17조
조은비, 박범창, 좌진우

한국어 기반 인공지능 생성 텍스트 탐지

최종발표

Contents 목차

INTRO 팀원 소개 및 역할 분담

인공지능학과
202202501 조은비

- 회의 일정 관리
- 데이터 수집 전략 및 실험 설계
- Gemini 생성 데이터셋 구축
- 모델 학습 및 모델 앙상블 적용

컴퓨터융합학부
202002494 박범창

- Copliot, GPT 생성 데이터셋 구축
- 데이터 전처리
- 발표 자료 작성 및 발표

컴퓨터융합학부
202002565 좌진우

- Claude 생성 데이터셋 구축
- 보고서 작성
- 모델 성능 평가

연구 개요 연구 배경 및 필요성

생성형 AI의 발전과 영향

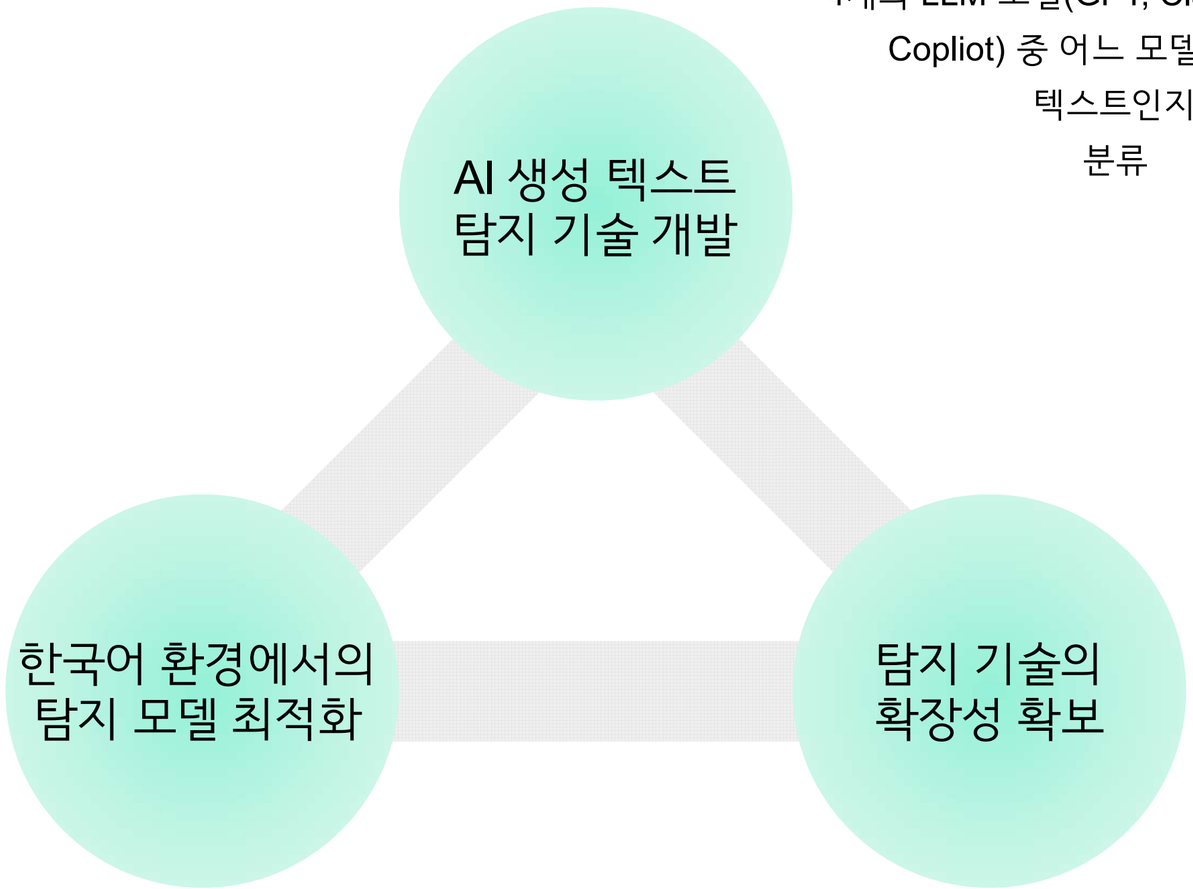
GPT, Gemini, LLaMA 등 대규모 언어 모델(LLM)이 등장하면서 인간 수준의 자연어 처리가 가능해짐
누구나 손쉽게 AI를 활용해 텍스트, 이미지, 오디오, 비디오 생성 가능

AI 악용 사례 증가

생성형 AI의 확산은 정보 조작, 허위 정보 생성, 저작권 침해, 부정행위 등 다양한 악용 사례로 이어짐
특히 인공지능이 생성한 텍스트는 사람이 작성한 것과 구별이 어려워 뉴스 기사 위조, 논문 대필, SNS 허위 계정
활동 등에서 악용 가능성이 높음.
이로 인해 생성된 텍스트가 사람이 작성한 것인지, 아니면 AI가 생성한 것인지를 판별하는 기술이 중요하며,
나아가 어떤 AI 모델에서 생성되었는지를 구분하는 것은 생성 콘텐츠의 신뢰도 분석, 부정행위 방지 등에
기여를 할 수 있음

연구 개요

연구 목표



한국어 기반 AI 생성 텍스트 탐지 모델
구축
4개의 LLM 모델(GPT, Claude, Gemini,
Copliot) 중 어느 모델로 생성된
텍스트인지
분류

기존 영어권 연구의 한계를 극복하고
한국어 환경에 특화된 모델 개발

단순 탐지가 아닌
언론/미디어, 교육, 법률, SNS 분야
등으로
적용이 가능하다.

사용자 분석 인터뷰 및 설문 & 문제 분석

문제점	이해 당사자	고충/니즈	이유	문제점 파악 방법
한국어 기반 AI 생성 텍스트 탐지의 어려움	NLP 연구자, 교수, 일반 사용자	AI가 생성한 글을 사람이 쓴 글과 구별하기 어려워 정확한 판별 도구 필요	영어 기반 도구는 한국어에서 탐지 정확도가 낮고, 실제 사용자들은 혼란을 겪고 있음	인터뷰 및 설문 조사
기존 AI 탐지 모델의 한계	언론 소비자 및 SNS 사용자	댓글이나 뉴스 제목 등 짧은 문장에서도 AI 탐지가 가능했으면 좋겠음	기존 탐지 모델은 긴 문장 위주로 학습되어 짧은 문장에서 탐지 성능이 매우 낮음	AI 관련 논문 조사 및 설문 응답 분석
AI 생성 가짜 뉴스의 확산 위험	일반 뉴스 소비자, 커뮤니티 사용자	신뢰할 수 있는 정보와 가짜 정보를 구분하고 싶어함	가짜 뉴스가 AI로 쉽게 생성되어 실제 뉴스처럼 유포되고 있어 혼란 발생	AI 뉴스 사례 분석 및 사용자 인터뷰
생성형 AI 기술에 대한 사회적 불신	일반 사용자, 교육기관 관계자	AI 생성 정보의 출처나 신뢰도를 표시해 주는 시스템을 원함	사용자들은 정보의 진위를 판별하기 어려워 AI 콘텐츠에 대한 신뢰가 낮아지고 있음	SNS, 커뮤니티 사례 및 뉴스 댓글 분석

사용자 분석 인터뷰 및 설문 & 문제 분석

기존 탐지 기술 언어적 한계

한국어 학습 데이터 부족

짧은 텍스트에서 탐지성능 감소

사용자 경험 및 실생활 적용 한계

핵심 아이디어 기존 문제 해결 방법

한국어 데이터셋 구축 및 탐지 모델 구현

- AI Hub의 “뉴스 기사 기계독해 데이터”와 이를 활용한 AI 생성 텍스트 데이터셋 수집 및 정제

- KoBART, KoELECTRA, KLUE-BERT 기반 모델 개발 및 학습

한국어 기반 데이터셋 확보와 한국어 기반 모델 사용을 통해 기존의 영어 기반의 탐지 기술의 언어적 한계 극복 가능

다양한 타입의 학습 데이터 확보

- 정형화된 데이터셋이 아닌 블로그 글, SNS 등 실생활에서 볼 수 있는 글들에 대한 데이터 수집 및 학습

형식이 자유로운 실제 텍스트 환경에서도 탐지 모델이 유연하게 작동될 수 있도록 성능 강화

핵심 아이디어 기존 연구 분석

AI 생성 텍스트 탐지 분야에서는 대부분의 최신 연구가 PLM(Pretrained Language Model)을 기반으로 모델을 설계

PLM의 장점:

- 전이학습이 가능하여 비교적 적은 양의 탐지용 데이터로도 빠른 학습 및 적용이 가능

- 기존 연구들에서 RoBERTa, BERT 등 PLM을 기반으로 한 분류기가 효과적으로 사용된 사례 다수 존재
왜 KoBART, KoELECTRA, KLUE-BERT를 사용했는가?

(Z. M. Kim, K. H. Lee, P. Zhu, V. Raheja, and D. Kang, “Threads of Subtlety: Detecting Machine-Generated Texts Through Discourse Motifs,”)

(Y. Li, Q. Li, L. Cui, W. Bi, Z. Wang, L. Wang, L. Yang, S. Shi, and Y. Zhang, "**MAGE: Machine-generated Text Detection in the Wild**," *arXiv preprint arXiv:2305.14960*, 2023.)

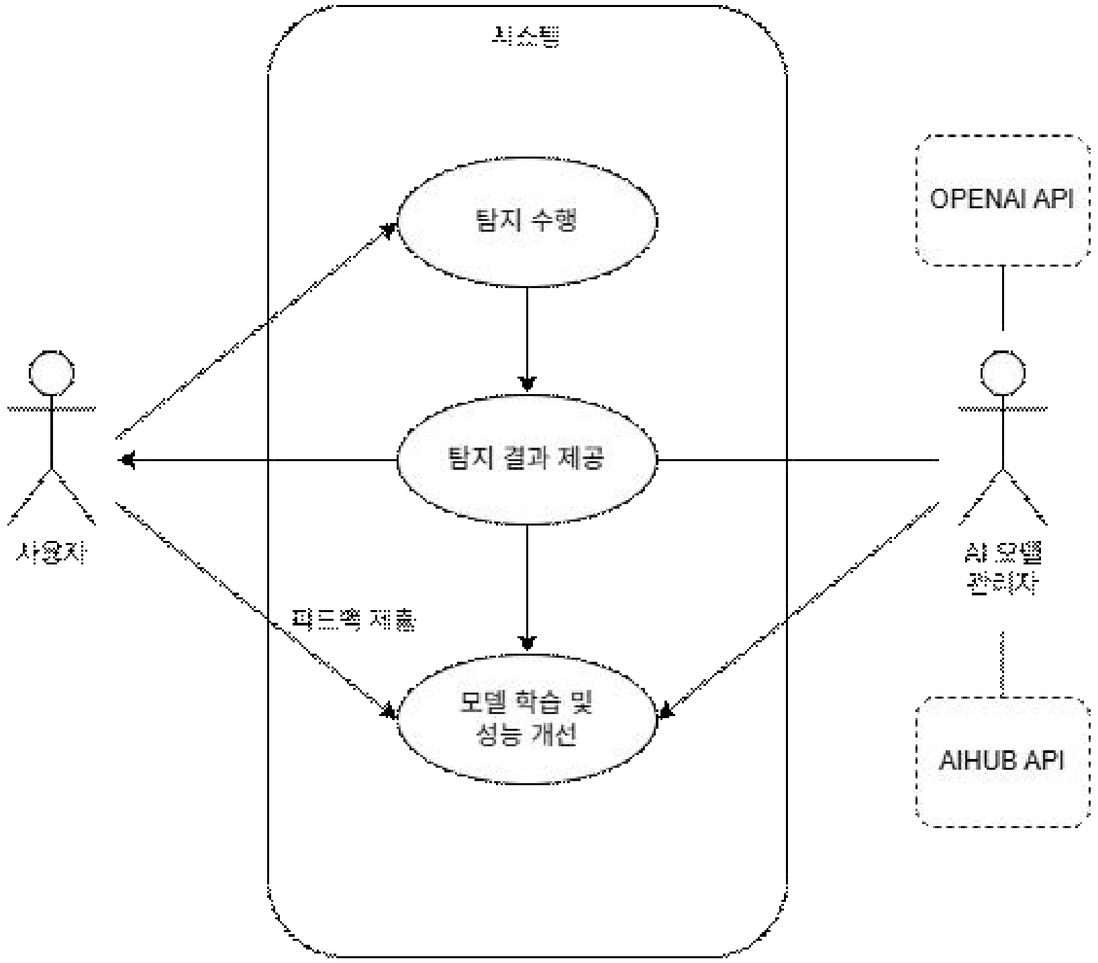
(S. Park, S. Kim, D.-K. Kim, and Y.-S. Han, "**Detecting LLM-Generated Korean Text through Linguistic Feature Analysis**," *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Toronto, Canada, 2025.)

한국어 기반 탐지 모델을 구축하기 위해서는 한국어 문법과 표현에 특화된 PLM이 필요

기존의 영어 기반 BERT나 RoBERTa는 구조적으로 매우 우수하지만, 한국어를 다루는 데에는 한계가 존재

→ 한국어에 최적화된 PLM인 KoBART, KoELECTRA, KLUE-BERT를 베이스모델로 사용하여 탐지 모델을 구축함으로써,
기존 영어 중심 탐지 기술의 언어적 한계를 극복하고, 한국어 텍스트의 문법적, 의미적 특성을 반영한 탐지 성능을 기대할 수 있음

데모 유스케이스 다이어그램



1. [사용자 입력] : 사용자가 AI 생성 여부를 판별하기 원하는 글을 입력

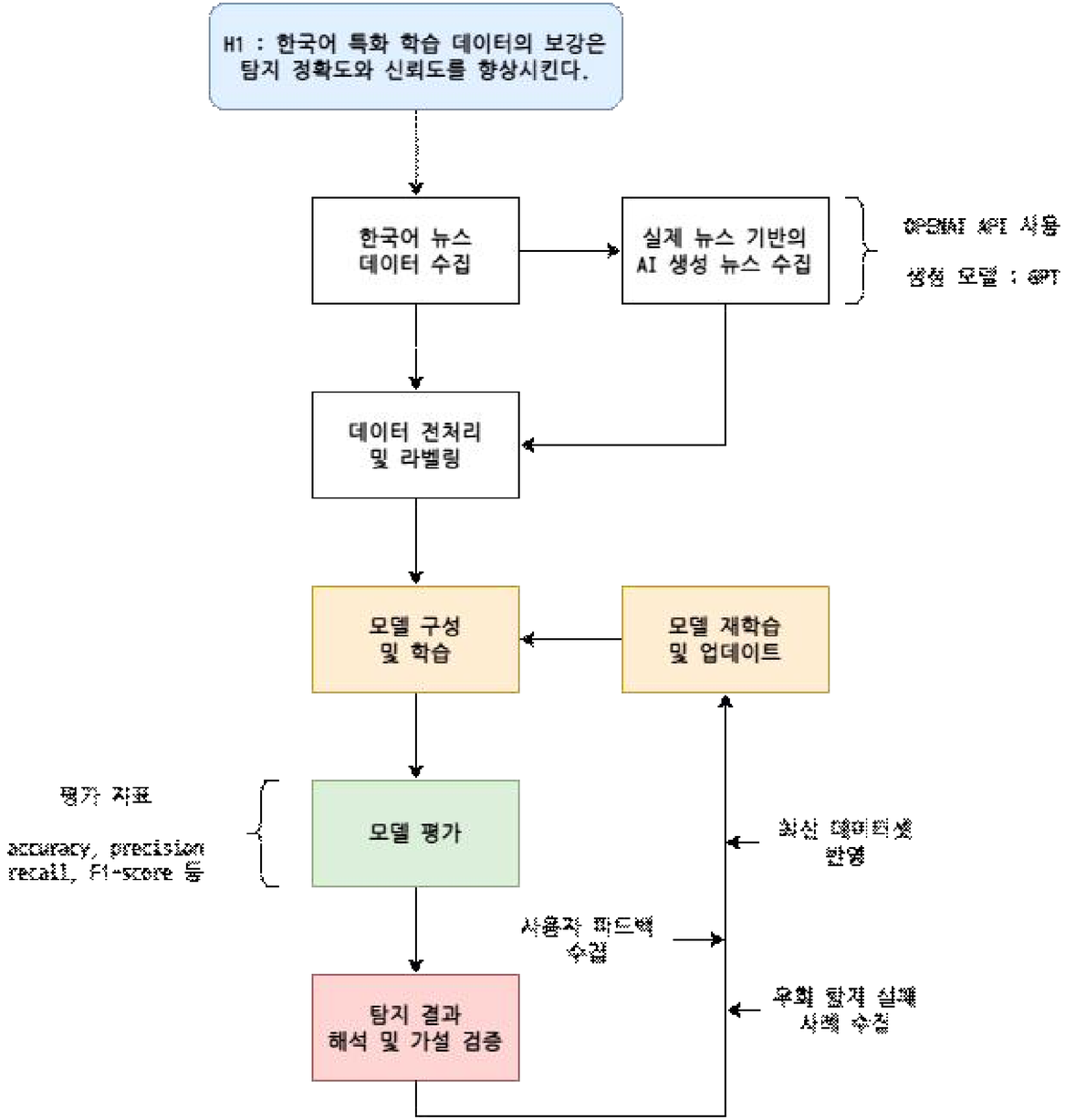
2. [탐지 요청 전송] : “탐지” 버튼 클릭 시, 입력 텍스트가 백엔드로 전송

3. [탐지 모델 분석] : 모델은 해당 텍스트가 어떤 AI로 작성되었는지를 판별

(GPT, Claude, Gemini, Copliot)

4. [결과 산출] : 시스템이 예상 생성 모델을 출력

데모 시퀀스 다이어그램


H1 : 한국어 특화 학습 데이터의 보강은
탐지 정확도와 신뢰도를 향상시킨다.
한국어 뉴스
데이터 수집
실제 뉴스 기반의
AI 생성 뉴스 수집
OPENAI API 사용
생성 모델 : GPT
데이터 전처리
및 라벨링
모델 구성
및 학습
모델 재학습
및 업데이트
평가 지표
accuracy, precision
recall, f1-score 등
모델 평가
최신 데이터셋
반영
사용자 피드백
수집
탐지 결과
해석 및 가설 검증
유효 탐지 실제
사례 수집

테스트 테스트 계획

실험 대상 / 환경

- 데이터 : 동일 프롬프트에 대해 각 모델(GPT-3.5, GPT-4, Claude, Gemini)에서 수집한 텍스트 데이터셋 (총 2000개)

- 실험 환경: Python 3.x, Google Colab Pro, PyTorch, Huggingface Transformers

- 모델 구성: KoBERT, KoELECTRA, klue/roberta-base 기반 다중 클래스 분류기

측정 지표 및 도구

정량 지표 : Accuracy, Precision, Recall, F1-score, Confusion Matrix

정성 지표 : 오분류 사례, 문체 패턴 차이 분석

도구 : scikit-learn, pandas, matplotlib, seaborn

테스트 프로토타입

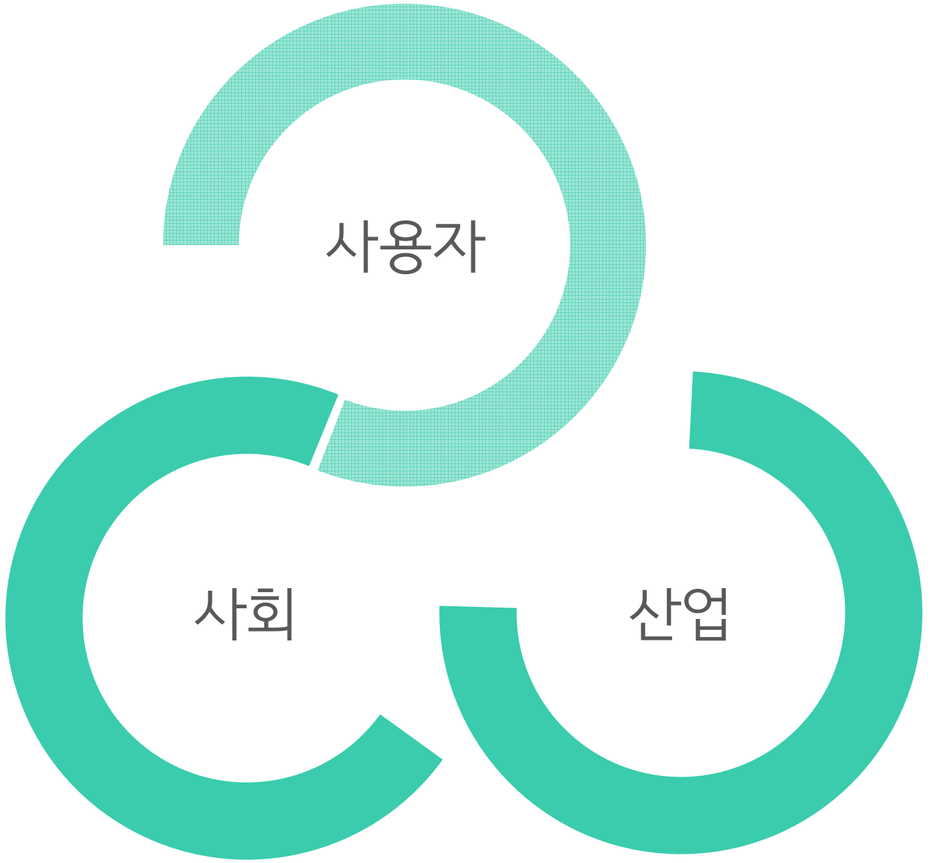
프로토타입

테스트 성능 평가


한국어 생성 모델 분류기
앙상블 모델 Classification Report:
precision recall f1-score support
claude 0.85 0.95 0.89 75
copilot 0.84 0.81 0.82 75
gemini 0.98 0.83 0.90 75
gpt4 0.84 0.89 0.86 75
accuracy 0.87 300
macro avg 0.88 0.87 0.87 300
ghted avg 0.88 0.87 0.87 300

기대 효과 및 추가 계획

기대 효과



허위 정보로부터 보호 / 올바른 정보 소비

뉴스, 블로그, 댓글 등 다양한 온라인
컨텐츠에서 정보의 신뢰성 검증 가능

가짜 뉴스, 조작 콘텐츠 유포 억제를 통해
정보 생태계의 신뢰도 향상

교육기관, 언론사 등 공공 분야에서 AI
콘텐츠
감별 도구로의 활용 가능성 제시

AI 오남용 견제 장치로 작동

콘텐츠 검수 자동화 시스템과 연계하여
운영 효율성 및 신뢰도 동시 확보

법률, 의료, 교육 등 전문 분야에서의
AI 콘텐츠 감시 도구로 확장 가능성

기대 효과 및 추가 계획

추가 계획

향후 추가 계획

짧은 텍스트 데이터에 대한 탐지 성능 향상
- 댓글, SNS 등 짧은 텍스트를 위한 탐지 성능 향상을 위한 모델 학습 및 개선

학습 데이터의 도메인 다양화
- 뉴스 기사 뿐만 아니라, 다양한 온라인 환경에서의 텍스트 데이터를 학습 데이터로 사용

사용자 피드백 기반 탐지 결과 개선
- 사용자로부터 받은 피드백을 탐지 모델 학습에 이용하여 시스템 개선

Thank You