영어 객관식 문제 풀이 서비스 개발

부산대학교 정보컴퓨터공학부

지도교수 탁성우

팀명 Re:Fresh

팀원 202055509 강형원

202055538 박규태

202055605 주연학

목차

1. 서론

- 문제 상황

- 최근 AI의 발달과 함께 AI를 사용한 서비스들이 대중적으로 퍼지기 시작하면서, 일상생활의 많은 영역에서 AI를 활용한 서비스들이 늘어나고 있는 추세이다.

- 이런 상황에서, AI 서비스를 이용하여 문제를 풀이하고 이를 학습에 사용하는 사람들의 수 역시 늘어나고 있다.

- 이러한 취지에서, 졸업과제 주제로 “AI 기술을 활용한 영어 객관식 문제 풀이”가 가능한 서비스를 개발하는 것을 주제로 삼게 되었다.

- 과제 목표

1. AI 기술을 활용하여 영어 객관식 문제를 분석해 답을 구하는 모델을 개발한다.
   개발하는 모델은 특정 분야(영어 문제 풀이)에서 기존 대형 LLM와 비교해 근접한 성능을 내나 적은 자원을 소비하되 근접한 성능을 가지는 것을 목표로한다.

2. 개발한 모델을 바탕으로 웹기반 서비스를 개발한다.

- 요구조건 분석

● 중고등학생 수준의 영어 객관식 문제 풀이 서비스 개발

➔ 문제 유형은 여러가지가 있음

■ 빈칸에 오는 알맞은 단어 맞추기

■ 문단 순서 올바르게 배열

■ 주어진 글에 맞는 올바른 주제문 찾기 등

해당 문제 유형들에 있어 1문장 단위의 간단한 문제들부터 수능 문제 등의 어려운 문제들도 있음

➔ 학습을 통해 가벼운 문제들부터 해결해야 함

2. 본론

- SLM(Small Language Model)이란?

- SLM에 대해 이야기 하기 전 우리는 먼저 **LLM**(Large Language Model)에 대해 이야기를 할 필요가 있다. **LLM이란 방대한 양의 데이터로 사전 학습된 초대형 딥 러닝 모델을 말하는 것**으로, 현재 우리에게 친숙한 GPT가 대표적인 LLM이다.
- SLM은 **LLM에 비해 사전 학습된 데이터의 양이 적은 소형 모델**을 의미한다. 사전 학습된 데이터의 양이 적으므로, **특정 분야에 대해 전문적으로 학습**하는 경우가 많다.
- SLM은 범용성 면에서는 LLM에 밀리나, LLM에 비해 **모델이 경량화**되어 있다는 장점이 있으며 **학습을 통해 특화된 분야에서는 LLM보다도 우수한 성능을 보일 수 있다.**

- 문제 유형에 따른 적합한 SLM 조사

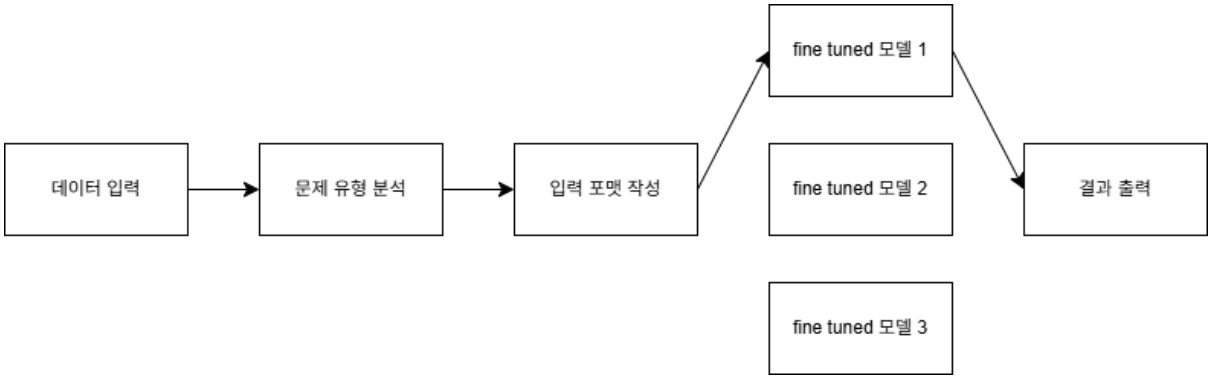
모델 종류		특징	관련 링크
비생성형 SLM	DistilBERT	BERT-base에 비해 가벼움	https://huggingface.co/docs/transformers/en/model_doc/distilbert
	TinyBERT	작고 빠르며, NLP에서도 경쟁력 있음	https://huggingface.co/huawei-noah/TinyBERT_General_4L_312D
	MobileBERT	가벼운 동시에 NLP 처리 성능은 강력	https://huggingface.co/docs/transformers/main/en/model_doc/mobilebert
	MiniLM	작은 크기, 뛰어난 성능	https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

	ALBERT	매우 가벼움, weight-sharing 주의	https://huggingface.co/albert/albert-base-v2
생성형 SLM	MobileLLM	모바일에서 효율	https://huggingface.co/facebook/MobileLLM-125M
	Phi-3.5	긴 컨텍스트, 다국어 지원	https://huggingface.co/bartowski/Phi-3.5-mini-instruct-GGUF
	GPT-2	프롬프트에서 텍스트를 생성하는 데 특화	https://huggingface.co/openai-community/gpt2
	Qwen	다국어 지원, 긴 컨텍스트	https://huggingface.co/Qwen

[표 1] 모델 비교표

- 시스템 구조

1. 문제를 입력 받는다. 문제를 입력 받는 형식은 문제의 지문 / 실제 문제 내용 / 문제에 대한 보기 3개로 나누어 입력 받게 된다.
2. 실제 문제 내용에 따라 문제 유형을 분석한다.
3. 문제 유형이 분석되면 해당 모델에 input으로 들어갈 수 있도록 입력된 데이터를 재가공한다.
4. 실제 매칭된 모델에 데이터를 입력한다.
5. 모델은 결과를 출력한다. 출력된 결과를 사용자에게 제공한다.



[그림 1] 시스템 구조도

- 도출된 결과와 문제의 실 답을 비교하여, 정확도 분석에 활용할 예정

- 개발 환경 및 사용 기술

개발 환경

- visual studio code (웹 기반 서비스 개발 및 로컬 학습)
- google colab
  - Colab에서 학습 가능한 모델들 -> Colab 환경 기반 학습
  - Colab에서 학습 불가능한 모델들 -> 학습이 가능한 로컬을 준비하여 로컬 학습

사용 기술

1. Fine tuning

- ‘미세 조정’이라고도 하며, 사전 학습된 모델에 대해 특정 데이터셋을 사용하여 추가적인 학습을 수행하는 작업을 말한다.
- fine tuning을 통해 우리가 처리하고자 하는 일들에 대해 모델들을 더 적합하게 조정할 수 있다.

2. pdfplumber

- pdf 파일을 읽어와 텍스트를 추출해주는 python 모듈
- pdf 형식으로 된 문서들에 대해 텍스트를 추출하여 txt로 변환 가능
- 데이터셋 제작 시 활용하여 편리한 데이터셋 제작에 활용 가능

3. 결론

- 최종적으로 개발할 서비스 및 프로그램에 대한 개요

- SLM을 활용하여 문제를 적합하게 풀어내는 서비스의 개발
    - 문제 유형에 따라 적합한 SLM 들을 선정해 모델들에 대해 데이터셋을 학습시킴
    - 공개 데이터셋을 적법하게 조정 + 실제 평가원 문제들을 데이터셋화 시켜 학습 및 검증, 테스트에 활용
    - 학습된 모델들을 기반으로 실제로 문제를 풀게 하고, 그 결과들을 비교하여 최적의 모델 선정

- 필요 시 문제를 읽어 오는 방식에 대한 다원화
    - 기본 베이스는 text-reading
        - 문제 및 보기를 모델이 읽을 수 있는 형태로 재가공하여, 모델에 파라미터로 넘겨주면 모델은 해당 내용들을 바탕으로 답을 추론함
    - 이미지에서 문제를 읽어 내어 텍스트로 넘기는 image-reading 방식도 고려 가능
        - 이미지에서 글자를 추출하여 text-reading이 가능한 형태로 문제를 재가공

- 제약 사항 분석

● 학습에 필요한 충분한 양의 데이터셋을 확보 가능?

  ■ 구할 수 있는 데이터셋들부터 우선 적용하여 최대한의 퀄리티를 유도함

● 학습에 필요한 자원이 충분한가? (AI 학습에 필요한 환경)

  ■ colab과 같은 원격 자원을 최대한 활용할 수 있도록 함

- 일정 및 역할 분담

(일정 -> 착수 보고서 제출 후부터)

5월		6월				7월					8월				9월				10월	
3	4	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2
모델 조사																				
		데이터셋 제작																		
					모델 파인 튜닝																
							중간 보고서 작성													
									서비스 개발											
														테스트 및 수정						
															최종 보고서 작성					
																		발표 준비		

이름	역할

강형원	계획 수립 모델에 대한 학습 데이터셋 자료 조사 및 제작
박규태	데이터셋 자료 조사 및 제작 모델에 대한 학습 웹 기반 서비스 제작 (프론트엔드)
주연학	데이터셋 자료 조사 및 제작 모델에 대한 학습 웹 기반 서비스 제작 (백엔드)