

연구팀명 BTS(Be the Talented Students) (부경대)

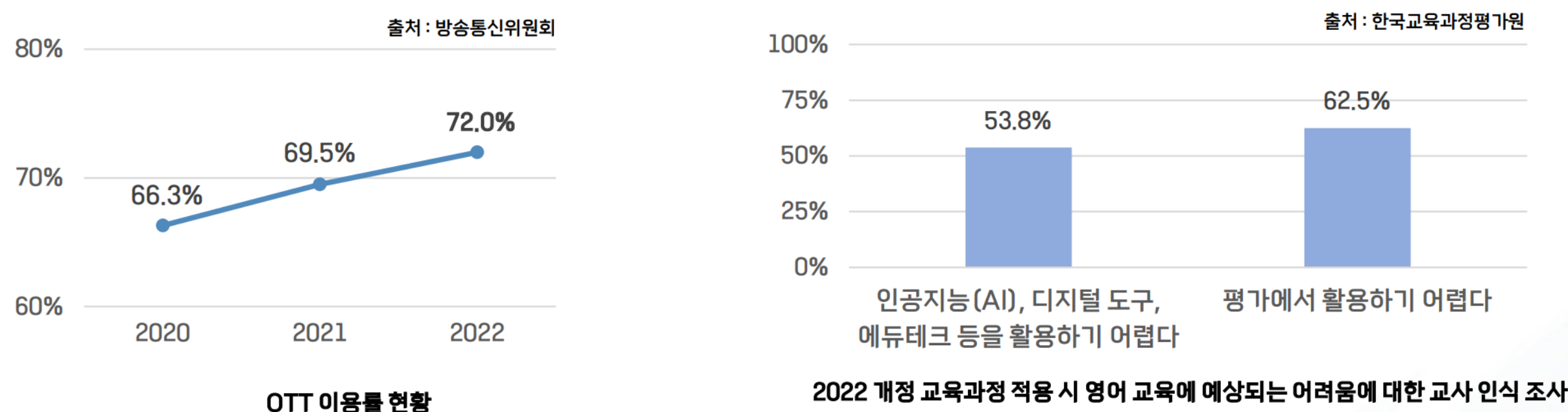
- + 지도교수 | 신인철
- + 참여업체 | (주)유림인포텍
- + 팀장 | Lan Dan
- + 팀원 | 나지현, 박효정, 이채혁

지능형 디지털 콘텐츠 관리 및 제작 기술 도구 연구

01 연구 개요



연구주제 선정 배경 및 타당성



- OTT 서비스의 확장, 문화 콘텐츠 소비 증가
- 문화 콘텐츠에 교육 모델 접목 → 흥미를 이끌어 내며 쉽게 사용 가능한 모델의 필요성

- 영어 학습의 필요성 증가
- 영어 어휘와 문법의 다양성, 발음의 다른 특성, 문화적 차이로 인한 학습자들의 어려움
- 학습자들의 개별적인 학습 스타일과 능력 수준 차이에 따른 맞춤형 학습 접근 필요
- 디지털 도구를 활용한 영어 학습 방법에 대한 관심 증가



연구 목적

자연스럽게 영어 능력을 향상시킬 수 있는 교육 모델 구축

- 주요 목표
- 경제적이며 가벼우면서도 GPT와 같은 LLM 모델과 유사한 성능을 가진 AI 모델의 개발
- 연구 범위
- 다양한 AI 알고리즘과 자연어 처리 기술을 활용한 핵심 기능 개발
(대치 문장 생성, 철자 오류 생성, 주요 주제 추출)
- 예상 결과 및 영향
- 저렴한 운영 및 개발 비용을 갖춘 AI 모델로, 하드웨어 및 인프라 리소스 절약
 - 자연어 처리 작업 속도 향상, 사용자 요청에 대한 빠른 응답 시간 제공
 - 모델 크기와 메모리 사용 최소화
 - 학문, 교육 도구 등 다양한 분야에서 활용할 수 있는 범용성 제공

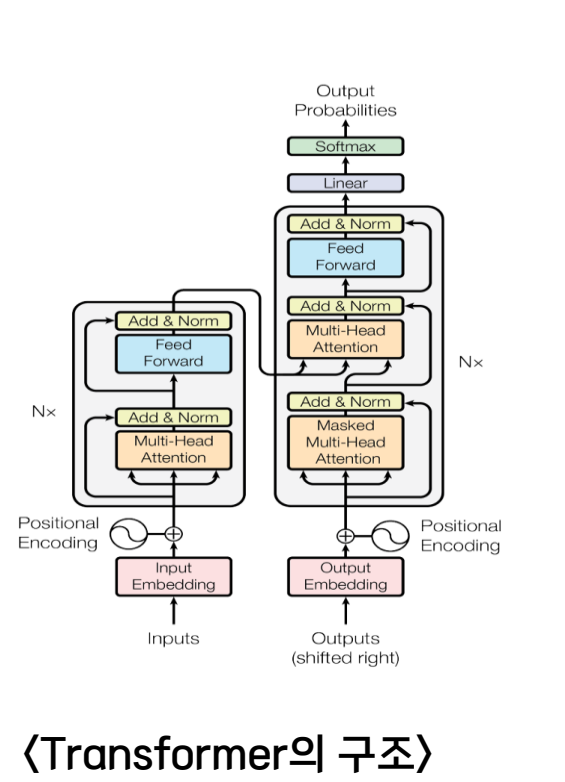
Transformer

2017 논문 'Attention Is All You Need'에서 발표된 신경망 구조
병렬처리가 어려운 RNN(Recurrent Neural Network)의 단점을 보완
자연어 처리 분야에서 기존 신경망 대비 높은 성능을 보임

크게 위치 인코딩, 멀티헤드 어텐션, 순방향 신경망으로 이루어져 있음
단어의 순서 정보를 위치 인코딩을 통해 인코딩

인코더 내의 멀티헤드 어텐션으로 문맥 내의 단어의 중요도 파악

순방향 신경망으로 학습



<Transformer의 구조>



자가 언어 학습 지원 모델

02 연구 내용

Transformer 기반 요약 모델 A

BERT(Bidirectional Encoder Representations from Transformers) 전이 학습 모델을 기반
사전 훈련이 완료된 BERT 모델을 대화 요약에 특화하기 위해 미세 조정(fine-tuning) 수행
미세 조정에는 Facebook 게시글 및 댓글 / Twitter 게시글 및 댓글 / CNN 뉴스 기사를 사용
미세 조정 후 모델의 하이퍼 파라미터(hyper parameter) 조정을 통해 추가적으로 모델 조정 (토큰 개수 최적화, Beam search 최적화)

전체 매개변수	레이어	self-attention head	hidden unit	크기
3억 5천만	24	16	1024	1.51GB

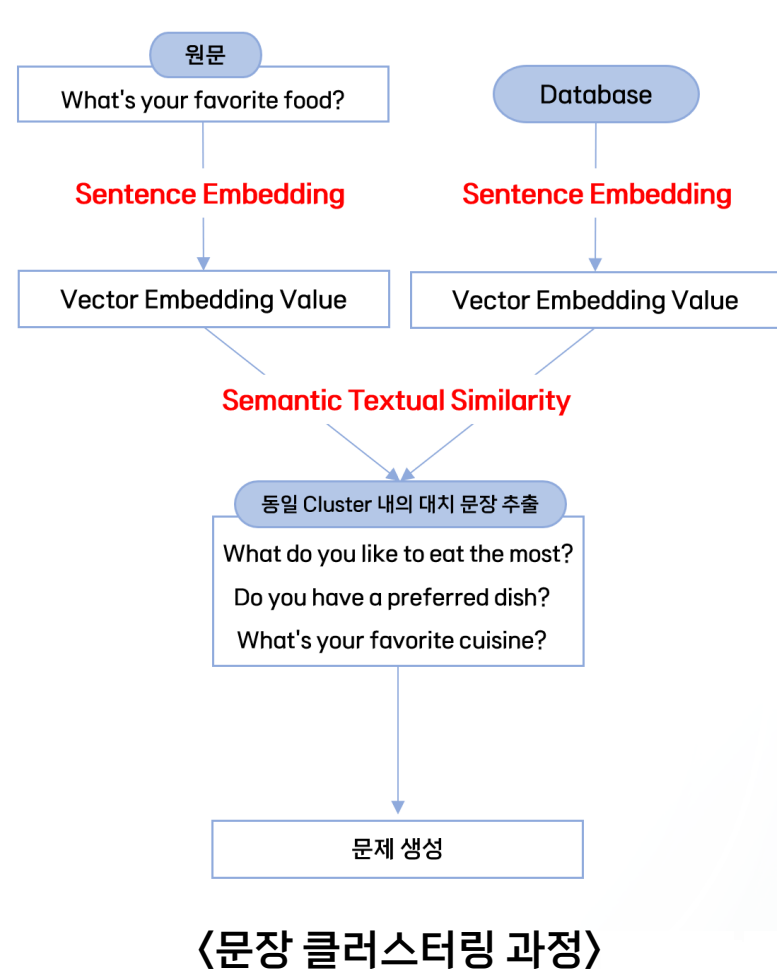
<모델 A의 제원>

Transformer 기반 클러스터링 모델 B

기존의 고전적인 DB 기반 클러스터링 방식에 Transformer를 접목
문장 임베딩 / 클러스터링 과정에서 Transformer를 사용
코사인 유사도(Cosine Similarity)를 통해 문장 간 의미론적 텍스트 유사성 판별
모델은 문장 임베딩 → 의미론적 텍스트 유사성 판별 과정을 통해 문장 클러스터링 수행
DB 구축에는 코넬 대학교의 'Cornell Movie-Dialogs Corpus'를 사용

Cornell Movie-Dialogs Corpus

영어 대화 말뭉치, 영화, TV 프로그램, 비디오 게임 등에서 수집된 대화
617편의 영화, 10,292장의 영화 캐릭터, 304,713개의 대사로 구성
데이터셋에서 캐릭터, 영화 ID 등 영화를 식별할 수 있는 라벨을 모두 제거한 후 사용



<문장 클러스터링 과정>

모델 A/B와 GPT-3.5 간 성능 비교 수행

모델 A의 대화 요약 성능 비교

Cornell Movie-Dialogs Corpus 데이터셋에서 대화를 대화 용치로 묶음
83,097개의 대화 용치 중, 15줄 이상 발화된 대화 용치 991개를 대상으로 비교
요약문의 길이는 문장 500자 이내로 제한
GPT-3.5의 경우 대화 용치의 내용을 500자 이내로 요약한 뒤 반환
원문과 요약문 사이의 ROUGE Score, METEOR Score, 정성적 평가로 성능 비교

모델 B의 클러스터링 성능 비교

사전 클러스터링 된 모델 B에 임의의 문장 1개를 넣어, 4개의 대치 문장 추출
GPT-3.5의 경우 임의의 문장 1개에 대한 대치 문장 4개를 생성한 뒤 반환
문장의 최소 길이는 10자로 제한, 임의의 영화 A의 대사 100개를 대상으로 비교
문장과 대치 문장 4개에 대한 코사인 유사도의 평균값으로 성능 비교

비교군 및 비교 지표

- GPT-3.5
gpt-3.5-turbo API를 사용(4,097 토큰, 2021년 9월)
GPT 모델은 별도의 미세 조정(fine-tuning)을 거치지 않음
- ROUGE score
ROUGE-1, ROUGE-2, ROUGE-L 지표를 사용
ROUGE score는 원문과 요약문 간의 일치성, 단어 출현 빈도, 단어 순서 등을 고려하여 평가 점수 산출
- METEOR score
요약문의 정확도와 유사도의 조화 평균을 통해 점수 산출
- 코사인 유사도(cosine similarity)
문장을 벡터화하여 벡터 간 코사인 값을 이용해 유사도 측정

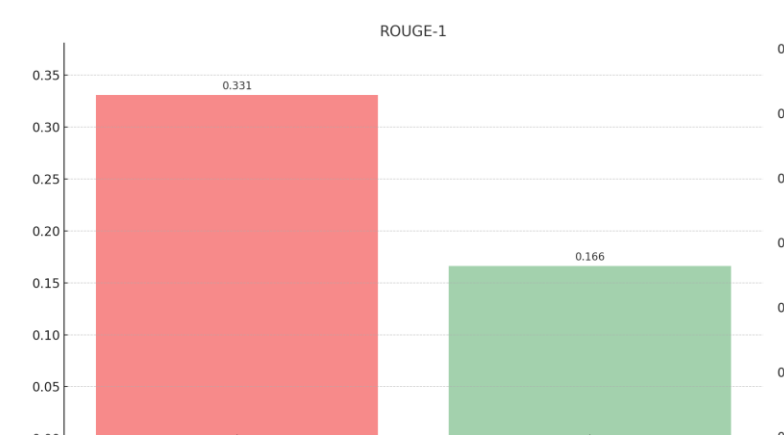
03 연구 결과

모델 A 분석 및 비교

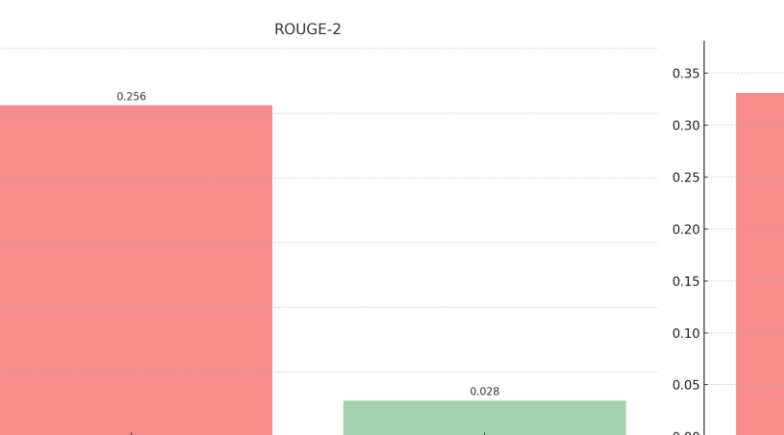
They won't find it because no such telegram exists. I and the Berlin Philharmonic toured the Third Reich, played in countries the Nazis had conquered. Are you saying that conducting in occupied territories from 1939 on was a commercial for Adolf and all he stood for? We never, never officially represented the regime when we played abroad. We always performed as a private ensemble. Art and politics must be separated.

The conversation revolves around the existence of a telegram and allegations regarding Wilhelm's association with the Nazis during World War II. Accusations of conscripting a music critic named Von der Null into the army due to bad reviews are discussed, but Wilhelm denies it. The conversation touches on the separation of art and politics.

<모델 A의 대화 용치 요약 예제>



<GPT-3.5의 대화 용치 요약 예제>



<모델 A와 GPT-3.5 간 비교>

결론

- 정성적 평가 결과 모델 A의 경우 원문을 최대한 유지하며 요약, GPT-3.5의 경우 대화 내용을 재조합하여 요약함을 알 수 있었음
- GPT-3.5 대비 모델 A의 ROUGE Score 및 METEOR Score가 2~9.15배 가량 높게 나옴
- 해당 평가 지표는 주로 기계 번역 분야에 사용되는 지표 → 원문의 단어를 사용하는 모델 A의 점수가 높게 나옴
- 정성적 평가 시에는 GPT-3.5의 성능이 모델 A의 성능 대비 상대적으로 높게 도출됨

모델 B 분석 및 비교

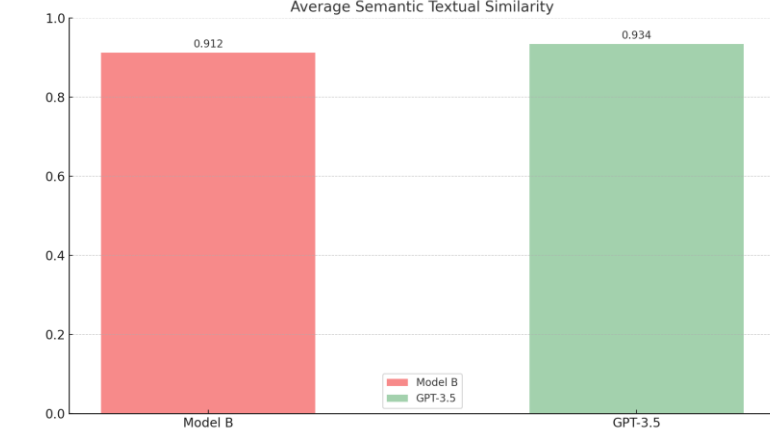
- I thought I never sold cigarettes to kids.

<원문 샘플>

- But I didn't sell cigarettes to any kids! Hey!
- I didn't sell cigarettes to any kids! I swear!
- But I didn't sell cigarettes to any kid!

- I never sold cigarettes to kids, just like I thought.
- It turns out that I didn't sell cigarettes to kids as I had thought.
- As it turns out, I didn't sell cigarettes to kids, just as I thought.

<모델 B가 클러스터에서 추출한 문장>



<GPT-3.5가 생성한 문장>

<모델 B와 GPT-3.5 간 코사인 유사도의 평균>

결론

- GPT-3.5 대비 모델 B의 코사인 유사도의 평균이 근소하게 낮게 도출됨
- 모델 B의 경우, 충분한 데이터베이스가 구축되면 생성형 AI 대비 적은 비용과 모델 크기로 유사한 성능을 낼 수 있을 것으로 기대됨
- 교육적인 목적으로 사용하기 위해서는 데이터베이스의 보완 및 추가적인 조정을 통해 0.95 이상의 평균 코사인 유사도가 필요할 것으로 사료됨

04

활용 방안 및 기대 성과



활용방안

온라인 교육 플랫폼

- MOOC (대규모 온라인 개방형 강좌) 제공업체에 적용
- 학습자들에게 더 많은 학습 도구, 자원 제공 가능

언어 평가 도구

- AI 모델은 학습자들의 언어 능력 평가, 진전 추적에 활용 가능
- 학습자들의 성장 모니터링, 평가에 도움

자기 학습 도구

- 학습자들의 개별적인 활용으로 자기 학습 지원
- 개인 학습 스타일에 따라 영어 능력 향상을 위한 자신만의 학습 경로 생성



기대성과

영어 학습

언어 능력 향상

- 학습자들의 자연스러운 언어 습득에 기여
 - 실제 대화와 문맥 경험 → 언어 이해, 말하기, 읽기, 쓰기 능력 모두 향상
 - 다양한 어휘, 표현 이해
 - 발음 및 억양 연습의 기회

학습 동기 부여

- 다양한 학습 활동과 상호작용을 통한 학습자의 흥미 유발

맞춤형 학습 경험

- 학습자의 능력 수준과 학습 스타일에 따른 맞춤형 학습 경험 제공

평가 및 피드백

- 학습자의 진행 추적, 개인화된 피드백 제공 → 학습자가 개선할 수 있는 방향 제시

콘텐츠

교육 콘텐츠 사업자

- 쉽고 편리하게 지능형 교육 콘텐츠 제작 및 제공

산업계 적용

- B2B, B2C 등 다양한 비즈니스 모델에 맞춤형 교육 시스템 공급

K-콘텐츠 세계화

- OTT를 통한 콘텐츠 활용에 교육 모델을 접목하여 긍정적 효과