BERT를 활용한 아동 기계독해 체험용

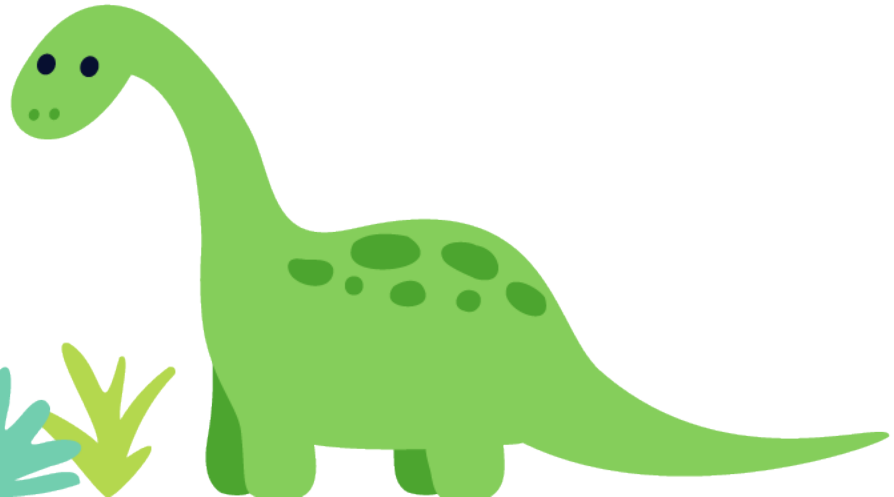
질의응답 모델/시스템

Contents

기계학습 체험을 위한 공룡관련

질의응답 모델/ 시스템

1. 제안동기 및 모델 선정
1. 실현 서비스 및 개요
1. 데이터 수집 및 학습
1. 최종결과 및 비전



1 제안동기 및 모델선정

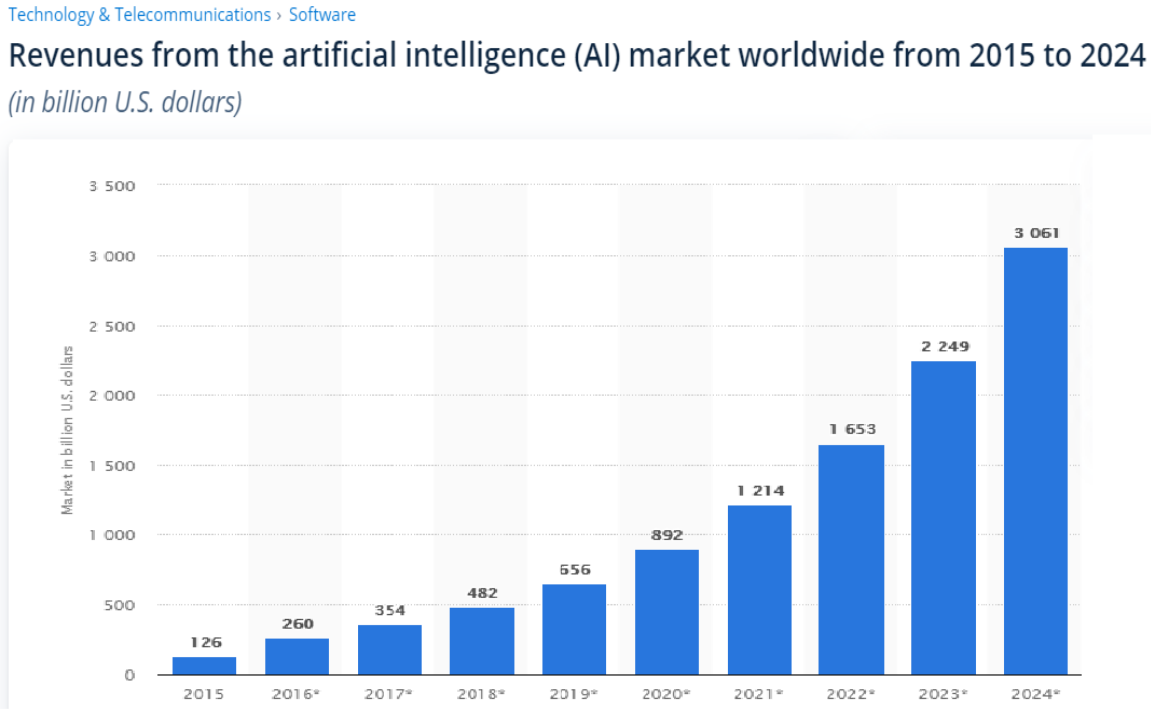
제안동기

AI의 성장

지속적인 인공지능 관련 업종들의 성장

- 빅데이터 등 AI와 접목시킨 분야들이 새롭게 대두되기 시작하면서 **시장 규모**도 점점 더 커짐.
- 4차 산업혁명 시대에 필요한 창의적 문제해결력을 갖춘 인재 양성을 위해 소프트웨어관련 교육이 늘어나고 있다.
- 초등학교 교육과정으로 '**코딩**' 교과목이 도입
- 하지만 그에 반해 우리나라 AI시장의 규모는 뒤처짐

  -> **AI에 친근해지는것이 중요하다고 생각**

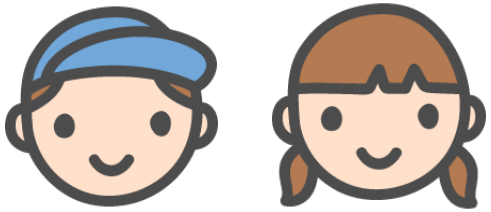


1 제안동기 및 모델선정

아이디어

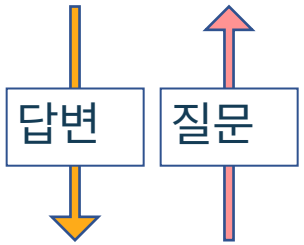
어린아이들에게 기계독해를 체험

어린 아이들을 대상



어린이가 흥미를 느낄만한 주제인

‘**공룡**’을 주제로 **기계독해**를 체험



기계독해

- 인공지능(AI) 알고리즘이 스스로 문제를 분석하고 질문에 최적화된 답안을 찾아내는 기술.
- **기계독해**를 이용하면 사람이 텍스트를 읽고 질문에 답변을 추론하듯이 인공지능(AI)이 문장 속에서 의미를 찾고 답변

핵심기술

자연어 처리(Natural Language Processing)

Ai타임스

MS-뉘앙스 인수합병 배경에는 자연어처리가있다

MS, 197억달러(약 22조 1200억원)에 뉘앙스 인수 발표
뉘앙스, 애플 시리 개발 참여한 AI 음성인식 개발업체
자연어처리 기술, 2025년 화이트칼라 내 50% 차지한다

•2021.03.17

구글, 특허 출원에 NLP 알고리즘 적용 제안

BERT 비롯한 AI · ML 모델 적용...특허 산업 혜택 볼 것
특허의 범주화와 기록물 관리 등에 도움 줄 것
미국 특허청, AI 모델 구축해 특허 출원 업무 처리

•2021.03.04

CIO FROM IDG

"자연어 처리 시장, 2026년 연평균 20.3% 성장"

김달훈 | CIO KR

전 세계 자연어 처리 시장 규모가 2020년 116억 달러에서 2026년이 되면 351억 달러로 성장할 전망이다. 2020년부터 2026년까지 자연어 처리 시장의 연평균 성장률은 20.3%에 이를 것으로 예측됐다. 자연어 처리 시장의 성장을 이끄는 주요 요인으로는 스마트 장치의 사용 증가, 다양한 업종에서 자연어 처리 기반 애플리케이션 채택 증가, 클라우드 기반 자연어 처리 솔루션 증가 등이 지목됐다.

•2021.03.04

1 제안동기 및 모델선정

NLP모델 비교

- 통계기반 -

TF-IDF, 단어-문맥 행렬

- 각 단어에 대한 중요도를 계산
- 문서의 핵심어를 추출하거나, 검색엔진 에서 검색 결과의 순위를 결정
- 문서들 사이의 비슷한 정도를 구하는 등의 용도

- 단어기반 -

Word2Vec, Fastest …

- 희소 표현(1,0 벡터화)
- 분산 표현 방법(비슷한 위치에서 등장하는 단어들은 비슷한 의미)
- 각 단어에 대한 중요도를 계산
- 사용자가 지정한 주변 단어의 개수에 대해서만 학습이 이루어지기 때문에 데이터 전체에 대한 정보를 담기 어려움

- 문장기반 -

ELMo(LSTM), GPT, BERT

ELMo

- 전이학습(Fine Tunning)
- LSTM은 재귀적이기 때문에 상당히 느림
- 양방향 모델로 구현하기 위해선 순방향과 역방향 두가지 Layer 모델이 필요

GPT

- 병렬계산이 가능한 Transformer사용
- 단방향성, 언어를 이해보다는 언어를 생성하는데 특화

1 제안동기 및 모델선정

사용 모델

문장형 BERT 사용

- 2018년 11월 구글이 공개한 인공지능(AI) 언어 모델로써 모든 자연어 처리 분야에서 좋은 성능을 내는 범용 Language Model

- **양방향성**을 지원, 문맥을 고려한 단어의 뜻을 파악
  - 자연어 처리(NLP)에서 **자연어 이해(NLU)부분에 특화**
    - 자연어 이해 ex) **QA**, intent classify

- **전이학습**(Fine tuning)을 지원하는 자연어 처리모델 중 성능이 가장 우수

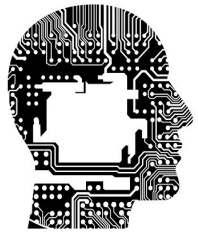
- **Transformer Encoder**(Self-attention)을 사용해 토큰 전체를 병렬적으로 계산함으로써 **속도가 매우 빠름**

#	Model	SST-2 Acc
1	$\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2018)	94.9
2	$\text{BERT}_{\text{BASE}}$ (Devlin et al., 2018)	93.5
3	OpenAI GPT (Radford et al., 2018)	91.3
4	BERT ELMo baseline (Devlin et al., 2018)	90.4
5	GLUE ELMo baseline (Wang et al., 2018)	90.4

핵심아이디어

AI의 꽃 '기계독해'
Bert 기반 질의응답 모델/ 시스템

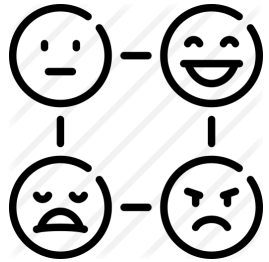
데이터 수집 및 전처리



Pre-training

Fine-Tunning

QA 웹페이지 구축



감성분석을 통한
반복학습 적용

2 실현 서비스 개요

FLOWCHART

데이터 수집 및 전처리

<Context 데이터>

- dino_1_갈리미무스.txt
- dino_2_게르마노닥틸루스.txt
- dino_3_고르고사우루스.txt
- dino_4_그나토사우루스.txt
- dino_5_길모레오사우루스.txt

Fine Tuning 을 위한 데이터

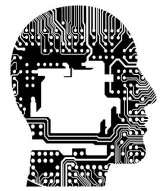
훈련

Pre-Training 을 위한 데이터

훈련

QA 모델 생성

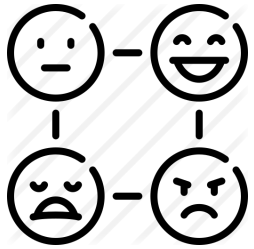
Fine-Tuning



Pre-training

출력

서비스의 이용



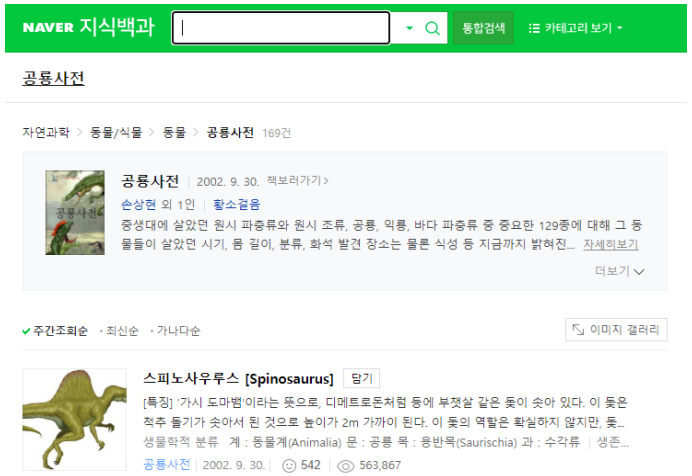
사용자의 리뷰분석

Feedback

3 데이터 수집 및 학습

데이터 수집 → 전처리 → Pre-training → Fine-tuning → 서비스 구현

Context Data



- 밝혀진 공룡의 종류: 169여종
- 네이버 지식백과 + 위키피디아 기준으로 Context 마련

Pre-trainning

- 제공 - ETRI
- 배포 모델 : KorBERT
- 세부 모델 : Korean_BERT_Morphology
- 모델 파라미터 : 30349 vocabs, 12 layer, 768 hidden, 12 heads,
- 학습데이터: **23GB 원시 말뭉치** (47억개 형태소)
- 딥러닝 라이브러리: tensorflow
- 소스코드: tokenizer 및 기계독해(MRC), 문서분류

Fine-Tuning

- KorQuAD 1.0에서 지원하는 한국어 QA_train 파일 (66,181쌍의 질의응답) +
- 주제와 관련된 500쌍의 Data자료
- (Id + Context + Question + Answer + Startpoint 이 한 쌍으로 포함된 json형식의 파일)
- KDinoQuAD.json

3 데이터 수집 및 학습

데이터 수집 → 전처리 → Pre-training → Fine-tuning → 서비스 구현

데이터 수집

- 공룡이름에 따른 context 수집 -> **txt저장**
- Pretraining 데이터 (label이 되지 않은 많은 양의 데이터가 필요) -> **말뭉치 데이터** 수집
- Fine Tuning 데이터 -> 질문에 대한 답을 담은 Context 데이터를 수집 + **task에 맞는 context**를 추가
- -> **Context**와 context에대한 **질문, 질문에 대한 답**, **답의 시작지점** 쌍으로하여 json형식으로 저장

전처리

- Pre-training을 위해 수집한 데이터를 전처리
- **Vocab** 파일을 생성

  -> **tf-record** 형식의 데이터 생성

  -> 형식을 통일화

  -> 학습의 **효과를 극대화** 하기 위함

  (모델의 성능에 영향을 주기때문)

3 데이터 수집 및 학습

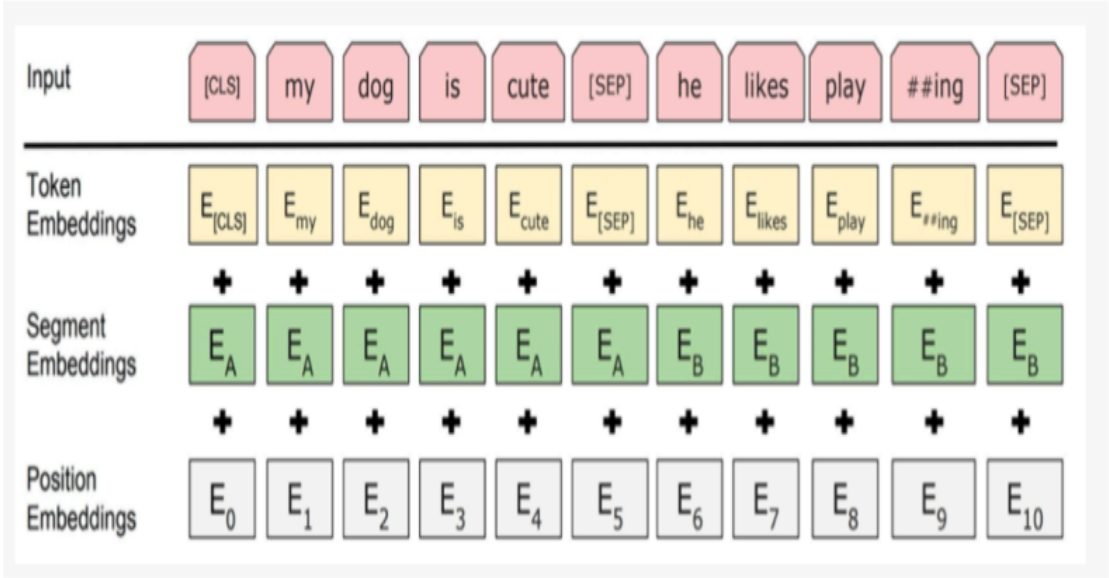
데이터 수집 → 전처리 → **Pre-training** → Fine-tuning → 서비스 구현

Pre-training의 목적

- **문장 내 단어의 의미를 문맥을 고려하여 파악하기 위함**

Pre-training 훈련 방법

- 마스크드 언어 모델(Masked Language Model) – **단어의 의미**
- 다음 문장 예측(Next sentence prediction, NSP) - **문맥**
- -> **Vocab 파일** 생성
- Vocab을 참고하여 **Tfrecord생성**(토큰화)
- **하이퍼 파라미터를 적용**, 훈련 후 최종 모델생성 pretrained.**ckpt(tensorflow)**



- [CLS] : 문장의 의미가 함축된 토큰
- [SEP] : 문장 A와 문장 B를 구분해주기 위한 토큰

3 데이터 수집 및 학습

데이터 수집 → 전처리 → **Pre-training** → Fine-tuning → 서비스 구현

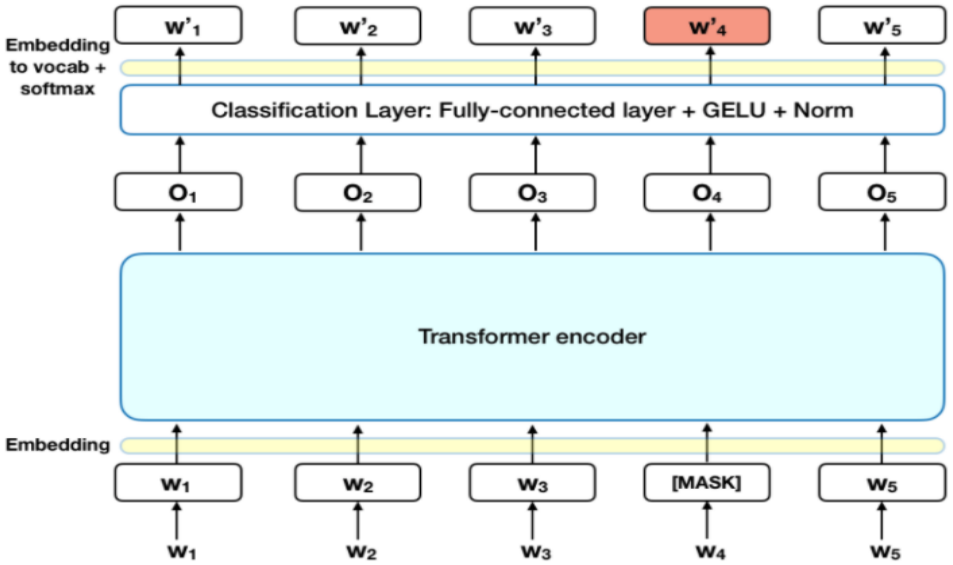
[Task 1] Masked Language Model

- 단어 중의 일부를 **[MASK] token** 으로 치환 (15%)
- 단어 중의 일부는 랜덤한 단어로 치환

-> [MASK] token 만을 예측하는 task

[Task 2] Next Sentence Prediction

- **랜덤으로 두 문장을 이어 붙여** 이것이 원래의 문장에서 바로 이어 붙여져 있던 문장인지를 맞추는 task
- 50% : sentence A, B가 실제 next sentence
- 50% : sentence A, B가 corpus에서 random으로 뽑힌(관계가 없는) 두 문장



Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP] LABEL = IsNext

Input = [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP] Label = NotNext

- 이어지는 문장의 경우

Sentence A = Sentence B

-> Label = IsNextSentence

- 이어지는 문장이 아닌 경우

Sentence A != Sentence B

-> Label = NotNextSentence

3 데이터 수집 및 학습

데이터 수집 → 전처리 → Pre-training → **Fine-tuning** → 서비스 구현

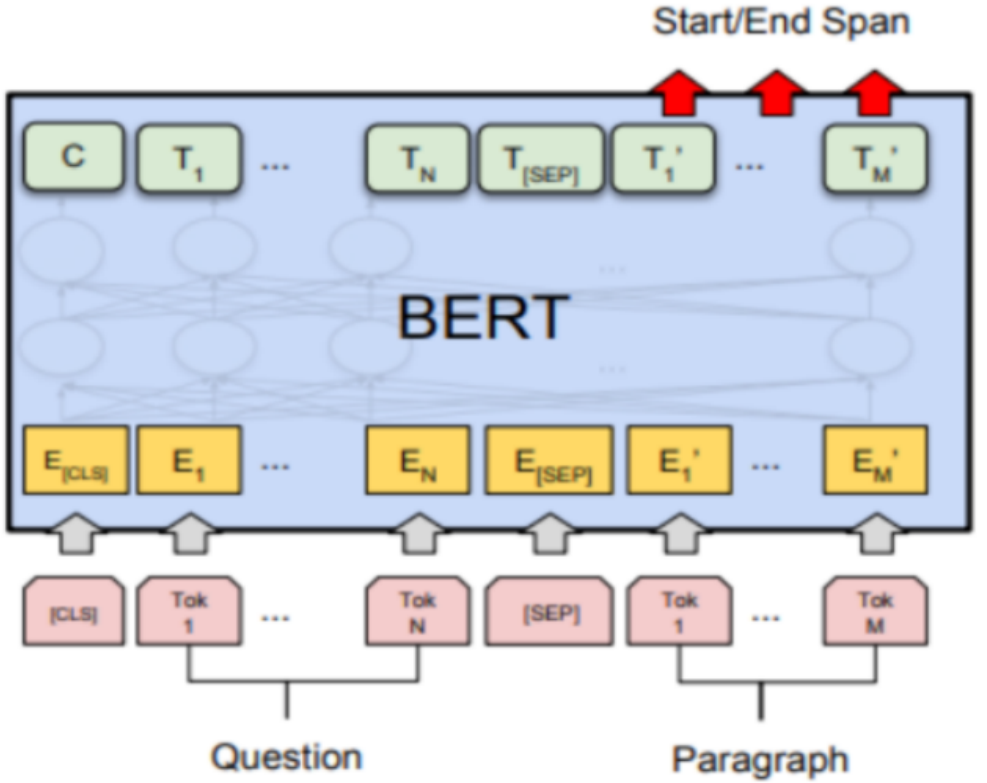
Fine-Tuning의 목적

- 최종적으로 **task의 목적에 최적화된 모델**을 만들기 위함

QA시스템 Fine-tuning 과정

- Pretrain이 완료된 ckpt파일을 **전이**받아 QA모델에 최적화된 모델을 만들기 위한 추가적인 훈련.
- **본문(Context)**과 각 본문 안에서의 **예상질문**, 그에 대한 **답**, 그리고 답의 **시작위치**를 한 쌍으로 훈련
- 질문과 지문이 주어지고, 그 정답의 위치를 맞추게 하는 훈련

{'paragraphs': [{'context': '에오랍토르는 원시적인 육식 공룡이다. '새벽의 약탈자'
'qas': [{'answers': [{'answer_start': 22, 'text': '새벽의 약탈자'}],
'id': 'dino_441_1',



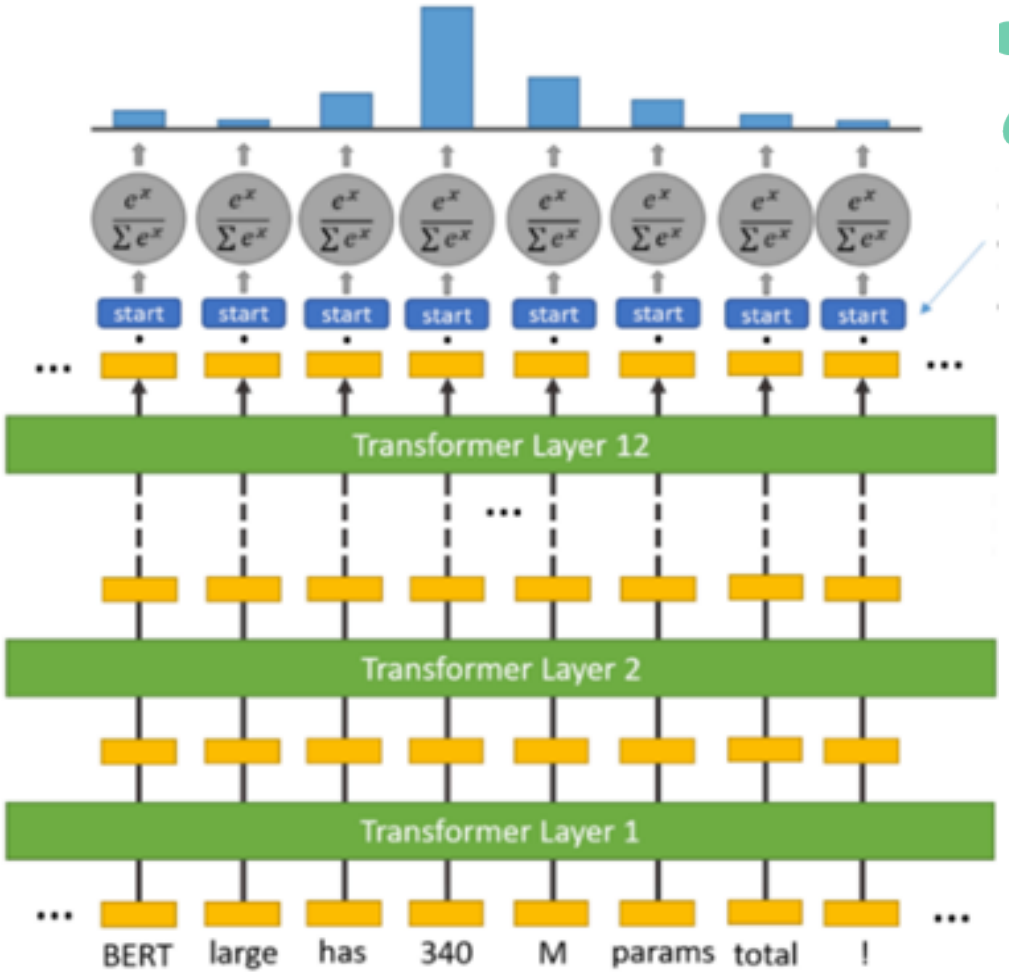
EM[Exact_Match]: 55.17 **"f1": 81.51**

3 데이터 수집 및 학습

데이터 수집 → 전처리 → Pre-training → **Fine-tuning** → 서비스 구현

Substring[정답 텍스트]를 찾는 과정

- [SEP] 토큰 이후부터 다음 [SEP]토큰 까지 각 단어 토큰들에게 질문에 맞는 **768** dimension의 **start vector**와 **end vector**를 **내적** 후 그 값을 산출
- Question과 가장 관련있는 **답변의 시작점과 끝점**을 찾기 위함
- 내적된 각 토큰에 **Softmax함수**를 적용
- Start벡터부분에서 Softmax적용 후 최대값을 가진 토큰을 **Start지점**, End벡터와 내적하여 최대값을 가진 토큰을 **End지점**으로 지정
- 본문과 Start, End지점을 **Mapping**하여 **Answer_text**를 산출



3 데이터 수집 및 학습

데이터 수집 → 전처리 → Pre-training → Fine-tuning → 서비스 구현

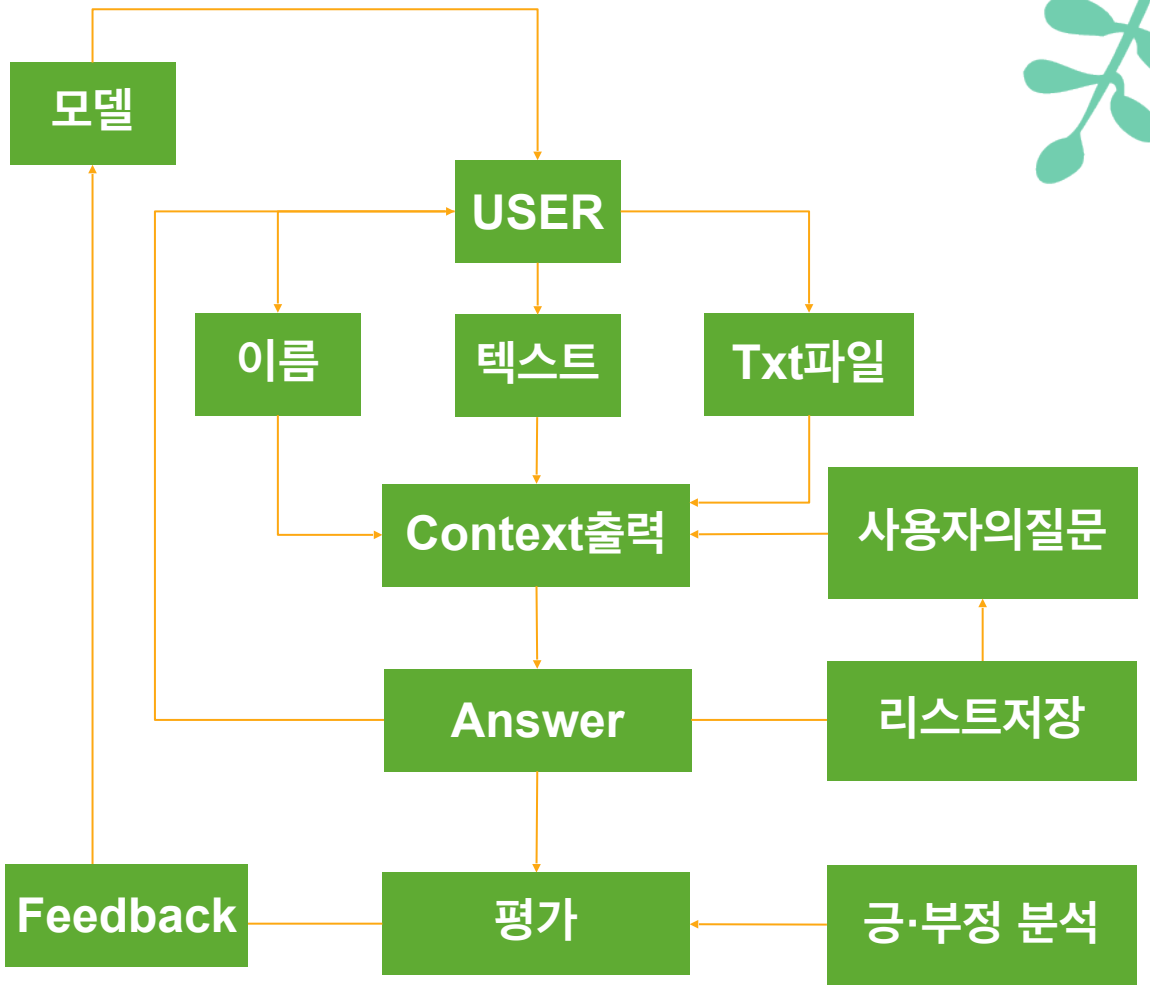
1. 공룡의 이름을 입력 or Text를 입력 or 공룡에 관한 context 가 담긴 txt를 불러와 **context를 출력**

1. 사용자가 관련 context에 관하여 **질문.**

1. Output으로 **질문에 대한 답**과 context에서 답부분의 위치를 **highlight**함과 동시에 히스토리를 리스트로 저장

1. 서비스 **종료시** 사용자에게 해당 프로그램에 관한 **리뷰**를 작성하도록 요구

1. 작성된 리뷰를 긍정, 부정분석(감정분석)실시 후 **Feedback**



3 데이터 수집 및 학습

데이터 수집 → 전처리 → Pre-training → Fine-tuning → 서비스 구현

1. 긍정, 부정 판별을 위해 사용된 모델

긍정, 부정 **레이블이 포함된** 네이버 영화 리뷰 20만개

LSTM을 활용하여 **긍정과 부정을 판별**할 수 있도록 학습

1. 모델의 활용

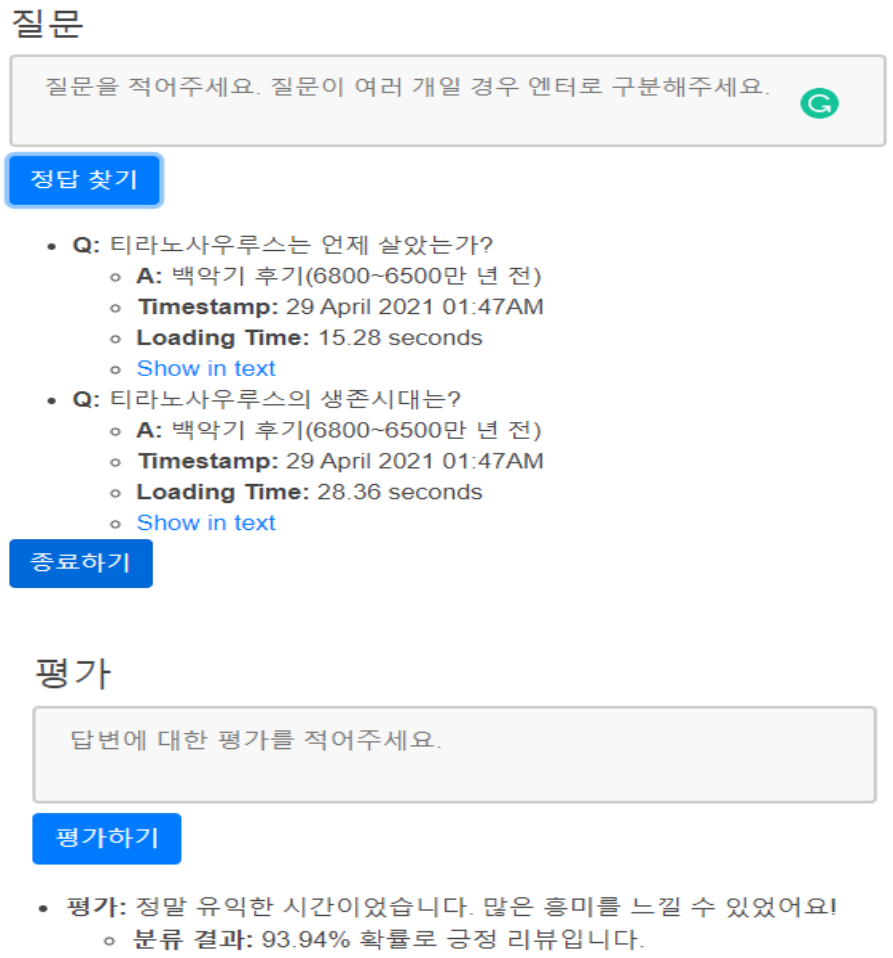
사용자의 리뷰에 LSTM을 활용한 모델을 이용 및 적용

리뷰에 대한 긍정과 부정을 판별 - 해당사용자에게 제공한 서비스(Answer)와 질문, context를 **한 쌍으로 판별된 리스트에 저장**

3. 각 리스트의 활용

긍정 리스트에 저장된 데이터는 **재학습**(Fintuning)

부정 리스트에 저장된 데이터는 **재검토** 및 분석 후 재학습



4 최종결과 및 비전

VISION

어린아이 뿐만아니라 AI에 관심있는 사람 모두에게 AI를 간접적으로 체험

TARGET

기계독해 기능을 통해 사용자의 요청에 문맥적요소를 고려, 정확성을 바탕으로 정확한 답을 제시

문맥적 파악

다양한 분야

기계 독해

개선방안

버전비교

KorQuAD 1.0		KorQuAD 2.0
	Html tag가 포함된 텍스트입력?	
	문장형식의 답변?	
	문단과 같이 긴 답변 가능?	

4 최종결과 및 비전

개선방안

KorQuAD 2.0을 통해 개선될 수 있는 방안

- **html tag 텍스트포함이 가능** – 훨씬 길이가 긴 context를 활용가능

  -> 사용자의 질문이 다양해질 수 있음, Crawling을 적극적으로 활용가능

  -> context를 미리 준비하지 않아도 됨

  -> 주제가 공룡에 한정된 것이 아닌 다양해질 수 있음.

- **질문에 대한 답의 제한이 완화** – 표, 문장, 문단 형식으로 답변이 가능

  -> 챗봇 서비스 등의 서비스와 결합가능

  -> 비즈니스적 활용방안 등으로 폭넓게 사용이 가능해짐(**실제 서비스화**가 가능해짐)

  -> 기계독해에 특화된 질의응답 모델이 필요한 도메인에 활용가능

감사합니다.