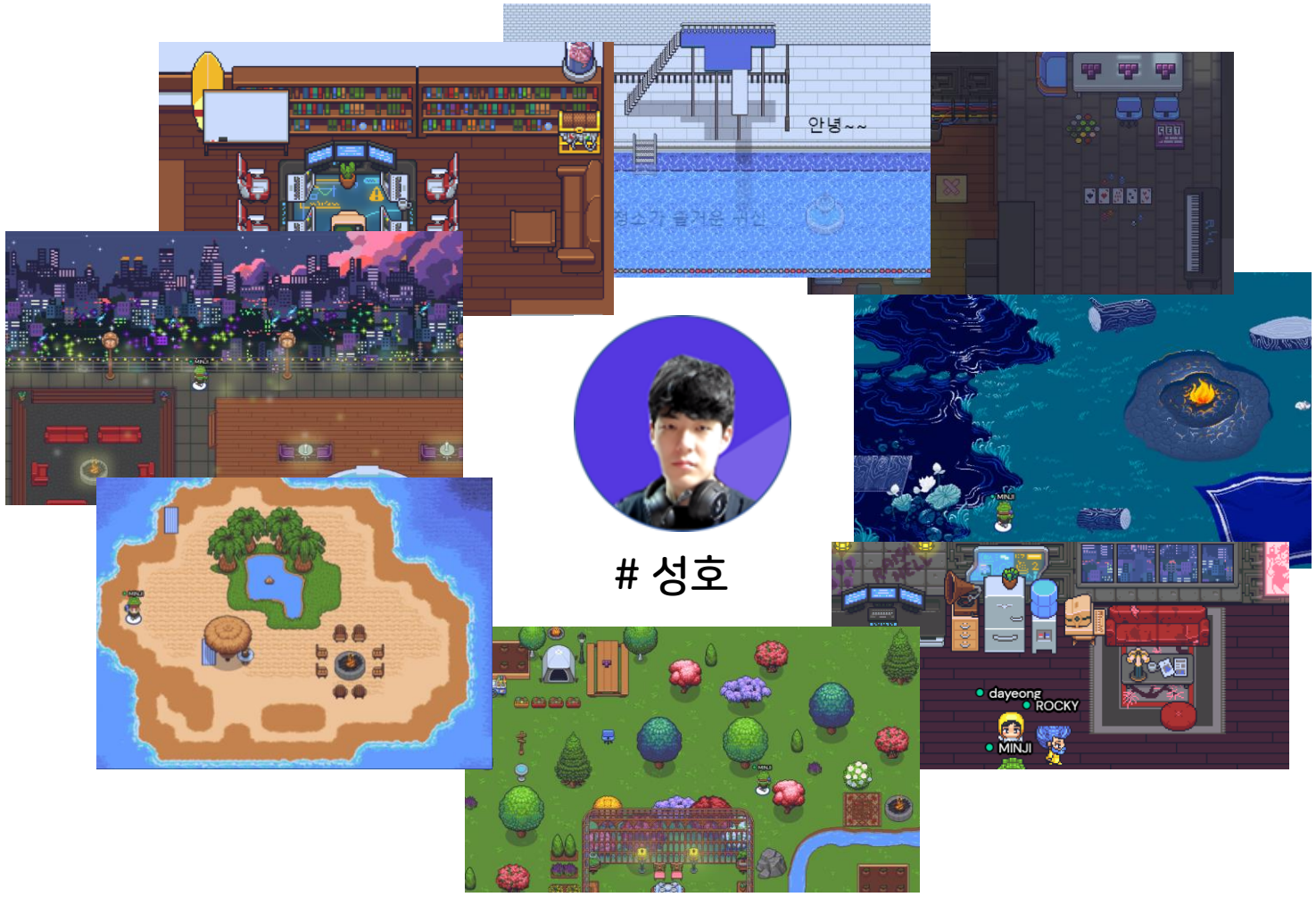
동건

다영

석민

민지

다인



성호

감사합니다

쿼터백 조

김다영 * 김다인 * 박성호 * 박재형 * 서동건 * 정민지 * 최석민