키워드를 활용한 기계 독해 모델

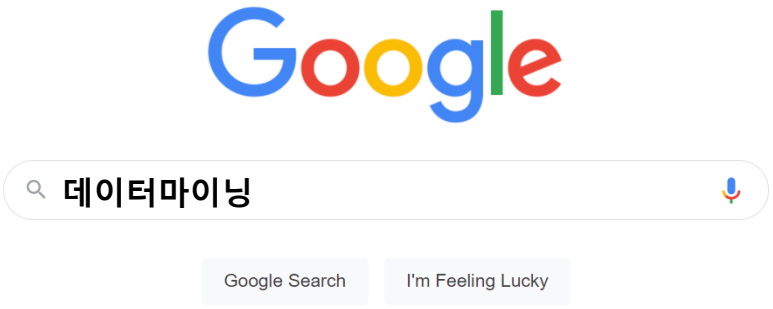
1007
이예진, 한미래

목차

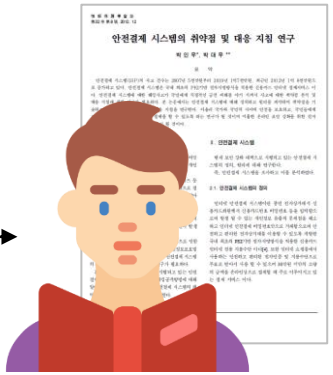
1. 문제 정의 및 해결 방법

- 문제 정의

  - 검색어로 전문적인 지식을 검색할 경우, 검색 결과가 광범위함
  - 논문과 같은 전문적인 문서에서 직접 문서를 읽고 이해하기 어려움
  - 사람이 긴 논문 텍스트에서 직접 원하는 정보를 찾기에는 시간이 오래 걸림



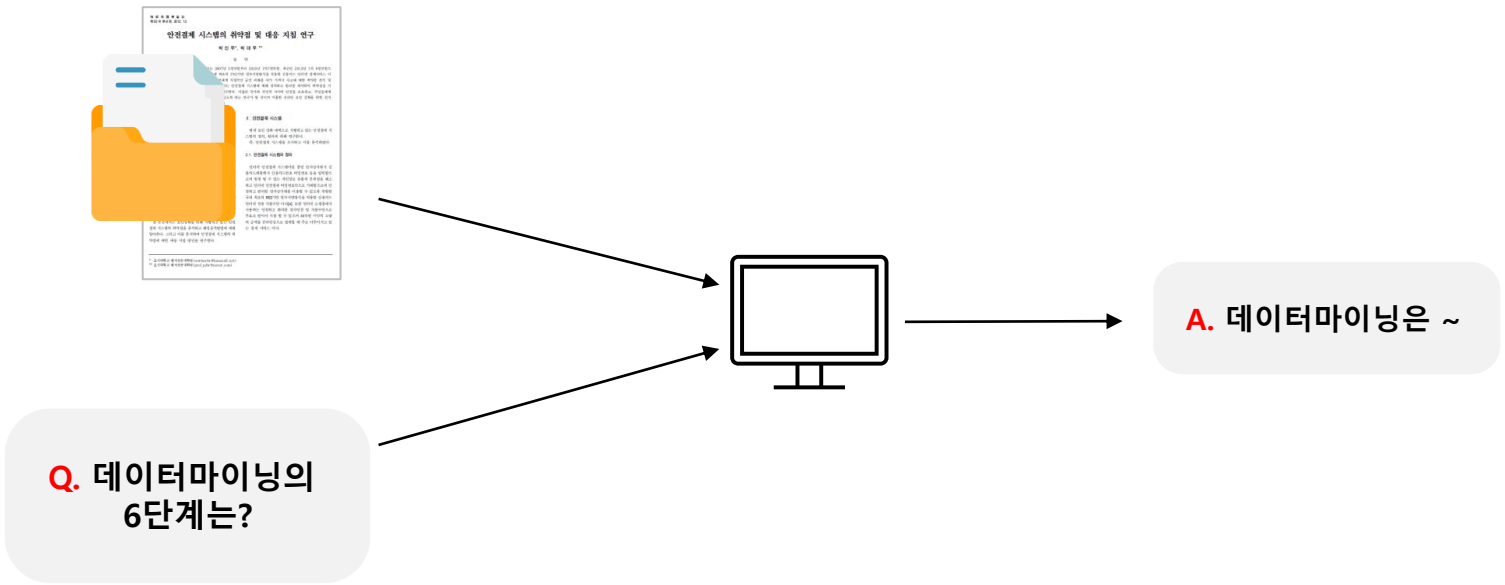
Q. 데이터마이닝의 6단계는?



A. ?

1. 문제 정의 및 해결 방법

- 기계독해 (Machine Reading Comprehension; MRC)
  - 기계가 주어진 문서를 이해하고 입력 받은 질문에 대한 답변을 추출하는 질의응답 작업

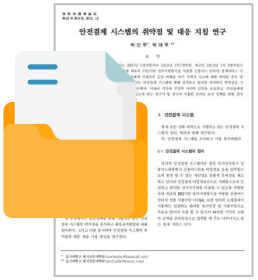


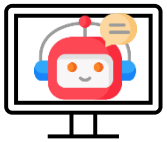
1. 문제 정의 및 해결 방법

- 해결 방법
  - 데이터의 66% (질문 난이도 중, 하)의 경우, 정답과 키워드가 동일한 문장 내에 존재
  - 키워드와 키워드를 포함한 단서 문장(Evidence sentence) 활용

    → 기존의 MRC 모델 보완

    → 검색어 매칭의 문제 해결





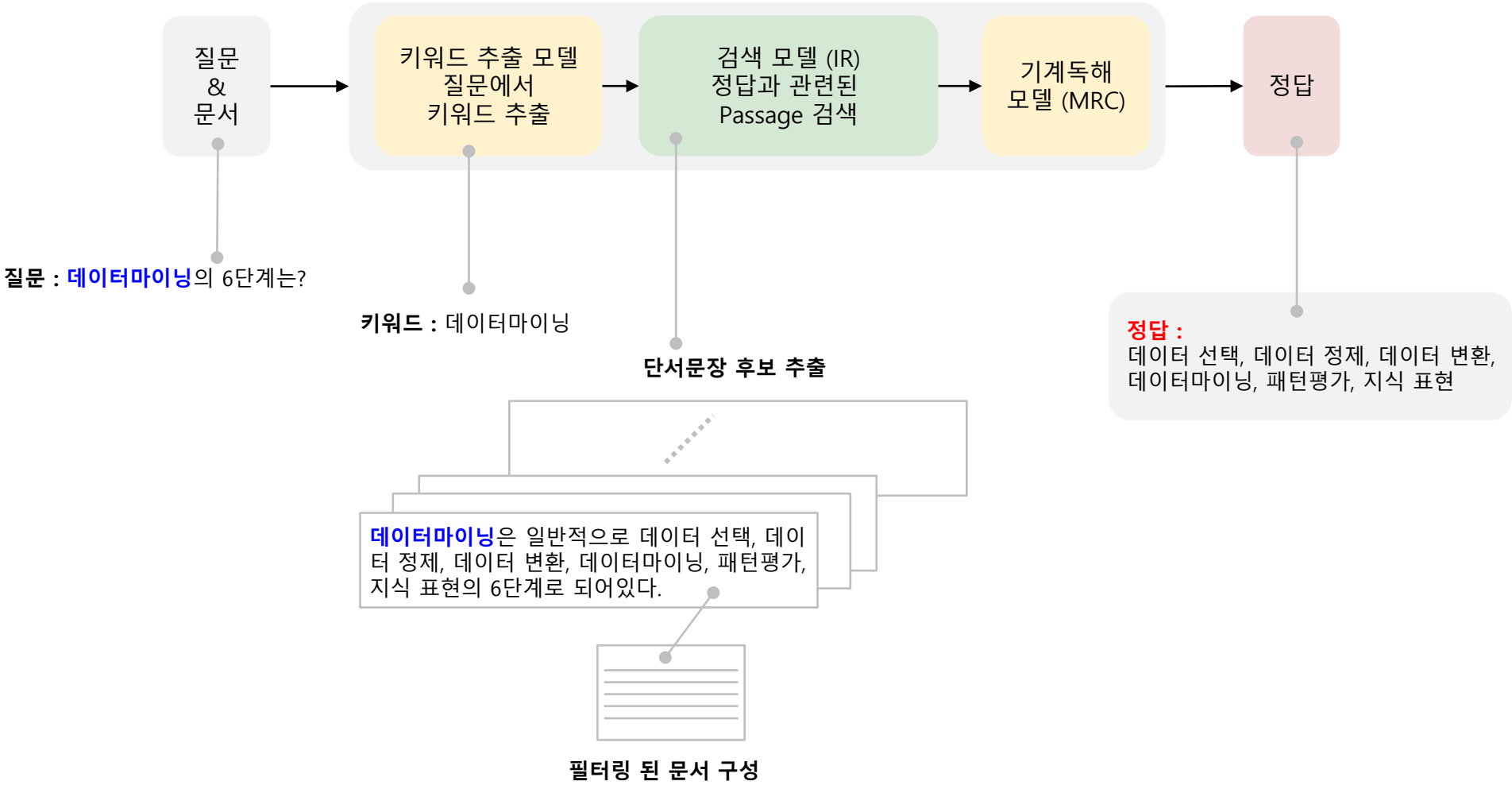
Q. 데이터마이닝의 6단계는?

A. 데이터 선택, 데이터 정제, 데이터 변환, 데이터마이닝, 패턴평가, 지식 표현

단서 문장 : **데이터마이닝**은 일반적으로 데이터 선택, 데이터 정제, 데이터 변환, 데이터마이닝, 패턴평가, 지식 표현의 6단계로 되어있다.

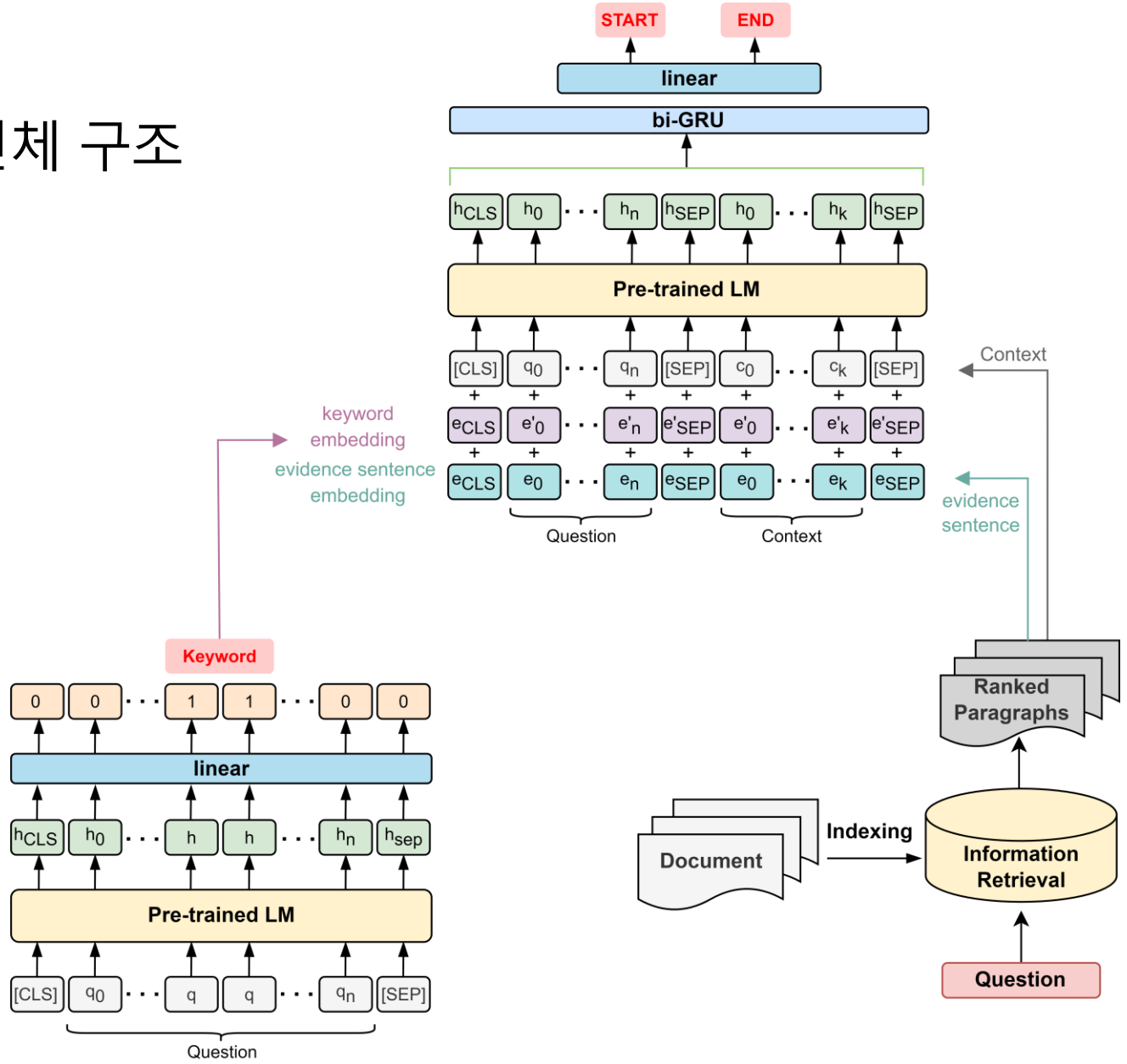
2. 제안 모델

- Pipeline

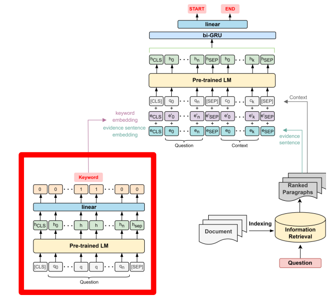


2. 제안 모델

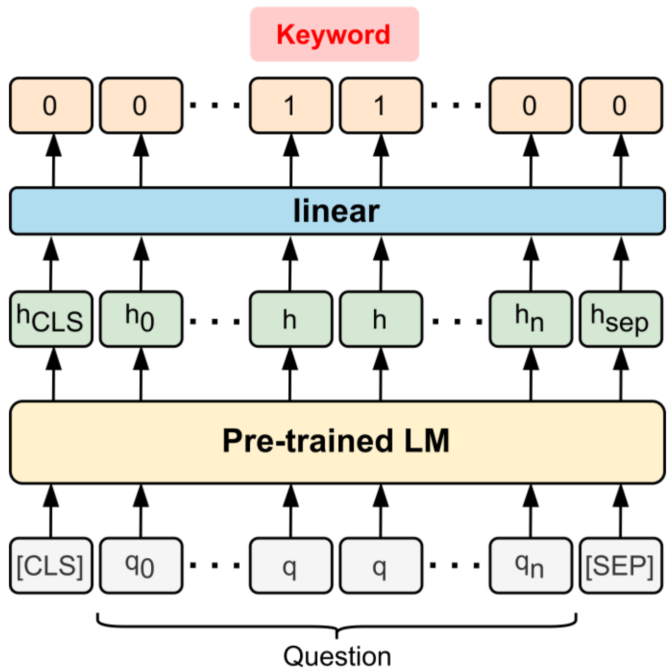
- 모델 전체 구조



2. 제안 모델



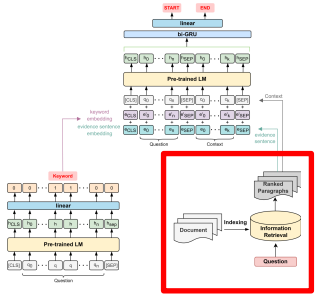
- 키워드 추출 모델 : 입력 시퀀스에 레이블 시퀀스를 부여 (Sequence Labeling)



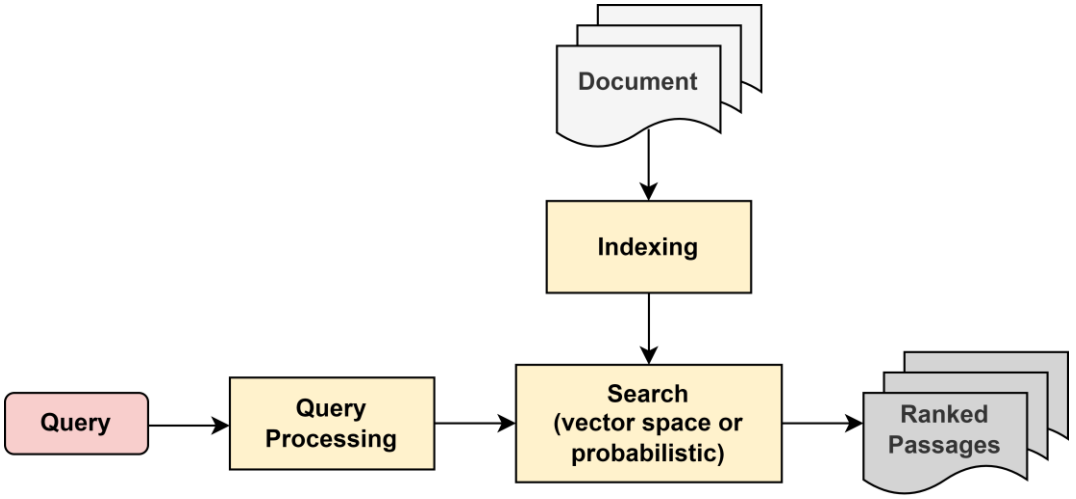
- 질문에서 키워드에 해당하는 토큰 추출

Query : 한국 정부에서 늘어나는 폐기물 발생량을 줄이기 위해 수립한 것은?
Label : 0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

2. 제안 모델



- 검색 모델 (Information Retrieval) : Lucene을 사용해서 데이터로부터 관련된 정보를 추출



- 데이터의 85%는 키워드가 질문에 대한 정답 앞에 존재
- 전체 논문을 3문장씩(Passage) 색인 (현재 문장 + 다음 2문장)
- 질문(Query)을 입력하여 질문에 대한 정답이 될 수 있는 상위 10개의 Passage 검색
  → 검색된 Passage를 필터링 된 문서(Filtered context)로 사용

2. 제안 모델

- 검색 모델 (Information Retrieval)에서의 인덱싱(Indexing) 방법

① 본 연구에서는 UC 의 서비스를 선정하는데 있어서 보다 사용자의 관점에서 접근함으로써 사용자에게 실제적인 편익을 줄 수 있는 방법을 제안하고자 한다.

② 일찍이 Jeff Moore 는 Crossing the Chasm 이라는 책을 통하여 많은 IT 기업이 좋은 기술과 아이템을 가지고도 90%에 이르는 실패율을 보이고 있는 것을 아래와 같이 설명하려 하였다.

③ 일반적으로 어떤 기술이 개발되면 초창기에는 혁신을 추구하거나 기술 매니아들 혹은 소위 Early Adopter 들이 구입을 하게되고 이들에 의해 편익성(convenience)가 검증되고 이의 결과에 따라 개방적인 대중들, 보수적인 대중들이 순차적으로 구입하게 된다는 것이다.

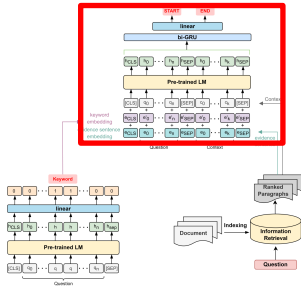
④ 그러나 많은 기술의 경우 여기서 말하는 기술과 성능(Performance)에 중점을 두어 실제로 대중이 원하는 해결안(Solution)/편익성(Convenience)을 간과하게 되는 경우가 많고 이것은 90%의 실패율을 보이는 것으로 설명하였다.

⑤ <그림 1> Moore 의 Chasm 곡선 이러한 결과는 u-서비스에서 더욱 확장이 될 수 밖에 없는데 그 이유는 u-서비스라는 것이 아직 존재해보지 않은 것이기 때문에 검증이 되어 있지 않고 더더구나 눈에 보이지 않으며 조용한 기술(Calm Technology)를 추구하고 있기 때문이 그 하나의 이유이다.

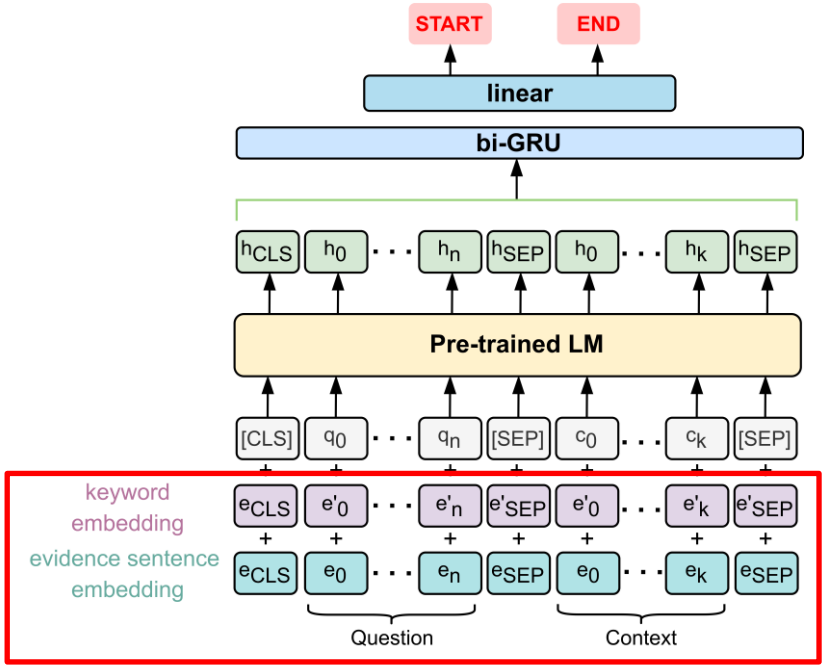
⑥ 이렇게 이루어진 서비스가 실제 사용환경에 적용되었을 때 사용자가 정말 편리함을 느끼고 자신이 원하던 문제가 해결되는가는 다르다.

Indexing : ①,②,③ / ②,③,④ / ③,④,⑤ / ④,⑤,⑥ …

2. 제안 모델



- 기계독해 (Machine Reading Comprehension)



기계독해 모델 고도화

- **키워드 임베딩(Keyword Embedding) :** 질문에 있는 핵심 키워드가 답변 추론에 잘 반영되도록 사용
- **단서 문장 임베딩(Evidence Sentence Embedding)** : 키워드를 포함하거나 단서 문장이 답변 추론에 반영되도록 사용
- 답변 길이가 긴 논문 데이터의 특성을 반영하여 답변 길이에 제한을 두지 않고 예측

3. 실험 및 평가

- 국내 논문 데이터 질의응답 셋
  - 논문 : 279,143개
  - 논문 QA쌍 : 831,182개

```
"doc_id": "논문ID",
"title": "제목",
"authors": "저자",
"journal": { "ko": "국문 학술지/학술대회 제목", "en": "영문 학술지/학술대회 제목" },
"year": "발행연도",
"context": "질의응답문장이 포함된 논문 풀텍스트",
"keywords": { "ko": "국문 키워드", "en": "영문 키워드" },
"qas": [
  {
    "level": "난이도 (1:하, 2:중, 3:상)",
    "id": "질의응답 셋 ID",
    "question": "질의",
    "answer": {
      "answer_text": "응답에 해당하는 텍스트",
      "answer_start": "응답 시작 인덱스"
    },
    "keyword": {
      "keyword_text": "핵심 어휘",
      "keyword_start": "키워드 시작 인덱스"
```

3. 실험 및 평가

- 실험 데이터
  - 학습 데이터셋
    - 전체 논문 데이터의 5% 사용
    - 논문 약 14,000 개
    - 논문 QA 쌍 약 41,500 개
  - 검증 데이터셋
    - 논문 약 3,350 개
    - 논문 QA쌍 약 10,000 개

3. 실험 및 평가

- 성능 평가

  – 평가 지표 : EM, F1 사용

  - **Exact Match (EM)**
    정답 텍스트의 어절과 예측 텍스트 어절 간의 단순 비교
    정답 1, 오답 0으로 계산

  - **F1 Score**
    정답 텍스트와 예측 텍스트 어절 간의 정밀도(precision)와 재현율(recall)을 구해서 F1 점수 계산

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

3. 실험 및 평가

- 성능 평가

키워드 추출 모델

Model	F1	Recall	Precision
RoBERTa-base	83.1	86.97	79.57
RoBERTa-large	82.33	87.46	77.77

IR 모델

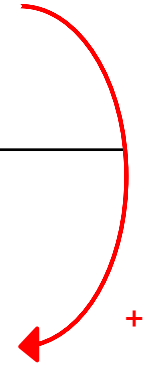
Rank	Recall
Top 1	59.67
Top 2	68.64
Top 3	72.74
Top 4	76.12
Top 5	77.85
Top 6	79.49
Top 7	80.65
Top 8	81.66
Top 9	82.19
Top 10	82.87

3. 실험 및 평가

- 성능 평가

MRC 모델

Model	EM	F1
RoBERTa-base (Our Implements)	20.91	46.25
RoBERTa-base w/o 답변길이 제한	17.51	73.67
RoBERTa-base + keyword 임베딩	16.92	74.99
Proposed model RoBERTa-base + keyword 임베딩 + evidence 임베딩	18.81	77.40
Proposed model RoBERTa-base + keyword 임베딩 + evidence 임베딩 **+ filtered context**	17.21	72.24



+ 31.15%p

+ 3.73%p

3. 실험 및 평가

- 성능 평가

MRC 모델

	시간
Proposed model RoBERTa-base + keyword 임베딩 + evidence 임베딩	1330초
Proposed model RoBERTa-base + keyword 임베딩 + evidence 임베딩 **+ filtered context**	101초

$\frac{7}{100}$ 로 시간 단축

- 검증 데이터 논문 약 600개, QA쌍 1800개에 대해서 추론 시 걸리는 시간
- 검색 모델로 필터링 된 context 사용시 기존 MRC 모델들 보다 7/100 의 추론 시간 단축

3. 실험 및 평가

- 성능 평가

MRC 모델

예시 1	**질문**	**감정의 색인과 검색 과정은 어떤 특징을 가지고 있는가?**
	정답	색인가와 이용자 사이의 주관적인 판단과 함께 이를 표현하는 용어 사용의 차이로 인해 검색 결과의 불일치로 이어지는 경향이 있다
	예측	감정의 색인과 검색은 색인가와 이용자 사이의 주관적인 판단과 함께 이를 표현하는 용어 사용의 차이로 인해 검색 결과의 불일치로 이어지는 경향이 있다.
예시 2	**질문**	**PC 재질의 튜브램프의 단점은?**
	정답	저온(?35 ℃ 이하)에서는 사용 환경 조건에 따라 파손되는 단점
	예측	저온(?35 ℃ 이하)에서는 사용 환경 조건에 따라 파손되는 단점이 있어 냉동 창고와 같은 저온용으로는 사용하기가 적합하지 않다.

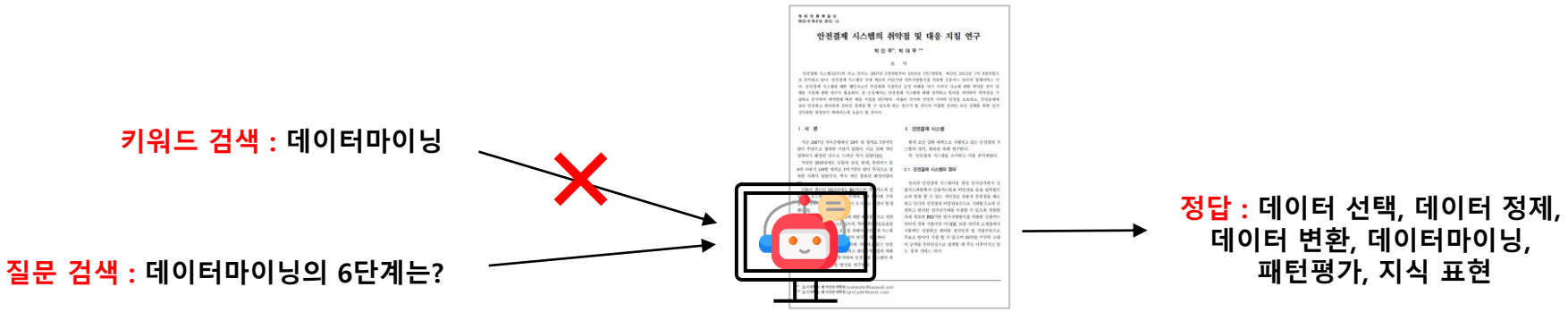
- 정답과 예측 답변의 길이가 길기 때문에 정량 평가 점수가 낮지만 정성 평가 시 예측 답변이 정답과 같은 문장임을 확인

4. 결론

- 키워드 검색이 아닌 질문(Query) 검색 가능
  - 질문(Query)으로 전문적인 지식에 대한 구체적인 답변 획득

- 검색 시간 단축
  - 사람이 직접 긴 텍스트를 읽고 이해하지 않아도 원하는 정보 추출

- 키워드가 존재하지 않는 경우에도 검색 가능
  - 키워드가 포함된 데이터가 구축되어 있지 않아도 질문에서 키워드를 추출하여 검색



5. 향후 연구

- 키워드 모델 및 검색모델의 성능 향상
- 전처리 및 후처리
  - 현재는 어떠한 전처리 및 후처리도 하지 않음
- 단서문장이 여러 개인 경우 고려
  - Multi-hop QA 적용