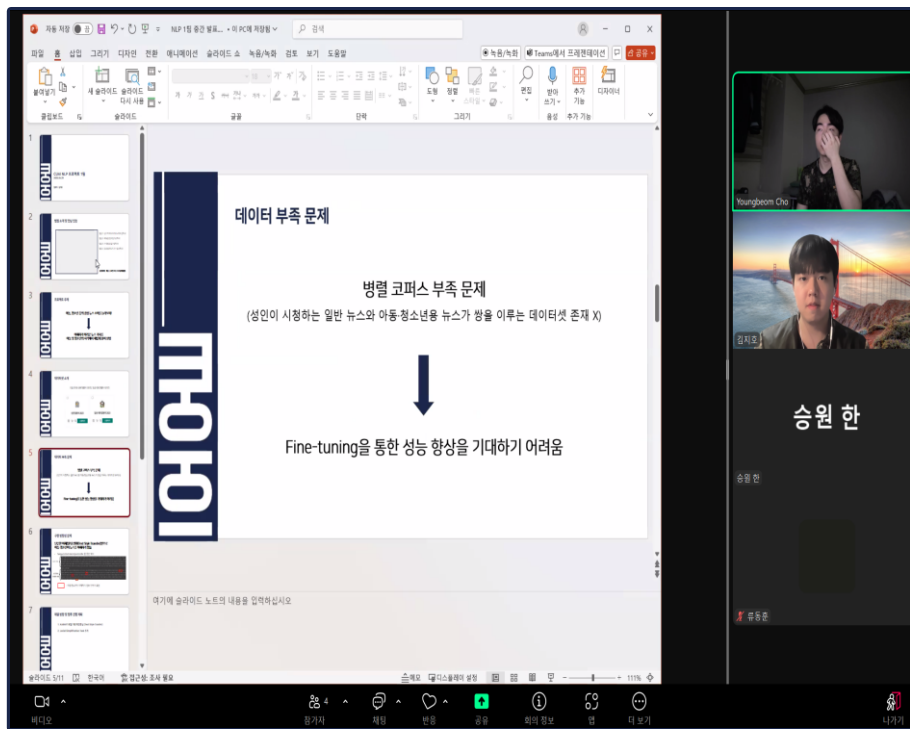


## CUAI NLP 프로젝트 1팀

2025.07.01

발표자 : 김지호

## 팀원 소개 및 만남 인증



팀원 1:  
김지호(미디어커뮤니케이션학부)

팀원 2: 류동훈(전자전기공학부)

팀원 3: 조영범(산업보안학과)

팀원 4: 한승원(에너지시스템공학부)

정기회의 - 매주 수요일 오후 11시  
(비대면)

## 프로젝트 주제

**아동, 청소년 층의 경성 뉴스 소비를 늘려보자!**




**이해하기 어려운 뉴스 기사를  
아동 및 청소년의 시각에서 쉽게 풀어 설명**


## 데이터셋 소개


국립국어원 신문 말뭉치 데이터, 일상 대화 말뭉치 데이터


☐



신문 말뭉치 2023










신청하기


☐



일상대화 말뭉치 2023







신청하기

## 구현 방향성 문제

# 단순한 어체(말투) 변환(Text Style Transfer)만으로 아동·청소년이 뉴스를 이해하기 힘들

- heegyu/kobart-text-style-transfer 를 통한 예시

뉴스 기사

'크립토 대통령'을 자처한 도널드 트럼프 미국 대통령이 집권하자 글로벌 가상자산 시장은 외레 움츠러들었다. 비트코인은 사상 최고가를 찍은 뒤 트럼프 대통령 취임 직후 급락했고, 알트코인 시  
트럼프 대통령 취임식이 열린 지난 1월 20일 글로벌 가상자산 시가총액은 3조5300억달러에 달했다. 그러나 취임 100일을 목전에 둔 이달 28일 약 3조287억달러로 14.20%가량 증발했다. 가상화폐  
미국 연합정부 차원의 추가 비축 기대감에 9만5000달러 선까지 올랐던 비트코인은 회담 직후 8만6000달러 선까지 떨어졌다. 이후 트럼프 대통령이 관세 압박 수위를 높이자 7만6873달러까지 말리  
특히 밌코인 시장에서는 트럼프 대통령이 논란의 당사자로 등장하며 시장 혼란을 부추겼다. 그는 취임 직전 공식 밌코인인 '오피셜 트럼프'를 출시했다. 이를 만에 44달러대까지 급등했으나 10여  
트럼프 행정부의 정책 방향이 구체화하기 전까지 변동성이 이어질 것으로 전망된다. 홍성욱 NH투자증권 책임연구원은 "가상자산 시장이 트럼프 대통령 취임 이후 하락한 건 시장의 기대감이 높았

초등학생  
어체로 변환

초등학생 스타일로 변환  
○○○  
비트로 대통령, 도널드 트럼프 미국 대통령 집권하자마자 세계 가상자산 시장은 급격하게 흔들렸음 비트코인은 사상 최고가 찍은 뒤 트럼프 대통령 취임 직후부터 급락했고, 알트코인 시장은 더 큰  
○○○  
트럼프트 대통령 취임식이 열렸던 지난 달 19일 이래로 세계 가상자산의 시가총액은 약 3조달러에 달했음. 그러나 취임 100일 뒤인 지난 달 말에는 미국 정부로부터 비트코인 추가 매입 금지 발표  
○○○  
미국 정부 정부 차원에서 추가 비축 기대감 9만달러 선까지 올라갔던 비트코인은 담날에 8만달러 선까지 떨어짐. 이후 도지코인은 미국 정부 차원에서 비트코인 비축을 위해 가격 조절에 나서면서  
○○○  
특히 밌코인 시장에서는 트럼프가 시장 개입의 당사자로 나서서 시장 혼란 조장하고 있음. 그는 취임 직후에 밌코인 시장에서의 맹활약으로 유명했는데, 이틀만에 44달러까지 폭등했지 ㅋㅋ,,,,,  
○○○  
트럼프 행 정부 방향이 구체화하기 전까지 변동성 커질 듯

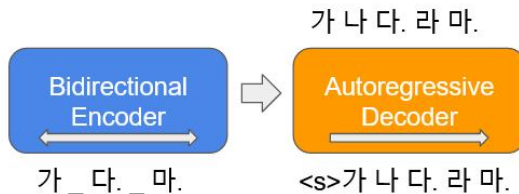


: 아동·청소년이 이해하기 힘든 단어가 포함

## 모델 소개

### KoBART

- BART(Bidirectional and Auto-Regressive Transformers)는 입력 텍스트 일부에 노이즈를 추가하여 이를 다시 원문으로 복구하는 autoencoder의 형태로 학습
- 한국어 BART(KoBART)는 논문에서 사용된 Text Infilling 노이즈 함수를 사용하여 40GB 이상의 한국어 텍스트에 대해서 학습한 한국어 encoder-decoder 언어 모델



<https://github.com/haven-jeon/KoBART>

## 모델 소개

## 데이터셋 구축

- OpenAI API를 이용하여 병렬 데이터셋 구축
- 사용 모델: GPT 4o mini
- 원문 기사 <-> 어휘 및 문장 단순화 기사의 병렬 데이터셋 구축
- 총 약 16000쌍의 병렬 코퍼스 생성
- 방법 1: 어려운 단어를 괄호를 통해 설명
- 방법 2: 어려운 단어를 쉬운 단어로 치환하여 설명

```
def simplify_text_for_youth(original_text, client, max_retries=3):
    """
    OpenAI API를 사용하여 원본 텍스트를 아동/청소년이 이해하기 쉬운 형태로 변환합니다.
    """
    prompt = f"""
    다음 문장을 초등학교 고학년~중학생이 쉽게 이해할 수 있도록 바꿔주세요.
    원래 의미는 유지하되, 쉬운 단어를 사용하고, 복잡한 개념은 간단히 풀어서 설명해주세요.

    원문: {original_text}

    아동/청소년용 문장:
    """
```

```
PROMPT = textwrap.dedent("""
    다음 문장에서 어려운 단어 (경제 전문 용어 등)를 ++초등학생도 이해할 만한 쉬운 단어++로
    - 어려운 단어 옆에 괄호를 열고 간단하게 그 단어에 대해 설명을 해줘, 예를 들어 연체율(%)
    - 원문의 의미는 유지해 줘,
    - 사실관계에 중점을 줘야 해, 인물이름 등의 팩트가 바뀌지 않게 주의해 줘
    - 사람의 말을 인용한 인용문을 바꿀 때는 인용문 형식을 유지해야 해, 내용만 쉽게 바꿔줘
    문장: "{sent}"
    """)
###
```

## 모델 소개

## 모델 학습

- gogamza/kobart-base-v2 모델을 baseline으로 모델 학습
- 2번째 데이터셋 구축 방법을 통해 생성한 예시

**원문:** 'ISO22301'은 국제표준화기구(ISO)가 제정한 국제표준 인증이다. 업무 수행이 불가능한 위기 상황이 발생했을 때 사전에 구축한 기준과 훈련된 행동에 따라 핵심 기능을 빠르게 복구하는 총체적 경영 능력을 평가해 수여한다.

**생성:** 'ISO22301'은 국제표준화기구(ISO)가 정한 국제표준 인증(국제적으로 인정받는 인증)이야. 일을 할 수 없는 상황이 생겼을 때 미리 정해진 기준(기준은 어떤 규칙이나 기준)과 잘 훈련된 행동(잘 행동하는 것)에 따라 중요한 기능을 빠르게 복구하는 총체적 경영 능력(일을 잘 하는 능력)을 평가해줘.

**문제점:** 괄호 때문에 오히려 가독성이 떨어지며, 어렵지 않은 단어에도 괄호가 생성되는 문제



## 모델 소개

피드백: KoBART는 너무 과거의 모델로  
최신 동향과는 동떨어진 모델



최신 LLM을 프로젝트에 적용하여 비교해보자

## 모델 소개

### GEMMA

- 구글에서 공개한 경량 오픈 LLM 모델
- Gemini의 연구, 기술을 이어받아 누구나 사용할 수 있도록 라이선스 제공
- 가벼운 크기(2B, 7B) 모델이 존재해 제한적인 환경에서도 가볍게 돌릴 수 있음
- Text-to-Text 디코더 only 구조



## 모델 소개

### GEMMA 2B 모델 파인튜닝

- 데이터를 이용해 GEMMA 2B 모델 파인튜닝
- GEMMA 모델 파인튜닝 시 프롬프트와 함께 병렬 데이터셋 제공
- QLoRA(Quantized Low-Rank Adapter) 적용 – 파라미터 학습 최소화, 경량화

```
def generate_prompt(sample):
    return (
        "<bos><start_of_turn>user\n"
        "다음 뉴스 기사를 청소년이 쉽게 이해할 수 있도록 변환.\n"
        "어려운 용어는 쉬운 말로 풀어쓰고, 복잡한 문장은 간결하게 바꾸되, "
        "중요한 정보는 유지.\n\n"
        f"{sample['original']}<end_of_turn>\n"
        "<start_of_turn>model\n"
        f"{sample['simplified']}<end_of_turn><eos>"
    )
```

```
lora_config = LoraConfig(
    r=8, #텍스트 변환 태스크는 단순 생성보다 더 복잡할 수 있어 약간 높은 랭크
    target_modules=["q_proj", "o_proj", "k_proj", "v_proj", "gate_proj", "up_proj", "down_proj"],
    task_type="CAUSAL_LM",
    lora_alpha=16, # 스케일링 파라미터 추가
    lora_dropout=0.1,
    bias="none"
)
```

## 모델 소개

### 파인튜닝 모델을 통한 추론

**원문:** 박태봉 시 자치행정과장은 “미처 신청하지 못한 시민 분들은 서둘러 신청해 주시고, 미사용 잔액은 시로 귀속되는 만큼 반드시 기한 내에 사용해 줄 것”을 당부했다.

**생성:** 박태봉 시 자치행정과장은 “미리 신청하지 못한 시민들은 빨리 신청해 주세요. 그리고 사용하지 않은 돈은 시에 돌아가니까, 꼭 기한 내에 사용해 주세요.”라고 부탁했습니다.

**원문:** 임헌경 맑은물사업소장은 “쾌적한 생활환경 조성과 안정적 하수처리를 위해 하수처리시설을 개선하겠다”고 밝혔다.

**생성:** 임헌경 맑은물사업소장은 “사람들이 편하게 살 수 있는 환경을 만들고, 하수를 안전하게 처리하기 위해 하수처리를 더 좋게 만들겠다”고 말했습니다.

감사합니다