2024 한국코퍼스언어학회
가을 전국학술대회

인간과 기계의
언어 소통

일시 : 2024년 10월 11일
장소 : 네이버 그린팩토리 & 1784

주최 한국코퍼스언어학회 | 한국정보과학회 언어공학연구회

후원 LG AI 연구원 | NCSOFT | VUNO | 한국전자통신연구원 | NEVER | Upstage | Saltlux | 국립국어원 | 마인즈솔루션 | 메트릭스 | 연세대학교 언어정보연구원 | 테디썸 | 튜터러스랩스 | 싸인랩 | IIRTECH

2024년도 한글 및 한국어 정보처리 & 한국코퍼스언어학회 공동 학술대회

한국코퍼스언어학회 프로그램

일시 2024. 10. 11(금) 09:00~17:50 **장소** 네이버 그린팩토리 & 1784

09:00~09:30	등록	1784 3층
09:30~11:00	[세션 1] 사회: 봉미경(연세대) 1784 #4 **말뭉치와 언어 분석**	[세션 2] 사회: 남신혜(경희대) 1784 #5 **글쓰기와 AI**
	한·중 병렬 말뭉치의 개체명 정렬 관계 연구 발표: 비립(숙명여대), 토론: 최재웅(고려대)	**복합양식 텍스트 교육의 효과 분석** 발표: 주민재(명지대), 토론: 윤영민(인하대)
	통역 품질 향상을 위한 감정 분석 기반 RAG 시스템 개발: 구어 코퍼스를 중심으로 발표: 송지현(이화여대), 이용훈(충남대), 토론: 한진영(중앙대)	**LLM을 활용한 한국어 글쓰기 평가와 생활기록부 생성 모델의 실제** 발표: 임경태(서울과기대), 토론: 이진(연세대)
	근현대 텍스트의 재발견과 어휘개념사 발표: 김일환(성신여대), 토론: 이성우(한림대)	**학습자 글쓰기 자동 평가 모델: 피쳐 기반 모델** 발표: 최지명(이화여대), 토론: 노영빈(NCsoft)
11:00~11:45	[초청강연] **하이퍼클로바X 개발 현황과 네이버의 소버린전략** _ 유강민 리더(네이버)	GF커넥트홀
11:45~12:30	[초청강연] **진실하고 안전한 언어모델 개발 (Toward Truthful and Safe Language Model)** _ 이환희 교수(중앙대)	GF커넥트홀
12:30~14:00	점심 식사	
14:00~15:00	[개회사 및 초청강연] **Towards Human-like Conversational AI: A Cognition-oriented Framework** _ 최진호 교수(Emory Univ.)	GF커넥트홀
15:00~15:50	[초청강연] **역사말뭉치 분석 및 활용** _ 옥철영 명예교수(전 울산대)	GF커넥트홀
15:50~16:10	커피 브레이크	
16:10~17:40	[수상 논문 발표] 사회: 임경태(서울과기대) 1784 #4 **국립국어원 '인공지능(AI)말평' 활용 연구**	[세션 3] 사회: 송영숙(Sionic AI) 1784 #5 **언어모델과 말뭉치**
	한국어 이미지 캡셔닝 향상을 위한 유창성 개선 모듈 발표: 유용상, 이기훈 ,임형준(롯데이노베이트)	**인간 가치 정렬(Human Alignment)를 위한 한국어 지시 이행(Instruction Following) 말뭉치 설계를 위한 기초 연구** 발표: 한지윤(업스테이지), 토론: 조원익(삼성전자)
	Tabular-TX: In-Context Learning을 통한 주제-설명 구조 기반 표 요약 발표: 곽태윤, 김지수, 정기용(성균관대학교), 이동건(포항공과대학교), 박희선(성균관대학교)	**LLM을 이용한 수학문제 합성데이터 구축** 발표: 이숙의, 장지현, 강수희, 홍채은(마인즈솔루션), 토론: 신서인(한림대)
	프롬프팅과 미세 조정 모델의 의미적 앙상블을 활용한 한국어 Table-to-Text 성능 향상 발표: 강어진, 홍은진, 김이서, 김주애(한국외국어대학교)	**Bllossom 프로젝트: 한국어 언어 모델의 꽃을 피우다** 발표: 함영균(테디썸), 토론: 김일근(HP)
	한국어 표 설명 능력 향상을 위한 전처리 및 학습방법론 탐구 발표: 김창현, 김승희, 김태욱(한양대학교)	
17:40~17:50	폐회사	

후원 LG AI 연구원, NCSOFT, VUNO, 한국전자통신연구원, NAVER, Upstage, Saltlux, 국립국어원, 마인즈솔루션, 메트릭스, 연세대학교 언어정보연구원, 테디썸, 튜터러스랩스, 싸인랩, IIRTECH

2024 한국코퍼스언어학회 가을 전국학술대회

인간과 기계의 언어 소통

목 차

2024 한국코퍼스언어학회 가을 전국학술대회

KACL-HCLT 초청 강연 및 국립국어원 '인공지능(AI) 말평' 활용 연구

|| KACL-HCLT 초청 강연

○ 하이퍼클로바X 개발 현황과 네이버의 소버린전략 - 유강민 리더 (네이버)

○ 진실하고 안전한 언어모델 개발 (Toward Truthful and Safe Language Model)

- 이환희 교수 (중앙대)

○ Towards Human-like Conversational AI: A Cognition-oriented Framework

- 최진호 교수 (Emory University)

발표자료 다운로드 링크 :

https://sites.google.com/view/hclt2024/%ED%94%84%EB%A1%9C%EA%B7%B8%EB%9E%A8/invited-talk

|| 국립국어원 '인공지능(AI) 말평' 활용 연구

○ 한국어 이미지 캡셔닝 향상을 위한 유창성 개선 모듈

- 유용상, 이기훈 ,임형준(롯데이노베이트)

○ Tabular-TX: In-Context Learning을 통한 주제-설명 구조 기반 표 요약

- 곽태윤, 김지수, 정기용(성균관대학교), 이동건(포항공과대학교), 박희선(성균관대학교)

○ 프롬프팅과 미세 조정 모델의 의미적 앙상블을 활용한 한국어 Table-to-Text 성능 향상

- 강어진, 홍은진, 김이서, 김주애(한국외국어대학교)

○ 한국어 표 설명 능력 향상을 위한 전처리 및 학습 방법론 탐구

- 김창현, 김승희, 김태욱(한양대학교)

발표자료 다운로드 링크 :

https://drive.google.com/file/d/1hrJjJK4clrJ3v0XHH-7Jt2Kt65kEyOD1/view

2024 한국코퍼스언어학회 가을 전국학술대회

후 원 사

2024 한국코퍼스언어학회 가을 전국학술대회

인간과 기계의 언어 소통

Session 1

한·중 병렬 말뭉치의 개체명 정렬 관계 연구 (비립, 숙명여대)

—

통역 품질 향상을 위한 감정 분석 기반 RAG 시스템 개발 :
구어 코퍼스를 중심으로 (송지현 이화여대, 이용훈 충남대)

—

근현대 텍스트의 재발견과 어휘개념사 (김일환, 성신여대)

Asymmetric Alignments of Mulitlingual Named Entities: Focus on the CJK Parallel Corpus

Dylan Fei @ Sukmyung Univ.
KACL 2024 Fall

Background

Analysis

Introduction

Experiment

Conclusion

One Word but Many Tongues

- Lexical Taxonomy
- Conceptual Metaphor
- Semantic Representation

Topics

Hypothesis

Multilingual Named Entities Aligment

- Linguistics:
  - Word-sense disambiguation
  - Semantic Mapping
- NLP:
  - Named-entity Recognition
  - Entity Linking

Challenges

Connections

Difficulties in NE Alignment

- In-language:
  - Ambiguity
  - Absence
  - In-context Information
  - .....
- Between-language:
  - Diverse Names
  - Missing Concepts
  - Inconsistent Terminologies
  - ...

Multilingual Named Entities Aligment

- Linguistics:
  - Word-sense disambiguation
  - Semantic Mapping
- NLP:
  - Named-entity Recognition
  - Entity Linking

Challenges

Connections

CJK's Similaries

- Naming Convention
- Chinese-character Lexicons
- Oriental Lifestyle
- Historical Communication
- ...

Multilingual Named Entities Aligment

- Linguistics:
  - Word-sense disambiguation
  - Semantic Mapping
- NLP:
  - Named-entity Recognition
  - Entity Linking

Challenges

Connections

One Word but Many Tongues

- Lexical Taxonomy
- Conceptual Metaphor
- Semantic Representation

Topics

Hypothesis

Reserach Question

- How does CJK's NEs differ from each other?
- Which kinds of NEs are most difficult to match up?
- Is it a way to unify CJK's NEs?

One Word but Many Tongues

- Lexical Taxonomy
- Conceptual Metaphor
- Semantic Representation

Topics

Hypothesis

Asymmetric Alignments of Mulitlingual Named Entities: Focus on the CJK Parallel Corpus

Dylan Fei @ Sukmyung Univ.
KACL 2024 Fall

Background
Analysis
Introduction
Experiment
Conclusion

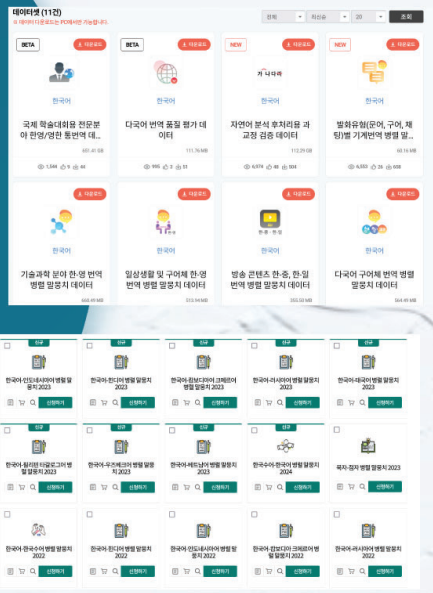
Background

- CJK Parallel Text Sourcing
- CJK NE Classification
- CJK NE Alignment

Multilingual Dataset
Entity Scheme
Cross-lingual Alignment

Absence of Multilingual Parallelled Data

- Lack of Multilingual Textsets across Korean and other Languages
- Lack of Alignment Methods for Bilingual Parallel Corpora
- Lack of Multilingual Metadata Resources for NE Pairing



데이터셋 (11건)

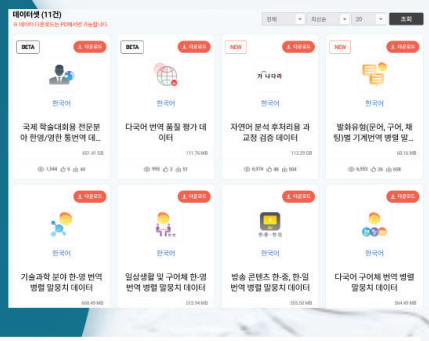
※ 데이터 다운로드는 PC에서만 가능합니다.

전체 | 최신순 | 20 | 조회

BETA · 다운로드 · 한국어 · 국제 학술대회용 전문분야 한영/영한 통번역 데... · 651.41 GB · 1,544 9 44	BETA · 다운로드 · 한국어 · 다국어 번역 품질 평가 데이터 · 111.76 MB · 995 3 51	NEW · 다운로드 · 한국어 · 자연어 분석 후처리용 과교정 검증 데이터 · 112.29 GB · 6,974 48 504	NEW · 다운로드 · 한국어 · 발화유형(문어, 구어, 채팅)별 기계번역 병렬 말... · 60.16 MB · 6,553 26 658
다운로드 · 한국어 · 기술과학 분야 한-영 번역 병렬 말뭉치 데이터 · 660.49 MB	다운로드 · 한국어 · 일상생활 및 구어체 한-영 번역 병렬 말뭉치 데이터 · 513.94 MB	다운로드 · 한국어 · 방송 콘텐츠 한-중, 한-일 번역 병렬 말뭉치 데이터 · 355.50 MB	다운로드 · 한국어 · 다국어 구어체 번역 병렬 말뭉치 데이터 · 564.49 MB

Absence of Multilingual Parallelled Data

- Lack of Multilingual Textsets across Korean and other Languages
- Lack of Alignment Methods for Bilingual Parallel Corpora
- Lack of Multilingual Metadata Resources for NE Pairing



Background

- CJK Parallel Text Sourcing
- CJK NE Classification
- CJK NE Alignment

Multilingual Dataset

Entity Scheme

Cross-lingual Alignment

Various NE Classifications

Korean:

- NIKL: LC, DT, TI, OG, PS
- NAVER: PER, ORG, LOC, POH, DAT, TIM, DUR, MNY, PNT, NOH
- Kmounlp: PER, FLD, AFW, ORG, LOC, CVL, DAT, TIM, NUM, EVT, ANM, PLT, MAT, TRM
- .....

Chinese:

- MSRA: Person, Location, Organization
- People's Daily: Person, Organization, Geopolitical, Date
- CLUENER 2020: Address, Book, Company, Game, Government, Movie, Name, Organization, Position, Scene
- .....

Japanese:

- KWDLC: Artifact, Optional, Organization, Date, Location, Money, Person, Time, Percent
- ner-wikipedia: Person, Corporation, Political Organization, Other Organization, Location, Facility, Product, Event
- .....

Background

- CJK Parallel Text Sourcing
- CJK NE Classification
- CJK NE Alignment

Multilingual Dataset

Entity Scheme

Cross-lingual Alignment

subset	language	duration	size	entities
train	DE	224.5 h	86,410	97,492
	ES	141.5 h	47,611	66,482
	FR	186h	65,952	80,255
	NL	38.5 h	16,533	19,566
dev	DE	4h	1,610	1,880
	ES	4h48	1,529	2,094
	FR	4h22	1,527	1,884
	NL	2h16	963	1,074
test	DE	5h	1,966	2,061
	ES	5h	1,512	2,198
	FR	4h30	1,656	2,004
	NL	2h30	1,120	1,272

Upsurging Concerns on Multilingual NER

Representative Datasets on European Languages:

- WikiNEuRal(2021)
- MultiNERD(2022)
- Semeval-MultiCoNER v2(2023)
- MSNER(2024)

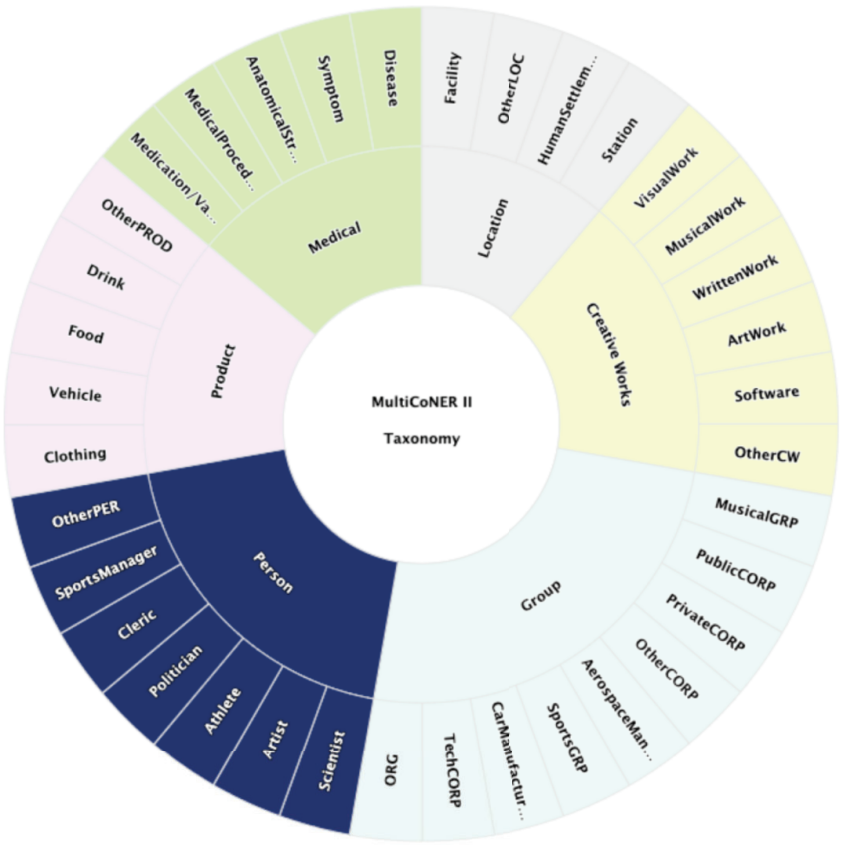
Dataset Version	Sentences	Tokens	PER	ORG	LOC	MISC	OTHER
WikiNEuRal EN	116k	2.73M	51k	31k	67k	45k	2.40M
WikiNEuRal ES	95k	2.33M	43k	17k	68k	25k	2.04M
WikiNEuRal NL	107k	1.91M	46k	22k	61k	24k	1.64M
WikiNEuRal DE	124k	2.19M	60k	32k	59k	25k	1.87M
WikiNEuRal RU	123k	2.39M	40k	26k	89k	25k	2.13M
WikiNEuRal IT	111k	2.99M	67k	22k	97k	26k	2.62M
WikiNEuRal FR	127k	3.24M	76k	25k	101k	29k	2.83M
WikiNEuRal PL	141k	2.29M	59k	34k	118k	22k	1.91M
WikiNEuRal PT	106k	2.53M	44k	17k	112k	25k	2.20M
WikiNEuRal EN DA (CoNLL)	29k	759k	12k	23k	6k	3k	0.54M
WikiNEuRal NL DA (CoNLL)	34k	598k	17k	8k	18k	6k	0.51M
WikiNEuRal DE DA (CoNLL)	41k	706k	17k	12k	23k	3k	0.61M
WikiNEuRal EN DA (OntoNotes)	48k	1.18M	20k	13k	38k	12k	1.02M

Dataset Version	Sentences	Tokens	PER	ORG	LOC	ANIM	BIO	CEL	DIS	EVE	FOO
MultiNERD	164.1K	3.6M	75.8K	33.7K	78.5K	15.5K	0.2K	2.8K	11.2K	3.2K	11.(
README	License										
ES	172.2K	4.3M	70.9K	20.6K	90.2K	10.5K	0.2K	2.4K	8.9K	6.8K	7.(
MultiNERD NL	171.7K	3.0M	56.9K	21.4K	78.7K	34.4K	0.1K	2.1K	6.1K	4.7K	5.(
MultiNERD DE	156.8K	2.7M	79.2K	31.2K	72.8K	11.5K	0.1K	1.4K	5.2K	4.0K	3.(
MultiNERD RU	129.0K	2.3M	43.4K	21.5K	75.2K	7.3K	0.1K	1.2K	1.9K	2.8K	3.(
MultiNERD IT	181.9K	4.7M	75.3K	19.3K	98.5K	8.8K	0.1K	5.2K	6.5K	5.8K	5.(
MultiNERD FR	176.2K	4.5M	89.6K	28.2K	90.9K	11.4K	0.1K	2.3K	3.1K	7.4K	3.(
MultiNERD PL	195.0K	3.0M	66.5K	29.2K	100.0K	19.7K	0.1K	3.3K	6.5K	6.7K	3.(
MultiNERD PT	177.6K	3.9M	54.0K	13.2K	124.8K	14.7K	0.1K	4.2K	6.8K	5.9K	5.(
MultiNERD ZH	195.3K	5.8M	68.3K	20.8K	49.6K	26.1K	0.4K	0.8K	0.1K	5.1K	1.(

Dataset Version	Sentences	Tokens	PER	ORG	LOC	MISC	OTHER
WikiNEuRal EN	116k	2.73M	51k	31k	67k	45k	2.40M
WikiNEuRal ES	95k	2.33M	43k	17k	68k	25k	2.04M
WikiNEuRal NL	107k	1.91M	46k	22k	61k	24k	1.64M
WikiNEuRal DE	124k	2.19M	60k	32k	59k	25k	1.87M
WikiNEuRal RU	123k	2.39M	40k	26k	89k	25k	2.13M
WikiNEuRal IT	111k	2.99M	67k	22k	97k	26k	2.62M
WikiNEuRal FR	127k	3.24M	76k	25k	101k	29k	2.83M
WikiNEuRal PL	141k	2.29M	59k	34k	118k	22k	1.91M
WikiNEuRal PT	106k	2.53M	44k	17k	112k	25k	2.20M
WikiNEuRal EN DA (CoNLL)	29k	759k	12k	23k	6k	3k	0.54M
WikiNEuRal NL DA (CoNLL)	34k	598k	17k	8k	18k	6k	0.51M
WikiNEuRal DE DA (CoNLL)	41k	706k	17k	12k	23k	3k	0.61M
WikiNEuRal EN DA (OntoNotes)	48k	1.18M	20k	13k	38k	12k	1.02M

README License

Dataset Version	Sentences	Tokens	PER	ORG	LOC	ANIM	BIO	CEL	DIS	EVE	FOC
MultiNERD	164.1K	3.6M	75.8K	33.7K	78.5K	15.5K	0.2K	2.8K	11.2K	3.2K	11.
ES	173.2K	4.3M	70.9K	20.6K	90.2K	10.5K	0.3K	2.4K	8.6K	6.8K	7.
MultiNERD NL	171.7K	3.0M	56.9K	21.4K	78.7K	34.4K	0.1K	2.1K	6.1K	4.7K	5.
MultiNERD DE	156.8K	2.7M	79.2K	31.2K	72.8K	11.5K	0.1K	1.4K	5.2K	4.0K	3.
MultiNERD RU	129.0K	2.3M	43.4K	21.5K	75.2K	7.3K	0.1K	1.2K	1.9K	2.8K	3.
MultiNERD IT	181.9K	4.7M	75.3K	19.3K	98.5K	8.8K	0.1K	5.2K	6.5K	5.8K	5.
MultiNERD FR	176.2K	4.3M	89.6K	28.2K	90.9K	11.4K	0.1K	2.3K	3.1K	7.4K	3.
MultiNERD PL	195.0K	3.0M	66.5K	29.2K	100.0K	19.7K	0.1K	3.3K	6.5K	6.7K	3.
MultiNERD PT	177.6K	3.9M	54.0K	13.2K	124.8K	14.7K	0.1K	4.2K	6.8K	5.9K	5.
MultiNERD ZH	195.3K	5.8M	68.3K	20.8K	49.6K	26.1K	0.4K	0.8K	0.1K	5.1K	1.



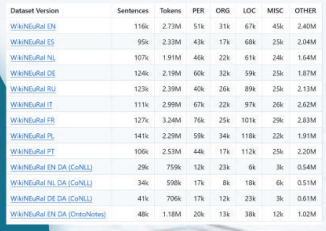
subset	language	duration	size	entities
train	DE	224.5 h	86,410	97,492
	ES	141.5 h	47,611	66,482
	FR	186h	65,952	80,255
	NL	38.5 h	16,533	19,566
dev	DE	4h	1,610	1,880
	ES	4h48	1,529	2,094
	FR	4h22	1,527	1,884
	NL	2h16	963	1,074
test	DE	5h	1,966	2,061
	ES	5h	1,512	2,198
	FR	4h30	1,656	2,004
	NL	2h30	1,120	1,272

subset	language	duration	size	entities
train	DE	224.5 h	86,410	97,492
	ES	141.5 h	47,611	66,482
	FR	186h	65,952	80,255
	NL	38.5 h	16,533	19,566
dev	DE	4h	1,610	1,880
	ES	4h48	1,529	2,094
	FR	4h22	1,527	1,884
	NL	2h16	963	1,074
test	DE	5h	1,966	2,061
	ES	5h	1,512	2,198
	FR	4h30	1,656	2,004
	NL	2h30	1,120	1,272

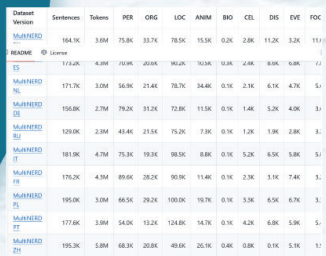
Upsurging Concerns on Multilingual NER

Representative Datasets on European Languages:

- WikiNEuRal(2021)
- MultiNERD(2022)
- Semeval-MultiCoNER v2(2023)
- MSNER(2024)



Dataset Version	Sentences	Tokens	PER	ORG	LOC	MISC	OTHER
WikiNEuRal EN	116k	2.73M	51k	31k	67k	45k	2.40M
WikiNEuRal ES	95k	2.33M	43k	17k	68k	25k	2.04M
WikiNEuRal NL	107k	1.91M	46k	22k	61k	24k	1.64M
WikiNEuRal DE	124k	2.19M	60k	32k	59k	25k	1.87M
WikiNEuRal RU	123k	2.39M	40k	26k	89k	25k	2.13M
WikiNEuRal IT	111k	2.99M	67k	22k	97k	26k	2.62M
WikiNEuRal FR	127k	3.24M	76k	25k	101k	29k	2.83M
WikiNEuRal PL	141k	2.29M	59k	34k	118k	22k	1.91M
WikiNEuRal PT	106k	2.53M	44k	17k	112k	25k	2.20M
WikiNEuRal EN DA (CoNLL)	29k	759k	12k	23k	6k	3k	0.54M
WikiNEuRal NL DA (CoNLL)	34k	598k	17k	8k	18k	6k	0.51M
WikiNEuRal DE DA (CoNLL)	41k	706k	17k	12k	23k	3k	0.61M
WikiNEuRal EN DA (OntoNotes)	48k	1.18M	20k	13k	38k	12k	1.02M



Dataset Version	Sentences	Tokens	PER	ORG	LOC	ANIM	BIO	CEL	DIS	EVE	FOO
MultiNERD	164.1K	3.6M	75.8K	33.7K	78.5K	15.5K	0.2K	2.8K	11.2K	3.2K	11.1
README	License										
ES	171.2K	4.3M	70.9K	20.6K	90.2K	10.3K	0.3K	2.4K	8.6K	6.8K	7.7
MultiNERD NL	171.7K	3.0M	56.9K	21.4K	78.7K	34.4K	0.1K	2.1K	6.1K	4.7K	5.4
MultiNERD DE	156.8K	2.7M	79.2K	31.2K	72.8K	11.5K	0.1K	1.4K	5.2K	4.0K	3.4
MultiNERD RU	129.0K	2.8M	43.4K	21.5K	75.2K	7.3K	0.1K	1.2K	1.9K	2.8K	3.1
MultiNERD IT	181.9K	4.7M	75.3K	19.3K	98.5K	8.8K	0.1K	5.2K	6.5K	5.8K	5.8
MultiNERD FR	176.2K	4.3M	89.6K	28.2K	90.9K	11.4K	0.1K	2.3K	3.1K	7.4K	3.1
MultiNERD PL	195.0K	3.0M	66.5K	29.2K	100.0K	19.7K	0.1K	3.3K	6.5K	6.7K	3.1
MultiNERD PT	177.6K	3.9M	54.0K	13.2K	124.8K	14.7K	0.1K	4.2K	6.8K	5.9K	5.4
MultiNERD ZH	195.3K	5.8M	68.3K	20.8K	49.6K	26.1K	0.4K	0.8K	0.1K	5.1K	1.1

Background

- CJK Parallel Text Sourcing
- CJK NE Classification
- CJK NE Alignment

Multilingual Dataset

Entity Scheme

Cross-lingual Alignment

Asymmetric Alignments of Mulitlingual Named Entities: Focus on the CJK Parallel Corpus

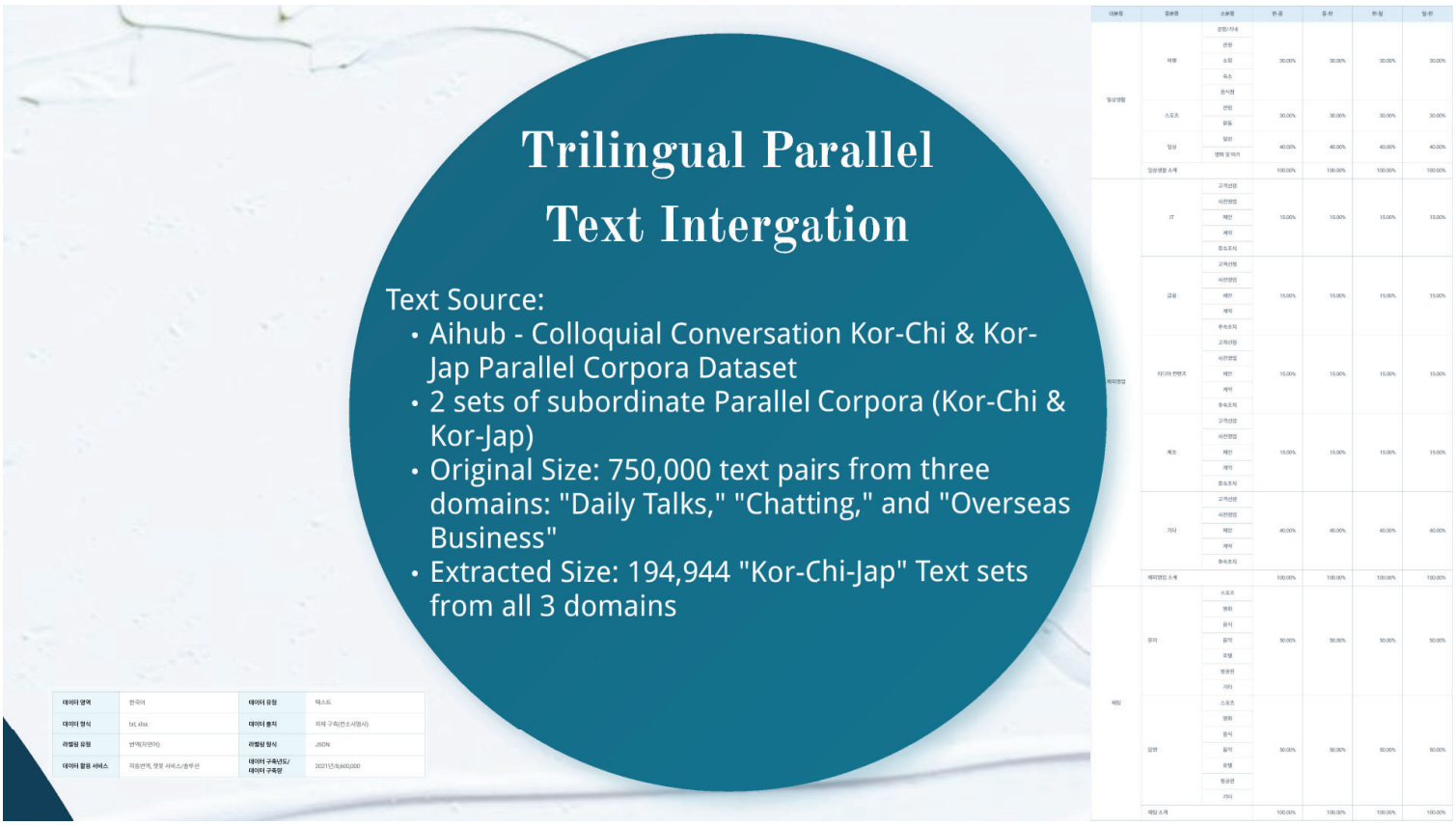
Dylan Fei @ Sukmyung Univ.
KACL 2024 Fall

Background
Analysis
Introduction
Experiment
Conclusion

Methods & Steps

- CJK Parallel Corpus Construction
- Statistical Test on Frequencies
- Qualitative Comparison on NE linking Relations

Data Source
Preprocessing
Entity Distribution



데이터 영역	한국어	**데이터 유형**	텍스트
데이터 형식	txt, xlsx	**데이터 출처**	자체 구축(컨소시엄사)
라벨링 유형	번역(자연어)	**라벨링 형식**	JSON
데이터 활용 서비스	자동번역, 챗봇 서비스/솔루션	**데이터 구축년도/ 데이터 구축량**	2021년/8,600,000

Trilingual Parallel Text Intergation

Text Source:

- Aihub - Colloquial Conversation Kor-Chi & Kor-Jap Parallel Corpora Dataset
- 2 sets of subordinate Parallel Corpora (Kor-Chi & Kor-Jap)
- Original Size: 750,000 text pairs from three domains: "Daily Talks," "Chatting," and "Overseas Business"
- Extracted Size: 194,944 "Kor-Chi-Jap" Text sets from all 3 domains

Methods & Steps

- CJK Parallel Corpus Construction
- Statistical Test on Frequencies
- Qualitative Comparison on NE linking Relations

Data Source

Preprocessing

Entity Distribution

Auto-NER Labeling on CJK Text Sets

Auto-NER Modules & Models:

- Chinese:
  - Spacy - "zh_core_web_trf"
  - F1: 0.76
- Japanese:
  - Spacy - "ja_core_news_trf"
  - F1: 0.83
- Korean:
  - Transformers - "KoELECTRA-small-v3-modu-ner"
  - F1: 0.83

Filtered Size after Auto-NER:

- 51.54% (100,466/194,944)

Methods & Steps

- CJK Parallel Corpus Construction
- Statistical Test on Frequencies
- Qualitative Comparison on NE linking Relations

Data Source

Preprocessing

Entity Distribution

NE Tagset Comparison

Chinese (OntoNotes 5.0)	Japanese (OntoNotes-Jap)	Korean (ETRI-TTA)	Examples
CARDINAL	CARDINAL	QT	Numerals that do not fall under another type.
DATE	DATE	DT	Absolute or relative dates or periods.
EVENT	EVENT	EV	Named hurricanes, battles, wars, sports events, etc.
FAC	FAC	LC	Buildings, airports, highways, bridges, etc.
GPE	GPE	LC	Countries, cities, states.
LANGUAGE	LANGUAGE	CV	Any named language.
LAW	LAW	TM	Named documents made into laws.
LOC	LOC	LC	Non-GPE locations, mountain ranges, bodies of water.
MONEY	MONEY	CV	Monetary values, including unit.
	MOVEMENT		
NORP	NORP	CV	Nationalities or religious or political groups.
ORDINAL	ORDINAL	QT	"first", "second", etc.
ORG	ORG	OG	Companies, agencies, institutions, etc.
PERCENT	PERCENT	QT	Percentage, including "%".
PERSON	PERSON	PS	People, including fictional.
	PET_NAME		
	PHONE	QT	
PRODUCT	PRODUCT	AF	Objects, vehicles, foods, etc. (Not services.)
QUANTITY	QUANTITY	QT	Measurements, as of weight or distance.
TIME	TIME	TI	Times smaller than a day.
	TITLE_AFFIX		
WORK_OF_ART	WORK_OF_ART	AF	Titles of books, songs, etc.

분류	표기	정의
ARTIFACTS	AF	사람에 의해 창조된 인공물로 문화재, 건물, 악기, 도로, 무기, 운송수단, 작품명, 공산품명이 모두 이에 해당
ANIMAL	AM	사람을 제외한 짐승
CIVILIZATION	CV	문명/문화
DATE	DT	기간 및 계절, 시기/시대
EVENT	EV	특정 사건/사고/행사 명칭
STUDY_FIELD	FD	학문 분야, 학파 및 유파
LOCATION	LC	지역/장소와 지형/지리 명칭 등을 모두 포함
MATERIAL	MT	원소 및 금속, 암석/보석, 화학물질
ORGANIZATION	OG	기관 및 단체 명칭
PERSON	PS	인명 및 인물의 별칭 (유사 인물 명칭 포함)
PLANT	PT	꽃/나무, 육지식물, 해초류, 버섯류, 이끼류
QUANTITY	QT	수량/분량, 순서/순차, 수사로 이루어진 표현
TIME	TI	시계상으로 나타나는 시/시각, 시간 범위
TERM	TM	타 개체명에서 정의된 세부 개체명 이외의 개체명
THEORY	TR	특정 이론, 법칙 원리 등

Methods & Steps

- CJK Parallel Corpus Construction
- Statistical Test on Frequencies
- Qualitative Comparison on NE linking Relations

Data Source

Preprocessing

Entity Distribution

Asymmetric Alignments of Mulitlingual Named Entities: Focus on the CJK Parallel Corpus

Dylan Fei @ Sukmyung Univ.
KACL 2024 Fall

Background

Analysis

Introduction

Experiment

Conclusion

Analysis

Language	Label	tokens	types
Chinese	DT	18526	2738
	EV	305	191
	LC	1430	501
	OG	3777	2034
	PS	2140	1318
	TI	6073	1253
Sum		32251	8035
Japanese	DT	10829	1709
	EV	577	254
	LC	1109	313
	OG	2915	1798
	PS	2709	1511
	TI	4732	616
Sum		22871	6201
Korean	DT	21365	3545
	EV	831	545
	LC	9525	1390
	OG	2754	1480
	PS	2851	1617
	TI	7478	1629
Sum		44804	10206

- Frequency-based Quantitative Anaylsis
- Concordance-based Qualitative Anaylsis

Quantitative

Qualitative

Statistical Magnitudes

- TTR (Type-Token Ratio)
- K-W Test (Kruskal–Wallis test)

TTR

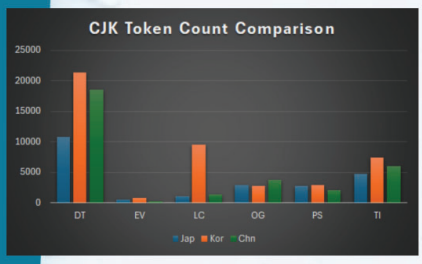
K-W Test

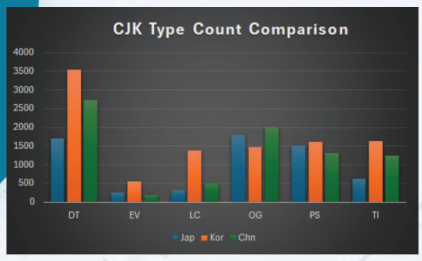
TTR Calculation & Comparison

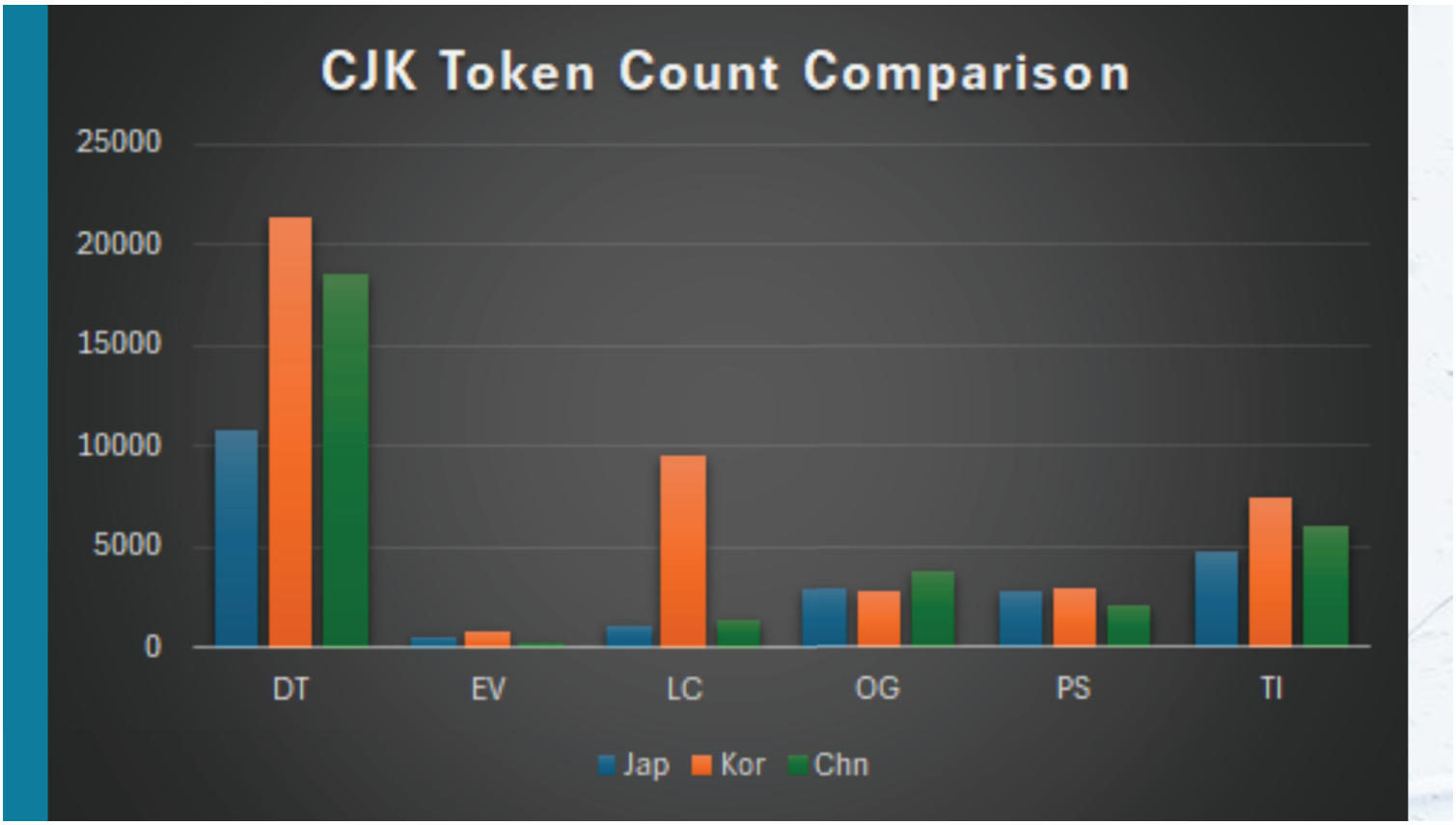
TTR	Kor	Jap	Chn(±0.09)
DT	0.166	0.158	**0.134~0.161**
EV	0.656	**0.440**	0.569~0.682
LC	0.146	**0.282**	**0.318~0.381**
OG	0.537	**0.617**	0.490~0.586
PS	0.567	0.558	0.560~0.671
TI	0.218	**0.130**	0.187~0.224

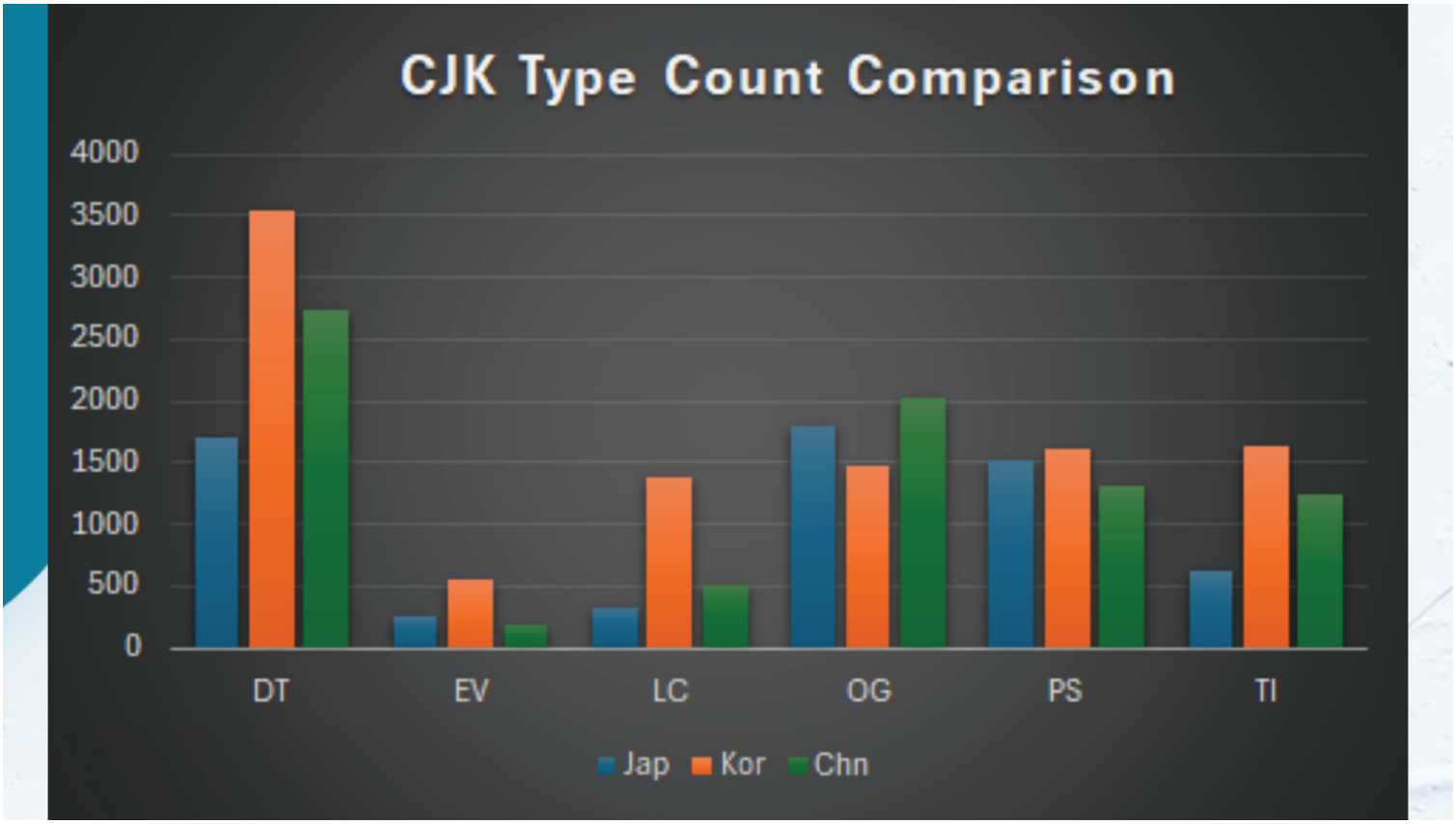
TTR Formula:
- (Type/Token)*(100% ± |F1$_{kor}$ - F1$_{jap\ or\ Chn}$|)

TTR Comparison:
- Token Count:
  - LC, DT, TI
- Type Count:
  - LC, DT
- TTR:
  - Kor-Jap: EV, LC, TI, OG
  - Kor-Chn: DT, LC
  - Chn-Jap: TI, EV, OG






CJK Token Count Comparison
25000
20000
15000
10000
5000
0
DT
EV
LC
OG
PS
TI
Jap
Kor
Chn


CJK Type Count Comparison
4000
3500
3000
2500
2000
1500
1000
500
0
DT
EV
LC
OG
PS
TI
Jap
Kor
Chn

TTR	Kor	Jap	Chn(±0.09)
DT	0.166	0.158	**0.134~0.161**
EV	0.656	**0.440**	0.569~0.682
LC	0.146	**0.282**	**0.318~0.381**
OG	0.537	**0.617**	0.490~0.586
PS	0.567	0.558	0.560~0.671
TI	0.218	**0.130**	0.187~0.224

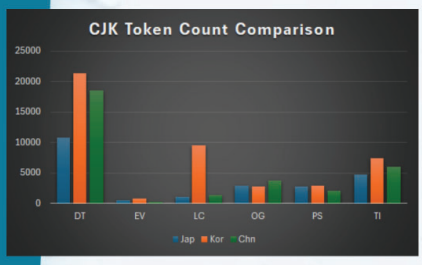
TTR Calculation & Comparison

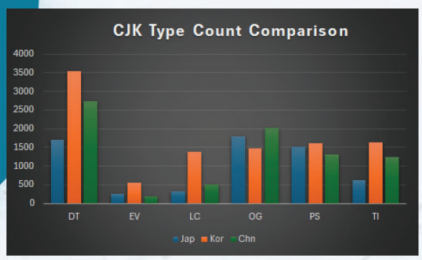
TTR Formula:

- (Type/Token)*(100% ± |F1_kor - F1_jap or Chn|)

TTR Comparison:

- Token Count:
  - LC, DT, TI
- Type Count:
  - LC, DT
- TTR:
  - Kor-Jap: EV, LC, TI, OG
  - Kor-Chn: DT, LC
  - Chn-Jap: TI, EV, OG





Statistical Magnitudes

- TTR (Type-Token Ratio)
- K-W Test (Kruskal–Wallis test)

TTR

K-W Test

Nonparametric Test Results

Steps:

- Pick up 5 main NE types
- Random Sampling (100 samples out of each types by languages)
- Normality tests for 5 NE types(retained)
- Multigroup comparison through K-W Tests
- Further pairwise comparisons across CJK groups

Results:

- Chn-Jap: None
- Kor-Chn: OG
- Kor-Jap: DT, TI

		df	Kolmogorov-Smirnov		Shapiro-Wilk	
			Statistic	Sig.	Statistic	Sig.
DT	Chinese	100	0.421	0.000	0.249	0.000
	Japanese	100	0.396	0.000	0.251	0.000
	Korean	100	0.442	0.000	0.119	0.000
EV	Chinese	100	0.441	0.000	0.120	0.000
	Japanese	100	0.446	0.000	0.112	0.000
	Korean	100	0.420	0.000	0.189	0.000
LC	Chinese	100	0.425	0.000	0.173	0.000
	Japanese	100	0.441	0.000	0.123	0.000
	Korean	100	0.441	0.000	0.121	0.000
OG	Chinese	100	0.480	0.000	0.349	0.000
	Japanese	100	0.433	0.000	0.376	0.000
	Korean	100	0.391	0.000	0.447	0.000
PS	Chinese	100	0.452	0.000	0.487	0.000
	Japanese	100	0.460	0.000	0.381	0.000
	Korean	100	0.434	0.000	0.133	0.000
TI	Chinese	100	0.421	0.000	0.249	0.000
	Japanese	100	0.396	0.000	0.251	0.000
	Korean	100	0.442	0.000	0.119	0.000

Independent-Samples Kruskal-Wallis Test	DT	EV	LC	OG	PS	TI
Total N	300	300	300	300	300	300
Test Statistic	**15.721**	3.109	0.016	**7.562**	4.902	**8.061**
Degree Of Freedom	2	2	2	2	2	2
Asymptotic Sig.(2-sided test)	0.000	0.211	0.992	0.023	0.086	0.018

Sample 1-Sample 2	NE type	Test Statistic	Std. Error	Std. Test Statistic	Adj. Sig.
Chinese-Japanese	DT	-22.490	10.678	-2.106	0.106
	OG	-11.450	8.568	-1.336	0.544
	TI	-19.500	10.566	-1.846	0.195
Korean-Chinese	DT	19.820	10.678	1.856	0.190
	OG	-23.560	8.568	-2.750	0.018
	TI	9.990	10.566	0.946	1.000
Korean-Japanese	DT	42.310	10.678	3.962	0.000
	OG	-12.110	8.568	-1.413	0.473
	TI	29.490	10.566	2.791	0.016

		df	Kolmogorov-Smirnov		Shapiro-Wilk	
			Statistic	Sig.	Statistic	Sig.
DT	Chinese	100	0.421	0.000	0.249	0.000
	Japanese	100	0.396	0.000	0.251	0.000
	Korean	100	0.442	0.000	0.119	0.000
EV	Chinese	100	0.441	0.000	0.120	0.000
	Japanese	100	0.446	0.000	0.112	0.000
	Korean	100	0.420	0.000	0.189	0.000
LC	Chinese	100	0.425	0.000	0.173	0.000
	Japanese	100	0.441	0.000	0.123	0.000
	Korean	100	0.441	0.000	0.121	0.000
OG	Chinese	100	0.480	0.000	0.349	0.000
	Japanese	100	0.433	0.000	0.376	0.000
	Korean	100	0.391	0.000	0.447	0.000
PS	Chinese	100	0.452	0.000	0.487	0.000
	Japanese	100	0.460	0.000	0.381	0.000
	Korean	100	0.434	0.000	0.133	0.000
TI	Chinese	100	0.421	0.000	0.249	0.000
	Japanese	100	0.396	0.000	0.251	0.000
	Korean	100	0.442	0.000	0.119	0.000

Step
- P
- R
  c
- N
  t
- N
  k
- F
  a

Independent-Samples Kruskal-Wallis Test	DT	EV	LC	OG	PS	TI
Total N	300	300	300	300	300	300
Test Statistic	**15.721**	3.109	0.016	**7.562**	4.902	**8.061**
Degree Of Freedom	2	2	2	2	2	2
Asymptotic Sig.(2-sided test)	0.000	0.211	0.992	0.023	0.086	0.018

Sample 1-Sample 2	NE type	Test Statistic	Std. Error	Std. Test Statistic	Adj. Sig.
Chinese-Japanese	DT	-22.490	10.678	-2.106	0.106
	OG	-11.450	8.568	-1.336	0.544
	TI	-19.500	10.566	-1.846	0.195
Korean-Chinese	DT	19.820	10.678	1.856	0.190
	OG	-23.560	8.568	-2.750	0.018
	TI	9.990	10.566	0.946	1.000
Korean-Japanese	DT	42.310	10.678	3.962	0.000
	OG	-12.110	8.568	-1.413	0.473
	TI	29.490	10.566	2.791	0.016

Nonparametric Test Results

Steps:

- Pick up 5 main NE types
- Random Sampling (100 samples out of each types by languages)
- Normality tests for 5 NE types(retained)
- Multigroup comparison through K-W Tests
- Further pairwise comparisons across CJK groups

Results:

- Chn-Jap: None
- Kor-Chn: OG
- Kor-Jap: DT, TI

		df	Kolmogorov-Smirnov		Shapiro-Wilk	
			Statistic	Sig.	Statistic	Sig.
DT	Chinese	100	0.421	0.000	0.249	0.000
	Japanese	100	0.396	0.000	0.251	0.000
	Korean	100	0.442	0.000	0.119	0.000
EV	Chinese	100	0.441	0.000	0.120	0.000
	Japanese	100	0.446	0.000	0.112	0.000
	Korean	100	0.420	0.000	0.189	0.000
LC	Chinese	100	0.425	0.000	0.173	0.000
	Japanese	100	0.441	0.000	0.123	0.000
	Korean	100	0.441	0.000	0.121	0.000
OG	Chinese	100	0.480	0.000	0.349	0.000
	Japanese	100	0.433	0.000	0.376	0.000
	Korean	100	0.391	0.000	0.447	0.000
PS	Chinese	100	0.452	0.000	0.487	0.000
	Japanese	100	0.460	0.000	0.381	0.000
	Korean	100	0.434	0.000	0.133	0.000
TI	Chinese	100	0.421	0.000	0.249	0.000
	Japanese	100	0.396	0.000	0.251	0.000
	Korean	100	0.442	0.000	0.119	0.000

Independent-Samples Kruskal-Wallis Test	DT	EV	LC	OG	PS	TI
Total N	300	300	300	300	300	300
Test Statistic	**15.721**	3.109	0.016	**7.562**	4.902	**8.061**
Degree Of Freedom	2	2	2	2	2	2
Asymptotic Sig.(2-sided test)	0.000	0.211	0.992	0.023	0.086	0.018

Sample 1-Sample 2	NE type	Test Statistic	Std. Error	Std. Test Statistic	Adj. Sig.
Chinese-Japanese	DT	-22.490	10.678	-2.106	0.106
	OG	-11.450	8.568	-1.336	0.544
	TI	-19.500	10.566	-1.846	0.195
Korean-Chinese	DT	19.820	10.678	1.856	0.190
	OG	-23.560	8.568	-2.750	0.018
	TI	9.990	10.566	0.946	1.000
Korean-Japanese	DT	42.310	10.678	3.962	0.000
	OG	-12.110	8.568	-1.413	0.473
	TI	29.490	10.566	2.791	0.016

Statistical Magnitudes

- TTR (Type-Token Ratio)
- K-W Test (Kruskal–Wallis test)

TTR

K-W Test

Analysis

Language	Label	tokens	types
Chinese	DT	18526	2738
	EV	305	191
	LC	1430	501
	OG	3777	2034
	PS	2140	1318
	TI	6073	1253
Sum		32251	8035
Japanese	DT	10829	1709
	EV	577	254
	LC	1109	313
	OG	2915	1798
	PS	2709	1511
	TI	4732	616
Sum		22871	6201
Korean	DT	21365	3545
	EV	831	545
	LC	9525	1390
	OG	2754	1480
	PS	2851	1617
	TI	7478	1629
Sum		44804	10206

- Frequency-based Quantitative Anaylsis
- Concordance-based Qualitative Anaylsis

Quantitative

Qualitative

CJK Entites of Various Metions

- Lexical Variation
- Semantic Variation
- Pragmatic Variation
- Intercultural Variation

→

- DT: 금일, 공휴일
- TI: 오전, 오후
- PS: 신, 부처님
- EV: 한국 전쟁, 수능
- OG: 현대, 식약처
- LC: 미국, 남북한

DT

TI

PS

EV

OG

LC

Date Entity Examples

금일:

금일 오후 일정이 명일 오전으로 변경되었는데요.

本日午後の日程が明日の午前に変更されましたが。

今天下午的日程变为明天上午了。

공휴일:

RNN LSTM 알고리즘을 사용하여 **공휴일** 택배 접수량을 예측합니다.

RNN LSTMアルゴリズムを使用して、**祝日**の宅配受付量を予測します。

使用RNN LSTM算法预测**假日**包裹接收数量。

Chinese		Japanese		Korean	
count	entity	count	entity	count	entity
2890	今天	341	金曜日	3657	오늘
1551	明天	331	一週間	1654	내일
752	昨天	329	一日	801	어제
339	一天	284	月曜日	449	하루
333	下周	274	週末	356	다음 주
298	一个月	272	水曜日	282	다음주
289	一周	271	明日	258	여름
282	下个月	268	土曜日	239	일주일
233	上周	219	日曜日	227	주말
180	上个月	215	1年	194	수요일
175	这周	204	2週間	184	월요일
171	周末	183	1週間	177	금요일
164	3天	176	火曜日	166	토요일
157	明年	167	一ヶ月	165	일요일
135	夏天	167	木曜日	161	그날
134	今年	144	3日	153	이번 주
130	去年	138	来週	152	지난주
121	下周一	114	1ヶ月	150	화요일
114	3个月	99	3ヶ月	147	목요일
113	这个月	91	末	146	가을

CJK Entites of Various Metions

- Lexical Variation
- Semantic Variation
- Pragmatic Variation
- Intercultural Variation

→

- DT: 금일, 공휴일
- TI: 오전, 오후
- PS: 신, 부처님
- EV: 한국 전쟁, 수능
- OG: 현대, 식약처
- LC: 미국, 남북한

DT

TI

PS

EV

OG

LC

Time Entity Examples

오전:

가장 빠른 비행기는 **오전** 다섯 시 입니다.

一番早い飛行機は**午前**五時です。

最快的飞机是**早上**五点。

오후:

저희 매장 오늘 **오후** 10시까지 영업합니다.

私たちの競技場今日の**午後**10時まで営業します。

本店营业至今天**晚上**10点。

Chinese		Japanese		Korean	
count	entity	count	entity	count	entity
278	晚上	261	30分	679	저녁
269	30分钟	247	1時間	377	아침
211	10分钟	208	10分	293	오후
176	一个小时	165	5分	264	점심
149	凌晨	127	3時	261	밤
148	5分钟	120	2時	233	오전
138	下午	117	2時間	208	새벽
130	早上	116	7時	87	3시
104	20分钟	102	午後3時	77	낮
99	1小时	100	午後2時	72	7시
92	上午	93	20分	71	오후 3시
88	1个小时	89	8時	69	오후 2시
79	今晚	85	6時	68	##시
73	两个小时	82	9時	65	30분 정도
70	昨晚	81	10時	62	2시
61	7点	75	5時	62	1시간
59	今天下午	72	一時間	61	30분
59	9点	71	4時	52	9시
58	15分钟	63	30分後	51	8시
51	明天上午	61	15分	49	10분

CJK Entites of Various Metions

- Lexical Variation
- Semantic Variation
- Pragmatic Variation
- Intercultural Variation

→

- DT: 금일, 공휴일
- TI: 오전, 오후
- PS: 신, 부처님
- EV: 한국 전쟁, 수능
- OG: 현대, 식약처
- LC: 미국, 남북한

DT

TI

PS

EV

OG

LC

Person Entity Examples

신:

나의 존재만으로도 **신**께 감사해야 해요.

私の存在だけでも**神様**に感謝しなければなりません。

只要我的存在本身,就要感谢**上帝**。

부처님:

그 후 제자들도 **부처님**을 따라 네 개의 발우를 써서 공양하는 전통이 생겨났다고 합니다.

その後、弟子たちもお**釈迦**さまを見習い、四つの鉢盂を使って供養する伝統が生まれたそうです。

此后弟子们也跟着**佛祖**用四个钵盂供养的传统诞生了。

Chinese		Japanese		Korean	
count	entity	count	entity	count	entity
131	A	65	J	62	J
41	上帝	51	A	48	M
28	黄政民	50	キム	47	이
23	金	48	ダーリン	46	김
20	赵寅成	44	コロナ	38	H
19	麦克风	31	ファン・ジョンミン	37	##구
14	辛奇	25	ブラック・ウィドウ	35	황정민
12	普拉提	23	チョ・イソン	33	아이언맨
11	米其林	20	お水	30	바이킹
11	AAA	17	ハニー	26	성합
11	高尔夫	16	ワンちゃん	25	탕쭝
11	J	15	シャネル	25	하나님
10	陈列	15	チキン	25	S
10	OO	14	パパ	22	K
9	M	13	ジム	20	크루엘라
9	刘海	13	シノブシス	19	조인성
8	李	13	イエス	14	오
8	马卡龙	13	K	14	예수
8	约翰	12	M	14	하느님
8	B	11	イ	13	이상형

CJK Entites of Various Metions

- Lexical Variation
- Semantic Variation
- Pragmatic Variation
- Intercultural Variation

→

- DT: 금일, 공휴일
- TI: 오전, 오후
- PS: 신, 부처님
- EV: 한국 전쟁, 수능
- OG: 현대, 식약처
- LC: 미국, 남북한

DT

TI

PS

EV

OG

LC

Event Entity Examples

Chinese		Japanese		Korean	
count	entity	count	entity	count	entity
66	奥运会	81	映画祭	75	올림픽
5	世界杯	52	コロナウイルス	25	PO
5	第二次世界大战	51	オリンピック	17	수능
4	韩国战争	30	コロナ禍	12	결승전
4	毕业典礼	24	新型コロナウイルス感染症	9	월드컵
3	日本帝国主义	14	新型コロナウイルス	8	기말고사
3	平昌冬奥会	7	ワールドカップ	7	코로나 사태
3	第一次世界大战	7	アイスアメリカーノ2杯	5	준결승전
3	冰雪奇缘	6	コロナ19	5	LC
3	韩国国内电影节	5	公務員試験	4	플레이오프
2	东京奥运会	5	アジア大会	4	준결승
2	奥运	5	韓国戦争	4	한일전
2	光州	5	大統領選挙	4	##경기
2	戛纳电影节	4	朝鮮戦争	4	아시안게임
2	足球赛	4	アイスアメリカーノ二杯	4	축제
2	巴黎奥运会	4	五輪	4	##A
2	排球	3	お盆	4	중간고사
2	美国职业棒球大联盟	3	チャンピオンズリーグ	4	박창동계올림픽
2	夏令营	3	第1次世界大戦	4	FTA

한국전쟁:

한국전쟁이라고 북한과 남한이 치뤘던 6.25전쟁 등을 기리는 공간이야.

朝鮮戦争といって、北朝鮮と韓国が行った朝鮮戦争などを追悼する空間だ。

叫做**韩国战争**，是纪念北韩和南韩经历的6.25战争的空间。

수능:

우리나라는 **수능**이라는 중요한 시험을 치르고 대학에 입학해요.

我が国は、**修学能力試験**という大事な試験を受けて大学に入学します。

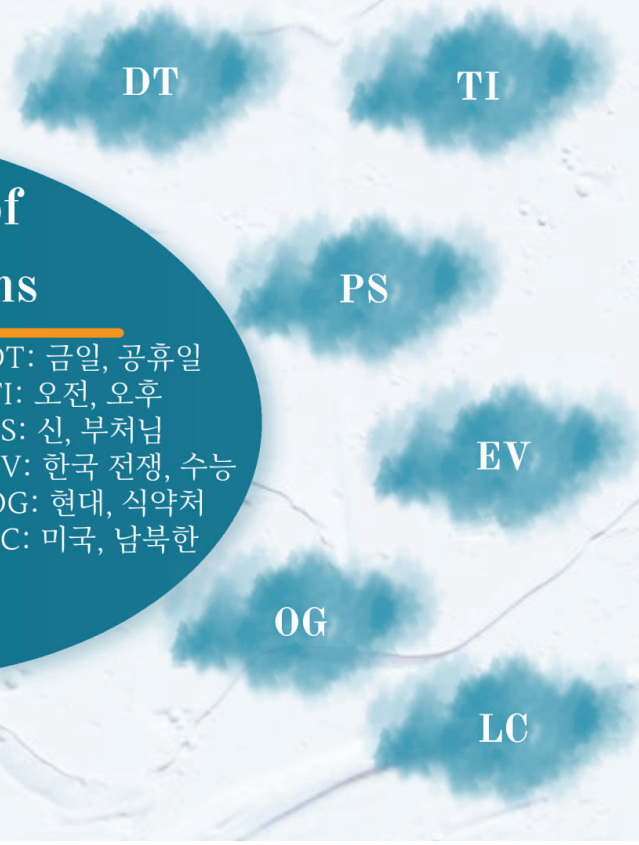
在韩国，学生通过参加一项名为**CSAT**的重要考试进入大学。

CJK Entites of Various Metions

- Lexical Variation
- Semantic Variation
- Pragmatic Variation
- Intercultural Variation

→

- DT: 금일, 공휴일
- TI: 오전, 오후
- PS: 신, 부처님
- EV: 한국 전쟁, 수능
- OG: 현대, 식약처
- LC: 미국, 남북한



Organization Entity Examples

현대:

나는 **현대 중공업**에 1998년도에 입사했습니다.

私は**現代重工業**に1998年に入社しました。

我于1998年加入**现代公司**的**重工业**。

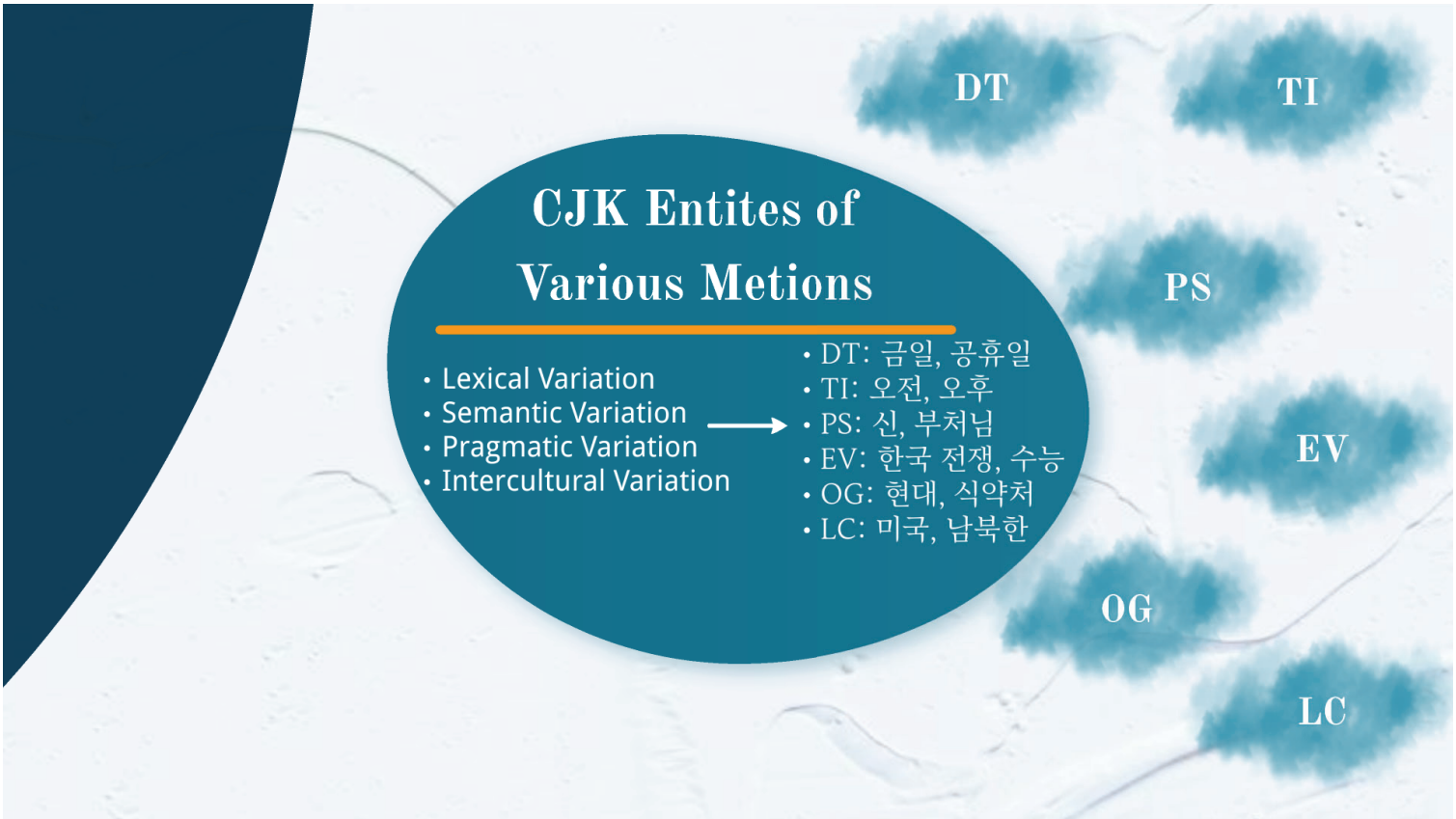
식약처:

국내 **식약처** 및 FDA 승인도 받은 상태입니다.

国内の**食品医薬品安全処**およびFDAからも承認を受けた状態です。

已获得**国内食品医药品安全厅**及FDA的批准。

Chinese		Japanese		Korean	
count	entity	count	entity	count	entity
343	A	101	AAA	96	정부
97	AAA	73	A社	61	경찰
90	A公司	36	A	54	군대
59	B	28	御	46	마블
49	APP	21	BTS	37	AAA
33	YouTube	20	マーベル	33	G
27	M	16	大韓航空	29	인보이스
21	C	15	アマゾン	27	나이키
21	S	15	ユーチューブ	27	군
20	Instagram	13	D社	22	우체국
17	SNS	13	A大学	22	아마존
17	H	12	弘大	22	법원
16	A和B	12	コーラ	18	샤넬
16	B公司	12	G	18	성함
15	PPL	11	アイスアメリカーノ	17	맥도날드
14	J	11	ペイパル	17	스타벅스
13	KakaoTalk	11	GOC	17	디즈니
13	Netflix	10	ランボルギーニ	14	나사
13	SUV	10	H	13	세관
12	通信公司	10	貴社	13	AA



Locatioin Entity Examples

대한민국:

저는 **대한민국**에 사는 사진작가 지망생 20살 AAA입니다.

私は**韓国**に住んでいる写真作家志望20歳のAAAです。

我是生活在**韩国**的摄影师志愿生20岁的AAA。

남북한:

남북한 민족의 정을 느낄 수 있는 영화?

韓国と北朝鮮の民族の情を感じる映画ですか。

能感受到**南北**民族感情的电影?

Chinese		Japanese		Korean	
count	entity	count	entity	count	entity
222	济州岛	235	済州島	1697	한국
140	欧洲	142	ヨーロッパ	504	미국
50	大海	34	漢江	342	중국
31	南山塔	27	東南アジア	285	일본
28	汉江	26	アフリカ	270	서울
26	东南亚	20	江華島	258	필리핀
24	非洲	20	北欧	250	제주도
19	江原道	20	コロナ	175	부산
19	南山	18	漢拏山	146	유럽
18	江南	16	南山	120	프랑스
17	西欧	14	アジア	117	영국
17	江华岛	13	西ヨーロッパ	115	베트남
15	东欧	11	東ヨーロッパ	104	캐나다
14	亚洲	11	中東	92	독일
13	京畿道	10	鬱陵島	90	태국
12	北欧	10	西欧	83	호주
12	江陵	9	東欧	79	이탈리아
11	东海	8	東海	77	인천
11	中东	8	清渓川	76	홍콩
10	大田	8	南怡島	69	뉴욕

CJK Entites of Various Metions

- Lexical Variation
- Semantic Variation
- Pragmatic Variation
- Intercultural Variation

→

- DT: 금일, 공휴일
- TI: 오전, 오후
- PS: 신, 부처님
- EV: 한국 전쟁, 수능
- OG: 현대, 식약처
- LC: 미국, 남북한

DT

TI

PS

EV

OG

LC

Analysis

Language	Label	tokens	types
Chinese	DT	18526	2738
	EV	305	191
	LC	1430	501
	OG	3777	2034
	PS	2140	1318
	TI	6073	1253
Sum		32251	8035
Japanese	DT	10829	1709
	EV	577	254
	LC	1109	313
	OG	2915	1798
	PS	2709	1511
	TI	4732	616
Sum		22871	6201
Korean	DT	21365	3545
	EV	831	545
	LC	9525	1390
	OG	2754	1480
	PS	2851	1617
	TI	7478	1629
Sum		44804	10206

- Frequency-based Quantitative Anaylsis
- Concordance-based Qualitative Anaylsis

Quantitative

Qualitative

Asymmetric Alignments of Mulitlingual Named Entities: Focus on the CJK Parallel Corpus

Dylan Fei @ Sukmyung Univ.
KACL 2024 Fall

Background

Analysis

Introduction

Experiment

Conclusion

Conclusion

- Multilingual parallel corpora significantly facilitate mulitingual entities identification and linking.
- Asymmetrical linkages attribute to not just linguistic Features but also entity categories.
- Chinese-character-based entities interact in more complex mapping relations.

Discussion

Reference

What's next

- Univesal NER for CJK's NE Classification?
- Genearal Alignment Framework for CJK's NEs?
- Multi-NER Data Augmentation via Large Language Models?

Conclusion

- Multilingual parallel corpora significantly facilitate mulitingual entities identification and linking.
- Asymmetrical linkages attribute to not just linguistic Features but also entity categories.
- Chinese-character-based entities interact in more complex mapping relations.

Discussion

Reference

Alshammari, N., & Alanazi, S. (2021). The impact of using different annotation schemes on named entity recognition. Egyptian Informatics Journal, 22(3), 295-302.

Alves, D., Kuculo, T., Amaral, G., Thakkar, G., & Tadic, M. (2020). UNER: Universal Named-Entity RecognitionFramework. arXiv preprint arXiv:2010.12406.

Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., & Clematide, S. (2022, August). Overview of HIPE-2022: named entity recognition and linking in multilingual historical documents. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 423-446). Cham: Springer International Publishing.

Gan, C., Zhang, Q., & Mori, T. (2023, June). Sentence-to-label generation framework for multi-task learning of japanese sentence classification and named entity recognition. In International Conference on Applications of Natural Language to Information Systems (pp. 257-270). Cham: Springer Nature Switzerland.

Han, L., Jones, G. J., & Smeaton, A. F. (2020). MultiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. arXiv preprint arXiv:2005.10583.

Kulkarni, M., Preoţiuc-Pietro, D., Radhakrishnan, K., Winata, G. I., Wu, S., Xie, L., & Yang, S. (2023, May). Towards a unified multi-domain multilingual named entity recognition model. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2210-2219).

Kvapilıková, I., Artetxe, M., Labaka, G., Agirre, E., & Bojar, O. (2021). Unsupervised multilingual sentence embeddings for parallel corpus mining. arXiv preprint arXiv:2105.10419.

Liu, P., Guo, Y., Wang, F., & Li, G. (2022). Chinese named entity recognition: The state of the art. Neurocomputing, 473, 37-53.

Luo, H., Jin, Y., Liu, X., Shang, T., Chen, R., & Liu, Z. (2024). GEIC: Universal and Multilingual Named Entity Recognition with Large Language Models. arXiv preprint arXiv:2409.11022.

Malmasi, S., Fang, A., Fetahu, B., Kar, S., & Rokhlenko, O. (2022). MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. arXiv preprint arXiv:2208.14536.

Malmasi, S., Fang, A., Fetahu, B., Kar, S., & Rokhlenko, O. (2022, July). Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022) (pp. 1412-1437).

Mayhew, S., Blevins, T., Liu, S., Šuppa, M., Gonen, H., Imperial, J. M., ... & Pinter, Y. (2023). Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. arXiv preprint arXiv:2311.09122.

Meeus, Q., & Moens, M. F. (2024). MSNER: A Multilingual Speech Dataset for Named Entity Recognition. arXiv preprint arXiv:2405.11519.

Naraki, Y., Yamaki, R., Ikeda, Y., Horie, T., & Naganuma, H. (2024). Augmenting NER Datasets with LLMs: Towards Automated and Refined Annotation. arXiv preprint arXiv:2404.01334.

Naseer, S., Ghafoor, M. M., bin Khalid Alvi, S., Kiran, A., Rahmand, S. U., Murtazae, G., & Murtaza,

Thank you

Conclusion

- Multilingual parallel corpora significantly facilitate mulitingual entities identification and linking.
- Asymmetrical linkages attribute to not just linguistic Features but also entity categories.
- Chinese-character-based entities interact in more complex mapping relations.

Discussion

Reference

Asymmetric Alignments of Mulitlingual Named Entities: Focus on the CJK Parallel Corpus

Dylan Fei @ Sukmyung Univ.
KACL 2024 Fall

이 연구는 한·중·일 병렬 말뭉치를 바탕으로 한국어, 중국어, 일본어 개체명의 비대칭적 분포와 매칭 현상에 초점을 맞추고 통계 실험과 용례 분석을 통해서 앞으로 다국어 개체명 분류 및 식별을 위한 통합적인 프레임워크 제안에 의의가 있다.
다국어 개체명의 정렬 관계는 인지의미론적으로 개체명 개념(Concept)과 표현(Mention) 간의 매핑 양상과 깊게 연계되어 있으며, 정보추출 및 언어 모델링의 입장에서도 개체명 관련 의미적 표현(semantic representation)의 정확성 및 다양성 평가에도 중요한 연관성이 존재한다. 특히 한중일 삼국은 역사적으로 언어 교류가 많고 지리적으로 지식 공유의 속도도 매우 빠른 편이라서 각종 전문용어에 기반한 개체명 관련 개념적 체계가 상당히 유사할 것으로 예상된다. 그러나 언어 유형과 문화 배경에서 비롯된 다양한 차이점 때문에 개체명을 표현하는 방식과 범위는 겹치는 부분도 있지만 달라지는 부분도 없지 않다.
위 현상을 재조명하기 위해서, 이 연구는 한중일 병렬정렬 말뭉치를 재정비하고 개체명 자동 추출 기술을 활용함으로써, 기존의 한중일 3개국어의 개체명 분류 체계 중 공통되는 카테고리를 먼저 추출하고 이를 바탕으로 카테고리별 개체명 분포의 차이성 유무를 추론 통계 검증으로 확인했으며 개별적 용례 분석을 토대로 비대칭적 개체명 분포의 구체적인 발생 조건을 분석하고 있다.
그런데 이 연구에 도입한 말뭉치 자원은 구어성 강한 텍스트로 구성되어 있다. 따라서 구어 발화의 장르적 특성이 개체명 매칭 관계에 영향을 끼칠 수 있을지는 추후 연구로 밝혀야 할 문제로 남는다. 이외에 새로 정비한 3개국어 병렬 말뭉치에서 추출한 개체명 데이터는 규모가 방대한데 도입된 개체명 자동 식별 모듈의 성능 한계가 분명히 존재하므로 분석 결과를 해석하는 과정에서 특별히 유의할 필요가 있다. 그리고 정량적 분석에 도입된 데이터 양에 비해서 정성분석에 선정된 용례의 범위와 수량이 상대적으로 작은 편인데 한·중·일 비대칭적 개체명 분포의 세부 유형을 더 밝히기 위해서 언어학 각 층위 이론에 기반한 심층적인 용례 분석으로 계속 이어질 여지가 작지 않다고 본다.
종합적으로, 이 연구의 한계점과 향후 연구의 확장 가능성을 감안해서 다음과 같은 질문을 제기해 볼 수 있다.

1. 실험용 데이터셋의 한계점 및 추후 보완 방안
2. 한중일 비대칭 개체명의 세부 유형 분류 기준
3. 통계 분석 결과와 용례 분석 결과 간의 일치성
4. 앞으로 한·중·일 개체명을 위한 통일된 개체명 분류 및 식별 프레임워크 구현 가능성

통역 품질 향상을 위한 감정 분석 기반 RAG 시스템 개발

- 구어 코퍼스를 중심으로 -

송지현(이화여자대학교)
이용훈(충남대학교)

차 례

1. 머리말

번역과 통역은 출발어(Source Language)를 도착어(Target Language)로 전환한다는 점에서 동일하지만, 글을 옮기는 번역에 비해 통역은 목소리나 감정과 같은 비언어적 요소들이 복합적으로 영향을 미친다. 그중에서도 감정은 대표적인 비언어적 요소로서, 동일한 내용이라도 감정에 따라 표현 방식이 달라질 수 있으며, 이에 따라 통역 방식도 달라져야 하는 경우가 발생한다. 따라서 감정에 따른 구어 코퍼스의 특성을 분석함으로써 감정 표현이 발화 상황에 어떤 영향을 미치는지 파악하고, 통역 상황에서 능동적으로 대처할 수 있는 수단으로 활용할 수 있다.

본 연구에서는 통역과 유사한 구어 코퍼스에서 감정 표현을 분석함으로써, 연사의 감정 변화가 연설이나 대화에 어떤 영향을 미치는지 파악하고자 한다. 이를 위해 인공지능 모델을 파인 튜닝하여 감정 분석기를 생성하고, TED 강연 자막을 대상으로 감정 분석을 수행하였다. 또한, 검색 증강 생성(Retrieval-Augmented Generation, 이하 RAG) 시스템을 구축하여 감정 분석 결과를 컨텍스트로 제공하고, 생성형 인공지능 모델을 통해 추가적인 분석을 수행하였다. RAG를 활용한 결과, 컨텍스트가 제공되지 않은 경우에 비해 보다 구체적이고 신뢰성 있는 응답을 얻을 수 있었으며, 이는 거대 언어 모델(Large Language Model, 이하 LLM)의 환각(Hallucination) 현상을 줄이고 실제 데이터에 기반한 정확한 답변을 생성하는 데 효과적임을 확인하였다.

2. 연구 동기 및 목적

의사소통은 단순히 전달되는 내용뿐만 아니라 상황과 환경을 아우르는 전반적인 문맥, 얼굴 표정, 감정, 목소리의 높낮이 등 다양한 비언어적 요소들이 복합적으로 영향을 미치는 복잡한 과정이다. 이러한 요소들에 따라 동일한 내용, 즉 같은 텍스트라도 전혀 다른 의미로 해석될 수 있음을 의미한다.

특히 통역은 글을 다른 언어로 옮기는 번역과는 달리, 말을 말로 전달하는 과정에서 언어적 요소와 비언어적 요소를 모두 고려해야 하는 고차원적인 활동이다. 통역사는 연사의 발화 내용뿐만 아니라 그 감정과 의도, 그리고 상황적 맥락까지도 즉각적으로 파악하여 대상 언어로 전달해야 한다. 이러한 과정에서 비언어적 요소인 감정 표현은 통역 품질에 직접적인 영향을 미친다.

그러나 기존의 통역 연구에서는 비언어적 요소에 대한 체계적인 분석이 부족한 실정이다. 이에 본 연구는 통역과 유사한 구어 코퍼스에서 감정 표현을 분석함으로써, 연사의 감정 변화가 연설이나 대화에 어떤 영향을 미치는지 분석하고자 한다. 이를 통해 통역사들이 연사의 감정 변화를 보다 정확하게 예측하고, 다음 발화에 대비함으로써 통역의 정확성과 품질을 향상시킬 수 있을 것으로 기대된다.

3. 선행연구

구어로 이루어지는 의사소통에서는 다양한 비언어적 요소들이 중요한 역할을 한다. 특히 시각적으로 인지되는 얼굴 표정과 감정 표현은 의사소통에 큰 영향을 미치는 대표적인 비언어적 요소로 알려져 있다(Koolagudi et al., 2012).

감정이 대화에 미치는 영향에 대한 연구로는, 인공지능 스피커가 감정에 따라 어떻게 다른 방식으로 발화하는지를 분석한 연구(장지혜 외, 2019)와 감정에 따른 발화자의 음성 변화를 분석한 연구(이수정 외, 1999) 등이 있다. 이러한 연구들은 감정이 발화 내용뿐만 아니라 발화 방식에도 영향을 미친다는 것을 보여준다.

그러나 감정 연구에는 여러 어려움이 존재한다. 가장 큰 이유는 감정이 매우 주관적이라는 점이다. 동일한 텍스트라도 개인의 경험, 가치관, 문화 등에 따라 서로 다른 감정으로 해석될 수 있으며, 감정을 몇 가지로 정의하고 분류할지에 대한 객관적인 기준을 설정하기 어렵다. 이에 따라 모든 분야와 학계에서 보편적으로 합의된 감정의 이론적 정의가 부족한 실정이다(El Ayadi et al., 2011).

그럼에도 불구하고 감정을 분석하기 위해서는 일정한 기준에 따라 감정을 분류하는 것이 필수적이다. Cowie와 Reddy(2001)는 모든 감정을 기본적인 1차 감정(Primary Emotion) 인 분노(Anger), 역겨움(Disgust), 공포(Fear), 기쁨(Joy), 슬픔(Sadness), 놀라움(Surprise)으로 축소할 수 있다는 팔레트 이론을 제시하였다. 또한 Russell(2005)은 감정을 Valence-Arousal 이라는 연속적인 값으로 표현하여 2차원 공간에 배치하는 방법을 제시하였다. 그러나 이러한 방법들은 감정의 개수를 연구자가 임의로 지정한 것으로, 학자나 개인에 따라 상이할 수 있다.

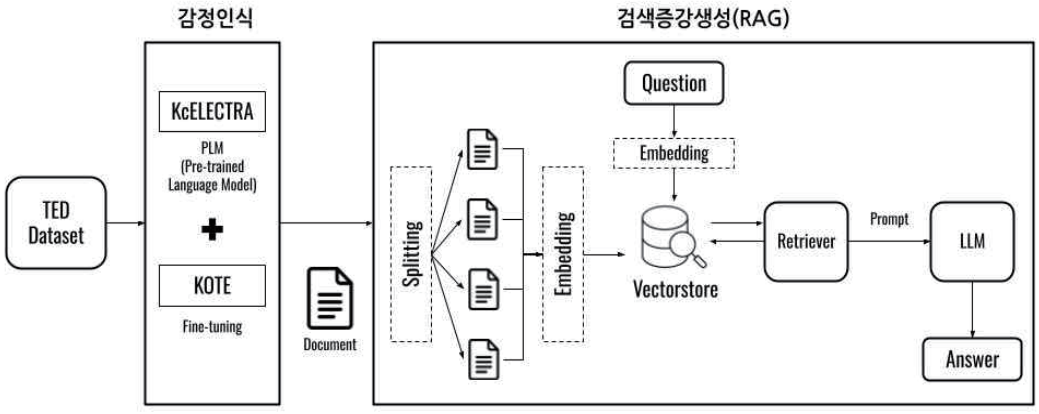
이러한 한계를 극복하기 위해 머신러닝의 클러스터링 기법을 도입하여 감정을 분류하는 방법이 제시되었다(Jeon et al., 2022). 해당 연구에서는 감정 표현이 담긴 1만 5천 개의 샘플을 전처리한 후, 사전 학습된 워드 벡터 모델인 FastText를 활용하여 1,787개의 단어를 추출하였다. 각 단어의 벡터 공간 상의 위치를 확인하고, 머신러닝 클러스터링 기법인 UMAP을 통해 단어들이 유의미하게 모여 있는 44개의 클러스터를 도출하였다. 이후 각 클러스터 내의 단어들을 분석하여 해당 클러스터가 의미하는 감정을 명명하였다. 이 과정을 통해 감정을 보다 객관적이고 데이터 기반으로 분류하였다.

최근 인공지능 기술의 발전으로 LLM이 등장하면서, 자연어 처리 분야에서 혁신적인 변화가 일어나고 있다(Brown et al., 2020). LLM은 방대한 양의 데이터를 학습하여 다양한 언어 처리 작업에서 우수한 성능을 보이고 있으나, 환각(hallucination) 현상이나 특정 도메인 지식의 부족 등의 한계가 존재한다(이훈희, 2023). 특히 감정 분석과 같은 특수한 작업에서는 일반적인 LLM만으로는 정확한 결과를 도출하기 어려울 수 있다.

이러한 한계를 극복하기 위해 RAG 방식이 주목받고 있다(Lewis et al., 2020). RAG는 외부 지식 베이스에서 관련 정보를 검색하여 LLM에 제공함으로써 모델의 한계를 보완하고, 보다 정확하고 신뢰성 있는 응답을 생성할 수 있게 한다. 감정 분석 분야에서도 RAG를 활용하여 대규모의 구어 코퍼스에서 감정 정보를 효과적으로 추출하고 활용할 수 있다.

예를 들어, 내부 보안 자료를 참조하거나(이훈희, 2023), 실시간으로 업데이트되는 뉴스 기사를 반영하는 질의응답 시스템을 개발하는 연구가 진행되었다(박성민 외, 2024). 또한 챗봇의 성능 향상을 위해 파인 튜닝과 RAG 방식을 비교한 연구에서는 RAG의 우수성을 확인하였다(손지원 외, 2024). 이러한 연구들은 RAG가 LLM의 성능 향상과 특정 도메인 지식의 활용에 효과적임을 보여준다.

4. 연구방법



<그림 1> 감정인식 및 검색증강생성(RAG) 구현 도식

4.1. 인공지능 모델을 통한 감정분석기 생성

감정을 분석하기 위해서는 구어 코퍼스와 그에 해당하는 감정에 대한 정보가 함께 필

요하였기 때문에, 코퍼스-감정이 병렬적으로 구성된 데이터를 최우선적으로 수집하고자 하였다. 데이터 수집의 기준은 1) 배우들이 연기를 해서 만들어진 데이터보다 실제 상황에서 오간 대화 데이터를 우선적으로 선정하고, 2) 감정의 종류가 적어도 5가지 이상이 되는 데이터셋을 선정하는 것으로 구상하였다. 그러나 구할 수 있는 코퍼스-감정 데이터셋의 대부분이 영어이며, 여러 데이터셋을 합칠 경우 각 데이터셋마다 분류된 감정의 개수가 다르며, 최근 데이터는 없다는 한계가 있었다.

이를 극복하기 위하여 Jeon(2022)이 제안한대로 대규모 한국어 데이터를 통해 사전학습된 감정분류 인공지능 모델을 활용하여 코퍼스에 대한 감정을 예측하여 코퍼스-감정 데이터를 생성하고, 이것을 분석하는 방법을 시도하였다. 인공지능 모델을 통한 감정 데이터 생성이 가능할 것으로 생각한 이유는 1) 모델의 성능 지표인 F1 점수가 모든 감정에 대해 대체로 60에서 70 안팎으로 비교적 높은 수치를 보여주고 있었으며, 2) 비록 생성된 감정이 정확하지 않더라도 모든 코퍼스에 대해 동일한 인공지능 모델, 즉 동일한 조건이 적용 되었으므로 결과를 분석하는 의의가 있을 것이라고 판단하였다.

사전학습 언어모델(Pre-trained Language Model, PLM)은 약 1억 8천개의 한국어 댓글 데이터(약 17GB)를 학습시킨 KcELECTRA를 사용하였으며, 추가적으로 약 5만개의 댓글 데이터에 <표1>에 제시된 감정 44개가 레이블링 되어있는 'KOTE Dataset(Korean Online Comments Emotions Dataset)'을 이용해 파인 튜닝(Fine-tuning)하였다. 결과적으로 이 인공지능 모델은 어떤 문장을 입력받으면 그에 해당하는 감정을 예측하여 출력하는 감정분류기로 동작하게 된다. 이를 통해 감정에 대한 정보가 없는 순수 코퍼스 데이터만으로 감정 분석이 가능해지고, 데이터 수집의 자유도 향상을 꾀할 수 있었다.

<표 1> 감정분류(총 44개)

연번	감정	연번	감정
1	불평/불만	23	짜증
2	환영/호의	24	어이없음
3	감동/감탄	25	없음
4	지긋지긋	26	패배/자기혐오
5	고마움	27	귀찮음
6	슬픔	28	힘듦/지침
7	화남/분노	29	즐거움/신남
8	존경	30	깨달음
9	기대감	31	죄책감
10	우쭐댐/무시함	32	증오/혐오
11	안타까움/실망	33	흐뭇함(귀여움/예쁨)
12	비장함	34	당황/난처
13	의심/불신	35	경악
14	뿌듯함	36	부담/안_내킴
15	편안/쾌적	37	서러움
16	신기함/관심	38	재미없음
17	아껴주는	39	불쌍함/연민
18	부끄러움	40	놀람
19	공포/무서움	41	행복
20	절망	42	불안/걱정
21	한심함	43	기쁨
22	역겨움/징그러움	44	안심/신뢰

4.2. 데이터 선정 및 감정 레이블 생성

분석 데이터로는 TED 연설의 자막을 선정하였다. 선정하게 된 이유는 1) 많은 정제를 거쳐 배포되는 타 연설문에 비해 구어 표현을 살리며 자연스럽게 전사된 경우가 많으며, 2) 연사의 말하는 방식과 강연 장소의 환경이 순차 통역과 유사하고, 3) 다양한 언어로 자막을 제공하므로 향후 타 언어에 대해 연구할 경우 확장성이 용이하다고 판단했기 때문이다.

연설 데이터, 즉 구어 코퍼스를 선정하기 위해 TED 연설문 검색창에서 제시하는 가장 인기가 많은 카테고리 6개(비즈니스, 엔터테인먼트, 디자인, 과학, 기술, 글로벌 이슈)에서 가장 시청수가 많은 연설 10개를 선택하였다. 총 60개의 연설문의 감정을 예측하고, 그것을 분석하였다.

감정 예측을 위한 코퍼스의 단위로는 순차통역의 길이와 유사하다고 판단되는 단위로 휴리스틱 메소드(Heuristic method)로 분리하였다. 그 결과 카테고리별 코퍼스 길이의 평균은 비즈니스 295.96자, 과학 231.69자, 엔터테인먼트 248.67자, 기술 222.62자, 디자인 288.47자, 글로벌 이슈 250.29자였고, 대체로 250자 내외에 분포되어 있었다.

4.3. 검색증강생성(RAG)을 통한 데이터 분석

4.3.1. 검색증강생성(RAG)의 정의와 이점

검색 증강 생성(Retrieval-Augmented Generation, RAG)은 LLM에 외부 데이터를 컨텍스트로 제공하여 응답을 생성하는 방식이다(손지원 외, 2023). <그림 1>에서 볼 수 있듯이, RAG는 수집한 외부데이터를 임베딩 모델을 통해 벡터화한 후 벡터 스토어(Vectorstore)라는 데이터베이스에 저장한다. 사용자가 질문하면, 질문도 벡터화하여 검색기(Retriever)를 통해 데이터베이스에서 가장 유사한 데이터를 검색하고, 이를 참고하여 보다 정확한 답변을 생성한다(박성민 외, 2024).

LLM을 활용한 오픈 도메인 질의응답 시스템은 외부 지식 없이 사전 학습된 대용량 데이터를 기반으로 답변을 생성한다. 이로 인한 가장 큰 한계점은 환각(Hallucination) 현상이 발생할 수 있고, 사전 학습된 지식의 갱신이나 확장이 어렵다는 점이다(박성민 외, 2024). 이러한 문제를 해결하기 위해 사전 학습된 LLM에 신규 데이터를 추가하여 학습하는 파인 튜닝(fine-tuning) 방식이나, 사용자가 프롬프트에 직접 컨텍스트를 삽입하는 방식을 사용할 수 있다(정천수, 2023). 그러나 손지원 외(2023)에 따르면 파인 튜닝 방식에 비해 RAG 방식이 환각과 답변 오류가 적고 정확도가 높았으며, 모든 컨텍스트를 일일이 프롬프트에 삽입하는 것은 현실적으로 어렵기 때문에 참고할 정보를 데이터베이스에 저장하고 프롬프트를 통해 LLM에 전달하는 RAG 방식이 보다 정확하고 효율적이라고 할 수 있다(전청수, 2023).

4.3.2. 연구방법

본 연구에서는 감정 분석기로 생성된 문장-감정 데이터를 컨텍스트(Context)로 제공하고, 사용자의 질문(Question)과 함께 프롬프트(Prompt)를 구성하여 LLM에 제공하는 RAG 시스템을 구축한 후 답변을 생성하였다. 예시 프롬프트는 아래와 같다.

(1) 프롬프트:
당신은 한-영 전문 통역사입니다.
주어진 컨텍스트의 문장과 감정을 참고하여 아래 질문에 답하세요.
#질문 {question}
#컨텍스트 {context}

그 결과, 컨텍스트가 주어지지 않은 경우에 비해 더 구체적이고 정확한 답변을 얻을 수 있었다.

5. 연구결과

본 연구에서는 각 카테고리에서 '깨달음' 감정이 가장 높은 빈도를 보이는 것을 확인할 수 있었다. 이는 'Ideas Worth Spreading(널리 알릴만한 가치가 있는 아이디어)'를 공유하는 TED 연설의 취지가 연설 내용에 반영된 결과로 해석된다. 동일한 이유로 '기대감', '감동/감탄', '신기함/관심' 등의 감정도 모든 카테고리에서 높은 빈도를 나타내었다.

특히 주목할 만한 점은 다른 카테고리에 비해 글로벌 이슈의 경우 '비장함', '의심/불신', '불안/걱정'과 같은 부정적인 감정의 빈도가 높게 나타난 것이다. 글로벌 이슈 카테고리에서 '비장함' 감정을 포함한 문장은 전체 328개 샘플 중 190개로, 매우 높은 비중을 차지하였다. 일부 문장을 분석한 결과, 국제 사회의 질병 심각성을 알리거나 탈북자의 경험 등을 다루는 등 비교적 심각한 소재의 내용이 주를 이루고 있음을 확인하였다. 이는 다른 카테고리에 비해 부정적이거나 심각한 감정의 빈도가 높은 이유로 추론할 수 있다.

RAG 시스템을 구축하여 감정 분석기로부터 생성된 문장-감정 데이터를 컨텍스트로 제공하고, 생성형 인공지능 모델을 통해 추가적인 분석을 수행하였다. RAG를 활용한 경우, 컨텍스트가 제공되지 않은 경우에 비해 보다 구체적이고 정확한 응답을 얻을 수 있었다.

예를 들어, 특정 감정이 나타나는 문장을 제시하고 해당 문장의 통역 방안을 요청하였을 때, RAG 시스템은 입력된 데이터에서 정확히 해당 문장들을 찾아내어 적절한 통역 방안을 제시하였다. 이는 감정 분석 결과가 RAG 시스템 내에서 효과적으로 활용되었음을 보여준다. 반면에 기존의 LLM에 동일한 프롬프트를 입력한 경우에는 일반적인 대답만을 얻을 수 있었다. 이는 RAG를 통해 모델의 환각 현상을 줄이고, 실제 데이터에 기반한 신뢰성 있는 응답을 생성할 수 있었음을 의미한다. 따라서 RAG는 거대 언어 모델의 한계를 극복하고, 특정 도메인에 특화된 응답을 제공하는 데 효과적임을 확인할 수 있었다.

(1) **질문(Question)**: 비장함이라는 감정이 나타나는 문장들을 제시하고, 감정을 살려 어떻게 통역할 수 있는 방안을 제안해줘.

(2) **답변(Answer)**

ㄱ. General LLM:

"나는 내 운명을 받아들이고 이 길을 갈 것이다."

통역: "I will accept my destiny and walk this path."

해설: 주인공의 결연한 의지와 운명을 받아들이는 비장함을 그대로 전달하기 위해 직역에 가까운 표현을 사용했습니다. "Accept my destiny"와 "walk this path"는 원문의 심정을 잘 담아냅니다. (후략)

ㄴ. RAG:

비장함이라는 감정이 나타나는 문장들을 다음과 같이 제시할 수 있습니다:

"게다가 이런 것들은 타인의 행복에까지 치명적입니다. 그러므로 우리 마음에 이들의 침범이 잦아질 수록 연쇄반응으로 우리는 더 비참해지고 고통받을 것입니다." (중략)

이 문장들을 감정을 살려 통역할 때는 다음과 같이 할 수 있습니다:

"Moreover, these things are fatal to the happiness of others. Therefore, as these intrusions into our minds become more frequent, we will become more miserable and suffer as a chain reaction."

6. 결론

본 연구에서는 인공지능 모델을 파인 튜닝하여 감정 분석기를 생성하고, 이를 활용하여 TED 강연 자막을 분석하였다. 감정 분석 결과, 모든 카테고리에서 '깨달음' 감정이 가장 높은 빈도를 보였으며, 이는 TED 연설의 취지인 '널리 알릴 만한 가치가 있는 아이디어'의 공유가 연설 내용에 반영된 것으로 해석된다. 특히 글로벌 이슈 카테고리에서는 '비장함', '의심/불신', '불안/걱정' 과 같은 부정적인 감정의 빈도가 높게 나타났다. 이는 해당 연설들이 국제 사회의 심각한 문제나 탈북자의 경험 등을 다루고 있기 때문으로 분석된다. 이러한 결과는 각 카테고리별로 연설 내용의 주제와 감정적 특성이 상호 연관되어 있음을 시사한다.

또한 검색 증강 생성(RAG) 시스템을 구축하여 감정 분석 결과를 컨텍스트로 제공하고, 생성형 인공지능 모델을 통해 추가적인 분석을 수행하였다. RAG를 활용한 결과, 컨텍스트가 제공되지 않은 경우에 비해 보다 구체적이고 신뢰성 있는 응답을 얻을 수 있었다.

그러나 본 연구는 분석 대상이 총 60개의 연설문으로 제한되어 있어 결과의 일반화에 한계가 있다. 향후 연구에서는 카테고리의 범위를 확대하고 각 카테고리별 연설문의 수를 늘려 더 많은 데이터를 분석함으로써 보다 일반적인 특성을 파악할 필요가 있다. 또한, 현재는 한국어 감정 분석 모델만을 사용하여 한국어 문장에 대한 감정 인식에 제한이 있었다. 영어 사전학습 모델과 데이터셋을 활용하여 파인 튜닝을 진행한다면 영어 문장에 대한 감정 인식 모델을 개발할 수 있으며, 이를 통해 한-영 양 언어의 감정을 병렬적으로 비교하여 통역 분야에 더욱 의미 있는 기여를 할 수 있을 것이다.

그럼에도 불구하고, 본 연구는 감정 레이블이 존재하지 않는 데이터에도 인공지능 감정 분석기를 통해 감정 분석을 수행할 수 있음을 보였다. 또한, RAG 방식을 활용하여 거대 언어 모델의 한계를 극복하고 생성형 인공지능의 응답 품질을 향상시킬 수 있음을 확인하였다. 이는 통역사의 생산성 향상과 통역 연구 분야에서 인공지능 활용의 가능성

을 제시했다는 점에서 학술적 및 실무적 의의가 있다.

참고문헌

이훈희. (2023). 내부 자료의 보안을 위한 RAG 검색-증강 방법의 생성형 AI 적용 방법. 한국항공우주학회 학술발표회 초록집. 강원.

손지원, 김민성, 김부건, 박상준, 표성민, 배지훈, & 이종혁. (2023). Fine-tuning과 RAG를 활용한 특정 도메인 챗봇 제작 방법 비교. Proceedings of KIIT Conference. 제주.

박성민, 이예빈, & 민덕기. (2024). AnswerLink: 검색증강생성(RAG)을 활용한 최신 정보 기반 자연어 질의응답시스템. 한국통신학회 학술대회논문집. 강원.

정천수. (2023). LLM 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현: RAG 모델과 LangChain 프레임워크 기반. 지능정보연구, 29(4), 129-164.

장지혜, & 주다영. (2019). 인공지능 스피커의 정서별 감정발화에 따른 사용성 평가. 한국HCI학회 학술대회 논문집, 705-712.

이수정, 김명재, & 김정수. (1999). 구어체 정서표현에 있어서의 음성 특성 연구. 감성과학, 2(2), 53-66.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877-1901).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, G., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32-80.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572-587.

Jeon, D., Lee, J., & Kim, C. (2022). User guide for KOTE: Korean Online Comments Emotions Dataset. arXiv preprint arXiv:2205.05300.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. International Journal of Speech Technology, 15(2), 99-117.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459-9474).

Majid, A. (2012). Current emotion research in the language sciences. Emotion Review, 4(4), 432-443.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17(3), 715-734.

근현대 디지털텍스트의 재발견과 어휘개념사

김일환
성신여대 국어국문학과
ilhwan52@sungshin.ac.kr

발표 개요

- 도입
- 근현대 디지털 텍스트 현황
- 근현대 디지털 텍스트의 가공
- 근현대 디지털 텍스트의 활용
- 마무리

근현대 텍스트에 주목하는 이유

- 인공지능, 대규모 언어모델 등의 등장
  - 학습용 데이터로서 대규모의 다양한 말뭉치 구축
    - ✓ 형태분석 말뭉치, 어휘의미분석 말뭉치, 구문분석 말뭉치
    - ✓ 상식추론 말뭉치, 함의분서 말뭉치, 개체명 말뭉치, 표 설명 말뭉치…
      - ➢ 최근 들어 양보다 질에 더욱 많은 관심

- 말뭉치에 대한 제한된 관심
  - 모두 현대국어만을 대상으로 하고 있음
    - ✓ 현대국어의 역사는 매우 짧음
  - 시기적인 확장은 필요하다
    - ✓ 19세기부터 20세기 초기의 텍스트 자료들
      - ➢ 기계적인 접근, 처리에 제약
      - ➢ 인공지능을 위한 필요성, 시급하지 않음

근현대 텍스트의 접근 가능성

- 근현대 텍스트에 접근하는 방법

  - 근현대 텍스트를 처리할 수 있는 도구를 개발하는 방법
    - ✓ 상업성이 있을까
    - ✓ 시대마다 다른 다양한 표기의 텍스트를 모두 처리할 수 있을까
      - ➢ 15세기 처리 도구, 16세기 처리 도구 등등
  - 근현대 텍스트를 가공하여 처리하는 방법
    - ✓ 자동 처리가 가능할까
    - ✓ 수동, 반자동 전처리를 통해 현대표기로 변환
      - ➢ 많은 시간과 노력이 필요함

근현대 텍스트 3대장

신문류	잡지류	소설류
✓ 동아일보 ✓조선일보 ✓시대일보 ✓독립신문 ✓…	✓개벽 ✓삼천리 ✓별건곤 ✓학지광 ✓동광 ✓…	✓ 혈의누 ✓장화홍련전 ✓금방울전 ✓옥중화 ✓…

5


東亞日報

答刑을 催廢

사원일부터 태형을폐지

그대신에난징역이나구류

廢止에伴하난施設

감옥과간수를만히느려서

태형대신에구금을식힌다

答刑廢止와

在監者增加

監獄擴張

九十度나太甚

嘉禮擇日

文化向上의産物

향진군무사정의말

다국에서

6

동아 디지털아카이브

笞刑을僅廢 廢止에伴하난施設

3면 사회

사월일일부터태형을폐지 그대신에난징역이나구류
감옥과간수를만히느려서 태형대신에구금을식힌다

笞刑을僅廢
사월일일부터태형을폐지
그대신에난징역이나구류
잇던것이라도 폐하지 안이하면 안될때에 새법령을뎡하야 일반
을곤난케하고 내외국인의게 비샹한 공격을밧으면서도 재졍이
부족하니 시긔가일흐니 벌핑계를다하며 폐지를 안이하랴던바
조선에서도 조선사람의게만한하야 쓰이던 태형제도도 변쳔하
는시세가 당국쟈를로라서 드듸여금일부터는 폐지하게 되얏다
이에대하야 총독부수야졍무총감은말하되
廢止에伴하난施設
감옥과간수를만히느려서
태형대신에구금을식힌다
笞刑廢止와
在監者增加
태형폐지의 결과재감자의 수가증가할것은 당연한리치라 최근
삼개년간의 평균태형밧은 수효를계산하면 재판사건과 범죄즉
결사건을합하야 일개년에 인원수효가 오만칠쳔삼백 이십사인
이요 그집행한태형의도수(재판사건일인평균 칠십일도、즉결사

7

원문 한글변환

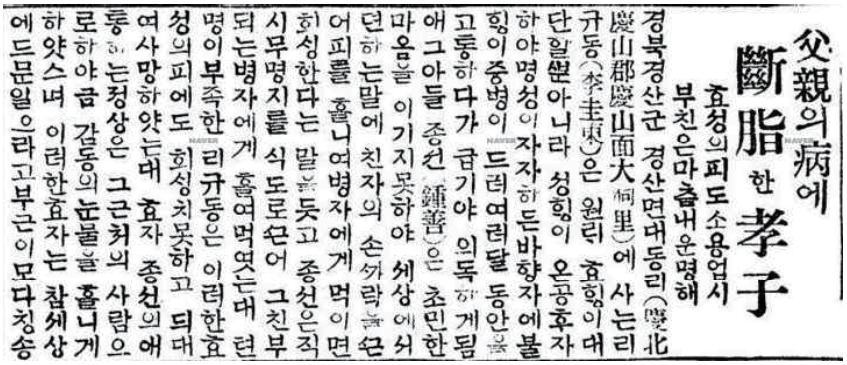
笞刑(태형)을僅廢(근폐) 廢止(폐지)에伴(반)하난施設(시설)

3면 사회

사월일일부터태형을폐지 그대신에난징역이나구류
감옥과간수를만히느려서 태형대신에구금을식힌다

笞刑(태형)을僅廢(근폐)
사월일일부터태형을폐지
그대신에난징역이나구류
잇던것이라도 폐하지 안이하면 안될때에 새법령을뎡하야 일반
을곤난케하고 내외국인의게 비샹한 공격을밧으면서도 재졍이
부족하니 시긔가일흐니 벌핑계를다하며 폐지를 안이하랴던바
조선에서도 조선사람의게만한하야 쓰이던 태형제도도 변쳔하
는시세가 당국쟈를로라서 드듸여금일부터는 폐지하게 되얏다
이에대하야 총독부수야졍무총감은말하되
廢止(폐지)에伴(반)하난施設(시설)
감옥과간수를만히느려서
태형대신에구금을식힌다
笞刑廢止(태형폐지)와
在監者增加(재감자증가)
태형폐지의 결과재감자의 수가증가할것은 당연한리치라 최근
삼개년간의 평균태형밧은 수효를계산하면 재판사건과 범죄즉
결사건을합하야 일개년에 인원수효가 오만칠쳔삼백 이십사인

신문기사



父親의病에 斷脂한 孝子

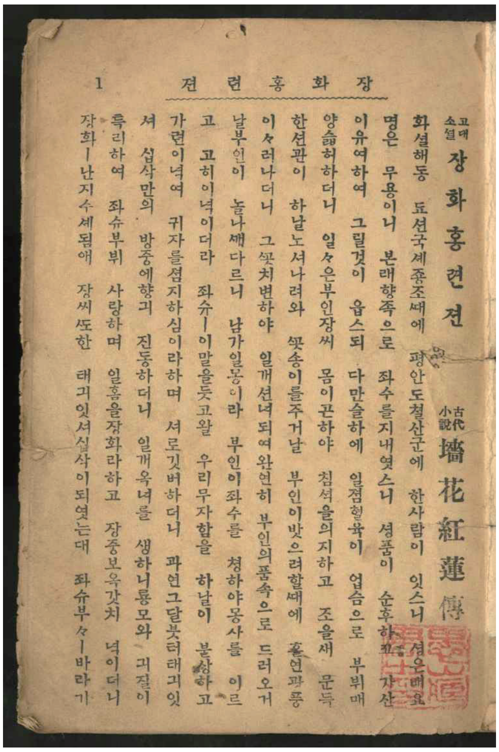
효성의피도소용업시 부친은마츰내운명해

경북경산군 경산면대동리(慶北慶山郡慶山面大洞里)에 사는리규동(李圭東)은 원래 효행이 대단할 뿐아니라 성행이 온공후자하야 명성이 자자하든바 향자에 불행이 중병이 드러여러달 동안을 고통하다가 급기야 위독하게 됨애그아들 종선(鍾善)은 초민한 마음을 이기지 못하야 세상에서 뎐하는 말에 친자의 손까락을 끈어 피를 흘니여 병자에게 먹이면 회생한다는 말을 듯고...

1920.6.15. 기사 "父親의 病에 斷脂한 孝子"
경북 경산군 경산면 대동리(慶北慶山郡慶山面大洞里)에 사는 리규동(李圭東)은 원래 효행이 대단할 뿐아니라 성행이 온공후자하야 명성이 자자하든바 향자에 불행이 중병이 드러여러달 동안을 고통하다가 급기야 위독하게 됨 애그아들 종선(鍾善)은 초민한 마음을 이기지 못하야 세상에서 뎐하는 말에 친자의 손까락을 끈어 피를 흘니여 병자에게 먹이면 회생한다는 말을 듯고...

8

활자본 소설


1 장화홍련전

고대소설 장화홍련전 古代小說 墻花紅蓮傳

```
<KCTML>
	<Header>
		<Title>장화홍련젼</Title>
		<Edition>경판본</Edition>
		<VariantInfo></VariantInfo>
		<Author>미상</Author>
		<Date>미상</Date>
		<Publisher>대성서림</Publisher>
		<PubPlace>서울</PubPlace>
		<Source>국립중앙도서관</Source>
		<RevisionDesc>원문 표기를 그대로 전사함</RevisionDesc>
	</Header>
	<Body><p><pb= "1" >화셜해동 됴션국세종조때에 평안도철산군에 한사람이 잇스니 셩은배요 명은 무용이니 본래 향족으로 좌수를지내엿스니 셩품이 순후하고 가산이유여하여 그릴것이 읍스되 다만슬하에 일졈혈육이 업슴으로 부뷔매 앙슯허하더니 일일은부인장씨 몸이곤하야 침셕을의지하고 조을새 문득 한션관이 하날노셔나려와 꼿송이를주거날 부인이밧으려할때에 홀연광풍이이러나더니 그꼿치변하야 일개션녀되여완연히 부인의품속으로 드러오거날부인이 놀나깨다르니 남가일몽이라 부인이좌수를 청하야몽사를 이르고 고히이녁이더라 좌슈ㅣ이말을듯고왈 우리무자함을 하날이 불상하고 가련이녁여 귀자를졈지하심이라하며 셔로깃버하더니 과연그달붓터태긔잇셔 십삭만의 방중에향긔 진동하더니 일개옥녀를 생하니룡모와 긔질이 특리하여 좌슈부뷔 사랑하며 일홈을장화라하고 장중보옥갓치 녁이더니 장화ㅣ난지수세됨애 장씨또한 태긔잇셔십삭이되엿는대 좌슈부부ㅣ 바라기를<pb= "2" > 아달생하기를 쥬야바라더니 일일은 또녀아를낫커늘마암에 셔운하나 할일업셔일홈을홍련이라 하엿더니 장화홍련형뎨 졈졈자라매 얼골이화려하고긔질이 긔묘함이 셰상에 무쌍하고 효행이 더욱특츌하니 좌슈부쳐 형뎨자람을보고 사랑함이비할때업는중 넘무숙셩함을 매양념려하더니 신운이 불행하여 장씨홀연득병하여 증세날노위중하여 좌수와장화형뎨주야로 약을 힘쓰되조곰도 효험이업는지라 장화형뎨초초하야 하날씌 축슈하여그 (중략)
```


활자본 소설

```
<KCTML>
	<Header>
		<Title>장화홍련전</Title>
		<Edition>경판본</Edition>
		<VariantInfo></VariantInfo>
		<Author>미상</Author>
		<Date>미상</Date>
		<Publisher>대성서림</Publisher>
		<PubPlace>서울</PubPlace>
		<Source>국립중앙도서관</Source>
		<RevisionDesc>원문 표기를 현대표기로 변환함</Revision
	</Header>
	<Body><p><pb= "1" >화설해동 조선국 세종조 때에 평안도 철산군
니 본래 항족으로 좌수를 지내었으니 성품이 순후하고 가산이 유여하여 그릴
로 부부가 매양 슬퍼하더니 일일은 부인 장씨 몸이 곤하야 침석을 의지하고
를 주거늘 부인이 받으려 할 때에 홀연 광풍이 일어나더니 그 꽃이 변하여
거늘 부인이 놀라 깨달으니 남가일몽이라 부인이 좌수를 청하여 몽사를 이르
무자함을 하늘이 불상하고 가련히 여겨 귀자를 점지하심이라 하며 서로 기뻐
중에 향기 진동하더니 일개 옥녀를 생하니 용모와 기질이 특이하여 좌수 부
여기더니 장화 난 지 수세됨에 장씨 또한 태기 있어 십삭이 되었는데 좌수
바라더니 일일은 또 여아를 낳거늘 마음에 서운하나 할 일 없어 이름을 홍
굴이 화려하고 기질이 기묘함이 세상에 무쌍하고 효행이 더욱 특출하니 좌수
는 중 너무 숙성함을 매양 염려하더니 신운이 불행하여 장씨 홀연 득병하여
약을 힘쓰되 조금도 효험이 없는지라 장화형제 초초하야 하늘께 축수하여 (중
```

```
<KCTML>
	<Header>
		<Title>장화홍련전</Title>
		<Edition>경판본</Edition>
		<VariantInfo></VariantInfo>
		<Author>미상</Author>
		<Date>미상</Date>
		<Publisher>대성서림</Publisher>
		<PubPlace>서울</PubPlace>
		<Source>국립중앙도서관</Source>
		<RevisionDesc>디지털 가공 텍스트에 형태소 정보를 부착함(분석 표지는 '21세기 세종계획' 기준</RevisionDesc>
	</Header>
	<Body>
화설해동	화설해동/NNP
조선국	조선국/NNP
세종조	세종조/NNG
때에	때/NNG+에/JKB
평안도	평안도/NNP
철산군에	철산군/NNP+에/JKB
한	한/MM
사람이	사람/NNG+이/JKS
있으니	있/VA+으니/EC
성은	성/NNG+은/JX
배요	배/NNP+이/VCP+오/EC
이름은	이름/NNG+은/JX
무용이니	무용/NNP+이/VCP+니/EC
본래	본래/NNG
항족으로	항족/NNG+으로/JKB
```

근대 잡지

11

근대잡지 표기의 다양성

<title>대한자강회월보 제1호</title>
<date>1906-07-31</date>
<text>
<head>
大韓自强會月報序 [三編]
</head>
日에 有客이 促膝附耳□고 叩之于余 曰 自强會之名稱과 及 其 趣旨□ 業已公佈大世矣라. 固不容再問이어니와 必其裏情이 或有□니 願一聞之□노라. 余ㅣ 曰 惡라 是何言也오 君之所謂 裏情은 表面으로 稱云自强이라□고 其 內容은 浮沈渾敦□야 阿私作俑을 是云者歟아 抑或外假華名□고 內挾僞雜□야 以營營於肥己之私□ 是云者歟아 抑 或 形式上으로 腐敗□ 攻擊□되 實地 情衷은 往藉美名□고 終無事行之可能結局者□ 是云歟아 於 此三者에 有一焉이년 其 已强者도 必衰弱乃己어늘 況如今萎 微儒弱之吾人이 奚暇에 有自强之希望哉아 本會 趣旨 及 名稱은 表裏 均一□야 無他

<title>개벽 제5호</title>
<date>1920-11-01</date>
<text>
<head>
권두언
</head>
널리봅시다. 크게 먹읍시다. 보기가 실커든 아조 눈을 감을지라도 보거든 반듯이 널리 봅시다. 널리 보는 者에게는 넓은 局量이 생기리이다. 안이 생기지는 못하리이다. 從來의 우리는 確實히 넓게 보지 못하얏나이다. 그리하야 近視眼이 된 셈이외다. 近視眼者에게야 辨識力의 充分을 바랄 수 잇슴닛가. 自國人 外에는 皆 寇敵으로 보앗스며 自己信認하는 思想 外에는 皆異端으로 보앗나이다. 이 誤見의 發動으로는 門닐고 다가티 살 살 자고 차자 오는 貧窘을 大砲로 迎하야(丙寅庚午의 洋亂은 무엇이며 江華條約은 그 原因이 무엇입닛가) 維新의 길을 우리 혼자 끈헛스며 「밤낫 남의 糟粕만 仰할 것이 안이라 우리에게도 이러한 獨創이 잇다.」고 일어서는 幾多 發明家, 思想家, 宗敎家(그 例는 無數함니다. 爲先 最近의 天道敎 元祖

<title>삼천리 제4호</title>
<date>1930-01-11</date>
<text>
<head>
將來十年에 쟈랄 生命!!, 言論界, 敎育界 等
</head>
十年에 자란 生命

중병 앓는 사람에게 요긴한 것은 발병 이후의 그 경과와 또 병 나을 때까지의 치료 방법 등에 잇다. 병석에 잇는 그에게 넷날의 건강과 몸 회복한 뒤의 할 일들을 천백 가지로의 론하는 것이 무에가 필요하랴. 문제는 현재의 구제에 잇다. 병든 몸이매 건강의 회복에 젼 목적이 잇는 것이다.

마찬가지로 우리에게 필요한 것은 사 천년 오 천 년하는 긴 력사가 아니요, 또한 금후의 백년지계 천년지계의 수립에 잇지 안타. 실로 자각의 운동이라 할 그미 이후의 과거 십 년동안과 그 십 년을 로대로 한 금후 십 년의 행진에 모든 것이 달닌 터이다. 생각건대 이미 못은 싹이니 소슬 것이요, 여무러진 꼿봉오리니 피고야 박이리라. 못기 십 년 피기 십 년이라 하면 력사 행진의 년대 측정에 과히 틀님이 엇슬진저.

12

한자 처리(독음)

334	施	시	이							
331	接	접	첩	삽	잡					
330	兒孩	아해	예해							
329	不絶	부절	불절	비절						
327	吳	오	호	화	우					
326	沈	침	심	담						
322	若	약	야							
322	夜	야	액							
320	責	책	채							
317	則	칙	즉	즉						
315	福	복	부							
311	鐵	철	쇠							
309	倍	배	패							
307	興	흥	흔							
307	差	차	치							
305	據	거	극							
303	投	투	두							
303	樣	양	상							
303	亞米利加	아미리가	아미이가	악미리가	악미이가					
303	旣	기	희							
298	充實	충실	충지							
293	弟	제	퇴							
292	李氏	리씨	리지	리시	리정	이씨	이지	이시	이정	
292	大端	대단	대천	대전	태단	태천	태전	다단	다천	다전
291	春	준	준							
290	朝鮮文學	조선문학	조선문교	조선문할						
288	壞太利	오태리	오태이	욱태리	욱태이					

13

한자 처리(이체자 정규화)

1483	不	不	
1357	六	六	
1290	理	理	
1017	女	女	
761	金	金	
727	兩	兩	
660	歷	歷	
609	勞	勞	
608	論	論	
505	李	李	
497	利	利	
496	樂	樂	
471	老	老	
429	留	留	
352	連	連	
339	立	立	
283	流	流	
276	來	來	
272	戀	戀	
269	復	復	
255	靈	靈	
220	落	落	
213	隷	隷	
207	列	列	
201	聯	聯	
198	例	例	
191	易	易	
190	露	露	
169	旅	旅	
154	臨	臨	
151	良	良	
148	倫	倫	
142	力	力	
138	年	年	
136	更	更	
136	=	=	
134	禮	禮	
133	龍	龍	
128	劣	劣	
126	念	念	
124	亂	亂	
120	陸	陸	
119	冷	冷	
114	綠	綠	

14

한자 변환 - 동음이체자의 처리

코드 번호	한자	코드 번호	한자	독음
0x2f928	獺	0x737a	獺	달
0x2f929	王	0x738b	王	왕
0x2f92a	玑	0x3eac	玑	공
0x2f92b	玥	0x73a5	玥	모
0x2f92c	坪	0x3eb8	坪	평
0x2f92d	坪	0x3eb8	坪	평
0x2f92e	琦	0x7447	琦	대
0x2f92f	瑜	0x745c	瑜	유
0x2f930	瑱	0x7471	瑱	전
0x2f931	璋	0x7485	璋	쇄
0x2f932	瓊	0x74ca	瓊	경
0x2f933	甁	0x3f1b	甁	형
0x2f934	甦	0x7524	甦	유

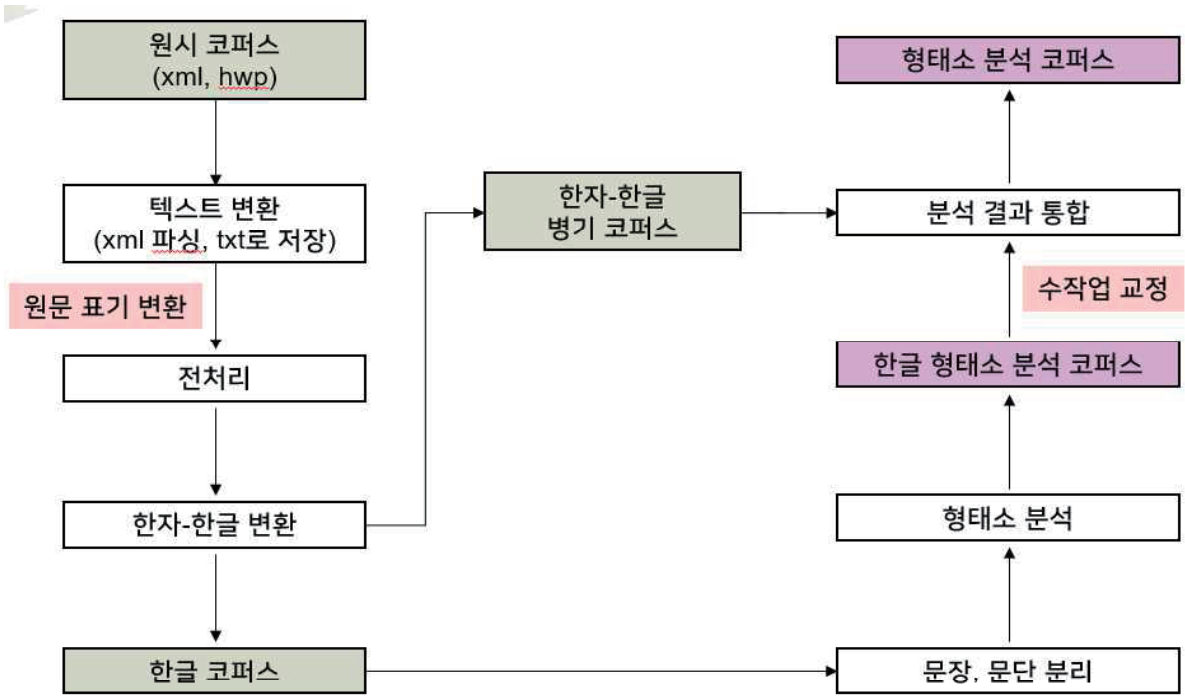
15

근대잡지 코퍼스

No.	자료명	어절수	발행연도
1	대조선독립협회회보	18,570	1896
2	대한자강회월보	89,009	1906
3	서우	94,024	1906
4	태극학보	207,342	1906
5	대한유학생회학보	30,861	1907
6	기호흥학회월보	74,524	1908
7	대동학회월보	61,102	1908
8	대한학회월보	67,631	1908
9	대한협회회보	91,146	1908
10	서북학회월보	122,720	1908
11	호남학보	43,683	1908
12	대한흥학보	100,103	1909
13	개벽	2,703,970	1920
14	동광	1,111,851	1926
15	별건곤	2,022,056	1926
16	삼천리	3,225,007	1929
17	만국부인	18,571	1932
18	삼천리문학	84,307	1938
19	대동아	46,915	1942
20	소년	220,010	1909
21	창조	257,592	1919
22	청춘(1915~)	359,742	1915
23	폐허, 폐허 이후, 백조	132,247	1922
24	학지광	314,050	1914
	총계	11,497,033	

16

한국 근대잡지 코퍼스 구축 과정



17

한국 근대 잡지 코퍼스

- 근대 문예잡지 원문과 전처리
  - 원문

```
<title>소년 제1년 제1권</title>
<date>1908.11.01</date>
<text>
창간사
나는 이 雜誌의 刊行하는 趣旨에 對하여 길게 말씀하지 아니하리라.
그러나 한마디 簡單하게 할 것은
```

  - 한자-한글 병기 코퍼스

```
<title>소년 제1년 제1권</title>
<date>1908.11.01</date>
<text>
창간사
나는 이 雜誌(잡지)의 刊行(간행)하는 趣旨(취지)에 對(대)하여 길게 말씀하지 아니하리라.
그러나 한마디 簡單(간단)하게 할 것은
```

  - 한글 코퍼스

```
<title>소년 제1년 제1권</title>
<date>1908.11.01</date>
<text>
창간사
나는 이 잡지의 간행하는 취지에 대하여 길게 말씀하지 아니하리라.
그러나 한마디 간단하게 할 것은
```

18

한글 표기 변환

✓ 각주, 표, 그림 등도 모두 제거
✓ 띄어쓰기 교정

원문 표기	변환한 현대 표기	비고
에셔	에서	조사 형태 변화
까지	까지	옛글자
하야	하여	어미 형태 변화
잇	있	어간 형태 변화
의게	에게	조사 형태 변화
깃치	같이	조사 형태 변화
가티	같이	조사 형태 변화
조흔	좋은	연철 표기
겟	겠	어미 형태 변화
업서서	없어서	연철 표기
느즌	늦은	연철 표기

19

전처리 예시: '학지광'

예시 1: 원본 입력 텍스트

학지광 제2호、1914년 5월 발행

二號之光이 出現

編輯人

春風이 淡蕩에 百花가 含艶ᄒᆞ야 吾人의 滿目恍惚을 提供함은 自然界의 美術이오、文明이 波動에 萬邦이 咸號ᄒᆞ야 人權問題의 鋒矢가 激烈함은 東西界의 政爭이오、思潮가 一新에 氣魄이 復活ᄒᆞ야 理想的 融合의 祥氣를 頗呈함은 現今 我學界의 狀況이라、茲에 學界의 理想을 綜合 又ᄂᆞᆫ 融和코저 ᄒᆞ야 本誌의 繼續 刊行할 必要가 有함으로 千障萬碍에 不拘ᄒᆞ고 **第二號의 光이 出現된 所以라** ᄒᆞ노라

龍騰虎嘯에 江海가 沸騰ᄒᆞ고 草木이 戰兢ᄒᆞᆷ은 動物界의 雄長이오、麟出鳳儀에 瑞雲이 曄然ᄒᆞ고 韶音이 和諧ᄒᆞᆷ은 聖治代의 徵祥이오、史論이 剛直에 陰譎이 潛形ᄒᆞ고 僞詐가 恐愒ᄒᆞᆷ은 社會上의 制裁力이오、正義에 基礎ᄒᆞ야 公正을 標本ᄒᆞ고 學界 動靜을 隨聞輒載ᄒᆞ야 或 敬意를 表ᄒᆞ며 或 批難을 加ᄒᆞᆷ은 本誌의 特色이라、茲에 愛讀ᄒᆞ신 兄弟의 要求에 應ᄒᆞ야 **第二護의 光이 出現된 所以라** ᄒᆞ노라

20

전처리 예시: '학지광'

예시 2: 표기법, 띄어쓰기 변환 및 마크업 부착

<text>
<doc>학지광제2호</doc>
<date>1914.05</date>
<head>二號之光이 出現</head>
<author>編輯人</author>
春風이 淡蕩에 百花가 含艶하야 吾人의 滿目恍惚을 提供함은 自然界의 美術이오, 文明이 波動에 萬邦이 咸號하야 人權問題의 鋒矢가 激烈함은 東西界의 政爭이오, 思潮가 一新에 氣魄이 復活하야 理想的 融合의 祥氣를 頗呈함은 現今 我 學界의 狀況이라, 茲에 學界의 理想을 綜合 또는 融和코저 하야 本誌의 繼續 刊行할 必要가 有함으로 千障萬碍에 不拘하고 第二號의 光이 出現된 所以라 하노라.
龍騰虎嘯에 江海가 沸騰하고 草木이 戰兢함은 動物界의 雄長이오, 麟出鳳儀에 瑞雲이 ▦然하고 韶音이 和諧함은 聖治代의 徵祥이오, 史論이 剛直에 陰譎이 潛形하고 僞詐가 恐慴함은 社會上의 制裁力이오, 正義에 基礎하야 公正을 標本하고 學界 動靜을 隨聞輒載하야 或 敬意를 表하며 或 批難을 加함은 本誌의 特色이라, 茲에 愛讀하신 兄弟의 要求에 應하야 第二護의 光이 出現된 所以라 하노라.

21

전처리 예시: '학지광'

예시 3: 한자 변환한 텍스트

<text>
<doc>학지광제2호</doc>
<date>1914.05</date>
<head>이호지광이 출현</head>
<author>편집인</author>
춘풍이 담탕에 백화가 함염하야 오인의 만목황홀을 제공함은 자연계의 미술이오, 문명이 파동에 만방이 함호하야 인권문제의 봉시가 격렬함은 동서계의 정쟁이오, 사조가 일신에 기백이 부활하야 이상적 융합의 상기를 파정함은 현금 아 학계의 상황이라, 자에 학계의 이상을 종합 또는 융화코저 하야 본지의 계속 간행할 필요가 유함으로 천장만애에 불구하고 제이호의 광이 출현된 소이라 하노라.
용등호소에 강해가 비등하고 초목이 전긍함은 동물계의 웅장이오, 인출봉의에 서운이 ▦연하고 소음이 화해함은 성치대의 징상이오, 사론이 강직에 음휼이 잠형하고 위사가 공명함은 사회상의 제재력이오, 정의에 기초하야 공정을 표본하고 학계 동정을 수문첩재하야 혹 경의를 표하며 혹 비난을 가함은 본지의 특색이라, 자에 애독하신 형제의 요구에 응하야 제이호의 광이 출현된 소이라 하노라.

22

〈개벽〉 텍스트

```
K======          ma_1920_013_002.tag      ======>
〈title〉
개벽 제2호
〈/title〉
〈date〉
1920-07-25
〈/date〉
〈text〉
〈head〉
01236201         권두시                   권두/NNG+시/NNG
〈/head〉
〈p〉
〈s〉
01236202         사람은                   사람/NNG+은/JX
01236203         누구나                   누구/NP+나/JX
01236204         죽을                     죽/VV+을/ETM
01236205         때까지                   때/NNG+까지/JX
〈/s〉
〈/p〉
〈p〉
〈s〉
01236206         무언지                   무엇/NP+이/VCP+ㄴ지/EM
01236207         늘                       늘/MAG
01236208         밤낮                     밤낮/MAG
01236209         求(구)함이               구하/VV+ㅁ/ETN+이/JKS
01236210         있다.                    있/VA+다/EM+./SF
〈/s〉
〈/p〉
〈p〉
〈s〉
01236211         圓滿(원만)코             원만/NNG+하/XSV+고/EM
01236212         無謬(무류)한             무류/NNG+하/XSV+ㄴ/ETM
```


오류 수정: 라인 대체

```
--------------------------------------------------------------------------------
01320056         過(과)코자               과하/VV+고자/EM
01320060         志(지)로소이다.          지/NNG+이/VCP+로소이다/EM+./SF
01320065         迎(영)하여               영하/VV+아/EM
01320068         謀(모)코자               모하/VV+고자/EM
01320082         張甲福(장갑복)에게       장갑복/NNP+에게/JKB
01320092         貽(이)함에               이하/VV+ㅁ/ETN+에/JKB
01320109         迷(미)치                 미하/VV+지/EM
01320114         愼(신)하여               신하/VV+아/EM
01320117         防(방)하여               방하/VV+아/EM
01320123         墮(타)치                 타하/VV+지/EM
01320131         月(월)로                 월/NNG+로/JKB
01320137         3島(도)                  3/SN+도/NNG
01320142         奪(탈)한                 탈하/VV+ㄴ/ETM
01320146         覆轍(복철)이             복철/NNG+이/JKS
01320149         殿下(전하)는             전하/NNG+는/JX
01320161         知(지)키                 지하/VV+기/ETN
01320167         知(지)하는               지하/VV+는/ETM
01320176         何邊(하변)에             하변/NNG+에/JKB
01320189         來(내)하여               내하/VV+아/EM
01320190         我(아)의                 아/NP+의/JKG
01320192         咬(교)하여도             교하/VV+아도/EM
01320195         感(감)치                 감하/VV+지/EM
01320199         我(아)를                 아/NP+를/JKO
01320200         咬(교)함인지도           교하/VV+ㅁ/ETN+이/VCP+ㄴ지/EM+도/JX
01320219         殿下(전하)는             전하/NNG+는/JX
01320228         腹心股肱(복심고굉)된     복심고굉/NNG+되/XSV+ㄴ/ETM
01320238         保(보)합니까.            보하/VV+ㅂ니까/EM+./SF
```

오류 수정: 패턴 1

구루마/NNG+꾼/NNG	구루마꾼/NNG
구루마/NNG+꾼/XSN	구루마꾼/NNG
구주/NNG+대전/NNG	구주대전/NNP
구주/NNG+전란/NNG	구주/NNP+전란/NNG
국민/NNG+교육/NNG	국민교육/NNG
국민전체/NNG	국민/NNG+전체/NNG
국제간/NNG	국제/NNG+간/NNB
국제연맹/NNP+협회/NNG	국제연맹협회/NNP
군비축소/NNG	군비/NNG+축소/NNG
균열/NNG+적/XSN	균열적/NNG
근로대중/NNG	근로/NNG+대중/NNG
근화/NNP+학교/NNG	근화학교/NNP
금강/NNP+상회/NNG	금강상회/NNP
기기/NNG+괴괴/NNG	기기괴괴/NNG
기에/NNG+가/NNG	기에가/NNG
깜장/NNG+전/NNG	깜장전/NNG
나가놀/NNG	나가놀/VV+ㄹ/ETM
내량/NNG+박물관/NNG	내량박물관/NNP
노국/NNG	노국/NNP
노농/NNG+파/NNG	노농파/NNG
노동/NNG+심/NNB	노동심/NNG
노동/NNG+심/NNG	노동심/NNG
노동/NNG+심/XSN	노동심/NNG
노동문제/NNG	노동/NNG+문제/NNG

오류 수정: 패턴 2

學(학)한	학하/VV+ㄴ/ETM
學(학)함이	학하/VV+ㅁ/ETN+이/JKS
學博士(학박사)가	학박사/NNG+가/JKS
學問硏究所(학문연구소)임을	학문/NNG+연구소/NNG+이/VCP+ㅁ/ETN+을/JKO
學海(학해)에	학해/NNG+에/JKB
學理的(학리적)이었다.	학리적/NNG+이/VCP+었/EP+다/EM+./SF
學生(학생)다웁게	학생/NNG+다웁/XSA+게/EM
宅(택)에서	댁/NNG+에서/JKB
守(수)키	수하/VV+기/ETN
安分(안분)이라면	안분/NNG+이/VCP+라면/EM
安商浩(안상호),	안상호/NNP+,/SP
安州(안주)에	안주/NNP+에/JKB
安州郡守(안주군수)	안주/NNP+군수/NNG
安心田(안심전)	안심전/NNP
安柄一商店(안병일상점)인데	안병일상점/NNP+이/VCP+ㄴ데/EM
安氏(안씨)께	안/NNP+씨/NNB+께/JKB
安福派(안복파)	안복파/NNP
宋今璇(송금선)씨가	송금선/NNP+씨/NNB+가/JKS
宋今璇(송금선)씨는	송금선/NNP+씨/NNB+는/JX
宋公(송공)이	송/NNP+공/NNB+이/JKS
宋古下(송고하)요	송고하/NNP+이/VCP+요/EM
完全(완전)이	완전이/MAG

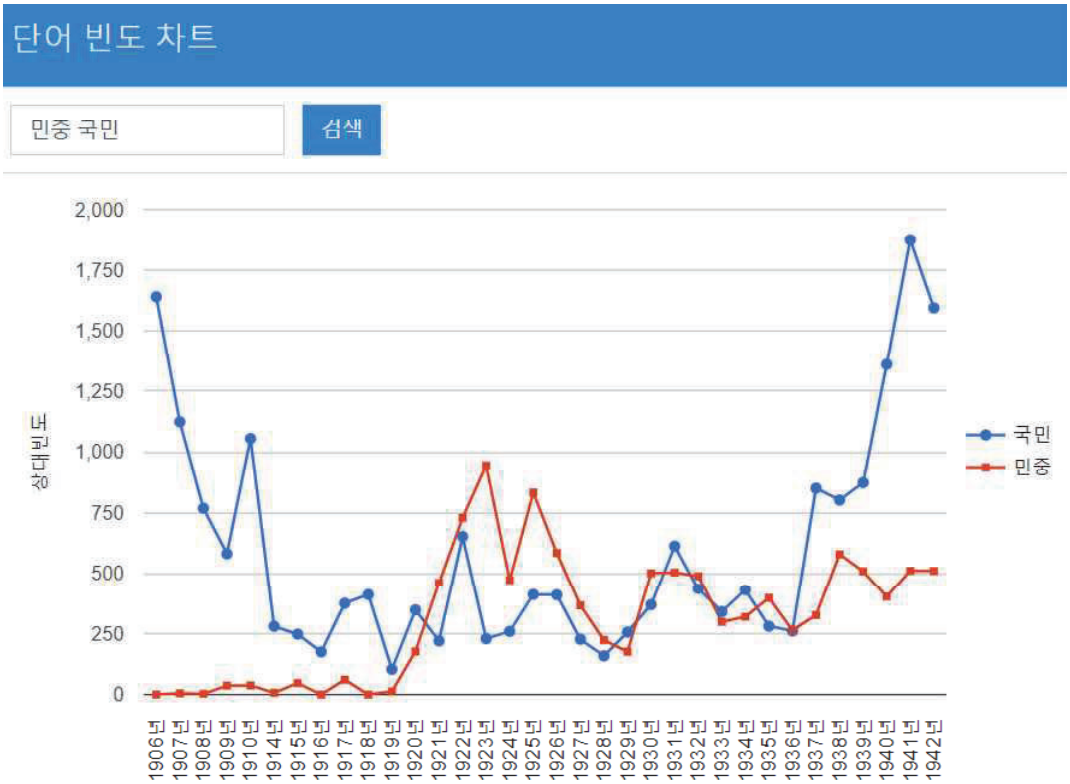
어휘개념사 연구의 전제

- 어휘사와 개념사의 관계
  - 국어사의 관점에서는 어휘나 단어의 역사에 내재하는 언어 변화 원리를 탐구하는 데 중점을 둔다면, 개념사에서는 역사적 의미를 갖는 어휘의 개념 형성과 변화에 관심을 기울임(허재영 2024).

- 정량적 기법의 적용과 분석
  - 어휘 사용 빈도의 양상, 문맥의 변화 등에 대한 계량적 분석
  - 분석 결과에 대한 어휘-개념사적인 해석과 상상력

27

단어 사용 빈도 예시



28

나오며

- 인공지능과 관련한 기술이 크게 주목을 받는 시기에 현대 이전의 텍스트 데이터에 대한 관심을 제고할 필요가 있음

- 한국어의 역사적 변화뿐 아니라 개념사적인 탐색을 위해 근현대 자료에 대한 계량적 접근이 필요하며, 이를 위해 해당 텍스트를 적절히 가공하고 이를 공유하는 과정이 요구됨

- 이 연구에서는 근현대 잡지를 대상으로 하여 표기법, 한자 문제, 분석 오류 수정 방법 등에 대해 구체적인 처리 방법을 제안해 보았음
  - 향후 집단지성에 의해 더욱 품질이 향상된 자료로 발전할 수 있을 것으로 기대

2024 한국코퍼스언어학회 가을 전국학술대회

인간과 기계의 언어 소통

Session 2

복합양식 텍스트 교육의 효과 분석 :
디지털 콘텐츠 크리에이션에 관한 학습자 인식과
텍스트 구성 전략 논의를 중심으로 (주민재, 명지대)

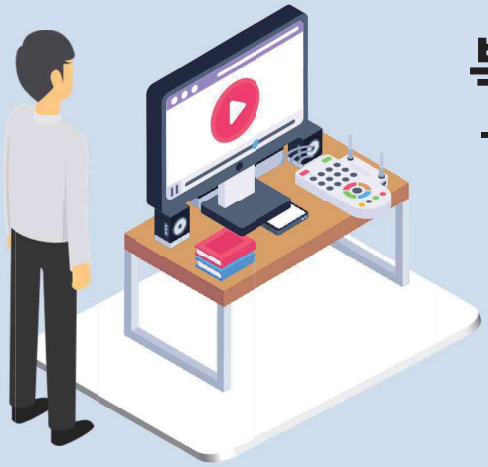
—

LLM을 활용한 한국어 글쓰기 평가와
생활기록부 생성 모델의 실제 (임경태, 서울과기대)

—

학습자 글쓰기 자동 평가 모델: 피쳐 기반 모델 (최지명, 이화여대)

KACL 한국코퍼스언어학회 2024년 가을 학술대회



복합양식 텍스트 교육의 효과 분석
- 디지털 콘텐츠 크리에이션에 관한 학습자 인식과 텍스트 구성 전략 논의를 중심으로

주민재(명지대학교)

INDEX

1
디지털 콘텐츠의 생산과 복합양식 텍스트의 관계

1. 디지털 콘텐츠 생산과 복합양식 텍스트의 관계

디지털 콘텐츠의 생산과 소비

다양한 디지털 플랫폼을 활용한 정보의 생산과 소비 경향이 더욱 강해지는 추세

1. 비판적 평가, 창의적 활용 능력을 향상시키는 방안에 대한 체계적인 인식과 교육 방법 및 교육 콘텐츠 개발
2. 정보의 양과 복잡성의 증가로 인해 신뢰성 평가와 윤리적 활용 능력의 중요성에 대한 인식 강화 필요성

대다수의 대학에서 디지털 리터러시 교육 필요성 인식 → 관련 교육 프로그램들은 여전히 부족

1. 디지털 콘텐츠 생산과 복합양식 텍스트의 관계

복합양식 텍스트

디지털 콘텐츠의 생산과 소비에서 중요한 역할

→ 다양한 미디어 요소를 결합하여 정보를 더욱 명확하게 전달할 수 있는 수단

- 복합양식 텍스트를 이해하고 생산하는 능력
  : 학습자들이 디지털 환경에서 효과적인 소통과 학습에 필수적 요인
- 복합양식 텍스트를 활용한 교육 콘텐츠 개발 정도
  : 매우 제한적이고 관련 교육에 통합하려는 시도가 부족한 상황

디지털 리터러시와 복합양식 텍스트에 대한 심층적 이해 기반 교육 필요

2 디지털 플랫폼과 콘텐츠 크리에이션

2. 디지털 플랫폼과 콘텐츠 크리에이션

◆ 디지털 리터러시(digital literacy)

디지털 환경에서 정보 검색, 평가, 활용, 창의적인 문제 해결을 가능하게 하는 능력

1. 최근 디지털 콘텐츠를 생산하여 자신의 전달하려는 정보의 전달력을 최대화하는 능력까지 포괄
2. 정보의 신뢰성을 평가하고 윤리적 의사결정을 내리는 것과 같은 고차원적 사고 능력 의미

학습자가 디지털 환경에서 효과적인 학습과 의사소통의 능동적 수행에 핵심적인 역할

2. 디지털 플랫폼과 콘텐츠 크리에이션

◆ 복합양식 텍스트(multimodal text)

문자, 이미지, 소리, 영상 등 다양한 양식(mode)들이 결합하여
하나의 의미를 전달하는 텍스트

디지털 플랫폼의 발전 : 복합양식 텍스트의 대중화를 촉진하는 매개체

- 소셜 미디어 플랫폼의 출현

: 텍스트, 이미지, 동영상 등을 결합하여 자신만의 독창적인 콘텐츠 생성 환경 제공

: 복합양식 텍스트 제작과 공유의 용이성

➜ 복합양식 텍스트는 소셜 미디어 플랫폼이 디지털 커뮤니케이션의 주요 형태로 자리 잡는데 핵심적인 역할

2. 디지털 플랫폼과 콘텐츠 크리에이션

인스타그램 릴스 → **이미지+텍스트**
사용자들이 자신의 이야기를 시각적으로 전달

유튜브 쇼츠 → **짧은 동영상**
간단한 메시지를 효과적으로 전달

디지털 플랫폼은 복합양식 텍스트를 직관적이고 효율적으로 생산 및 소비 환경 제공
디지털 시대의 커뮤니케이션 방식에서 중요한 역할

2. 디지털 플랫폼과 콘텐츠 크리에이션

디지털 플랫폼

디지털 플랫폼은 10-20대 일상 커뮤니케이션에서 빼놓을 수 없는 역할

정보 검색, 콘텐츠 소비, 사회적 상호작용, 자기 표현 등의 다양한 목적으로 활용

유튜브
- 자신만의 동영상 콘텐츠를 제작하고 공유하며 상호작용하는 공간
- 참여 문화가 구체적으로 실현되는 플랫폼

쇼츠 틱톡
- 정보 전달의 새로운 형태
- 짧고 간결한 콘텐츠를 통해 빠르게 정보를 전달하고 수용

2. 디지털 플랫폼과 콘텐츠 크리에이션

콘텐츠 크리에이션(contents creation)

디지털 플랫폼에서 다양한 양식들을 결합하여 새로운 콘텐츠를 창출하는 과정

- 복합양식 텍스트는 메시지의 효과적 전달에 필수적인 요소
  → 시각적 요소와 청각적 요소를 결합한 콘텐츠
  : 사용자들에게 더 직관적이고 감성적으로 다가갈 수 있기 때문
  : 최종적으로는 메시지의 전달력의 강화로 귀결

2. 디지털 플랫폼과 콘텐츠 크리에이션

콘텐츠 크리에이터

- 디지털 플랫폼에서 단순한 사용자 창작 콘텐츠를 생산하는 수준을 뛰어넘고
- 수용자들에게 상당한 영향력을 행사하는 인플루언서로 인식
- 복합양식 텍스트를 활용하여 강력한 메시지를 전달하고 수용자들의 행동과 인식 변화에 영향

시각적 요소, 음악, 문자 등이 결합된 콘텐츠

수용자들에게
복합양식 텍스트의 중요성과 효과를
자연스럽게 인식

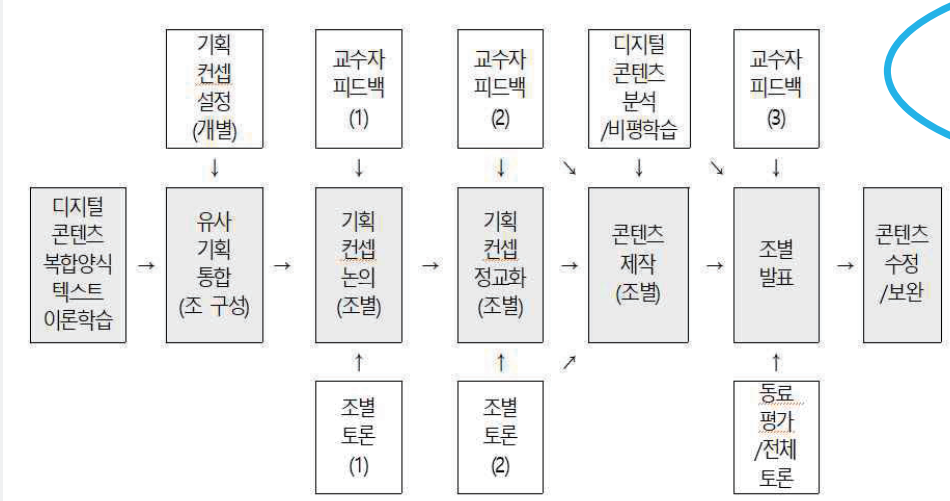
복합양식 텍스트는 정보 전달만이 아니라
감정적이고 심미적인 경험을 제공하는 도구로 활용

수용자들의 기대치를 변화시키는 핵심 요인

3

문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석



[그림 1] 문제중심학습 기반 콘텐츠 제작 수행 개념도

문제중심학습 기반 콘텐츠 제작
= 디지털 콘텐츠 제작 활동

→ 중심 활동 + 보조 활동

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석

◆ 콘텐츠 제작 수행 과정을 문제중심학습에 기반을 두는 이유

● 문제중심학습(PBL)

학습자가 실제 문제 상황에서 스스로 문제를 정의하고
해결 방안을 모색하는 과정에서 학습이 이루어지도록 설계된 교수법

- 문제중심학습 기반 설계
  → 학습자 중심 교육 방법이 창의성과 협력이 중요한 디지털 콘텐츠 제작 과정에 효과적
  → 학습자가 이론적 지식을 실제 상황에 적용할 수 있는 능력 함양에 유리

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석

● 연구참여자 정보

순번	학년	인원 수(명)
1	1학년	15
2	2학년	8
3	3학년	5
4	4학년	7
	합계	35

- 총 15주, 30차시 진행
- 피드백을 수행하면서 지속적 관찰
- 학습자들에게 10개의 반구조화된 질문들로 서면 인터뷰를 수행
- 일부 연구참여자들에게는 대면 인터뷰를 추가 진행

①수강이 미친 영향
②콘텐츠 제작 과정
③문제중심학습
(조별 활동)

에 대한 인식으로 범주화

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석

연구참여자 제작 디지털 콘텐츠 사례

- 인터뷰 방식(쇼츠)
- 웹툰 방식(쇼츠)
- 정보 분석 방식

- 특성
  - ① 1분 내외의 쇼츠 형식 제작 선호
  - ② 웹툰 방식 등 기존 영상 방식에서 탈피
  - ③ 양식들을 복합화한 텍스트 일반화 : 사진, 문자(자막), 인터뷰 시 얼굴 모형 활용 등

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석

서면 인터뷰 분석 방법

- **분석 방법**

질문별로 답변을 1차 분류 ➜ 유사한 답변 내용들을 묶어 2차 분류➜ 내용별로 코딩 ➜ 전체 분석

- 분석 범주
  - ① <콘텐츠 크리에이터의 이해와 콘텐츠 개발> 교과목 수강이 자신의 콘텐츠 크리에이션에 미친 영향에 대한 인식
  - ② 디지털 콘텐츠 제작 과정에 대한 인식
  - ③ 문제중심학습(조별 활동)에 대한 인식

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석 _ 교과목 수강이 자신의 콘텐츠 크리에이션에 미친 영향에 대한 인식

- 콘텐츠 분석 및 기획 능력의 강화 / 창의력과 문제 해결 능력의 배양

"독자적인 콘텐츠 개발"

'채널은 고급 레스토랑의 메뉴와 같아야 한다'는 교수님의 말씀이 지금도 기억에 남는다. 콘텐츠를 제작할 땐 채널 만의 시그니처 메뉴가 있어야 하고, 알맞은 메뉴를 구성해 놓아야 시청자 타겟팅이 성공한다는 당연한 사실을 본 수업을 통해 새삼 깨달았다.

연구참여자 8

콘텐츠를 제작할 때 그것을 소비하는 소비자들의 니즈를 분석한다든가 알맞은 플랫폼을 선택하는 것 등이 생각이 나는데 광고를 만들 때 소비자의 니즈를 분석하고 애드 브리프를 작성하는 등 하는 부분이 생각나서 어떻게 보면 광고도 콘텐츠 이기에 내 진로에 있어서도 많은 도움이 됐다고 생각한다.

연구참여자 20

"문제 해결 능력 향상"

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석 _ 교과목 수강이 자신의 콘텐츠 크리에이션에 미친 영향에 대한 인식

- 비판적 분석 능력의 향상

"전략적 의도 설정 및 분석력 향상"

평가하는 능력보다는 분석하는 능력이 향상된 거 같다. 콘텐츠를 제작하고 나서 '빵빵이의 일상'에 올라온 콘텐츠들을 다시 보니 정말 대단하다는 생각이 들었다. 콘텐츠 제작 전에 볼 때는 영상에 나오는 인물들이 너무 생생하고 역동적이어서 장면마다 인물들의 모습을 새로 그린 줄 알았다. 하지만 콘텐츠를 제작하고 나서 보니 같은 그림을 계속 사용했지만 편집의 기술로 같은 그림이 아닌 것처럼, 매 장면마다 다른 것처럼 보이게 했다는 것을 알 수 있었다.

연구참여자 13

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석 _ 디지털 콘텐츠 제작 과정에 대한 인식

- 수용자 요구와 선호도 파악의 중요성 / 기획과 전략 수립 과정의 중요성

"수용자 요구 파악과 적용"

콘텐츠를 제작하는 데에 중요한 것은 스스로의 기호와 대중의 선호 사이에서 균형을 잡는 일인 것 같았다. 결과적으로 콘텐츠는 누군가에게 보이기 위한 것이므로 크리에이터 개인의 일방적인 욕구 배출구가 되어서는 안 되겠다고 생각했다.

연구참여자 35

"기획, 전략 수립 능력 향상"

기획을 하기 위해서 레퍼런스를 서칭하던 와중 가성비 맛집을 추천해주는 롱폼이나 블로그 등은 많았지만 숏츠는 없었다는 것을 알고 이를 파악하여 숏츠의 특성을 활용한 짧은 영상으로 필요한 정보만 딱딱 보여주면 타겟팅한 2,30대들의 니즈를 충족시켜 줄 수 있을 것 같다는 생각에 기획을 하게 되었습니다.

연구참여자 6

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석 _ 디지털 콘텐츠 제작 과정에 대한 인식

- 세부 기술의 중요성 인식

"편집 기술의 중요성 인식"

편집기술을 배웠더라면.. 싶은 순간이 정말 너무 너무 많았다. 콘텐츠를 제작하며 처음 해본 편집이기에 프로그램을 어떻게 사용하는지 조차도 모르는 상태였다. 일일이 유튜브에서 찾아보며 콘텐츠를 제작함과 동시에 편집을 배우려고 하니 시간이 너무 빠듯했고, 원하는 퀄리티가 나오지 않았다. 편집기술을 먼저 배웠으면 좀 더 빠르고, 퀄리티 있는 콘텐츠를 만들 수 있지 않았을까라는 생각이 든다.

연구참여자 13

"콘텐츠 제작 기술 필요성 인식"

영상 편집 기술과 일러스트나 포토샵을 조금이라도 더 일찍 배워서 기본기를 탄탄하게 다지고 경험을 많이 했다면 영상을 편집하는 과정에서 훨씬 작업 속도가 빨랐을 것이다.

연구참여자 3

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석 _ 문제중심학습(조별 활동)에 대한 인식

- 협업을 통한 창의적 아이디어 도출

혼자서 콘텐츠 제작을 생각했을 때는 항상 혼자서 할 수 있는 촬영과 제작의 한계를 정해두고 아이디어를 생각하게 되었는데 조원들과 함께 이야기를 하니 혼자보다 더욱 다양한 아이디어가 나왔고 유튜브 콘텐츠 제작에 있어 영상 제작 뿐만 아니라 진행되어야 하는 썸네일 제작, 채널 프로필 구성, 영상 제목과 설명 등 부가적인 일 또한 함께 진행하고 객관적으로 수정할 수 있어 혼자 구상한 처음보다 더 나은 완성작이 나왔다고 생각한다.

연구참여자 16

"조원들과의 협력의 시너지 효과"

"협업을 통한 다양한 시도 가능, 완성도 향상"

왜냐하면 혼자 하기 힘든 작업을 함께 협동하여 해결할 수 있고, 주관적인 시선에 갇히지 않은 채로 더 많은 시도를 해볼 수 있기 때문입니다. 또 끊임없이 상의가 이루어져 전체적인 완성도가 향상됩니다.

연구참여자 18

3. 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석 _ 문제중심학습(조별 활동)에 대한 인식

- 의사소통의 중요성 / 피드백의 효용성

하지만 (조별 활동이) 소통에 있어서는 문제가 발생할 수 있고 시간이 충분하지 않다 보니 의견 충돌 등의 돌발 상황이 생겼을 때 신속하게 대처하지 못하고 오히려 콘텐츠의 질이 떨어질 수 있다.

연구참여자 17

"의사소통 및 의견 조율의 중요성

"피드백의 효용성"

자신이 만든 콘텐츠를 자신이 보는 것은 객관적이지 못한 부분이 분명히 있다고 생각한다. 데이트 코스 블로그를 제작하면서 독자들이 읽을 수 있도록 만드는 방법에 대해 피드백을 받았었는데 생각하지도 못했던 부분을 개선할 수 있어서 도움이 많이 되었던 것 같다.

연구참여자 12

4

문제중심학습 기반 디지털 콘텐츠 제작 교육 결과 정리

4. 문제중심학습 기반 디지털 콘텐츠 제작 교육 결과 정리

연구참여자들 인식 변화

디지털 콘텐츠 제작

- 시청자와의 소통과 감정적 연결을 전략적으로 이끌어내는 복합적인 과정이라는 인식 강화
- 다양한 요소들이 상호 보완적으로 작용, 하나의 강력한 메시지를 전달하는 과정으로 인식
- 창의적 사고와 전략적 접근이 필수적인 요소임을 명확하게 인식
- 복합양식 텍스트의 다양한 요소들을 통합적으로 활용하는 과정에서 시청자의 기대와 반응에 대한 예측에 부합하는 방향으로 메시지 전달 전략 구사

4. 문제중심학습 기반 디지털 콘텐츠 제작 교육 결과 정리

연구참여자들 능력 향상

문제중심학습 적용 결과

- 콘텐츠 제작을 위해 자기 주도적으로 문제 정의 및 해결에 필요한 구체적 방안 모색
  - ➜ 단순한 지식 습득을 넘어 실제 문제 해결 능력을 강화할 수 있도록 유도
- 협력의 중요성 인식 및 각자의 아이디어와 차별적 능력 결합을 통해 문제 해결 경험
  - ➜ 문제해결능력 향상
- 디지털 콘텐츠 크리에이션에 대한 학습자들의 인식 강화
- 문제중심학습 방식과 조별 활동을 통해 학습자들의 문제 해결 능력의 심화, 발전

감사합니다

복합양식 텍스트 교육의 효과 분석
- 디지털 콘텐츠 크리에이션에 관한 학습자 인식과 텍스트 구성 전략 논의를 중심으로
토론문

윤영민(인하대학교)

주민재 선생님의 발표 감사히 들었습니다.
디지털 콘텐츠의 리터러시뿐만 아니라 크리에이션이라는 부분에까지 착목하신 연구와 분석, 정말 흥미롭게 읽었고 많은 공부가 되었습니다.

한 가지 우려되는 부분은 제가 선생님과 같이 해당 분야의 연구를 지속적으로 해 온 사람은 아니라는 것입니다. 이에 선생님의 의도와 의중을 잘못 파악하고 이해도 낮은 질문을 드림으로써 무례를 범하는 것이 아닌지 조심스럽습니다.
그럼에도 불구하고 선생님의 옥고와 발표를 통하여 궁금한 몇 가지 내용을 여쭙고 확인하는 것으로 오늘 토론자로서의 소임을 갈음하고자 합니다.

먼저, 다양한 디지털 플랫폼을 활용한 정보의 생산과 소비 경향이 더욱 강해지는 추세라고 하셨으나 이와 같은 양상과 추세는 이른바 사회 관계망 서비스의 등장, 발전과 함께 꾸준하고 지속적으로 강화되고 다양화되어왔다는 것이 저의 관견입니다.
이와 관련하여 복합양식 텍스트를 활용한 교육 콘텐츠 개발 정도가 매우 제한적이고 관련 교육에 통합하려는 시도가 부족한 상황이라고 지적하셨는데 이는 선생님의 진단이신지 교육 현장 또는 선행 연구를 통한 종합적인 판단이신지 궁금합니다.

둘째, 문제중심학습 기반 디지털 콘텐츠 제작 교육 과정 연구 분석에서 조별 활동에 대한 인식을 보면 결국 컨텐츠의 선정, 전달할 내용에 대한 효과적인 방법 모색, 협업 및 이를 위한 의사소통과 피드백이 무척 중요하게 작용하고 있는 것을 알 수 있는데, 혹시 이 외에 다른 의견(인식)에는 어떤 것들이 있었는지 궁금합니다.

마지막으로 디지털 콘텐츠 크리에이션의 일면으로 그 윤리의식도 굉장히 중요한 요소라고 판단됩니다. 이를 함께 함양해 갈 수 있는 효과적인 방안이 무엇인지 여쭙고자 합니다.

감사합니다.

LLM을 활용한 한국어 글쓰기 평가와 생활기록부 생성 모델의 실제

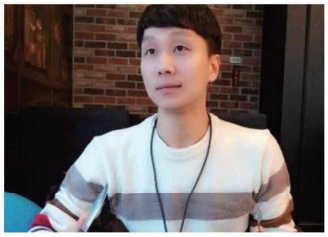
v 1.0

서울과학기술대학교 임경태
jujbob@gmail.com

Contents

연구개요

01 발표자 소개 및 연구 출처



Education

- **Doctoral: École Normale Supérieure (ENS),** Paris-France, Major : Language Science, Multilingual Dependency Parsing and Natural Language Understanding. Advisor : Prof. Thierry Poibeau (LATTICE, ENS)
- **Master:** Korea Advanced Institute of Science and Technology (KAIST), Korea, Major : Computer Science, Natural Language Processing. Advisor : Prof. Key-Sun Choi.
- **Bachelor: Dankook University,** South Korea. Computer Science.

연구개요

01 발표자 소개 및 연구 출처

Natural Language Engineering (2022), 1–23
doi:10.1017/S1351324922000298

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Neural automated writing evaluation for Korean L2 writing

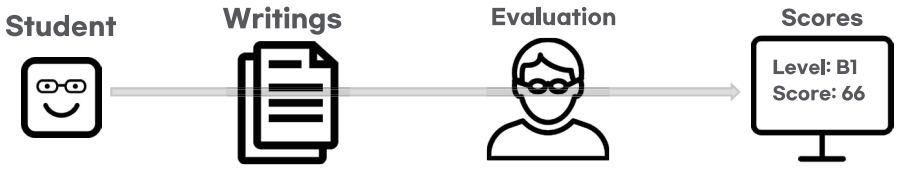
KyungTae Lim[1,†], Jayoung Song[2†] and Jungyeul Park[3,4,*]

[1]Hanbat National University, Daejeon 34158, South Korea,[2]Pennsylvania State University, State College, PA 16801, USA,[3]The University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and [4]University of Washington, Seattle, WA 98195, USA
*Corresponding author. E-mail: jungyeul@mail.ubc.ca

연구개요

02 Neural Automated Writing Evaluation (AWE) 이란?

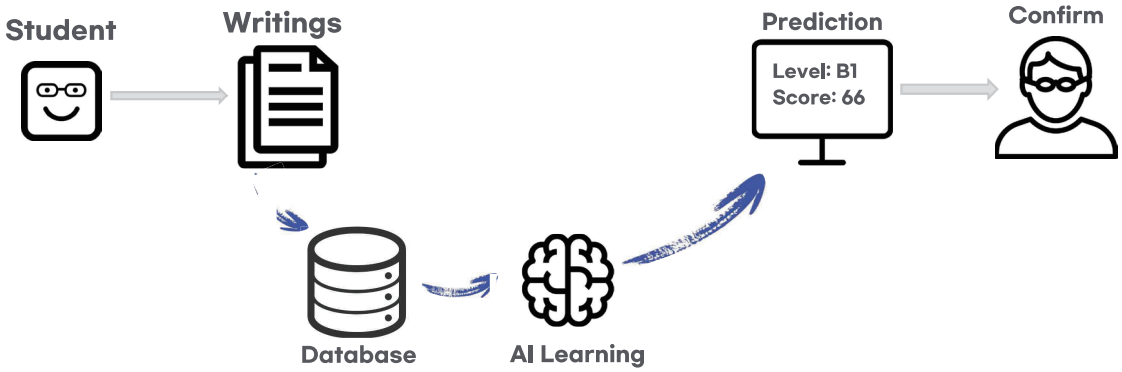
- AWE: 인간의 작문을 AI가 평가하는 기술
- 현실 세계의 글쓰기 평가 방식



연구개요

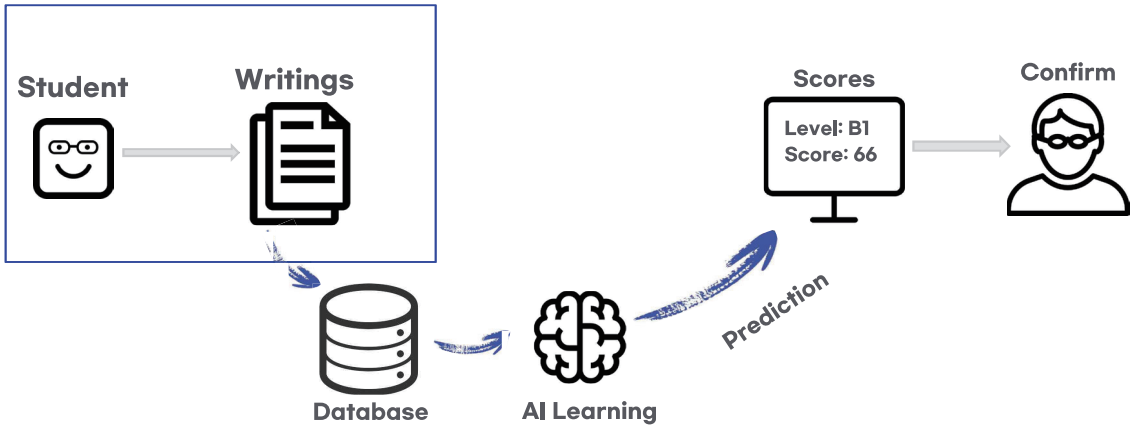
02 Neural Automated Writing Evaluation (AWE) 이란?

- AWE: 인간의 작문을 AI가 평가하는 기술
- 내가 하고싶은 자동 글쓰기 평가 방식



03 데이터 수집

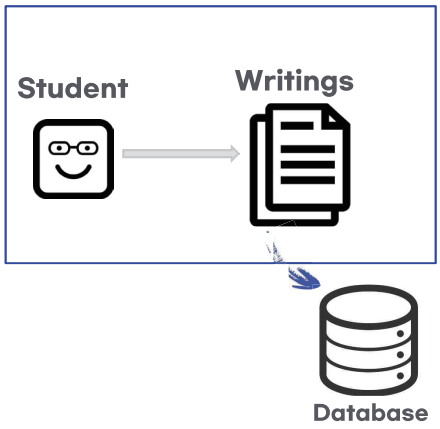
- AWE를 위해 어떻게 데이터를 구성해야 할까?



연구개요 - 데이터 수집

03 데이터 수집

- 경희대 한글 어학당에서 외국인이 작성한 한글 작문 문서를 데이터화




```
A100007_v02.xml
<lovol>초급1</lovol>
<nationality>중국</nationality>
<gender>여자</gender>
<term>기말고사</term>
<date>2013년 가을</date>
<topic>주말 이야기</topic>
<score>70</score>
<p><s>지는 일요일에 도시관에서 갔습니다.</s>
<s>저는 친구하고 도서관에서 갔습니다.</s>
<s>도서관에 책을 있었습니다.</s>
<s>도서관에서 공부했습니다.</s>
<s>저는 책을 읽었습니다.</s>
<s>그래서 일요일에 재미있습니다.</s></p>
```

Figure 1. Example of the Korean learner corpus: `<level>` = Level 1, `<nationality>` = Chinese, `<gender>` = female, `<term>` = final examination, `<date>` = Fall 2013, `<topic>` = my weekend, and `<score>` = 70. The present example of Korean writing can roughly be translated into *I went to the library on Sunday. I went to the library with a friend. There was a book in the library. I studied in the library. I read a book. So it's fun on Sunday.*

03 데이터 수집

- 경희대 한글 어학당에서 외국인이 작성한 한글 작문 문서를 데이터화
- 총 4094 개의 작문 데이터 수집

  - 6개의 주제에 대해 6개의 레벨: (A1, A2, B1, B2, C1, C2)과 및 점수가 존재

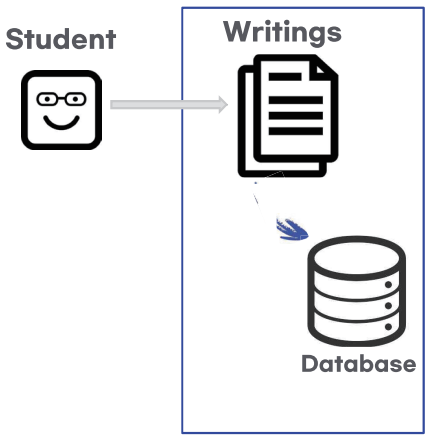
Table 1. Examples of most frequent prompts and their number of instances in the learner corpus

Prompts	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	(total)
My weekend	261	–	–	–	–	–	261
Seasons and weather in my country	–	171	–	–	–	–	171
The day that I remember the most	–	–	129	–	94	–	223
My future plans	24	96	–	–	–	–	120
My hobby	–	50	144	–	–	–	194
Other prompts	396	828	963	260	389	289	3125
(total)	681	1145	1236	260	483	289	4094

There are over 100 prompts which are used by only a small number of writing examples ("Other prompts" row in Table 1). Most prompts are only for specific proficiency levels, such as *My weekend*, *Seasons and weather in my country*.

04 데이터 가공

- 수집된 데이터를 딥러닝에 활용할 수 있도록 자질(Feature)을 추출하기 위해 데이터 가공
- 사람은 작문의 어떤 feature를 활용해서 평가를 할까??



04 데이터 가공

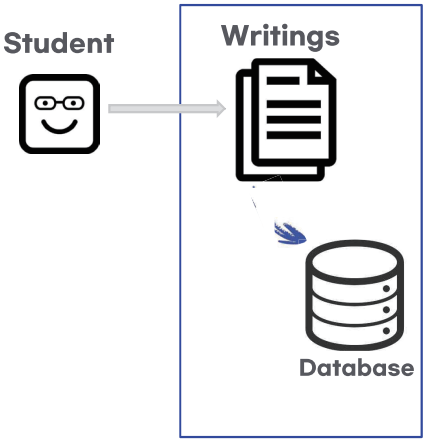
사람은 작문의 어떤 feature를 활용해서 평가를 할까??

- (Quantitative Complexity features): 문장길이, 문장을 구성하는 형태소 타입의 비율 등
- (Syntactic Complexity feature): 문장이 얼마나 복잡한가?, 의 활용
- (Fluency features): 문장이 얼마나 유창한가? 문법적 복잡도는?, 적절한 단어를 활용하는가?
- (Accuracy features): 문법적으로 정확한가?

연구개요 - 데이터 수집 - 데이터 가공

04 데이터 가공

- Quantitative Complexity feature 추출 (형태소 분석)
- 형태소 분석기를 직접 학습시켜 한국어 형태소 분석기 구현 (2021년 최고성능)

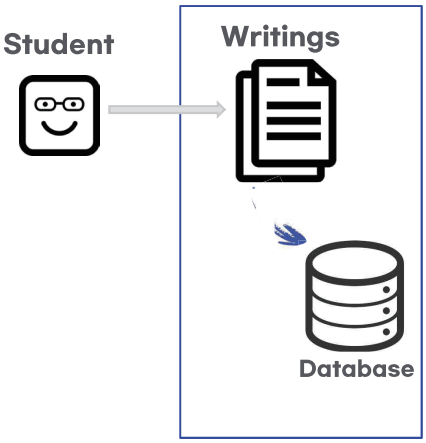


hajiman	('however')	*hajiman*/MAJ
bili	('Billy')	*bili*/NNP
ssi-hago	('Mr.-CONJ')	*ssi*/NNB+*hago*/JKB
naoko	('Naoko')	*naoko*/NNP
ssi-neun	('Ms.-TOP')	*ssi*/NNB+*neun*/JX
modu	('all')	*modu*/MAG
sajingi-ga	('camera-NOM')	*sajingi*/NNG+*ga*/JKS
eobs-eoss-eoyo.	('do_not_have-PAST-DECL')	*eobs*/VA+*eoss*/EP+*eoyo*/EF+./SF

Figure 2. Example of Sejong corpus-style POS tagging analysis: MA{J|G} are for adverbs, NN{P|D|G} for nouns, J{KB|X|KS} for postpositions, E{P|F} for verbal endings, VA for adjectives, and SF for punctuations.

04 데이터 가공

- Quantitative Complexity feature 추출 (문장 길이, 형태소 길이, 형태소의 중복 비율)



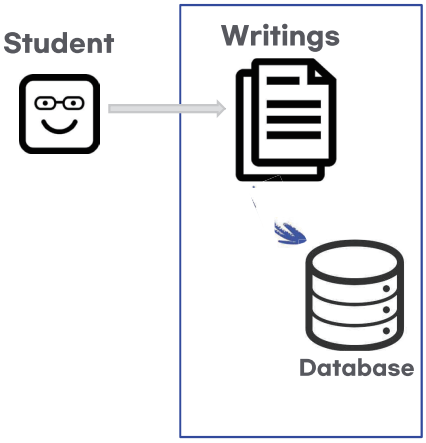
(1) a. 하지만 빌리 씨하고 나오코 씨는 모두 사진기가 없었어요.
hajiman bili ssi-hago naoko ssi-neun modu sajingi-ga eobs-eoss-eoyo.
However, Billy Mr.-CONJ Naoko Ms.-TOP all camera-NOM do_not_have-PAST-DECL.
"However, Mr. Billy and Ms. Naoko, both of them do not have a camera."
b. *hajiman bili ssi-hago naoko ssi-neun modu sajingi-ga eobs-eoss-eoyo.* (# of tokens by word = 8)
c. *hajiman bili ssi -hago naoko ssi -neun modu sajingi -ga eobs -eoss -eoyo.* (# of tokens by a morpheme = 13, punctuations excluded)

A type/token ratio is calculated using $\frac{\#\text{ of types}}{\#\text{ of tokens}}$, where the number of types represents the

연구개요 - 데이터 수집 - 데이터 가공

04 데이터 가공

- Syntactic Complexity feature 추출 (구문 분석)
- 직접 Transformer 기반의 한국어 구문분석기 구현 (2021년 최고성능)



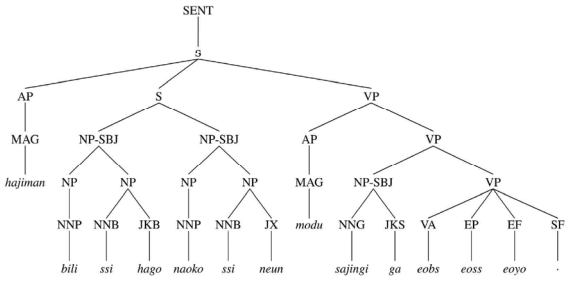
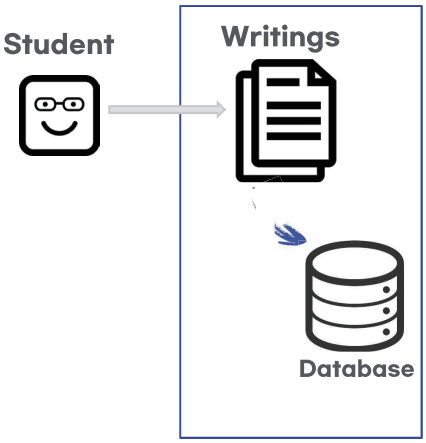


Figure 3. Example of phrase-structure analysis.

Non-terminal node가 얼마나 깊은가에 따라 문법적 복잡도 계산

04 데이터 가공

- Fluency feature 추출 (언어 모델): 얼마나 고품질의 표현을 구사했는가?
- 직접 Wikipedia 데이터를 이용해 언어모델 구현



흔한 단어가 많을수록 값이높음 / 유니크한 단어가 많을수록 값이높음

1. Fluency by Asano, Mizumoto, and Inui (2017): $f(h) = \frac{\log P_m(h)}{|h|} \quad \frac{\log P_u(h)}{|h|}$
2. Fluency by Ge *et al.* (2018): $f(h) = \frac{1}{1+H(x)}$ where $H(x) = -\frac{\log P_m(h)}{|h|}$
3. Fluency by the unigram language model: $f(h) = \frac{\log P_u(h)}{|h|}$

Where
P_m(h)는 문장 h가 순서를 고려하여 나올 확률을 의미,
P_u(h)는 문장 h의 모든 단어가 동시에 출현할 확률
|h|는 총단어 개수

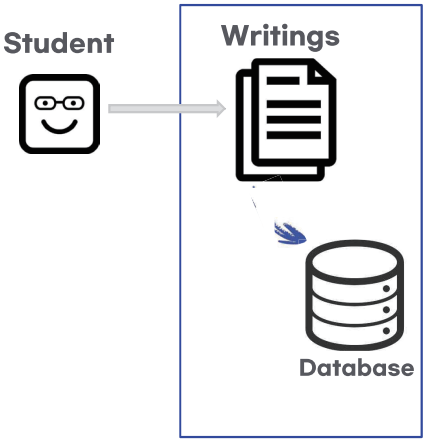
0.17 = -(log(0.2)/4), 0.85 = 1 / (1 + (-(log(0.2)/4)))
0.42 = -(log(0.02)/4) 0.70 = 1 / (1 + (-(log(0.02)/4)))
1.92 = -(log(0.00000002)/4) 0.34= = 1 / (1 + (-(log(0.00000002)/4)))

$P(Today\ is\ Wednesday)$=0.001
$P(Today\ Wednesday\ is)$=0.0000000001

연구개요 - 데이터 수집 - 데이터 가공

04 데이터 가공

- Accuracy feature 추출 (언어 모델): 문법적으로 얼마나 정확한가?
- 한국어 학습데이터가 없어 구현 불가..



04 데이터 가공

- 데이터 가공 결과물

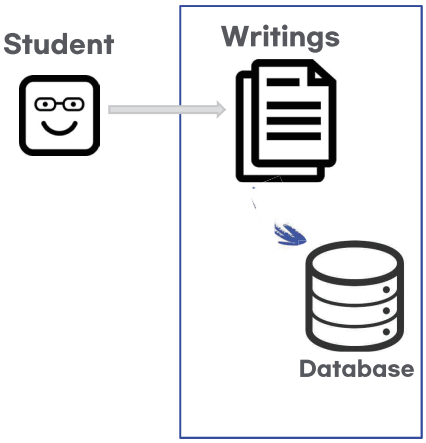
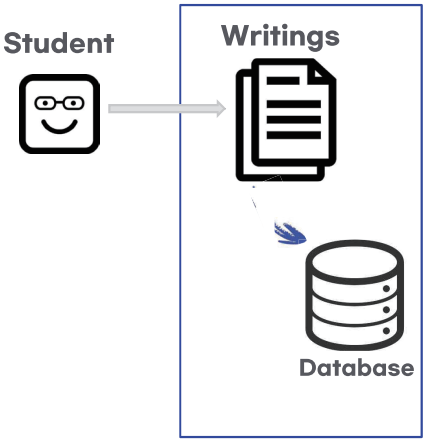


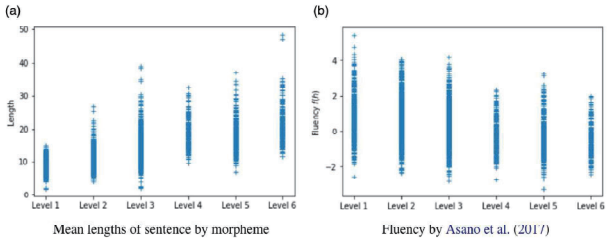
Table 2. Example of features and teir values for te learner's writing in Figure 1

	Features	Values
	bag of morp	저/PRON +는/JX 일요일/NNG +에/JKB 도서관/NNG +에서/JKB 가/VV +았/EP +습니다/EF 저/PRON +는/JX 친구/NNG +하고/JC 도서관/NNG +에서/JKB 가/VV +았/EP +습니다/EF 도서관/NNG +에/JKB 책/NNG +을/JKO 있/VV +었/EP +습니다/EF 도서관/NNG +에서/JKB 공부/NNG +하/XSV +았/EP +습니다/EF 저/PRON +는/JX 책/NNG +을/JKO 읽/VV +었/EP +습니다/EF 그래서/MAJ 일요일/NNG +에/JKB 재미있/VA +습니다/EF
complexity	# of sent	6
	# of para	1
	# of tok	43
	sent by morp	7.166666666666667
	wd by morp	2.263157894736842
	type/token ratio	0.46511627906976744
	bag of funct	+는/JX +에/JKB +에서/JKB +는/JX +하고/JC +에서/JKB +에/JKB +을/JKO +에서/JKB +는/JX +을/JKO +에/JKB
	# of vp eads	0
fluency	Asano *et al.* (2017)	0.16437636932893357
	Ge *et al.* (2018)	0.1491520664089971
	unigram LM	5.868943218961787
accuracy		not available
target	score	70
	proficiency level	Level 1

04 데이터 가공

- 데이터 가공 통계

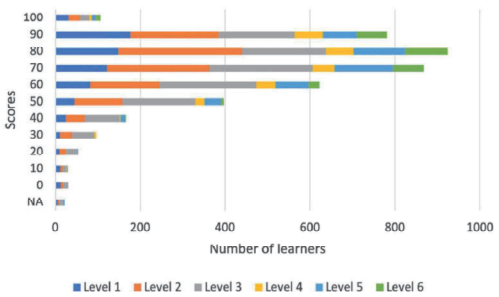




Mean lengths of sentence by morpheme

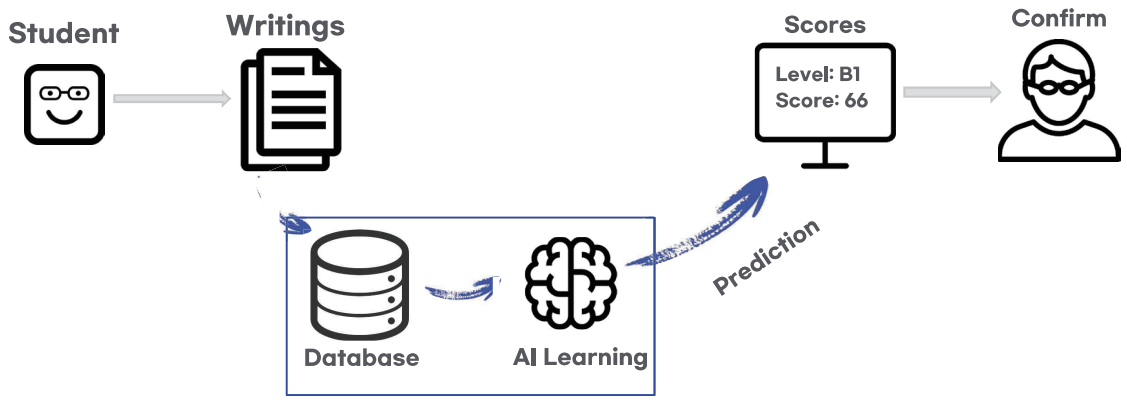
Fluency by Asano et al. (2017)

Figure 5. Distribution of sample features per level.



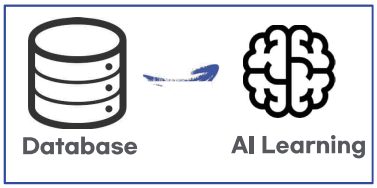
05 AI모델

- 추출된 feature 데이터를 활용해 지도학습 진행 (supervised-learning)



05 AI모델 (분류 기반)

- XLM-RoBERTa 기반의 다국어 사전학습 모델과 추출된 feature 정보를 활용하여 학습
  - 이때 AI모델이 어떤 feature에 집중했는지 확인하기 위해 Attention Score를 계산



Attention Score:
점수 예측을 위해 얼마나 AI가 해당 feature에 얼마나 집중했는가?

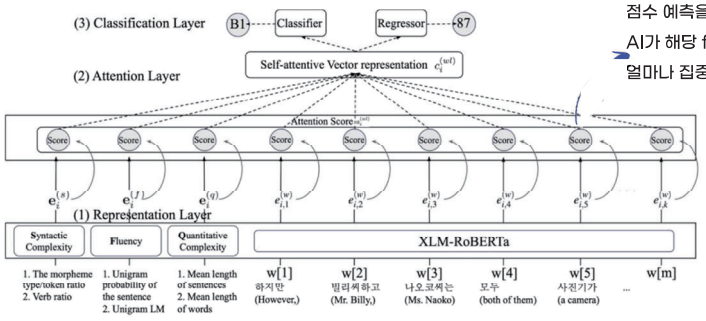
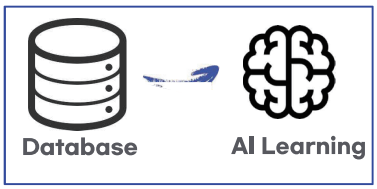


Figure 7. System structure of the proposed deep learning model. Three linguistic features are applied: syntactic complexity, fluency, and quantitative complexity, in addition to the sequence of token representations. Each token is transformed into a vector representation based on XLM-RoBERTa.

05 AI모델 (생성 기반)

- Bllossom 3.1-8B 모델 기반으로 평가를 진행
  - Instruction Tuning을 통해 글쓰기 레벨과 점수를 예측할 수 있도록 튜닝



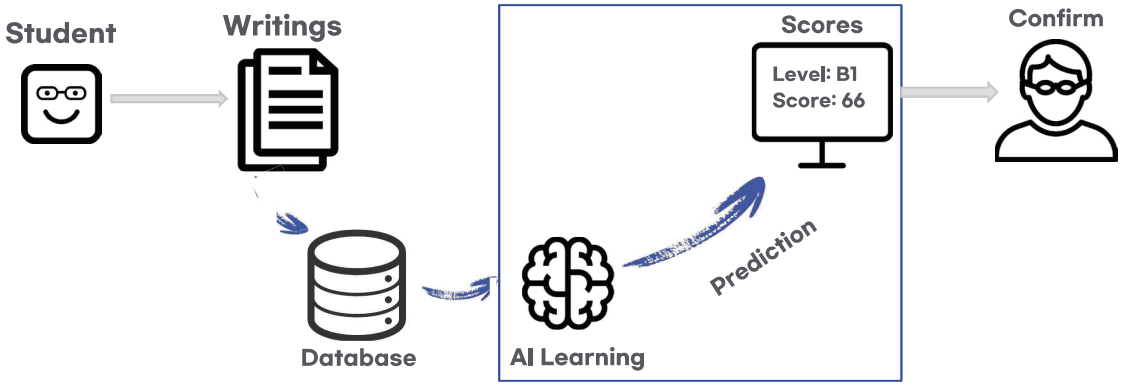
[Instruction] **AWE**
아래의 우리 가족에 대한 글의 종합적인 수준과 점수를 매겨줘.
(Please rate the overall quality and score of the article below on my family.)

[Input]
제 어머니는 의사입니다. 의사기 때문에 매일 매일 너무 바빱니다. 하지만…

[Output]
Level: 2
Score: 20

06 성능평가

- 제안한 AI 모델이 얼마나 잘 예측했어?
  - AI가 학습에 사용한 데이터 80%, AI가 한번도 본적 없는 나머지 20%로 성능평가 진행



06 성능평가 (분류형 모델 결과)

- 8가지 다른 feature조합을 이용하여 성능평가 진행

  - (M) BERT, (X) XLM-RoBERTa, (A) Attention, (S) Syntactic Complexity, (F) Fluency, (Q) Quantitative Complexity를 의미

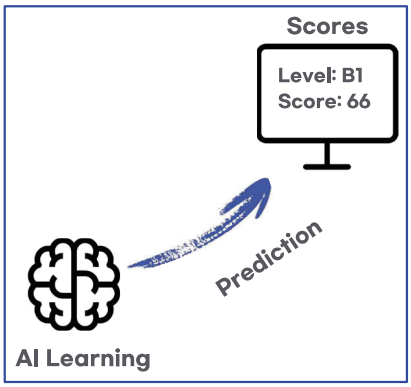


Table 5. Experiment results

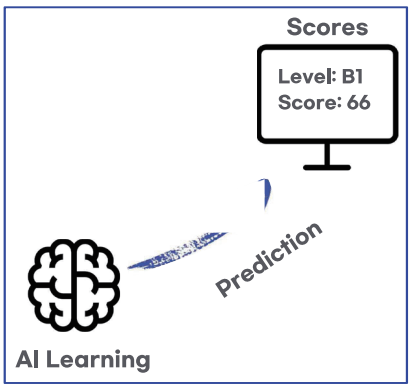
Model	ACC	MSE
(M)	95.83 (± 0.66)	12.11 (± 0.52)
(X)	96.16 (± 0.48)	12.46 (± 0.46)
(X) + (A)	96.14 (± 0.51)	11.98 (± 0.70)
(X) + (S)	**96.85** (± 0.91)	**11.4** (± 0.62)
(X) + (F)	96.06 (± 0.46)	12.01 (± 0.78)
(X) + (Q)	96.40 (± 0.17)	12.43 (± 0.73)
(X) + (A) + (S) + (F) + (Q)	96.71 (± 0.30)	11.96 (± 0.46)
(A) + (S) + (F) + (Q)	50.98 (± 1.27)	13.02 (± 1.02)

Accuracy for predicting a proficiency level and MSE for assigning a score for the learner's writing. (M) multilingual BERT only, (X) XLM-RoBERTa only, (X) + (A) XLM-RoBERTa and attention, (X) + (S) XLM-RoBERTa and syntactic complexity features, (X) + (F) XLM-RoBERTa and fluency features, (X) + (Q) XLM-RoBERTa and quantitative complexity features, (X) + (A) + (S) + (F) + (Q) XLM-RoBERTa and all features, and (A) + (S) + (F) + (Q) w/o pre-trained LMs.

06 성능평가 (생성형 모델 결과)

- 서로 다른 Instruction 및 학습 방법을 토대로 성능 평가

  - 생성 모델의 성능이 분류 모델에 비해 상당히 낮은 것을 확인함.



Model	AWE	QWK
llama3-8B-MUL	47.34	36.46
llama3-8B(5shot)	16.42	4.86
GPT4o(5shot)	41.55	39.25
BLLOSSOM-AWE	36.72	46.18
BLLOSSOM-GEC	-	-
BLLOSSOM-MUL	46.38	46.86
BLLOSSOM-GA	**60.39**	**64.16**
BLLOSSOM-AG	41.06	58.60

07 분석

- 과연 어떤 feature조합이 AWE 성능이 좋을까?

  - (M) BERT, (X) XLM-RoBERTa, (A) Attention, (S) Syntactic Complexity, (F) Fluency, (Q) Quantitative Complexity를 의미

Table 5. Experiment results

Model	ACC	MSE
(M)	95.83 (± 0.66)	12.11 (± 0.52)
(X)	96.16 (± 0.48)	12.46 (± 0.46)
(X) + (A)	96.14 (± 0.51)	11.98 (± 0.70)
(X) + (S)	**96.85** (± 0.91)	**11.4** (± 0.62)
(X) + (F)	96.06 (± 0.46)	12.01 (± 0.78)
(X) + (Q)	96.40 (± 0.17)	12.43 (± 0.73)
(X) + (A) + (S) + (F) + (Q)	96.71 (± 0.30)	11.96 (± 0.46)
(A) + (S) + (F) + (Q)	50.98 (± 1.27)	13.02 (± 1.02)

Accuracy for predicting a proficiency level and MSE for assigning a score for the learner's writing: (M) multilingual BERT only, (X) XLM-RoBERTa only, (X) + (A) XLM-RoBERTa and attention, (X) + (S) XLM-RoBERTa and syntactic complexity features. (X) + (F) XLM-RoBERTa and fluency features. (X) + (O) XLM-RoBERTa and quantitative complexity features, (X) + (A) + (S) + (F) + (Q) XLM-RoBERTa and all features, and (A) + (S) + (F) + (Q) w/o pre-trained LMs.

1. Syntactic Complexity와 다국어 사전 언어모델을 활용했을 때 최고점 즉 문법적 복잡도 정보가 성능향상 도움

2. Fluency정보는 성능에 영향이 없거나 오히려 하락시킬 수 있다.

3. 어찌되었든 단어정보가 가장 중요하다. (단어정보 없이 추출된 feature만 넣은경우)

연구개요 - 데이터 수집 - 데이터 가공 - 제안하는 모델 - 성능평가 - 분석

07 분석

- 그렇다면 AI 모델은 어떤 feature를 참조해서 의사결정을 했을까?

  - 우리가 추출한 feature들은 과연 의미가 있었을까? AI 모델이 집중한 feature를 attention score로 살펴보자

1. Syntactic Complexity를 많이 고려한걸 알 수있음

2. 상대적으로 Quantitative Complexity는 덜 고려함

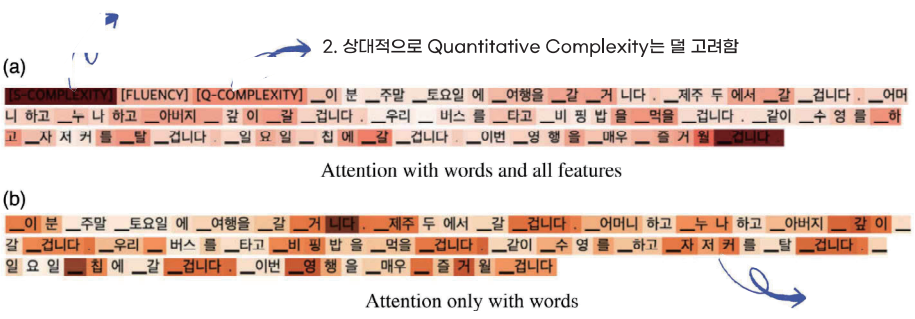


Figure 8. Visualization of the attention score proposed in Eq (8).

3. 추출된 feature 를 활용하지 않은 모델은 오타에 민감함

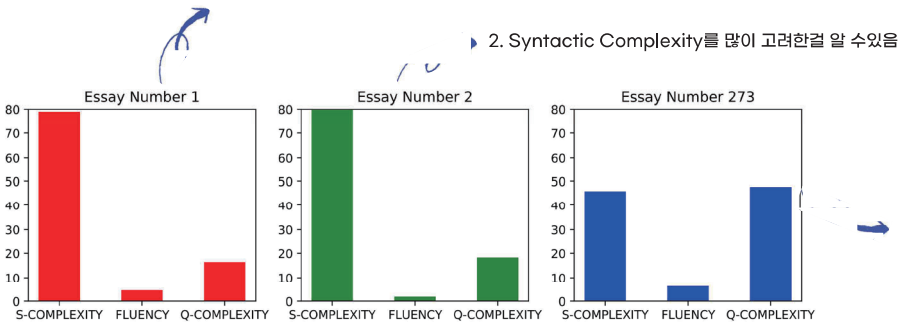
07 분석

- 그렇다면 AI 모델은 어떤 feature를 참조해서 의사결정을 했을까?

  - 각 글쓰기 마다 주요 참조 feature가 달랐음.

1. Syntactic Complexity를 많이 고려한걸 알 수있음

2. Syntactic Complexity를 많이 고려한걸 알 수있음



3. Quantitative Complexity를 많이 고려함

Figure 9. Visualization of the attention scores proposed in (8): [S-COMPLEXITY], [FLUENCY], and [Q-COMPLEXITY] for syntactic complexity, fluency, and qualitative complexity features, respectively.

연구개요 - 데이터 수집 - 데이터 가공 - 제안하는 모델 - 성능평가 - 분석

07 분석

- 각 레벨별 단어활용은 어떻게 다를까?

  - A1 ~ C1레벨은 활용하는 verbal ending 단어에서 확연한 차이가 드러남

1. 1~2 레벨의 경우 "~습니다". "~어요"와 같은표현이 주를 이룸

2. 활용 동사수 증가

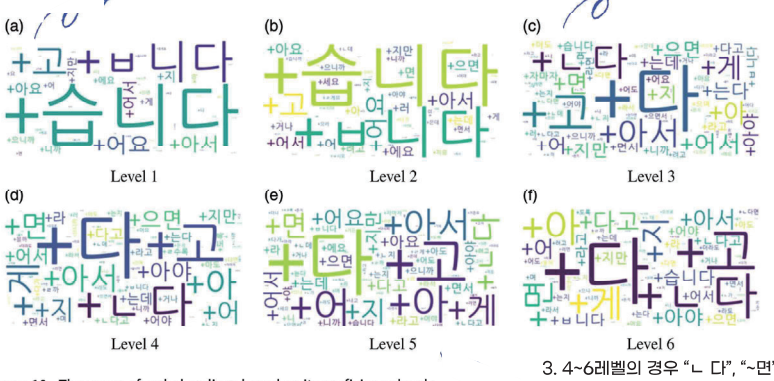


Figure 10. The usage of verbal endings based on its proficiency levels.

3. 4~6레벨의 경우 "ㄴ 다", "~면"와 같은 현상에 대한 설명문구가 증가

07 분석

- 혹시 추출 feature없이 단어에만 집중하면 더 좋은 예측 모델이 나오는거 아니야?

  - 현존하는 BERT 모델을 변화하여 성능평가 하였으나, 유의미한 성능향상이 없었음.

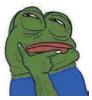
Table 6. Result comparison using different Korean monolingual BERTs

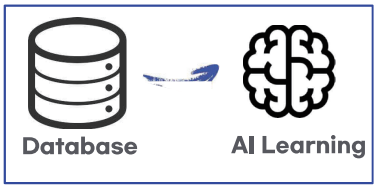
Model	ACC	Model size	Training data
Multilingual BERT (base)	95.83	641MB	–
XLM-RoBERTa (base)	96.16	1.2GB	–
KLUE-RoBERTa (base)	96.06	110MB	63GB
KoBERT (base)	95.94	351MB	10GB
DistilKoBERT	95.72	108MB	10GB
KoELECTRA (small-V2)	96.06	255MB	14GB
KoELECTRA (small-V3)	96.02	255MB	34GB

We evaluated the model using only the BERT model (i.e., we did not apply the proposed linguistic features). `Training Data` denotes the size of the Korean corpus used for training BERT.

08 활용

- 생성형 모델 (BIlossom 3.1-8B)에 글쓰기평가와 고등학교 생활기록부를 함께 학습시키면 특정 도메인 (생활기록부)에 대한 한국어 글쓰기를 더 잘할까?





[Instruction] **AWE**
아래의 우리 가족에 대한 글의 종합적인 수준과 점수를 매겨줘.
(Please rate the overall quality and score of the article below on my family.)

[Input]
제 어머니는 의사입니다. 의사기 때문에 매일 매일 너무 바쁩니다. 하지만…

[Output]
Level: 2
Score: 20

+

고등학교 생활기록부 특별활동 분야의 학습데이터
Instruction: 인수공통 모든 전염병의 열쇠 라는 책을 읽은 학생의 특별활동에 관련된 생활기록부를 작성해줘
Answer: '인수공통 모든 전염병의 열쇠'라는 책을 읽고, 동물 전염병에 관심을 갖게 되어대표적인 돼지 전염병인 아프리카돼지열병에 관해 조사하여 보고서를 제출함. 아프리카돼지열병은 바이러스성 출혈성 돼지 전염병으로…

표 1. 제안하는 Instruction Tuning 학습데이터의 구성 예제

08 활용

- 생성형 모델 (Bllossom 3.1-8B)에 글쓰기평가와 고등학교 생활기록부를 함께 학습시키면 특정 도메인 (생활기록부)에 대한 한국어 글쓰기를 더 잘할까? 

  - 방법1: 글쓰기 평가 + 생활기록부 생성을 multi-task로 학습 후 다음 Instruction을 활용

  "Task{학생부 생활기록부를 작성해줘}

  Level{이때 생성하려고 하는 글의 글쓰기에 대해 종합적인 글쓰기 점수와 레벨을 최상위 레벨로 작성해줘.}"

연구개요 - 데이터 수집 - 데이터 가공 - 제안하는 모델 - 성능평가 - 분석 - 활용

08 활용

- 생성형 모델 (Bllossom 3.1-8B)에 글쓰기평가와 고등학교 생활기록부를 함께 학습시키면 특정 도메인 (생활기록부)에 대한 한국어 글쓰기를 더 잘할까? 

  - 방법2: 다중 Agent 기반 Feedback

  글쓰기 평가 + 생활기록부 생성을 개별 task로 학습 후 각 모델이 의사소통 하도록 구성

감사합니다.

||토론문||

<LLM을 활용한 한국어 글쓰기 평가와 생활기록부 생성 모델의 실제>의 토론문

이진(연세대)

이 연구는 인공지능(AI) 모델을 활용하여 한국어 글쓰기 평가와 생활기록부 생성 모델을 구축하고 그 활용 가능성을 탐색하는 데 목적이 있습니다. 최근 다양한 인공지능 모델을 활용한 자동 평가 분야의 연구가 활발한 상황에서 쓰기 자동 평가를 위한 분류형 모델과 생성형 모델뿐 아니라 인스트럭션 튜닝(instruction tuning)과 다중 에이전트(multi-agent) 등과 같은 방법론도 함께 소개하고 있다는 점에서 본 연구의 의미가 있다고 생각합니다. 의미 있는 연구를 발표해 주신 발표자께 감사의 말씀을 드리며 쓰기 자동 평가 시스템의 교육적 활용의 관점에서 몇 가지 궁금했던 점에 대해 발표자의 의견을 구하는 것으로 토론자의 소임을 다하고자 합니다.

1. 현재 한국어 쓰기 자동 평가 연구에서는 자질 설계(feature engineering)를 통해 수동으로 추출한 자질을 활용하는 자질 기반 접근법과 단어 임베딩(word embedding)을 신경망(neural network)에 적용하여 자질을 자동으로 추출하는 딥러닝(Deep Learning) 기반 접근법, 두 가지 접근법을 결합한 하이브리드(hybrid) 접근법이 활용되고 있습니다. 그 외에 최근에는 거대언어모델(Large Language Model, LLM)을 활용한 쓰기 자동 평가도 이루어지고 있습니다. 자동채점 모델은 단순히 최신 접근법이고 평가 성능이 높다는 이유로 특정 접근법을 선호하기보다는 한국어 언어 자원 구축 현황과 교육적 활용을 고려해 선택할 필요가 있다고 생각됩니다. 한국어 특히, 한국어 학습자를 대상으로 하는 쓰기 자동채점의 경우 활용할 수 있는 공개된 데이터 세트가 거의 없고 교육적 활용 측면에서는 모델의 예측에 대한 해석 가능성(interpretability)도 고려할 필요가 있다고 봅니다. 본 연구에서도 다양한 모델을 활용하여 쓰기 자동 평가 모델을 구축하고 그 성능을 평가해 보셨는데 한국어 쓰기 자동 평가의 경우, 어떠한 접근법이 가장 효과적인 접근법이라고 생각하시는지 발표자의 의견을 여쭙습니다.

2. 자동채점의 정확성을 확보하기 위해서는 자질을 설계하는 과정이 가장 기초적이며 핵심적인 과정이라고 할 수 있습니다. 본 연구에서는 형태소 분석이나 구문 분석 등과 같은 자연언어처리(NLP) 기법을 활용하여 추출한 자질과 딥러닝 모델의 임베딩 결과를 결합하여 채점 자질로 활용한 것으로 보입니다. 본 연구에서 왜 유창성 자질은 모델의 성능 향상에 영향을 주지 못했으며 여러 딥러닝 모델 중에서 RoBERTa가 가장 좋은 성능을 보였는지 궁금합니다. 모델의 학습 방법이나 임베딩 기법, 풀링(pooling) 기법 등이 자동채점 성능에 어떤 영향을 미치는지 발표자의 의견을 여쭙습니다.

3. 연구에 활용한 채점 자질 중에서는 형태소 길이나 형태소 중복 비율과 같이 형태소 분석 결과를 활용하는 경우가 있었습니다. 한국어 학습자 특히, 본 연구에서 연구 대상으로 삼은 초급 학습자의 경우 작문에 철자 오류가 빈번하게 나타나 형태소 분석이 정확하게 되지 않을 것으로 보이는데 형태소 분석 결과를 어떻게 활용하셨는지 궁금합니다.

4. 마지막으로 본 연구에서는 생성형 모델(Bllossom 3.1-8B)에 글쓰기 평가와 고등학교 생활기록부를 함께 학습시켜 모델이 생활기록부의 평가를 생성하도록 하는 방안을 제시하셨습니다. 인스트럭션 튜닝(instruction tuning)과 다중 에이전트(multi-agent)를 활용하는 방안을 소개해 주셨는데 특히, 다중 에이전트 활용 방안이 흥미로웠습니다. 본 연구에서는 생성형 모델이 분류형 모델보다 성능이 상당히 낮은 것으로 나타났는데 이러한 방법론들이 생성형 모델의 성능을 향상시킬 것으로 보시는지 궁금합니다. 쓰기 자동 평가에서도 다중 에이전트를 토론(discussion)시켜 성능을 향상시키는 방법론이 활용되기도 하는데 이런 방법론이 적용된 다른 사례가 있는지 소개해 주시면 감사하겠습니다.

토론자가 본 연구에 대해 제대로 이해하지 못한 부분이 있다면 양해해 주시기 바라며 질문에 답변 부탁드립니다. 감사합니다.

학습자 글쓰기 자동 평가 모델: 피쳐 기반 앙상블 모델

최지명(이화여대)

차 례

1. 서론

글쓰기는 언어 능력 중 가장 종합적인 것으로, 여러 언어적 능력을 통합적으로 요구한다. 이는 모국어 화자뿐만 아니라 제2언어 학습자에게도 마찬가지이다. 학습자들은 글쓰기를 통해 습득한 언어 지식을 체계적으로 활용하고 표현하는 법을 배우고 이는 학습자의 언어적 표현 능력을 종합적으로 평가할 수 있는 자료 중 하나이다. 전통적으로 학습자의 글쓰기는 전문가들에 의해 평가되는데, 이때 문법, 어휘, 의미 구조 등 다양한 측면을 고려한다. 예를 들어 평가 요소를 복잡성(complexity), 정확성(accuracy), 유창성(fluency) 등으로 나눌 수 있다면 이를 위해 평가 요소들이 명확하게 구분 정의되고 일관성 있게 적용되어야 한다. 그러나 이러한 평가 방식은 평가자의 인지에 큰 부담으로 작용할 수 있으며(윤, 2021), 특히 평가 대상이 많을 경우 평가자의 주관적 판단에 의존하면 평가의 일관성과 객관성을 유지하기 어려울 수 있다.[1] 특히 한국어와 같이 전 세계적으로 학습자가 해마다 증가하는 경우(Lusin et al., 2023) 이러한 어려움은 더 커진다. 이러한 배경에서 자동 글쓰기 평가(AWE) 시스템에 대한 요구와 필요성이 커지는 것은 자연스러운 일이다(Fu, H., & Liu, 2022; Wang et al., 2022).

컴퓨터 기술과 자연어 처리(NLP)의 발전으로 자동 평가 시스템은 글쓰기 평가에 있어 중요한 도구로 인식되고 있다(Shi, & Aryadoust, 2022; Shen et al., 2023). 하지만 자연어 처리 도구와 모델이 다양하게 개발되고 그 성능 또한 우수한 영어를 제외하면 한국어를 포함한

1) 이러한 평가자의 평가 일관성 문제는 아래 3장에서 모델의 검증 데이터 중 하나인 외부 데이터 2에 대한 실험 과정에서도 발견되었다.

다른 언어로 된 글쓰기 자동 평가 연구는 많지 않고, 평가 모델의 개발과 검증을 위한 데이터세트도 부족하다(Horbach et al., 2017; Mizumoto et al., 2019; Song et al., 2020; 조 et al., 2021; Lee et al., 2022). 특히 한국어는 자유로운 어순, 복잡한 형태론적 구조, 문장 성분의 생략 등 고유의 특징으로 인해 텍스트를 분석 처리하는 데 어려움이 많다. 특히 학습자가 작성한 글은 명시적/비명시적 오류까지 존재하여 NLP 도구를 이용한 분석과 평가가 더욱 어려워진다.

이 연구는 한국어 학습자 글쓰기를 위한 자동 평가 시스템 모델 개발을 위한 중간 단계의 결과물로서 텍스트 자질(features)과 머신러닝에 기반하여 한국어 평가 모델의 가능성과 실제 적용 가능성을 테스트해 보기 위한 것이다. 따라서 글의 의미와 담화 구조 등 의미 기반 평가 요소는 도입하지 않았다. 여기에서 제안하는 모델은 서포트벡터머신(SVM)으로 구성된 앙상블 모델로서, 학습자의 글을 다양한 차원에서 분석하여 그것을 바탕으로 종합적 판단을 하는 멀티뷰(multi-view) 모델이다. 이 모델은 음절, 어휘 및 구문 사용, 담화 표지의 분포 등 여러 언어적 특질을 사용하여 글을 다각도로 분석하여 평가의 정확성과 신뢰성을 높일 수 있도록 하였다.

2. 분석 방법

2.1. 데이터 및 텍스트 전처리

이 연구에서는 국립국어원에서 구축하여 배포하는 한국어 학습자 코퍼스(이하 KL-Corpus로 칭함)를 사용하였다. KL-Corpus는 2015년부터 구축되기 시작하여, 2018년부터 순차적으로 공개된 것으로, 총 146개국의 101개의 모국어를 사용하는 다양한 학습자들의 글쓰기 및 말하기 데이터를 포함하고 있다.[2)]

KL-Corpus는 원시 코퍼스, 형태 주석 코퍼스 및 오류 주석 코퍼스로 구성되어 있는데 본 연구에서는 원시 코퍼스만을 사용하였다. 원시 코퍼스에 포함된 텍스트가 가장 많을 뿐 아니라, 본 모델의 목적이 데이터 처리 과정에서 형태 주석의 오류 수정이나 글의 오류 판단 등과 같이 사람의 개입이 필요한 단계가 없이 입력부터 출력까지 과정 전체를 자동화하는 것이기 때문이다.

여기서는 2023년에 배포된 코퍼스 중 2019년에 1차로 배포된 데이터를 이용하여 평가 모델을 구축하였다..[3)] 학습자의 등급은 '1급'부터 '6급 이상'까지 총 7개 등급으로 분류되어 있는데 '6급 이상'에 해당하는 텍스트 2개와 등급 '정보 없음' 즉 레이블이 없는 텍스트를 우선 제외하였다. 텍스트의 장르는 작문(시험 작문, 과제 작문, 기획 작문) 장르만 대상으로 하였으며, 텍스트 유형이 홍보문, 신문 기사문, 보고서 등과 같이 구체적인 시간, 장소 등 외부 자료를 참고한 것으로 판단되는 정보가 포함된 것은 제외하였다. 그리하여 최종적으로는 문어 데이터 14,775개 파일을 모델링을 위한 학습 및 테스트 데이터로 사용하였다. 구축된 모델의 일반화 성능을 검증하기 위해 연세대학교 한국어학당에서 2010년 및 2013년에 수집한 학습자 글쓰기 자료도 외부 데이터(unknown data 1, 2)로 사용하였다. 이 데이터를 통해 본 모델이 출처가 다른 데이터에 대해서도 잘 작동하는지를 알 수 있을 것이다.

2) https://kcorpus.korean.go.kr/service/goSummaryStatus.do (2024년 9월 30일 접속)

3) 이후에 배포된 데이터 중의 일부를 본 발표에서 시연할 데모 페이지에서 사용해 볼 것이다.

2.2. 텍스트의 언어적 자질

이 연구에서는 이러한 특징 자질로 크게 두 가지 종류의 언어적 자질을 이용하였다. 첫 번째는 n-gram으로, 음절부터 구문 패턴까지 다양한 단위의 n-gram 정보(N-grams)를 추출하였다. 두 번째는 텍스트의 계량적 특징으로서, 어휘 빈도 분포나 문장의 길이 등 텍스트의 형식적·구문적 특징을 나타내는 유형(quantitative textual complexity)이다. 여기에는 텍스트의 길이, 문장의 길이, 사용 어휘의 길이 등 길이 정보, TTR에 의해 대표되는 어휘 풍부성 지표, 문법적 품사별 사용 분포, 어휘 사용 프로파일(profile), 길이, 구둣점 사용 분포 등이 포함된다(<표 1>).

<표 1> 언어적 자질

유형	종류
quantitative textual complexity	TTR
	Bilog TTR
	Corrective TTR
	Root TTR
	품사(POS) 분포
	기능어 분포
	담화 표지(1): reference expression
	담화 표지(2): connectives
	문서 길이
	문장 길이
	어휘 프로파일(1): 어휘 수준(level)
	어휘 프로파일(2): 어휘 빈도 순위(rank)
N-grams (n = 1:4)	어절 n-grams
	형태소 n-grams
	품사 n-grams
	skeleton n-grams
	음절 n-grams

먼저 어휘 다양성을 측정하기 위한 지표로 TTR(Type/Token Ratio)과 그 변형인 Carroll's corrected ttr (Malven et al. 2004), Guiraud's Root TTR(Tweedie & Baayen, 1998; Malven et al. 2004), Bilog TTR(Tweedie & Baayen, 1998; Malven et al., 2004) 등 네 가지 측정값을 사용하였다. 품사 분포는 총 26개의 품사 사용 분포와 함께 명사와 동사의 비율 그리고 대명사와 명사의 사용 비율을 포함하였다.

어휘 프로파일은 텍스트에 사용된 어휘들의 수준을 나타내는 것으로 참조 어휘 목록을 이용하여 텍스트에 사용된 각 어휘가 어떤 레벨에 속하는지를 나타내고 이 레벨들의 분포에 따라 딕셔너리로 만든 것이다. 이 연구에서는 (김, 2017)에서 제시한 한국어 교육용 어휘 목록을 참조 목록으로 사용하였다. 이 목록은 동형 어휘를 포함하여 총 11,118개 어휘를 초급, 중급, 고급 이상 등 3개의 등급으로 나누어 제시하고 있다. 본 연구에서는 이 3개의 등급 외에 'not in the list'라는 범주를 추가하여 전체 어휘를 총 네 개의 등급으로 나타내었다.

담화 표지는 텍스트의 일관성(coherence)을 나타내는 지표로서 제2언어 습득론(SLA)에서 유창성(fluency)과 관련된 요인으로 볼 수 있다. reference expression은 앞선 문장에서 언급한 대상을 지시하는 기능(corefer)을 가진 어휘들을 말하는 것으로 영어에서는 대명사와 관사가 대표적이다. 본 연구에서는 지시대명사와 관형사 등 한국어에서 유사한 역할을 하는 어휘 23개를 선정하여 사용 분포를 측정하였다. connective는 텍스트의 응집성(cohesion)을 나타내는 데에 아주 중요한 역할을 하는 어휘들로서(Graesser et al., 2004) 한국어에서 많이 사용되는 connective 79개를 선정하여 이들의 분포를 측정하였다.

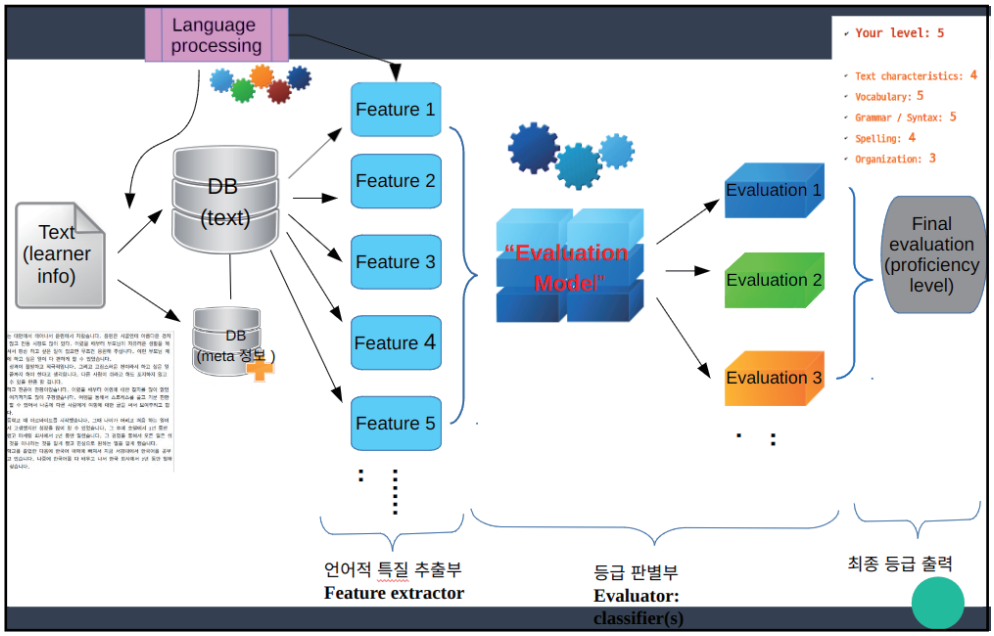
N-gram 자질은 음절부터 구문 패턴까지 다양한 단위의 n-gram 정보를 말한다. 추출이 용이하고 *n*의 값에 따라 길거나 짧은 토큰의 연속체를 얻을 수 있어 개별 단어를 자질로 하는 bag-of-words 모델의 단점 즉 단어들의 순서 즉 맥락 정보가 손실되는 단점을 보완할 수 있다. 이ngram 정보는 문장 단위로 추출할 수도 있고 문장 경계를 넘어서 문서 단위로 추출할 수도 있다. 여기서는 문서 단위에서 추출하였는데 이 단위에서 추출한 자질의 성능이 약간 더 뛰어났기 때문이다.

n-gram 중 skeleton n-gram은 구성하는 문법적인 어휘들로 구성된 구문 패턴을 추출하기 위한 것으로(최, 2023), 명사나 동사와 같은 내용어들을 대표 품사로 표기하고 이 내용어들을 연결하는 문법적 어휘들은 원래의 형태소로 그대로 표기하여 n-gram을 추출하는 방식이다. 이를 통해 문장의 구문적 골격 정보를 어느 정도 얻을 수 있어 구문 분석의 효과를 일정 수준 발휘할 수 있는 것으로 여겨진다. 형태소 n-grams과 skeleton n-grams는 하나의 문장을 어휘적 연결의 관점에서 그리고 구문적 연결의 관점에서 바라본 것으로, 서로 보완적 관계에 있다고 볼 수 있다(최, 2023). n-gram 자질은 학습 데이터에서 빈도 상위 0.001 ~ 0.003% 수준 이상인 것만 최종 자질로 선정 이용하였는데 이는 개별 모델당 평균 3,800여 개의 자질에 해당한다.

2.3. 자동 평가 모델 구조

이 연구에서는 각 자질을 이용하여 여러 개의 기본 모델(base learners)을 만들고 이 기본 모델들의 예측 결과를 이용하여 메타 모델(supralearner, meta-learner)이 최종적으로 판단하는 앙상블 모델을 도입하였다. 앙상블 모델은 모델의 일반화 성능을 향상시키고 개별 모델의 예측 오류를 다른 모델의 예측으로 보완할 수 있게 해 준다(Chali et al., 2009).

기본 모델(base learner)로는 support vector machine을 이용하였는데 이는 svm이 가진 특성 때문이다. Svm은 많은 수의 자질을 한 번에 다룰 수 있고 희소 벡터(sparse vector)를 잘 처리할 수 있으며(Cortes & Vapnik, 1995; Joachims, 1998), 커널 함수를 이용하여 비선형의 분류 작업도 잘 처리할 수 있다. SVM은 텍스트 분류, 감정 분석, 의료 기록 관리 등 자연어처리 분야 외에 이미지 분류, 음악 멜로디 예측 등 시각적 청각적 분야에 이르기까지 다양한 영역에서 활용되어 왔다. 하이퍼파라미터는 libsvm 라이브러리(v. 3.2.3)의 기본값을 이용하였는데, 텍스트 분류에서는 보통 하이퍼파라미터를 조정하는 것보다 텍스트 자질의 종류와 갯수가 모델의 성능에 더 큰 영향을 미치기 때문이다(Cueva Mora & Tierney, 2021). 기본 모델들의 예측 결과를 종합하여 최종적인 하나의 예측을 하는 메타 모델로는 randome forest를 이용하였다. 모델의 구조는 <그림 1>에 제시하였다.



<그림 1> 평가 모델의 구조

여기서 기본 모델들은 텍스트의 서로 다른 측면을 측정하여 정보를 얻게 되고 이 각각의 정보를 이용한 상위 모델이 최종적인 예측을 하게 된다는 점에서 이 모델은 일종의 멀티뷰 앙상블 모델(multi-view ensemble)이라고 볼 수 있다. 멀티뷰 모델은 고차원 데이터를 다루는 데 효과적이고(Sun, 2013, Kumar & Minz,2015), 하나의 데이터에 대해 서로 다른 종류의 정보를 이용하여 다양한 측면에서 볼 수 있게 해 주므로 모델의 일반화 가능성을 향상시킬 수 있다(Zhaoet al., 2017). 특히 텍스트의 분류에서는 딥러닝 모델과 비슷한 성능을 보이면서도 모델 구축이 상대적으로 용이하고 예측 결과를 해석하는 데에도 보다 유용하다(MacAvaney et al., 2019). 또한 각 관점(view)을 관장하는 분류기에 의한 예측에 따라 최종적인 등급을 산출하므로 메타모델(meta learner)에 의한 최종 등급과 함께 각 기본 모델이 예측한 요인별 등급도 같이 제시할 수 있다.

2.4. 모델의 훈련과 평가 방법

모델의 학습에 전체 데이터의 70%를 사용하고 나머지 30%를 따로 떼어 테스트 데이터로 사용하여 모델의 성능을 측정하였다. 학습 데이터는 계층적 샘플링(stratified sampling)을 수행하여 전체 데이터의 등급별 비율이 훈련 데이터에도 반영이 되도록 하였다. 등급별 샘플의 수는 1등급 25%, 2등급 20%, 3등급 17%, 4등급 16%, 5등급 13%, 6등급 9%로 다소 불균형하지만 언더샘플링이나 오버샘플링을 하지 않았는데, 이는 균형을 맞춰준 데이터로 학습했을 때와 비교해 볼 때 성능의 차이가 거의 없었기 때문이다.

기본 모델 및 앙상블 모델의 성능은 정확도(accuracy)와 F1 점수(macro averaged F1)로 측정하였다. 등급에 따라 예측 정확도가 약간 달라질 수 있으므로 보다 자세한 성능 평가를 위해 등급별 성능 점수도 같이 제시하였다.

3. 모델 성능 평가 결과

<표 2>는 기본 모델들 중에서 성능이 높은 것들을 제시한 것이다. 이 중에서도 성능이 가장 좋은 모델만을 취하여 앙상블 모델을 만들게 된다. 계량적 텍스트 복잡성(quantitative textual complexity) 자질의 경우 Boruta 알고리즘을 이용하여 성능이 가장 높은 80개의 자질을 선택하였다.

<표 2> 기본 모델의 성능

자질의 유형	모델 타입	자질	Accuracy	Precision	Recall	F1 score
Quantitative textual complexity	Model_01	80	0.625	0.601	0.599	0.600
eojeol n-gram	Model_02	unigrams	0.803	0.797	0.794	0.795
word n-gram	Model_03_1	unigrams	0.807	0.801	0.799	0.800
	Model_03_2	bigrams	0.809	0.806	0.802	0.803
pos n-gram	Model_04_1	4-grams	0.664	0.647	0.643	0.644
	Model_04_2	trigrams	0.641	0.625	0.620	0.621
skeleton -n-gram	Model_05_1	bigrams	0.743	0.726	0.724	0.724
	Model_05_2	trigrams	0.736	0.720	0.713	0.715
character n-gram	Model_06_1	bigrams	0.821	0.816	0.812	**0.814**
	Model_06_2	trigrams	0.821	0.814	0.812	0.813

기본 모델의 성능은 정확도를 기준으로 최저 0.625 – 최고 0.821로 자질의 종류에 따라 성능의 차이가 나타났다. 어절 n-gram, 형태소 n-gram, 음절 n-gram 모델은 모두 0.79 이상으로 상대적으로 높은 성능을, 그리고 계량적 텍스트 복잡성(quantitative textual complexity), 품사 n-gram, skeleton n-gram 모델은 0.625% - 0.743%의 상대적으로 낮은 성능을 나타냈다. 전자의 모델들은 사용 어휘나 문자열 등의 구체적(concrete) 자질을, 후자의 모델들은 패턴이나 문서 전체의 특징 등 다소 추상적(abstract) 자질에 속한다고 볼 수 있다.

기본 모델 중 가장 높은 성능을 보인 것은 Model_06_1(음절 bigrams) 모델로, 정확도 0.821, F1 점수 0.814를 기록했는데, 가장 성능이 낮은 계량적 텍스트 복잡성 모델(Model_01)과 비교하면 약 20%의 성능 차이를 보였다. 그러나 예측 정확도가 상대적으로 낮은 모델이라 하더라도 성능이 더 높은 모델이 잘못된 예측을 한 경우를 올바로 예측한 경우도 있으므로 이 모델들은 서로 보완적이라고(complementary)이다. 기본 모델들의 예측의 일치도는 Fleiss' kappa K= 0.694로, 각 모델이 학습한 자질의 종류에 따라 예측의 결과가 서로 조금씩 다른 경우들이 있다. 가장 낮은 성능을 보인 model_01과 가장 높은 성능을 보인 model_06_1은 K = 0.57의 일치도를 보였으나, model_01과 model_03_1간 예측이 서로 다른 경우에 오히려 model_01의 예측이 옳은 경우가 테스트 데이터의 전체 약 5.5%를 차지하였다. 즉 기본 모델로서 성능이 다소 낮다 하더라도 다른 모델에 반영되지 못한 텍스트의 다른 측면을 보완할 수 있음을 알 수 있다. 앙상블 모델은 이와 같이 서로 다른 모델들의 예측 결과를 종합적으로 판단할 수 있게 해 준다.

등급별 F1 점수의 분포(<표 3>)는 1등급이 모델 전체 평균 0.886으로, 가장 낮은 4등급 0.668과 약 22%의 차이를 보였고, 두 번째로 높은 2등급 0.773과도 약 11%의 차이를 보였다.

등급별 성능의 차이는 기본 모델의 종류에 따라 편차가 다소 큰데, 전체 성능이 가장 높은 model_06_1과 두 번째로 높은 model_03_1이 등급별 F1 점수에서는 가장 고르고 뛰어난 성능을 보였다. 그러나 이때도 등급별로 다소 편차가 여전히 존재한다.

<표 3> 기본 모델의 등급별 성능

모델 타입	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Model_01	0.852	0.690	0.541	0.485	0.501	0.527
Model_02	0.903	0.816	0.765	0.748	0.757	0.782
Model_03_1	0.906	0.830	0.777	0.753	0.770	0.765
Model_03_2	0.889	0.807	0.768	0.767	0.799	0.792
Model_04_1	0.842	0.695	0.612	0.553	0.581	0.584
Model_04_2	0.829	0.667	0.565	0.542	0.567	0.555
Model_05_1	0.901	0.787	0.710	0.649	0.641	0.659
Model_05_2	0.898	0.778	0.704	0.651	0.632	0.631
Model_06_1	0.916	0.837	0.773	0.767	0.786	0.805
Model_06_2	0.918	0.837	0.772	0.765	0.793	0.792
Avg.	**0.886**	**0.774**	**0.698**	**0.668**	**0.683**	**0.689**

등급별 예측 정확성을 평균값(Avg.)을 통해 살펴보면 전 모델에 걸쳐 1, 2, 3등급에 대한 F1 점수가 상대적으로 높게 나타나고(0.7 이상) 4등급부터 6등급까지는 상대적으로 낮게(0.7 미만) 나타났다. 등급간 샘플의 수가 불균형한 것이 이와 같은 등급별 예측 성능의 차이에 영향을 주었는지 확인해 보았으나 등급별 샘플의 수가 F1 점수에 영향을 거의 주지 않았다.

다음으로는 어휘, 구문, 형태론적 요소 등 텍스트의 다양한 측면을 나타내는 기본 모델들의 예측 결과들을 합하여 앙상블 모델을 만든 후 성능을 측정해 보았다. 이때 텍스트의 여러 정보를 측정할 수 있도록 여러 유형의 기본 모델들이 포함되도록 성능이 높고 서로 다른 종류의 자질을 측정한 5개 모델(model_01, model_03_1, model_06_1 등)을 조합하여 성능을 확인하였다. 앙상블 모델의 성능 측정 역시 기본 모델의 성능 측정에 사용되었던 것과 같이 테스트 데이터 30%를 이용하였다. <표 4>는 앙상블 모델의 성능과 등급별 F1 score를 제시한 것이다. 앙상블 모델 역시 등급별 샘플의 수의 균형을 맞추는 과정(undersampling이나 oversampling)을 도입해 보았으나 성능의 차이는 거의 없었으므로 테스트 데이터를 그대로 사용한 결과이다.

<표 4> 앙상블 모델 및 등급별 F1 score

	Accuracy	Precision	Recall	F1 score
ensemble model	0.827	0.824	0.820	0.821

metric	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
F1 score	0.904	0.834	0.789	0.789	0.801	0.811
Precision	0.866	0.871	0.783	0.786	0.800	0.840
Recall	0.891	0.800	0.800	0.790	0.800	0.783

앙상블 모델의 성능은 정확도 0.827, F1 점수 0.821로, 가장 높은 성능을 보인 기본 모델과 비교해 볼 때 정확도와 F1 점수 모두에서 조금 더 높은 성능을 보였다. 등급별로 확인해 보면

앙상블 모델이 전 등급에 걸쳐 기본 모델과 비교해 조금 더 고르고 향상된 성능을 보이는 것을 알 수 있다. 기본 모델에서는 중간 수준인 3 ,4, 5등급의 예측에서 오류가 상대적으로 많이 나타나는 반면 앙상블 모델에서는 그 오류가 줄었고 전 등급에 걸쳐 비교적 고른 예측 정확성을 보인다. 따라서 기본 모델에 비해 모델의 성능이 좋아지고 안정성이 높아졌다고 할 수 있다.

앙상블 모델에서 발생한 오류의 양상을 살펴보면 대부분이 바로 인접한 등급으로 잘못 분류한 경우였는데, 사람 평가자와 마찬가지로 바로 인접한 등급들 사이에서 판단의 어려움을 겪음을 알 수 있었다. 예측 오류가 바로 인접 등급(예를 들면 3등급을 2등급이나 4등급으로 판단)간 혼동으로 발생한 경우의 수를 살펴보면 총 553건으로, 전체 예측 건수의 약 12.5%에 해당한다. 이 중 약 7.4%는 실제 등급보다 한 등급 아래로 잘못 예측하였고, 나머지 5.3%는 한 등급 위로 잘못 예측한 경우였다. 실제 등급과 예측 등급간 차이가 큰 경우, 예를 들면 5등급을 1등급으로 보거나 1등급을 6등급으로 잘못 예측하는 등 예측의 편차가 극단적인 경우는 약 0.14%에 불과하였다.[4] 예측 오류가 발생한 경우를 등급별로 분류해 보면, 가장 많은 예측 오류가 발생한 것은 2등급의 글을 1등급으로 잘못 예측한 경우로, 2등급 전체 890건 중 약 11.3%, 다음으로 많은 예측 오류는 4등급 글을 3등급으로 잘못 예측한 것으로, 전체 700건 중 약 10%를 차지하였다.

외부 데이터 즉 학습 및 테스트 데이터로 사용하지 않았던 새로운 데이터(unknown data)로 최종 모델의 성능을 확인해 보면 기본 모델에 비해 앙상블 모델의 유의미성을 확인할 수 있다. 이 외부 데이터는 KL Corpus에 포함되지 않았고, 실제 교육 현장에서 입수한 데이터이므로 모델의 현장 적용 가능성을 보다 객관적으로 확인해 볼 수 있는 데이터라고 할 수 있다. 두 개의 외부 데이터는 모두 모델링용 데이터와 동일한 방식으로 전처리하였다. 먼저 외부 데이터1에 대한 앙상블 모델(without parameters)의 전체 성능과 등급별 성능을 제시하면 <표 5>와 같다.

<표 5> 외부 데이터 1에 대한 성능

	Accuracy	Precision	Recall	F1 score
ensemble model	0.866	0.891	0.861	0.868

metric	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
F1 score	0.934	0.727	0.845	0.942	0.867	0.811
Precision	0.880	0.906	0.763	0.945	0.918	0.840
Recall	0.994	0.607	0.946	0.940	0.821	0.783

외부 데이터 1에 대한 예측 정확도는 0.866, F1 점수는 0.868로 나타나 테스트 데이터보다 높은 정확도를 보여주었다. 등급별 F1 점수를 보면 2등급을 제외하고 나머지 등급들에서는 모두 0.845 이상의 고른 성능을 보인다. 테스트 데이터에서는 가장 낮은 예측 성공률을 나타낸 것이3등급과4등급(F1 0.789)이었으나 외부 데이터 1에서는 2등급에서 가장 낮은 예측 정확성을 보였다. 그러나 정확도가 0.906을 기록한 것은 2등급이라고 일단 예측만 되면 그것이 실제 2등급일 가능성이 90% 이상 확실하다는 것을 의미한다. 정확도가 가장 낮은 것은 오히려 3등급으로(0.763) 테스트 데이터에서와 비슷한 수준이다.

다음으로 외부 데이터2에 대해서도 예측을 실시하여 성능을 확인해 보았다.

4) 만약 top-2 accuracy로 평가하게 되면 이 앙상블 모델의 정확도는0.95 이상의 예측 정확도를 보이는 것으로 볼 수 있다.

<표 6> 외부 데이터 2에 대한 성능

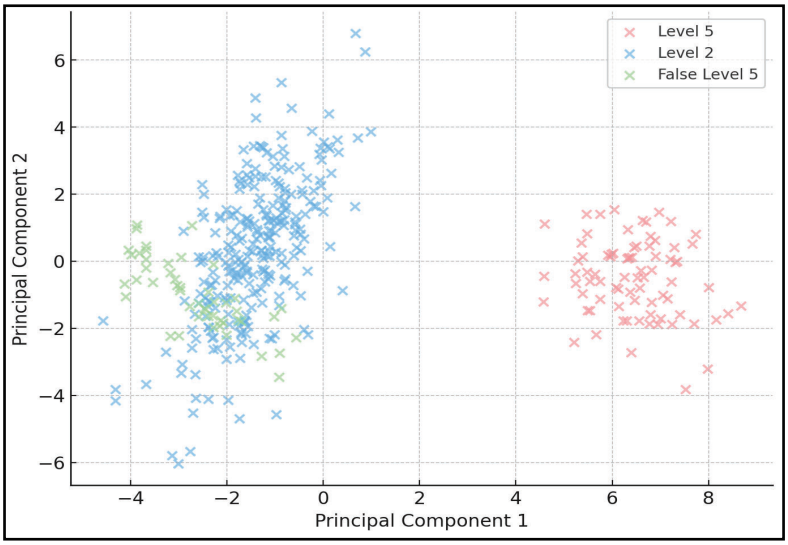
	Accuracy	Precision	Recall	F1 score
ensemble model	0.786	0.788	0.777	0.770

metric	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
F1 score	0.853	0.786	0.856	0.759	***0.554***	0.810
Precision	0.800	0.708	0.919	0.678	***0.754***	0.874
Recall	0.921	0.885	0.801	0.862	***0.438***	0.754

외부 데이터 2에 대한 예측 정확도는 0.786, F1 점수는 0.75로, 테스트 데이터에 대한 성능에 비해 약4-5%, 외부 데이터1에 대한 성능 대비 약8-10%의 성능 하락이 발생하였다. 그러나 등급별 F1 점수에서 볼 수 있듯이 성능 하락의 대부분은 4등급과 5등급, 특히 5등급의 예측 결과로 인한 것이었다. 외부 데이터1에서 가장 판정이 어려웠던 2등급은 외부 데이터 2에서 오히려 약 6%의 성능 향상을 보였고 그 외 다른 등급들의 성능 하락도 약 8% 수준이었다. 5등급의 예측에서 문제가 발생했음을 보여주는 지표는 재현율(recall)인데 그 값이 0.438로 다른 등급들에 비해 월등히 낮아 5등급으로 레이블링된 글에 대한 민감도(sensitivity)가 많이 낮음을 알 수 있다. 반면 2등급을 제외하면 나머지 등급에서의 예측 정확도는 테스트 데이터에서의 등급별 F1 점수와 큰 차이가 나지 않는다.

이에 대한 두 가지 원인을 생각해 볼 수 있는데, 첫째로는 모델 자체의 문제 특히 5등급 글을 찾아내는 민감도가 낮은 원인, 다음으로는 입력 자료 중 5등급으로 레이블링된 데이터 자체에 문제로 인한 원인을 생각해 볼 수 있다. 모델링 과정과 테스트 데이터 및 외부 데이터1에서의 성능 검증 결과를 살펴볼 때 5등급의 글을 찾아서 판정하는 데 특별한 문제가 없었다는 것을 알 수 있으므로 두 번째 원인 때문일 가능성이 더 크다고 짐작할 수 있다. 이를 확인하기 위해 실제 5등급이지만 잘못 분류된 사례들만 따로 떼내어 잘못 예측한 등급을 살펴보았는데 그 결과 2등급으로 잘못 예측한 것이 49건(57.6%)으로 가장 큰 부분을 차지하였다. 다음으로 33%는 4등급(19건) 내지 6등급(9건)과 같이 바로 인접한 등급으로 잘못 예측한 경우였다. 따라서 2등급으로 잘못 예측을 한 것이 예측 모델의 성능에 가장 큰 하락 요인으로 작용했음을 알 수 있다. 이는 외부 데이터 2에 포함된 5등급 텍스트에 대한 예측 오류가 특정한 패턴을 가진다는 것을 암시한다.

따라서 2등급으로 잘못 예측한 텍스트들이 실제 2등급과 유사한 특징을 보이는지 확인해 볼 필요가 있었다. 이를 위해 5등급 텍스트 중에서 올바로 예측한 79개 샘플과 2등급으로 잘못 예측한 것 49건 그리고 실제 2등급 텍스트 267개를 추출하여 PCA를 실시해 보았다. 변수로는 계량적 텍스트 복잡성(quantitative textual complexity) 자질을 이용하였다. 그 결과 <그림 2>에서와 같이 5등급으로 제대로 예측된 글들과, 레이블이 5등급이지만 2등급으로 잘못 예측된 글들은 차원 1을 기준으로 확연히 구분되고 후자의 글들은 오히려 2등급 글과 아주 유사한 것으로 나타났다. 즉 2등급에 속한 글들과 2등급으로 잘못 예측된 5등급 글들이 거의 동일한 집단으로 보인다. 이는 외부 데이터2에서의 전문가 평가가 5등급 일부 텍스트에 한해 일관적이지 못 했을 수도 있다는 것을 의미한다고 볼 수도 있다.



<그림 2> 2등급 *vs.* 5등급 *vs.*False 5등급(2등급으로 잘못 예측한 것)

4. 결론

이 연구에서는 머신러닝 앙상블(multi-view ensemble) 모델로 한국어 학습자의 글을 자동 평가할 수 있는 가능성을 입증하였다. 계량적 텍스트 복잡성(quantitative textual complexity) 자질과 n-gram 자질을 결합한 모델은 학습자 글의 수준을 정확하게 평가하는 데 유의미한 성능을 보였다. 특히 앙상블 모델은 개별 모델에 비해 한국어의 복잡한 문법 및 구문적 특징을 효과적으로 처리할 수 있음을 알 수 있었고, 오류가 포함된 다양한 형태의 학습자 텍스트를 안정적으로 평가할 수 있음을 확인하였다. 이러한 결과는 여러 텍스트 자질을 결합함으로써 개별 모델의 한계를 보완할 수 있음을 보여준다.

이 모델은 한국어에 국한되지 않고 다른 언어에도 적용할 수 있는 잠재력을 가지고 있다. 언어별 특성에 맞춰 자질의 추출 방식을 적절히 수정한다면 다국어 환경에서도 사용할 수 있을 것이다. 특히, 기본적인 문법적 지표부터 담화적 요소에 이르기까지 다양한 특성을 확장 가능하게 설정하였기 때문에, 다른 장르나 교육적 맥락에서도 적용 가능성이 크다.

여기서 제시한 모델의 주요 성과 중 하나는 외부의 새로운 데이터셋에 대해서도 안정적이고 높은 성능을 유지했다는 점이다. 이는 모델의 일반화 가능성을 보여주는 결과이며, 실제 교육 현장에서 즉각적인 피드백을 제공할 수 있는 도구로 활용될 수 있음을 시사한다. 평가자의 주관성이나 평가 시간의 한계를 극복할 수 있는 이러한 자동화된 평가 도구는 교사와 학습자 모두에게 유용할 것이다.

그럼에도 인접 등급 간의 오분류 문제는 여전히 해결해야 할 과제로 남아 있다. 특히, 서로 유사한 등급 사이의 오류를 줄이기 위해서는 더욱 정교한 자연어 처리 기법의 도입이 필요할 것이다. 향후 연구에서는 심층 학습 기반 모델과 같은 고도화된 기법을 결합하여 모델의 정확성을 더욱 향상시킬 수 있을 것이다.

결론적으로, 이 연구에서는 한국어와 같은 복잡한 언어 구조를 기계 학습을 통해 평가할 수 있는 새로운 접근법을 제시하였다. 다양한 언어적 특성을 앙상블 모델에 통합함으로써 자동 글쓰기 평가 시스템이 신뢰할 수 있고 확장 가능한 평가 도구로 기능할 수 있음을 보여주었다. 앞으로의 연구는 이러한 방법들을 더욱 정교하게 발전시키고, 다양한 언어 학습 환경에 적용할 수 있도록 그 적용 범위를 확대해 나가는 것이 바람직할 것이다.

다만, 이 연구는 최종 모델을 위한 예비 분석으로, 이어질 본 연구에서는 딥러닝 모델 구조를 도입하고 BERT와 XML-Roberta와 같은 대형 언어 모델(LLMs)의 기능을 통합한 하이브리드 평가 모델을 개발할 계획이다. 이를 통해 텍스트 내용적 측면을 추가로 반영하여 더욱 종합적이고 세밀한 분석이 가능할 것으로 기대한다. 이와 같은 하이브리드 접근은 기존의 기계 학습 기법과 딥러닝의 장점을 결합함으로써 모델의 전반적인 성능과 다양한 텍스트 장르 및 학습자 수준에 대한 적응력을 크게 향상시킬 수 있을 것으로 기대된다.

참고문헌

김중섭, (2017). 국제 통용 한국어 표준 교육과정 적용 연구. *국립국어원*.

윤금준. (2021). 국어교사의 쓰기평가 신뢰도에 영향을 미치는 인지 과정 특성 분석. *청람어문교육*, *79*, 127-158.

조희련, 이유미, 임현열, 차준우, & 이찬규. (2021). 딥러닝 기반 언어모델을 이용한 한국어 학습자 쓰기 평가의 자동 점수 구간 분류-KoBERT 와 KoGPT2 를 중심으로. *한국언어문화학*, *18*(1), 217-241.

최지명. (2023). **샴 네트워크를 이용한 한국어 저자 판별 모델 연구**. [박사학위논문, 연세대학교]. dCollection@yonsei. http://www.dcollection.net/handler/yonsei/000000545652

Chali, Y., Hasan, S. A., & Joty, S. R. (2009). A SVM-based ensemble approach to multi-document summarization. In *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22* (pp. 199-202). Springer Berlin Heidelberg.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*.

Cueva Mora, A., & Tierney, B. (2021). Feature Engineering vs Feature Selection vs Hyperparameter Optimization in the Spotify Song Popularity Dataset.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

Fu, H., & Liu, X. (2022). EFL learner engagement in automatic written evaluation. *Frontiers in Psychology*, *13*, 871707.

Horbach, A., Scholten-Akoun, D., Ding, Y., & Zesch, T. (2017, September). Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings Of The 12th workshop on innovative use of NLP for building educational applications* (pp. 357-366).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features.

Kumar, V., & Minz, S. (2015, August). Multi-view ensemble learning*: a supervised feature set*

partitioning for high dimensional data classification. In Proceedings of the Third International Symposium on Women in Computing and Informatics (pp. 31-37).

Kumar, A., & Yadav, J. (2023). A review of feature set partitioning methods for multi-view ensemble learning. *Information Fusion*, 101959.

Lee, J. H., Park, J. S., & Shon, J. G. (2022). A BERT-Based Automatic Scoring Model of Korean Language Learners' Essay. *Journal of Information Processing Systems, 18*(2), 282-291.

Lusin, N., Peterson, T., Sulewski, C., & Zafer, R. (2023). Enrollments in languages other than English in US institutions of higher education, fall 2021. *Modern Language Association of America*.

Sun, S. (2013). A survey of multi-view machine learning. Neural computing and applications, 23, 2031-2038.

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one, 14*(8), e0221152.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development* (pp. 16-30). Palgrave Macmillan UK.

Mizumoto, T., Ouchi, H., Isobe, Y., Reisert, P., Nagata, R., Sekine, S., & Inui, K. (2019, August). Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 316-325).

Shen, C., Tian, J., & Qu, S. (2023, June). Examining the Effects of Automated Writing Evaluation (AWE) Feedback on EFL Learners' Revision and Writing Quality. In *2023 4th International Conference on Education, Knowledge and Information Management (ICEKIM 2023)* (pp. 558-566). Atlantis Press.

Shi, H., & Aryadoust, V. (2022). A systematic review of automated writing evaluation systems. *Education and Information Technologies, 28*, 771-795.

Song, W., Zhang, K., Fu, R., Liu, L., Liu, T., & Cheng, M. (2020, November). Multi-stage pre-training for automated Chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6723-6733).

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*, 323-352.

Wang, Y., Luo, X., Liu, C. C., Tu, Y. F., & Wang, N. (2022). An integrated automatic writing evaluation and SVVR approach to improve students' EFL writing performance. *Sustainability, 14*(18), 11586.

Zhao, J., Xie, X., Xu, X., & Sun, S. (2017). Multi-view learning overview: Recent progress and new challenges. *Information Fusion, 38*, 43-54.

토론문

〈학습자 글쓰기 자동 평가 모델: 피처 기반 앙상블 모델〉의 토론문

노영빈 (엔씨소프트/연세대)

본 연구는 학습자 글쓰기 자동 평가 모델 개발의 일환으로, 텍스트에 포함된 다양한 언어적 자질을 서포트 벡터 머신(SVM) 기반 앙상블 모델을 활용해 학습자의 글을 자동 평가하는 것을 목표로 합니다. 국립국어원 〈한국어 학습자 말뭉치〉를 사용하여 학습자 글의 표층에서 관찰 가능한 다양한 언어적 자질을 자동 평가의 근거로 삼되, 의미 및 담화 구조와 같은 표층에 드러나지 않거나 문맥적 판단이 필요한 내용은 평가 근거에서 제외합니다. 이러한 접근은 평가자나 평가 대상자에게 직관적인 근거를 제공할 수 있습니다. 또한 의미 기반의 문맥적 판단 시스템으로 나아가기 위한 필요조건이기에, 본 연구가 자동 글쓰기 평가에 크게 기여할 것으로 생각합니다. 발표자의 연구를 바탕으로 개인적 견해와 경험에서 비롯한 몇 가지 질문을 드리고자 합니다.

1. 본 연구는 국립국어원 〈한국어 학습자 말뭉치〉를 사용하며 작문 장르를 시험 작문, 과제 작문, 기획 장르로 한정했습니다. 이들은 외부 자료를 참고하지 않는 장르이기에 전체 장르와 비교했을 때 언어적 자질 분포에 차이가 예상됩니다. **본 연구에서 목표로 한 장르 외의 다른 장르에 이 모델을 적용해 보셨는지, 적용했다면 특정 자질에서 어떤 차이를 보였는지 궁금합니다. 적용해 보지 않았다면, 어떤 차이점이 예상되는지 의견을 말씀해 주시기 바랍니다.**
2. 〈한국어 학습자 말뭉치〉는 다양한 모국어 사용자의 글쓰기로 구성되어 있습니다. **혹시 연구 결과를 학습자의 국적, 모국어, 또는 언어군 별로 군집하여 비교, 분석해 보셨는지 궁금합니다.**
3. 본 연구는 다양한 언어적 자질을 종합적으로 고려한 앙상블 모델 구조를 가지고 있습니다. **제안하신 언어적 자질 중 평가 결과에 가장 유의미한 영향을 미친 자질과 그 이유, 그리고 예상보다 영향을 미치지 않은 자질과 그 이유에 대해 말씀해 주시기 바랍니다.**
4. 최근 생성형 AI 분야에서 다양한 글쓰기 평가를 위한 언어 모델 연구(Judge LLM)가 활발히 진행되고 있습니다. **발표자께서는 이 연구가 외국어 학습자의 제 2 언어 글쓰기가 아닌, 1) 모국어를 사용하는 일반인의 글과 2) 생성형 AI 가 생성한 글을 평가하는 모델로서 어떤 활용성이 있다고 보시는지 의견을 부탁드립니다.**
5. 본 연구의 다음 단계인 의미적 자질을 포함한 한국어 글쓰기 평가 모델로 나아가기 위해, **한국어에서 가장 중요한 의미적 자질은 무엇이며, 이를 기계적으로 평가하기 위해 해결해야 할 문제와 어려움에 대한 발표자의 의견을 듣고 싶습니다.**

2024 한국코퍼스언어학회 가을 전국학술대회

인간과 기계의 언어 소통

Session 3

인간 가치 정렬(Human Alignment)를 위한 한국어 지시 이행 (Instruction Following) 말뭉치 설계를 위한 기초 연구

(한지윤, 업스테이지)

—

LLM을 이용한 수학문제 합성데이터 구축

(이숙의, 장지현, 강수희, 홍채은 마인즈솔루션)

—

Bllossom 프로젝트: 한국어 언어 모델의 꽃을 피우다

(함영균, 테디썸)

인간 가치 정렬(Human Alignment)를 위한 한국어 지시 이행(Instruction Following) 말뭉치 설계를 위한 기초 연구

한지윤

upstage

Try Pitch

Index

Try Pitch

인간 가치 정렬(Human Alignment)

정의

AI가 인간의 가치와 목표에 맞추어 작동하도록 설계하는 과정.

중요성

AI의 신뢰성: AI가 인간의 기대와 윤리적 기준에 맞춰 상호작용해야 신뢰를 얻을 수 있음.

실생활 적용: 의료, 법률, 고객 서비스 등 다양한 분야에서 AI가 인간의 요구에 맞는 결과를 제공해야 함.

AI에 특정한 페르소나를 부여하는 것도, 이러한 정의 과정을 통해 이루어질 수 있음

인간 가치 정렬(Human Alignment) 관련 연구

"Concrete Problems in AI Safety" (2016) - Dario Amodei et al.

- AI 안전성의 다섯 가지 주요 문제 제시: 부정적 부작용 방지, 보상 해킹 방지, 확장 가능한 감시 체계, 안전한 탐색, 분포 변화에 대한 견고성

"Deep Reinforcement Learning from Human Preferences" (2017) - Paul F Christiano et al.

- 인간 피드백을 바탕으로 AI가 선호를 학습하는 강화 학습 방법론 소개

"Scalable Agent Alignment via Reward Modeling" (2018) - Jan Leike et al.

- 보상 모델링을 통해 AI 에이전트의 정렬 문제 해결을 위한 확장 가능한 접근 방식 제안

"AI Safety via Debate" (2018) - Geoffrey Irving et al.

- AI 시스템 간의 토론을 통해 정렬을 이루고 위험을 평가하는 방안 논의

"Cooperative AI: Machines Must Learn to Find Common Ground" (2021) - Allan Dafoe et al.

 AI가 인간과 협력하여 공통 목표를 찾고, 다중 에이전트 환경에서의 협력적 행동 탐구

인간 가치 정렬(Human Alignment)을 효율적으로 달성하려면?

Try Pitch

지시 이행(Instruction Following)

정의

AI가 명확한 명령을 받고 이를 정확히 수행하는 능력.

중요성

정확한 지시 수행: 명령을 올바르게 이해하고 처리하는 것은 AI의 핵심 기능.

실용적 응용: 문서 작성, 번역, 문제 해결 등 다양한 분야에서 유용.

Try Pitch

지시 이행(Instruction Following) 관련 연구

"InstructGPT: Training language models to follow instructions with human feedback" (2022) - Long Ouyang et al. (OpenAI)

- InstructGPT 모델 개발 시, 데이터 수집과 정제 방법을 상세히 설명하고, 인간 피드백(RLHF)을 통한 훈련 방식 제안

"FLAN: Finetuning Language Models with Natural Instructions" (2022) - Yuwei Fang et al. (Google)

- 다양한 자연어 처리(NLP) 작업을 위한 자연어 지시사항 데이터셋 구축 과정 설명

"Self-Instruct: Aligning Language Models with Self-Generated Instructions" (2023) - Yizhong Wang et al.

- 대규모 언어 모델을 활용하여 자동으로 지시 데이터를 생성하는 방법을 제안하며, 효율적인 정렬을 목표함

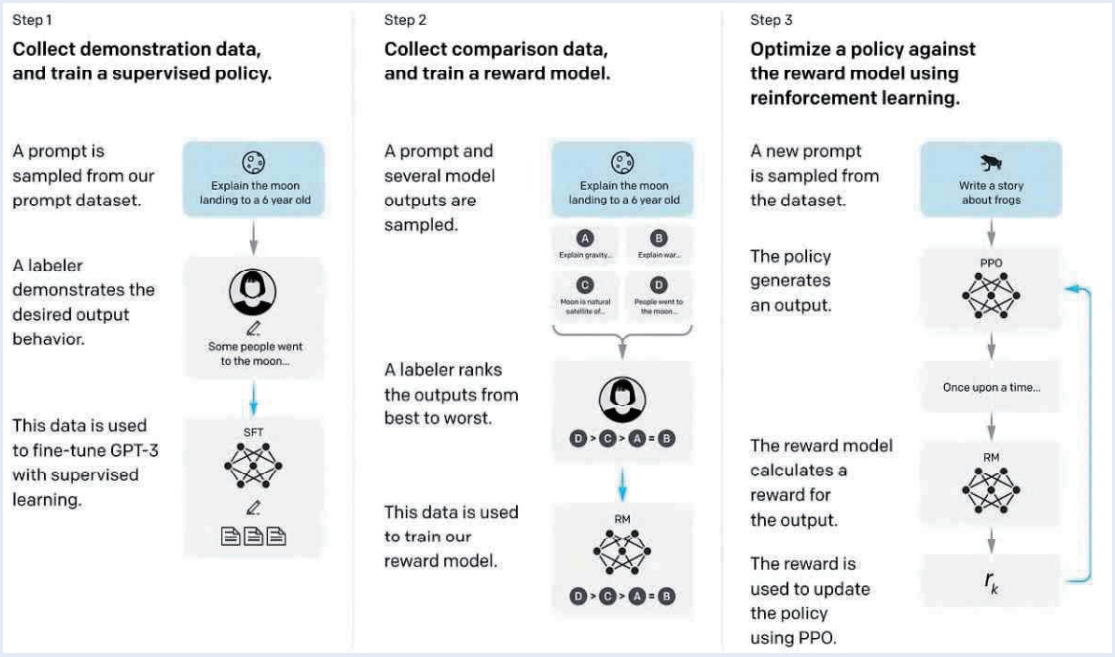
"Multitask Prompted Training Enables Zero-Shot Task Generalization" (2022) - Victor Sanh et al.

- T0 모델 학습을 위한 다중 작업 프롬프트 데이터셋 구축 과정과 그 효과 논의

"ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning" (2022) - Vamsi Aribandi et al.

- 107개의 다양한 NLP 작업을 대상으로 한 instruction 데이터셋 구축 방법을 상세히 설명

Aligning language models to follow instructions



https://openai.com/index/instruction-following/

RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback

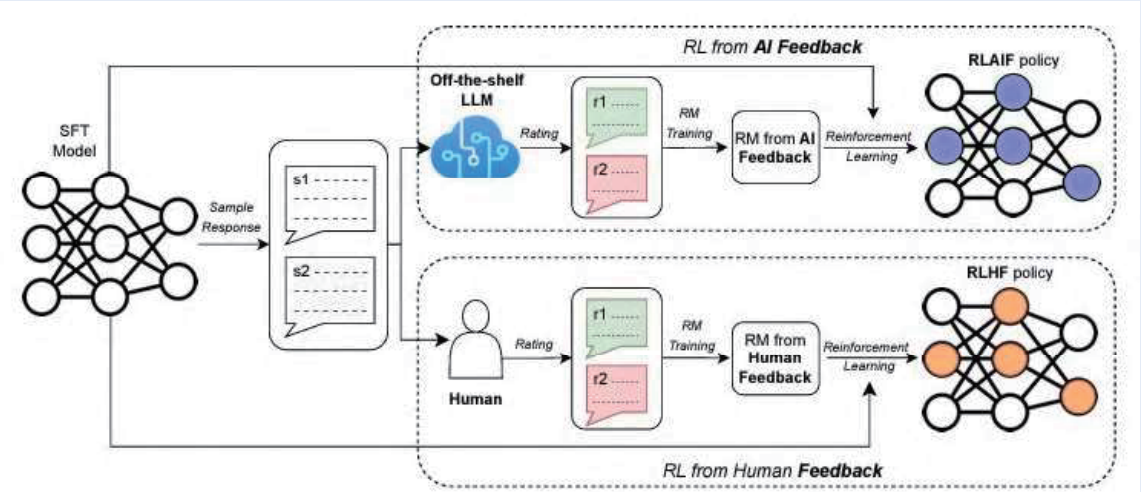
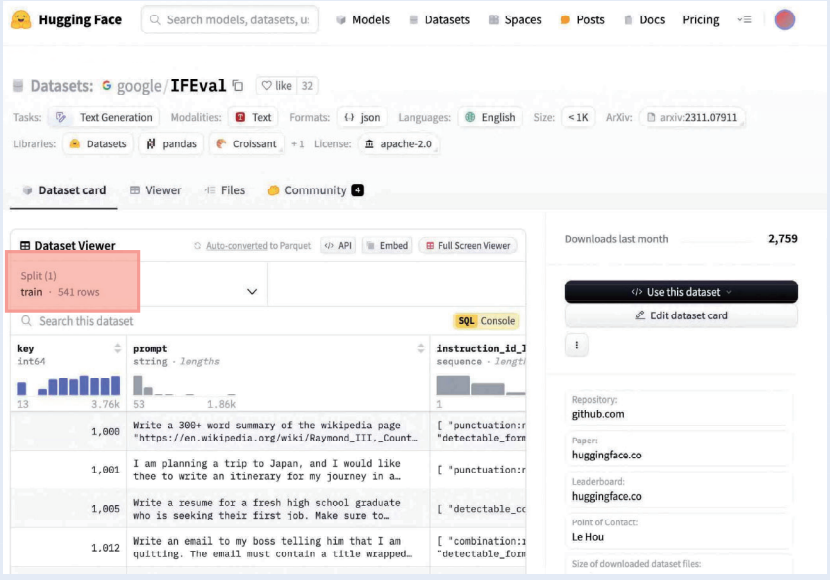


Figure 2: A diagram depicting RLAIF (top) vs. RLHF (bottom)

Try Pitch

https://openai.com/index/instruction-following/

IFEval



Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., ... & Hou, L. (2023). Instruction-following evaluation large language models. *arXiv preprint arXiv:2311.07911*.

Try Pitch

IFEval

Data Fields

The data fields are as follows:

- `key`: A unique ID for the prompt.
- `prompt`: Describes the task the model should perform.
- `instruction_id_list`: An array of verifiable instructions. See Table 1 of the paper for the full set with their descriptions.
- `kwargs`: An array of arguments used to specify each verifiable instruction in `instruction_id_list`.

key int64 763–1.14k 4.8%	prompt string · lengths 53–234 69.9%	instruction_id_list sequence · lengths 1–2 56.4%	kwargs list · lengths 1–2 56.4%
1,019	Given the sentence "Two young boys with toy guns and horns." can you ask a question? Please ensure that your response is in English, and in all lowercase letters. No capital letters are allowed.	["change_case:english_lowercase"]	[{ "num_highlights": null, "relation": null, "num_words": null, "num_placeholders": null, "prompt_to_repeat": null, "num_bullets": null, "section_spliter": null, "num_sections": null, "capital_relation": null, "capital_frequency": null, "keywords": null, "num_paragraphs": null, "language": null, "let_relation": null, "letter": null, "let_frequency": null, "end_phrase": null, "forbidden_words": null, "keyword": null, "frequency": null, "num_sentences": null, "postscript_marker": null, "first_word": null, "nth_paragraph": null }]

Try Pitch

IFEval

Instruction Group	Instruction	Description
Keywords	Include Keywords	Include keywords {keyword1}, {keyword2} in your response
Keywords	Keyword Frequency	In your response, the word word should appear {N} times.
Keywords	Forbidden Words	Do not include keywords {forbidden words} in the response.
Keywords	Letter Frequency	In your response, the letter {letter} should appear {N} times.
Language	Response Language	Your ENTIRE response should be in {language}, no other language is allowed.
Length Constraints	Number Paragraphs	Your response should contain {N} paragraphs. You separate paragraphs using the markdown divider: * * *
Length Constraints	Number Words	Answer with at least / around / at most {N} words.
Length Constraints	Number Sentences	Answer with at least / around / at most {N} sentences.
Length Constraints	Number Paragraphs + First Word in i-th Paragraph	There should be {N} paragraphs. Paragraphs and only paragraphs are separated with each other by two line breaks. The {i}-th paragraph must start with word {first_word}.
Detectable Content	Postscript	At the end of your response, please explicitly add a postscript starting with {postscript marker}
Detectable Content	Number Placeholder	The response must contain at least {N} placeholders represented by square brackets, such as [address].
Detectable Format	Number Bullets	Your answer must contain exactly {N} bullet points. Use the markdown bullet points such as: * This is a point.
Detectable Format	Title	Your answer must contain a title, wrapped in double angular brackets, such as <<poem of joy>>.
Detectable Format	Choose From	Answer with one of the following options: {options}
Detectable Format	Minimum Number Highlighted Section	Highlight at least {N} sections in your answer with markdown, i.e. *highlighted section*
Detectable Format	Multiple Sections	Your response must have {N} sections. Mark the beginning of each section with {section_splitter} X.
Detectable Format	JSON Format	Entire output should be wrapped in JSON format.
Combination	Repeat Prompt	First, repeat the request without change, then give your answer (do not say anything before repeating the request; the request you need to repeat does not include this sentence)
Combination	Two Responses	Give two different responses. Responses and only responses should be separated by 6 asterisk symbols: ******.
Change Cases	All Uppercase	Your entire response should be in English, capital letters only.
Change Cases	All Lowercase	Your entire response should be in English, and in all lowercase letters. No capital letters are allowed.
Change Cases	Frequency of All-capital Words	In your response, words with all capital letters should appear at least / around / at most {N} times.
Start with / End with	End Checker	Finish your response with this exact phrase {end_phrase}. No other words should follow this phrase.
Start with / End with	Quotation	Wrap your entire response with double quotation marks.
Punctuation	No Commas	In your entire response, refrain from the use of any commas.

Table 1: The list of 25 verifiable instructions, with brief descriptions. We use these instructions because we think they are either easy to verify or common in real-world applications. The list can be expanded trivially. For example, one can add "Language - Mixed Two Languages in Response" and "Detectable Format - XML Format".

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., ... & Hou, L. (2023). Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Try Pitch

IFEval

2,139	Compose a song with at least three sentences that can be sung by a professional singer in the style of a 1930s jazz standard. Include the keywords "rate" and "rte".	["length_constraints:number_sentences", "keywords:existence"]
2,787	What are the steps to get the GNSS timestamp on Android? Explain this to teenagers using at least 4 sentences and make sure the letter n appears at least 3 times.	["keywords:letter_frequency", "length_constraints:number_sentences"]

Try Pitch

단순히 문자적/표기적 특성만 반영하면 될까?

Try Pitch

언어적 특성을 반영한 지시 이행 학습의 필요성

언어의 문법적 차이

각 언어는 서로 다른 문법 구조를 가짐(예: SVO vs. SOV). 이러한 차이를 이해해야 AI가 정확하게 지시를 이행할 수 있음.

문화적 맥락의 중요성

언어는 그 문화와 깊이 연관되어 있으며, AI는 그 맥락을 이해해야 함 (예: 한국어의 존댓말/반말 구분).

어휘적 뉘앙스 차이

같은 단어도 언어와 문화에 따라 의미가 다르게 해석될 수 있음. AI는 이러한 미묘한 차이를 파악하고 대응해야 함.

정확한 의도 파악

뉘앙스와 간접 표현을 이해하는 AI는 인간의 의도를 더 정확히 파악하고, 신뢰할 수 있는 결과를 제공함.

교차 언어적 이해

다양한 언어 사용자에게 적절한 응답을 제공하기 위해 AI가 각 언어의 특성을 반영해야 함.

한국어로 소통을 더욱 잘 하려면?

한국어 지시 이행 말뭉치 설계 시 고려 요소

존비어 : 존댓말/반말
호칭어, 위계, 친소, 어투
문체와 장르, 사용역
문학적 표현, 비즈니스 메일, 일상 대화, 보고서, 학술적 문헌, 방언, 매체 등 언어의 사용 상황과 연관
한국 고유의 문화적 표현
속담과 관용어, 유행어
수 표현
서수와 기수, 단위 명사, 수량사

Try Pitch

한국어 지시 이행 말뭉치 설계 시 고려 요소

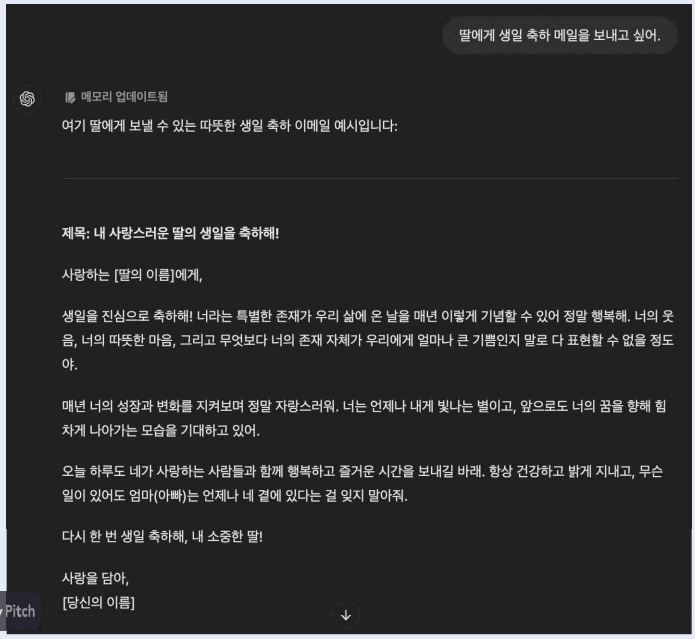
존비어 : 존댓말/반말
호칭어, 위계, 친소, 어투
문체와 장르, 사용역
문학적 표현, 비즈니스 메일, 일상 대화, 보고서, 학술적 문헌, 방언, 매체 등 언어의 사용 상황과 연관
한국 고유의 문화적 표현
속담과 관용어, 유행어
수 표현
서수와 기수, 단위 명사, 수량사

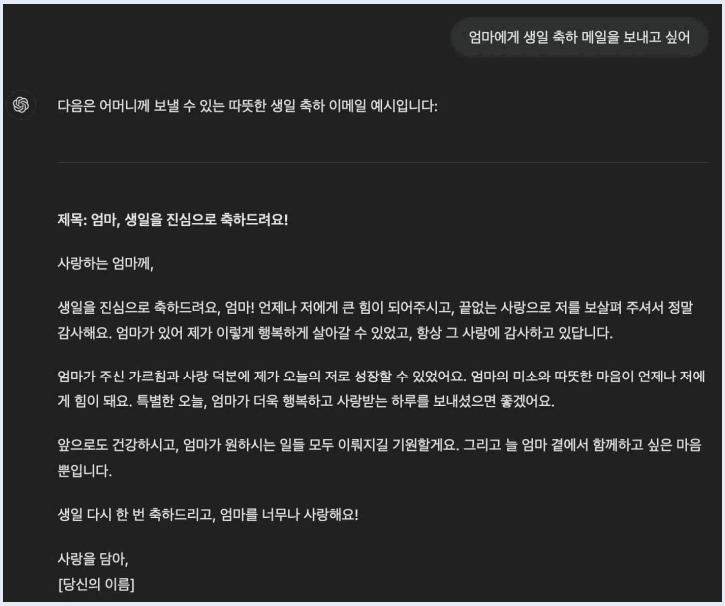
→ **누가, 누구에게, 어디에서**

Try Pitch

한국어 지시 이행 말뭉치 설계 시 고려 요소

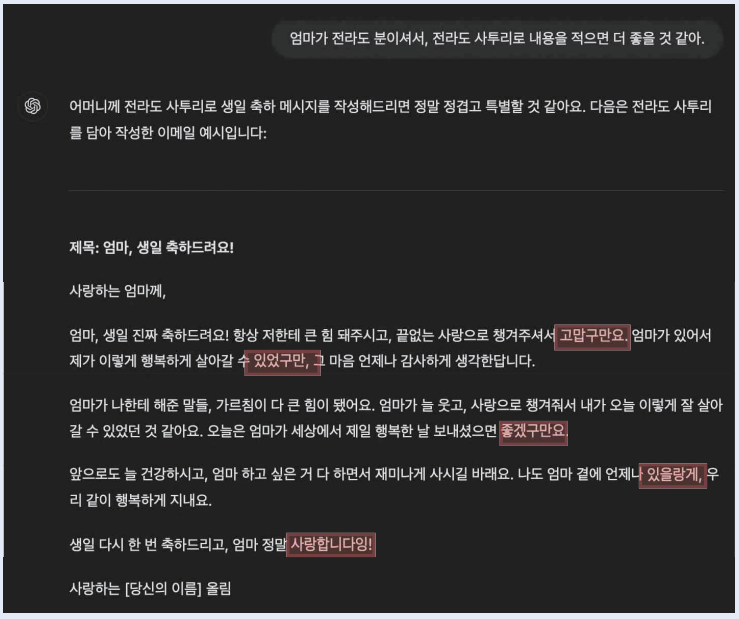
존비어 : 존댓말/반말





한국어 지시 이행 말뭉치 설계 시 고려 요소

문체와 장르, 사용역



한국어 지시 이행 말뭉치 설계 시 고려 요소

문체와 장르, 사용역

외부 거래처 대표님께 보낼 이메일 작성해줘. 존댓말로 아래 내용을 포함하도록 작성해줘.

미팅 일정 : 2024년 10월 11일

미팅 주제 : Kick-off

미팅 장소 :

참가자 :

다음과 같이 이메일을 작성하시면 좋겠습니다:

제목: Kick-off 미팅 일정 안내드립니다.

안녕하세요, 「대표님.

항상 저희 회사에 많은 관심과 성원 보내주셔서 감사드립니다.

다름이 아니오라, 저희의 프로젝트를 위한 Kick-off 미팅 일정을 아래와 같이 확정하여 안내드립니다.

- **일시:** 2024년 10월 11일 (금)
- **주제:** Kick-off 미팅
- **장소:** [장소를 입력해주세요]
- **참가자:** [참가자를 입력해주세요]

미팅에 대한 세부사항이나 추가 요청사항이 있으시면 언제든지 말씀해주시기 바랍니다.
그럼 미팅에서 뵙기를 기대하며, 좋은 하루 되시길 바랍니다.

감사합니다.

[당신의 이름]
[당신의 직책]
[당신의 회사명]
[당신의 연락처]

Try Pitch

한국어 지시 이행 말뭉치 설계 시 고려 요소

한국 고유의 문화적 표현

당나라식 계산법이 무슨 뜻이야?

'당나라식 계산법'이라는 용어는 일반적으로 잘 알려져 있지 않습니다. 이 용어가 특정한 맥락이나 분야에서 사용되는 것인지 추가 정보가 필요합니다.

하지만 당나라 시대의 수학에 대해 일반적으로 말씀드릴 수 있는 것은:

1. 당나라(618-907)는 중국 수학 발전의 중요한 시기였습니다.
2. 이 시기에 '십진법'과 '영(0)'의 개념이 더욱 발전했습니다.
3. 대표적인 수학서로 '산경십서(算經十書)'가 있었습니다.
4. 당시 인도 수학의 영향도 받았습니다.

'당나라식 계산법'이 이 중 어떤 것을 가리키는지, 또는 다른 특정한 방법을 의미하는지 확실치 않습니다. 이 용어를 들으신 구체적인 맥락이 있다면 알려주시면 더 정확한 정보를 제공할 수 있을 것 같습니다.

Claude can make mistakes.
Please double check response.

2024.08

당나라식 계산법이 무슨 뜻이야?

"당나라식 계산법"은 한국에서 사용되는 관용적 표현입니다. 이 표현의 핵심 의미는 다음과 같습니다:

1. 부정확하고 임의적인 계산 방식을 가리킵니다.
2. 주로 논리나 근거 없이 대충 어림잡아 계산하는 것을 비유적으로 표현합니다.
3. 자신에게 유리한 방향으로 수치를 조작하거나 해석하는 경우를 지칭하기도 합니다.
4. 정부나 기업의 불투명한 통계 발표나 예산 책정을 비판할 때 자주 사용됩니다.

이 표현은 실제 당나라의 계산 방식과는 관련이 없으며, 단지 멀고 오래된 과거를 상징하는 의미로 "당나라"가 사용된 것으로 보입니다.

더 자세한 설명이나 예시가 필요하시다면 말씀해 주세요.

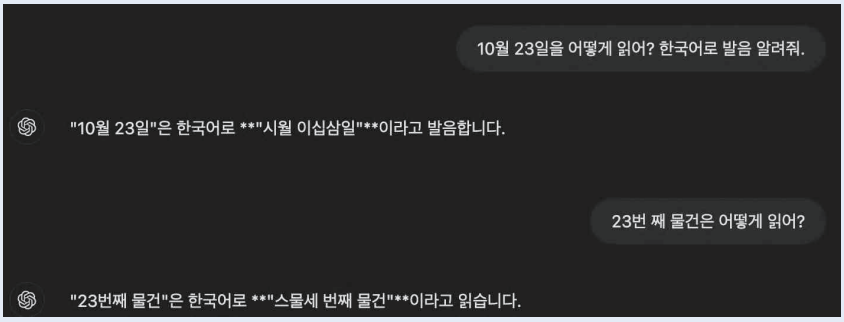
Claude can make mistakes.
Please double check responses.

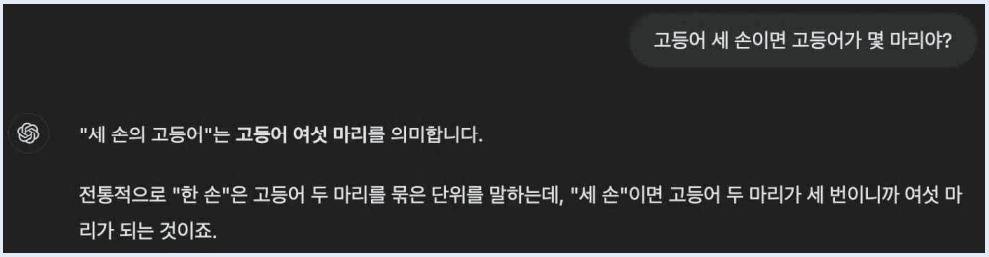
2024.10

Try Pitch

한국어 지시 이행 말뭉치 설계 시 고려 요소

수 표현





Try Pitch

요약 및 결론

인간 가치 정렬은 AI 시스템이 인간과 자연스럽게 개발자의 의도대로 상호작용하는 데 필수적이며,
한국어에 특화된 지시 이행 말뭉치 설계는 한국어 사용자 경험을 향상시키는 중요한 단계

Try Pitch

경청해 주셔서 감사합니다!

Try Pitch

참고 문헌

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565.*

Aribandi, V., Tay, Y., Schuster, T., Rao, J., Zheng, H. S., Mehta, S. V., Zhuang, H., Tran, V. Q., Bahri, D., Ni, J., Gupta, J., Hui, K., Ruder, S., & Metzler, D. (2022). ExT5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952.*

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Chatterjee, A., Gupta, N., Aggarwal, P., Narang, A., Chen, D., & Raychev, V. (2023). IF-Eval: An Evaluation Framework for Assessing Instruction-Following Capabilities of Language Models. *arXiv preprint arXiv:2305.11426.*

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

Fang, Y., Zhao, S., Huang, S., & Zhao, T. (2022). FLAN: Finetuning language models with natural instructions. *arXiv preprint arXiv:2210.02441.*

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786.*

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155.*

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Penguin.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., ... & Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207.*

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf.

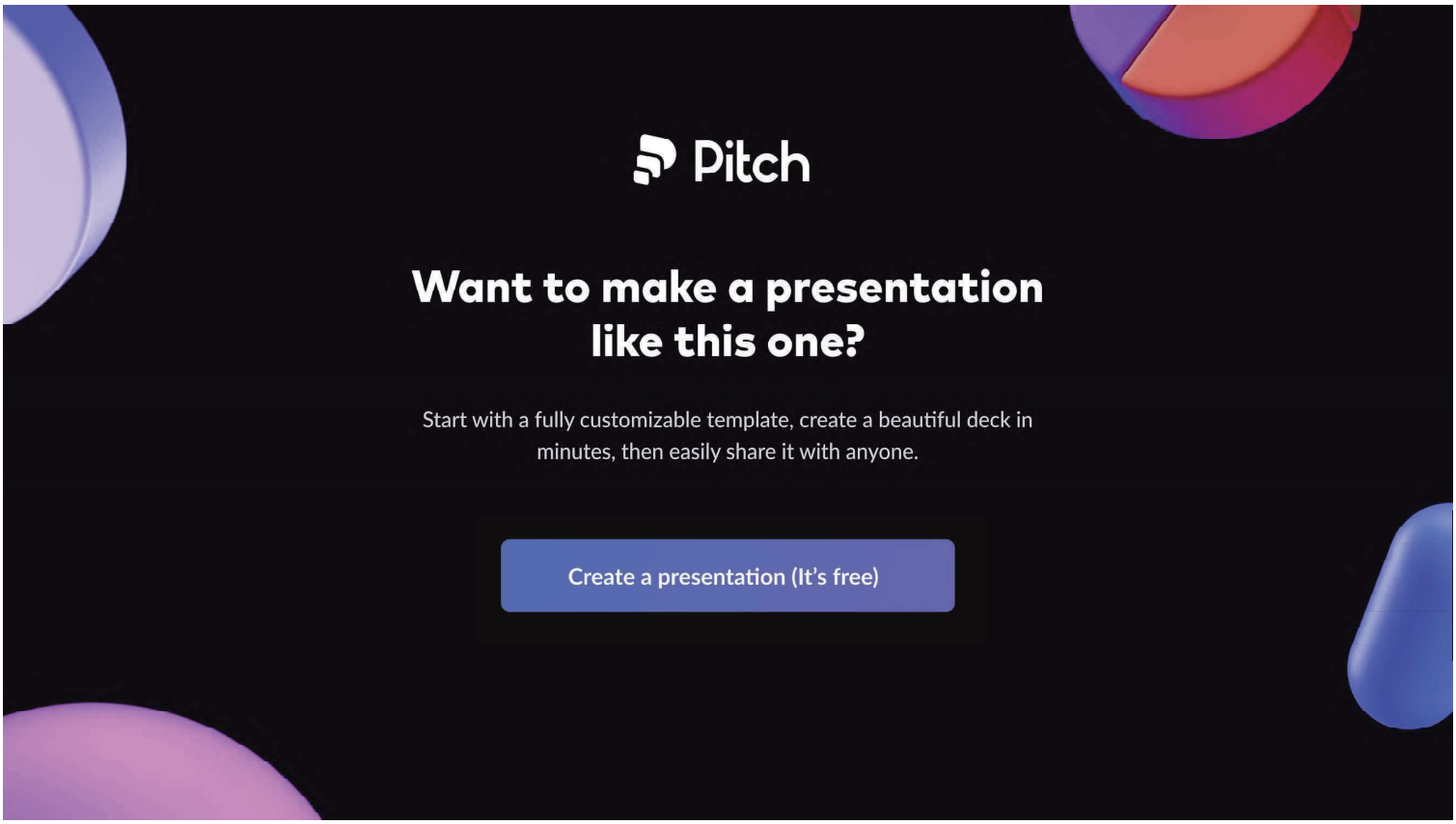
Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-Instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560.*

Wang, Y., Mishra, S., Alipoormolabashi, P., Zhang, Y., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H. A., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Deshpande, K., ... & Khashabi, D. (2022). Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. *arXiv preprint arXiv:2204.07705.*

Xie, T., Wu, N., Wang, X., Zhu, Y., & Jiang, Y. (2023). UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966.*

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., ... & Hou, L. (2023). Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911.*

Try Pitch


Pitch
Want to make a presentation like this one?
Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.
Create a presentation (It's free)

토론문 (발표: 한지윤, 토론: 조원익)

안녕하세요. 저는 본 발표문의 토론을 맡은 조원익입니다. 본 발표문은 AI가 인간의 가치와 목표에 맞추어 작동하도록 설계하는 인간 가치 정렬(human value alignment)에 대해 이야기하고 있으며, 이것이 첫 번째로 신뢰 가능하고 윤리적인 AI를 만들기 위해, 또 두번째로 실생활에 적용 가능한 유용한 AI를 만들기 위해 반드시 필요한 과정임을 역설하고 있습니다. 이러한 가치 정렬의 효율적 달성을 위해 지시 이행, 흔히들 instruction tuning으로 많이 들어보셨을 instruction following을 제시하고 있는데요, 이에 따른 두 가지 기대 효과로 정확한 지시 수행과 실용적 응용을 제시하며, InstructGPT와 RLHF 및 RLAIF, 그리고 Instruction Following 평가 데이터셋인 IFEval까지 관련 내용을 소개합니다.

저는 여기서 두 가지 주제로 토론 질의를 드려 보고자 합니다.

우선 인간 가치 정렬과 지시 이행에 대해 토론 나눠보겠습니다.

첫 번째로, Instruction Following은 주어진 발화의 의도와 목적을 모델 혹은 기계가 이해하고 그에 맞추어 동작하게 하기 위한 학습용 데이터셋으로 파악되는데, 이는 유용한 AI를 만드는 데에는 충분히 도움이 될 듯하나 처음에 이야기한 신뢰 가능하고 윤리적인 AI를 만드는 데에는 어떠한 방식으로 기여할 수 있을지에 대한 의문이 듭니다. 대표적으로 지시 이행 능력을 평가하는 IFEval의 경우 텍스트의 포맷팅이나 문장 형태 등이 원하는 대로 출력되었는지를 판단하는 것이 목적이며, 출력 결과 자체가 safe한지, 혹은 trustworthy한지 등에 관해서는 논하기 어려운 듯합니다. 혹시 이러한 부분까지 고려한 Instruction Following 연구 혹은 벤치마크가 있을까요? 나아가, 개념적으로 보았을 때, 인간 가치 정렬이 AI를 인간, 정확히는 개발하는 사람이 원하는 인간의 행동양식대로 작동하도록 설계하는 것이고, 여기에서 이야기하는 인간의 기대, 윤리적 기준, 그리고 활용 분야 등 '맥락'으로 부를 수 있는 요소가 다양하게 있다면, 이러한 맥락 정보가 어떤 형태로 Instruction Following시에 반영될 수 있는지 궁금합니다.

두 번째로 한국어에서의 지시 이행에 대해 토론 나눠보겠습니다.

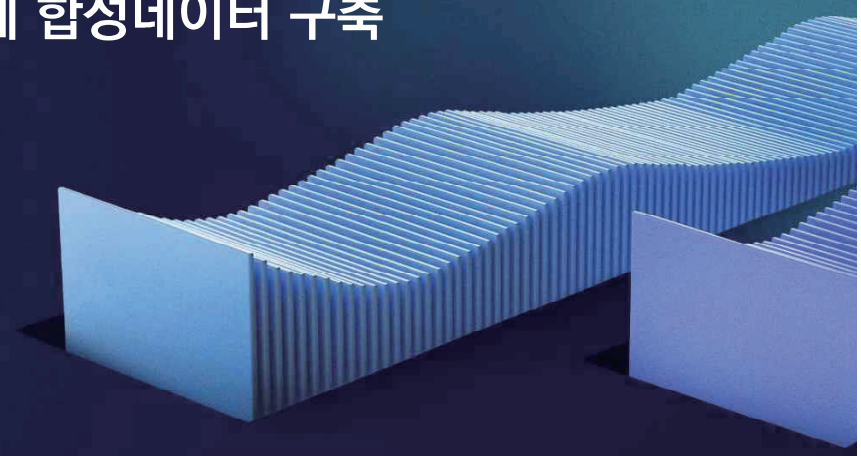
본 발표에서는 한국어 지시 이행 말뭉치를 구축할 때 고려할 사항들을 한국어의 특성을 반영하는 다양한 관점에서 소개하고 있습니다. 이로써, 한국어에서 고려할 점들이 IFEval에서와 같은 단순 형식적인 측면들 그 이상임은 충분히 확인이 되었습니다. 드리고 싶은 질문은, 이러한 요소들 중 한국어뿐 아니라 영어를 포함한 다른 언어에서들에서도 적용이 가능하거나 혹은 필요한 부분들을 어떤 과정을 통해 결정할 수 있을지에 대해서입니다. 예컨대 존비 혹은 language variety에 대한 부분은 타 언어에서도 politeness/formality나 지역별 언어(Britain vs American English 등)을 통해 나타날 수 있는 요소이고, 그 중에서도 한국어에 있어서 말씀하신 요소들을 우선적으로 택한 것이, 한국어가 한국이라는 문화권에서 주로 사용되기 때문인가에 대한 질문을 드리고 싶습니다. 또한, 영어 기반으로 대부분의 지시 이행 사전학습 데이터들이 공개되고 있는 요즈음, 공개된 영어 데이터에 말씀하신 요소들을 반영하여 개별어된 데이터로 재가공하는 자동화된 파이프라인이 가능할지, 가능하다면 어떤 주의사항이 있을지도 궁금합니다.

2024년도
한글 및 한국어 정보처리 & 한국코퍼스언어학회 공동학술대회

LLM을 이용한 수학문제 합성데이터 구축

발표 2024. 10. 11

이숙의, 장지현, 강수희, 홍채은



MINDS SOLUTION

목차

LLM을 이용한 수학문제 합성 데이터 구축

MINDS SOLUTION

1. 서론

수학 추론 능력과 LLM의 논리성

- 추론
  - 이미 알고 있는 사실이나 명제를 토대로 결론을 이끌어 내는 사고 과정, 알려진 사실에서 모르는 것을 이끌어 내는 과정
- 최근 LLM의 추론 능력 = AI의 '지능'
  - 수학 문제 해결에서 Scratchpad, Chain of Thought, Self-discover 등 LLM이 가진 추론 능력을 최대 활용하는 방안 모색 활발
  - 현 과제의 CoT는 구조화된 논리 흐름에 따라 각 단계가 명확하게 연결되며 논리적으로 정리된 추론 과정을 통해 답을 도출
  - 문제를 단계별로 나누어 서술형 문제 풀이의 추론과정 설명에 적합
  - 사람이 문제를 풀 때 문제 유형에 따라 구체적인 추론 방법을 설계하고, 그 후 실제 풀이 단계를 거치는 방식을 LLM에 적용한 방식에 따라 여러 LLM과 여러 task에서 전반적 성능향상의 효과가 있음을 제시
  - 이외에서 추론 방법 설계의 중요성과 구체적 방법에 관한 논의가 매우 활발하며, 이러한 LLM의 추론 능력을 통해 여러 목적의 합성데이터를 구축하려는 시도도 활발함
    (2024, by USC, Google DeepMind)

LLM을 활용한 수학문제해결

- LLM 최초 등장 당시 수학 추론 능력에서 약점을 보임
- 수학문제는 논리적 추론 과정을 적용한다는 점에서 LLM 논리성 검증의 좋은 도구
- 최근 GPT-4 이후 수학문제풀이에 관한 성능이 뛰어나다고 하나 영어 데이터 기반에 기반함
- 한국어 데이터 구축 필요

2. 관련연구

번역을 통한 수학문제 데이터 구축

- LLM 추론 능력 향상을 위한 '서술형 문제 – 답' 쌍으로 한국어 번역 코퍼스 구축
- 영어 데이터 활용
- GSM8k와 MetaMathQA를 활용한 한국어 번역 데이터 생성
  - 초등학교 수준의 수학문제, 8,000개 문제 포함
  - 산술, 기하, 소수점 계산 포함
  - 기초적 수학 능력 평가에 적합
  - LLM의 기본적인 수학 개념 이해도 평가에 사용
  - 다양한 문제 유형 포함, 범용적인 문제 해결 능력
- LLM을 통한 합성데이터 구축을 통해 한국어 수학문제데이터를 구축, 이를 통해 한국형 LLM 논리성 검증

2. 관련연구

영-한 번역 데이터 구축 유의점

- 문제 오류
  - 문화적 차이 고려
    - 수동 번역 작업 필요
    - 단순 번역으로 문제-풀이 쌍을 만들 경우 한국인이 이해하기 어려운 수학 문제 문맥으로 번역됨
- 풀이 오류
  - Solution의 풀이 중복
  - 풀이 과정 의미를 한국어법에 맞게 전달하기 위해서 어휘 삽입 필요, 이를 통해 단계별 추론이 명확하게 드러날 수 있음
- 원문 오류
  - 문제 내 오류 포함(의미 오류)
  - 수식 오류 포함

2. 관련연구

문제 오류

- 문화적 차이에 의해 단순 번역으로 문제-풀이 쌍을 만들 경우 한국인이 이해하기 어려운 수학 문제 문맥으로 번역됨
  - 배수를 나타내는 문제일 경우, "1년*0.5 + (1년*0.5)*0.5 + ((1년*0.5)*0.5)*0.5" 로 계산하는 방식과 "1년*0.5*3년치" 로 계산하는 방식의 차이가 있어서 한국인이 단순 접근하기가 어려움.

"problem": "Mrs. Tatiana owns a grocery store that sells different fruits and vegetables, which includes carrots. The price of carrots in the grocery store increases by 5% of the original price every year. What would be the price of carrots after three years if it was $120 initially? (Round to the nearest integer)",
"grade": "",
"type": "",
"solution": "How much does the price of carrots increase in the first year? ** In the first year, the price of carrots increases by 5/100*120 = $<<5/100*120=6>>6₩nHow much does the price of carrots increase after 3 years? ** After 3 years, the price of carrots will increase by $6*3 = $<<6*3=18>>18₩nWhat will be the total price of carrots after three years? ** The total price of carrots after three years will be 120+18 = $<<120+18=138>>138₩n#### 138",
"equation": "5/100*120=6₩n6*3=18₩n120+18=138",
"code": "a=5/100*120₩nb=a*3₩nc=120+b₩nanswer=c₩nprint(answer)",
"answer": "138"

"problem": "타티아나가 운영하는 식료품점의 당근 가격은 매년 원래 가격의 5%씩 상승합니다. 처음에 당근 가격이 120달러였다면 3년 후 당근 가격은 얼마일까요? (가장 가까운 정수로 반올림).",
"grade": "",
"type": "",
"solution": "첫해에 당근 가격은 얼마나 상승하나요? ** 첫해에 당근 가격은 5/100*120 = $<<5/100*120=6>>6 증가합니다. 3년 후 당근 가격은 얼마나 상승하나요? ** 3년 후 당근 가격은 $6*3 = $<<6*3=18>>18 증가합니다. 3년 후 당근의 총 가격은 얼마인가요? ** 3년 후 당근의 총 가격은 120+18 = $<<120+18=138>>138이 됩니다.#### 138",
"equation": "5/100*120=6₩n6*3=18₩n120+18=138",
"code": "a=5/100*120₩nb=a*3₩nc=120+b₩nanswer=c₩nprint(answer)",
"answer": "138"

풀이 오류

- Solution의 풀이 중복
- 풀이 과정의 의미를 명확히 전달하기 위해서는 '더', '총' 등의 어휘 삽입으로 단계별 추론의 근거 문장을 제시함

"problem": "A train has 172 people traveling on it. At the first stop 47 people get off and 13 more people get on, and at the next stop another 38 people get off. How many people are on the train?",
"grade": "",
"type": "",
"solution": "How many people are on the train at the first stop? ** At the first stop 172 people - 47 people = <<172-47=125>>125 people are on the train.₩nHow many people are on the train at the first stop? ** Then 13 more people get on increasing it to 125 people + 13 people = <<13+125=138>>138 people.₩nHow many people are on the train at the next stop? ** At the next stop 138 people - 38 people = <<138-38=100>>100 people remain on the train.₩n#### 100",
"equation": "172-47=125₩n13+125=138₩n138-38=100",
"code": "a=172-47₩nb=13+a₩nc=b-38₩nanswer=c₩nprint(answer)",
"answer": "100"

"problem": "열차에는 172명이 탑승하고 있습니다. 첫 번째 정거장에서 47명이 내리고 13명이 더 타고, 다음 정거장에서 38명이 더 내립니다. 기차에는 몇 명이 타고 있나요?",
"grade": "",
"type": "",
"solution": "첫 번째 정거장에 얼마나 많은 사람이 기차에 타고 있나요? ** 첫 번째 정거장에서 1/2멍 - 4/멍 = <<172-47=125>>125명이 열차에 타고 있습니다.첫 번째 정거장에서 기차에 몇 명이 타고 있나요? ** 13명이 더 탑승하여 125명 + 13명 = <<13+125-138>>138명으로 증가합니다.다음 정거장에는 몇 명이 기차에 타고 있나요? ** 다음 정거장에서 138명 - 38명 = <<138-38=100>>100명이 기차에 남아 있습니다.#### 100",
"equation": "172-47-125₩n13+125-138₩n138-38-100",
"code": "a=172-47₩nb=13+a₩nc=b-38₩nanswer=c₩nprint(answer)",
"answer": "100",
"id": 784

원문 오류

- 번역문 오류 단순 정제(영어 이름, 단순 실수) – 문제 이해 – 문제 풀이(수식, 답 확인) – 원문과 번역문 대조 및 확인 – 번역문 최종 수정

{
"problem": "Lauren is a cartoonist. She can draw 5 large-sized picture scenes per day, or she can draw 6 medium-sized picture scenes per day, or she can draw 7 small-sized picture scenes per day. She was hired for a big project to create 45 large-sized picture scenes, 36 medium-sized picture scenes, and 49 small-sized picture scenes. How many days will it take for her to create all of the picture scenes?",
"grade": "",
"type": "",
"solution": "How many days will it take to create 45 small-sized picture scenes? ** At 5 small-sized picture scenes per day, 45 small-sized scenes will take 45/5 = <<45/5=9>>9 days.₩nHow many days will it take to create 36 medium-sized picture scenes? ** At 6 medium-sized picture scenes per day, 36 medium-sized scenes will take 36/6 = <<36/6=6>>6 days.₩nHow many days will it take to create 49 large-sized picture scenes? ** At 7 large-sized picture scenes per day, 49 large-sized scenes will take 49/7 = <<49/7=7>>7 days.₩nHow many days will it take to create all of the picture scenes? ** Altogether, the entire project will take 9 + 6 +7 = <<9+6+7=22>>22 days to complete.₩n#### 22",
"equation": "45/5=9₩n36/6=6₩n49/7=7₩n9+6+7=22",
"code": "answer = 22₩nprint(answer)",
"answer": "22"
},

{
"problem": "로렌은 만화가입니다. 하루에 5개의 대형 그림 장면을 그릴 수 있고, 하루에 6개의 중형 그림 장면을 그릴 수 있으며, 하루에 7개의 소형 그림 장면을 그릴 수 있습니다. 그녀는 대형 그림 장면 45개, 중형 그림 장면 36개, 소형 그림 장면 49개를 만드는 대형 프로젝트에 고용되었습니다. 그녀가 모든 그림 장면을 만드는 데 며칠이 걸리나요?",
"grade": "",
"type": "",
"solution": "큰 크기의 그림 장면 45개를 만드는 데 며칠이 걸리나요? ** 하루에 5개의 작은 그림 장면을 만들면 45개의 작은 그림 장면은 45/5 = <<45/5=9>>9일이 소요됩니다.중간 크기의 그림 장면 36개를 만드는 데 며칠이 걸리나요? ** 하루에 6개이 중간 크기 그림 장면을 만들면 36개의 중간 크기 장면은 36/6 = <<36/6=6>>6일이 소요됩니다.소형 그림 장면 49개를 만드는 데 며칠이 걸리나요? ** 하루에 7개의 소형 그림 장면을 만들면 49개의 소형 장면은 49/7 = <<49/7=7>>7일이 소요됩니다.모든 그림 장면을 만드는 데 며칠이 걸리나요? ** 전체 프로젝트를 완료하는 데 9 + 6 +7 = <<9+6+7=22>>22일이 소요됩니다.#### 22",
"equation": "45/5=9₩n36/6=6₩n49/7=7₩n9+6+7=22",
"code": "answer = 22₩nprint(answer)",
"answer": "22",
"id": 1139
},

3. 수학문제 합성데이터 구축

수학문제 생성 프롬프트

- LLM을 통한 합성데이터 생성으로 한국어 수학문제데이터를 구축, 이를 통해 한국형 LLM 논리성 검증
- 〈수학문제 생성 프롬프트 작성에 관한 실험 (김재환, 2024)〉
  - 영어 번역 데이터 기반 LLM을 활용한 수학문제 생성을 위한 프롬프트 성능 실험
  - 규제 프롬프트 추가 – "The guidelines below are very strict and must be followed."
  - 이를 모방하여 추론 과정 및 성능을 유의미하게 향상시킴
  - GPT-3.5-Turbo 모델 기본 성능보다 17% 향상
  - Q-S-A 모두 제시한 답변 비율 96.67%

표 4 규제 프롬프트 추가 구조

기존 영문 프롬프트 도입부	추가된 규제 도입부
Integrate step-by-step reasoning and Python code to solve math problems using the following guidelines: - Every response should have bullet of question, solution, and answer. - Analyze the question and write functions to solve the problem; the function should not take any arguments.Here are some examples you may refer to:	Integrate step-by-step reasoning and Python code to solve math problems using the following guidelines: **The guidelines below are very strict and must be followed.** - Every response should have bullet of question, solution, and answer. - Analyze the question and write functions to solve the problem; the function should not take any arguments. Here are some examples you may refer to:

〈김재환(2024:18)의 표 발췌〉

3. 수학문제 합성데이터 구축

LLM을 활용한 합성데이터 구축

- LLM이 단계별로 문제를 해결하도록 유도하는 프롬프트 작성 방법
  - 수학문제 합성데이터 구축에서 CoT 작동방식에 따른 프롬프트 다양화 필요
  - 데이터의 정확성 면에서 Zero-Shot 불가
  - 다양성 면에서 One-Shot CoT 기반 Few-Shot의 적용
  - 모델이 추론하기 원하는 논리 과정 Few-Shot 학습으로 제공

- 활용 LLM
  - GPT-4o, Gemini, Claude 3.5 Sonnet

- Seed 문제 추출
  - 현행 초등학교 5-6학년 수학문제집의 서술형 활용
  - 저작권에 위반되지 않는 Seed 데이터 사용 필요
  - 현행 교과서와 주요 문제집 7종을 활용하여 Seed 문제를 선별하였으며, 저작권을 고려하여 말뭉치 정보에 원 데이터 출처를 반드시 기재함

3. 수학 문제 합성데이터 구축

프롬프트 작성 : 다양성과 원칙

- One-Shot 기반 Few-Shot 적용
  - 세 개 LLM을 통해 유사 유형 문제 생성으로 다양성 확보 , GPT-4o, Gemini, Claude 3.5 Sonnet
- 프롬프트 작성의 다양성
  - 예시문제 제시
  - 문장 구조 변형 요청
  - 문제 내용 변경 요청
  - 문제 길이 변형 요청
  - 특정 명사 제외 요청
  - 단위 명사 변경 요청
  - 답 변형 요청
  - 서술어 변형 요청
  - 혼합 요청 : 예) 문제 내용+길이, 문장 구조+ 내용

1차 프롬프트	2차 프롬프트	3차 프롬프트	...	N차 프롬프트
• 저는 5학년 수학 선생님입니다. '공약수와 최대 공약수 ' 를 가지고 수학 문제를 만들어야 하는데, 다음과 같은 예시 문제를 주겠습니다. 최대 몇 명에게 나누어 줄 수 있는지 구해 보세요." 이와 비슷한 문제들을 만들어 주시고 서술형 풀이 답안을 함께 써 주세요.	• 1번 문제와 2번 문제 모두 단일 답이 나와야 합니다. 복수 답이 아니라 단일 숫자의 답이 나올 수 있게 문제를 수정해 주세요.	• 2번 문제는 잘 했습니다. 문제 2에 대한 답은 6이 되는 거지요. 이처럼 문제 1번도 단일 답이 나오게 문제를 수정해 주세요.		• 어느 동물원에 사자 36마리, 호랑이 28마리, 코끼리 48마리가 있습니다. 동물원 직원들은 이 동물들을 모두 같은 크기의 우리에 가두려고 합니다. 가장 큰 우리를 만들기 위해 동물들을 최대한 많은 그룹으로 나눌 수 있을까요? 이 문제에 대한 서술형 답안을 다시 작성해 주세요.

LLM을 이용한
수학문제 합성 데이터 구축

3. 수학 문제 합성데이터 구축

프롬프트의 작성 :다양성과 원칙

단원	문제 유형	seed 문제	example_org	grade	llm	prompt1	설명	problem	solution	anwer	논의 사항
1. 자연수의 혼합 계산	1. 덧셈과 뺄셈이 섞여 있는 식의 계산	공원에 학생들이 26명 있습니다. 그중 여학생 9명과 남학생 13명은 축구를 하고 있습니다. 축구를 하고 있지 않은 학생은 몇 명인지 구하는 식으로 나타내고, 답을 구해 보세요.>의 예시 문제처럼 유사한 수학 문제를 만들고 서술형 풀이 답안을 써 주세요. 풀이 답안은 4~5문장 정도로 구성됩니다.	수학이 힘 증성출판사	5학년	ChatGPT4	당신은 초등학교 수학 선생님입니다. "자연수의 혼합 계산"이라는 과정에서 수학 문제를 내야 합니다. <공원에 학생들이 26명 있습니다. 그중 여학생 9명과 남학생 13명은 축구를 하고 있습니다. 축구를 하고 있지 않은 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.>의 예시 문제처럼 유사한 수학 문제를 만들고 서술형 풀이 답안을 써 주세요. 풀이 답안은 4~5문장 정도로 구성됩니다.	예시 문제 제시	하고 운동장에 학생들이 35명 있습니다. 그중 12명은 농구를 하고 있고, 8명은 달리기를 하고 있습니다. 농구나 달리기를 하고 있지 않은 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.	운동장에 있는 전체 학생 수는 35명입니다. 그중 12명은 농구를 하고 있고, 8명은 달리기를 하고 있습니다. 농구나 달리기를 하고 있지 않은 학생 수를 구하려면 농구와 달리기를 하고 있는 학생 수를 더한 다음, 전체 학생 수에서 빼야 합니다. 따라서 식은 35 - (12 + 8)이고, 답은 35 - 20 = 15명입니다. 농구나 달리기를 하고 있지 않은 학생은 15명입니다.		
						당신이 만든 문제는 이름과 숫자만 다를 뿐 문장 구조가 똑같은 문제입니다. 당신은 괄호 안의 예시를 참고만 하여 될 뿐, 똑같은 문장을 만들어내서는 안 됩니다. "자연수의 혼합 계산"이라는 출제 의도를 파악한 뒤 당신이 임의로 다시 문제를 만들어 주세요.	문장 구조 변형	식당에 손님이 48명 있습니다. 그중 18명은 저녁을 먹고 있고, 7명은 음료를 마시고 있습니다. 저녁을 먹거나 음료를 마시지 않는 손님은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.	식당에 있는 전체 손님 수는 48명입니다. 그중 18명은 저녁을 먹고 있고, 7명은 음료를 마시고 있습니다. 저녁을 먹거나 음료를 마시지 않는 손님 수를 구하려면 저녁을 먹고 있거나 음료를 마시고 있는 손님 수를 더한 다음, 전체 손님 수에서 빼야 합니다. 따라서 식은 48 - (18 + 7)이고, 답은 48 - 25 = 23명입니다. 저녁을 먹거나 음료를 마시지 않는 손님은 23명입니다.	23	
						위의 problem 문제를 배경 상황을 바꿔서 유사 문제를 만들어 주고, 서술형 3~5문장으로 풀이해 주세요.	문제 내용 변경	박물관에 관람객이 60명 있습니다. 그중 25명은 미술 전시관을 관람하고 있고, 18명은 역사 전시관을 관람하고 있습니다. 미술 전시관이나 역사 전시관을 관람하지 않는 관람객은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.	박물관에 있는 전체 관람객 수는 60명입니다. 그중 25명은 미술 전시관을 관람하고 있고, 18명은 역사 전시관을 관람하고 있습니다. 미술 전시관이나 역사 전시관을 관람하지 않는 관람객 수를 구하려면 미술 전시관과 역사 전시관을 관람하는 관람객 수를 더한 다음, 전체 관람객 수에서 빼야 합니다. 따라서 식은 60 - (25 + 18)이고, 답은 60 - 43 = 17명입니다. 미술 전시관이나 역사 전시관을 관람하지 않는 관람객은 17명입니다.	17	
						감사합니다. 이번엔 "자연수의 혼합 계산"의 개념을 이용하여 해결할 수 있는 수학 문제를 만들어 주세요. 단, 가상의 인물이 포함돼 있고 구체적인 상황이 연출되어야 합니다. 또한 최소 100자 이상을 사용하여 문제 만들어 주세요. 하지만 한 문제당 질문은 한 가지여야 합니다.	내용 및 문제 길이 변형	술속 마을에 사는 엘프 엘라는 오늘 아침에 과자를 250개 구웠습니다. 그중 75개는 이웃집에 선물로 주었고, 65개는 마을 축제에서 팔았습니다. 엘라는 남은 과자를 친구들과 나눠 먹으려고 합니다. 엘라가 친구들과 총 5명이서 과자를 똑같이 나누어 먹으려면, 한 사람당 몇 개씩 먹을 수 있을지 하나의 식으로 나타내고, 답을 구해 보세요.	엘라가 구운 전체 과자 수는 250개입니다. 그중 75개는 이웃집에 주었고, 65개는 마을 축제에서 팔았습니다. 남은 과자 수를 구하려면 전체 과자 수에서 이웃집에 준 과자와 축제에서 판 과자를 빼야 합니다. 따라서 식은 250 - (75 + 65)입니다. 75 + 65 = 140이므로, 250 - 140 = 110개입니다. 엘라와 친구들 5명이 과자를 똑같이 나누어 먹으려면, 110을 5로 나누어야 합니다. 110 ÷ 5 = 22이므로, 한 사람당 과자를 22개씩 먹을 수 있습니다.	22	
						감사합니다만 문제가 다소 짧은 느낌이 있습니다. 한 문제에 글자 수가 최소 70자를 넘기도록 다시 만들어 주세요	문제 길이 변형	해리와 그의 친구들은 매년 여름마다 모여 과일 나무를 가꾸고 있습니다. 올해는 사과나무 320그루와 배나무 180그루를 심었습니다. 그중 150그루의 사과나무와 100그루의 배나무가 이미 열매를 맺었습니다. 해리와 친구들이 열매를 맺지 않은 나무들을 동일한 비율로 나누어 돌보려면, 각 사람이 돌보아야 할 나무의 수는 몇 그루인지 하나의 식으로 나타내고, 답을 구해 보세요.	올해 해리와 친구들이 심은 사과나무는 320그루, 배나무는 180그루입니다. 그중 150그루의 사과나무와 100그루의 배나무가 이미 열매를 맺었습니다. 열매를 맺지 않은 나무 수를 구하려면 각각의 나무 수에서 열매를 맺은 나무 수를 빼야 합니다. 사과나무는 320 - 150 = 170그루, 배나무는 180 - 100 = 80그루입니다. 열매를 맺지 않은 나무는 총 170 + 80 = 250그루입니다. 해리와 친구들이 총 5명이므로, 각 사람이 돌보아야 할 나무 수는 250 ÷ 5 = 50그루입니다. 따라서 각 사람이 돌보아야 할 나무의 수는 50그루입니다.	50	
						퍼즐, 아이, 어린이가 문제에 들어가지 않는 다양한 문제를 만들어 주세요.	명사 제외	중세 왕국에서 열린 연회에서 총 650명의 귀족들이 참석했습니다. 그중 240명은 왕의 연설을 듣기 위해 대연회장에 모였고, 195명은 음악 연주를 듣기 위해 작은 연회장에 있었습니다. 나머지 귀족들은 무도회장에서 춤을 추고 있었습니다. 무도회장에서 춤을 추고 있는 귀족들은 몇 명인지 구하세요. 또한, 무도회장에 있는 귀족들을 10명의 시종들이 똑같이 돌보려면, 한 시종당 몇 명씩 돌봐야 하는지 계산해 보세요.	중세 왕국 연회에 참석한 전체 귀족 수는 650명입니다. 그중 240명은 왕의 연설을 듣기 위해 대연회장에 모였고, 195명은 음악 연주를 듣기 위해 작은 연회장에 있습니다. 무도회장에서 춤을 추고 있는 귀족들의 수를 구하려면 전체 귀족 수에서 대연회장과 작은 연회장에 있는 귀족 수를 빼야 합니다. 따라서 식은 650 - (240 + 195)입니다. 240 + 195 = 435이므로, 650 - 435 = 215명입니다. 무도회장에서 춤을 추고 있는 귀족들은 215명입니다. 이 귀족들을 10명의 시종들이 똑같이 돌보려면, 215를 10으로 나눠야 합니다. 215 ÷ 10 = 21.5이므로, 한	22	

3. 수학 문제 합성데이터 구축

단원	문제 유형	seed 문제	example_org	grade	llm	prompt1	설명	problem	solution	anwer	논의 사항
						위의 problem 문제를 단위 명사를 바꿔서 유사 문제를 만들어 주고, 서술형 3~5문장으로 풀이해 주세요.	단위 명사 변경	고속 모터보트입니다. 나머지 배들은 작은 낚싯배입니다. 작은 낚싯배는 몇 척인지 구하세요. 또한, 이 낚싯배들을 15개의 부두에 똑같이 배치하려면, 한 부두당 몇 척씩 배치해야 하는지 계산해 보세요.	명입니다. 고속 모터보트의 수를 빼야 합니다. 따라서 식은 750 - (300 + 210)입니다. 300 + 210 = 510이므로, 750 - 510 = 240척입니다. 작은 낚싯배는 240척입니다. 이 낚싯배들을 15개의 부두에 똑같이 배치하려면, 240을 15로 나눠야 합니다. 240 ÷ 15 = 16이므로, 한 부두당 16척씩 배치해야 합니다.	16	
						앞서 사용된 명사들을 중복해서 사용하지 않고, 3문장 이상의 이야기 형식으로 문제를 만들어주세요. 문제와 답변에 나올 수 있는 숫자는 200 이상 500 이하로 제한해주세요.	답 변형	대규모 농장에서 총 450개의 농작물이 수확되었습니다. 그중 180개는 옥수수이고, 120개는 감자입니다. 나머지 농작물은 토마토입니다. 수확된 토마토는 몇 개인지 구하세요. 또한, 이 토마토들을 10개의 바구니에 똑같이 나누려면, 한 바구니당 몇 개씩 넣어야 하는지 계산해 보세요.	대규모 농장에서 수확된 전체 농작물 수는 450개입니다. 그중 180개는 옥수수이고, 120개는 감자입니다. 토마토의 수를 구하려면 전체 농작물 수에서 옥수수와 감자의 수를 빼야 합니다. 따라서 식은 450 - (180 + 120)입니다. 180 + 120 = 300이므로, 450 - 300 = 150개입니다. 수확된 토마토는 150개입니다. 이 토마토들을 10개의 바구니에 똑같이 나누려면, 150을 10으로 나눠야 합니다. 150 ÷ 10 = 15이므로, 한 바구니당 15개씩 넣어야 합니다.	15	
						앞서 사용된 명사들이 문제에 나오지 않도록 하고, 서술어도 문제마다 다 다르게 표현해서 세 문제를 다시 만들어 주세요. 문제는 3문장 이상 이야기 형식으로 만들어 주세요.	서술어 변형	대형 카페에서 하루 동안 총 420잔의 음료를 판매했습니다. 그중 190잔은 커피였고, 150잔은 차였습니다. 나머지 음료는 주스였습니다. 판매된 주스는 몇 잔인지 구하세요. 또한, 이 주스를 7명의 바리스타가 똑같이 나누어 준비하려면, 한 사람당 몇 잔씩 준비해야 하는지 계산해 보세요.	대형 카페에서 판매된 전체 음료 수는 420잔입니다. 그중 190잔은 커피였고, 150잔은 차였습니다. 주스의 수를 구하려면 전체 음료 수에서 커피와 차의 수를 빼야 합니다. 따라서 식은 420 - (190 + 150)입니다. 190 + 150 = 340이므로, 420 - 340 = 80잔입니다. 판매된 주스는 80잔입니다. 이 주스를 7명의 바리스타가 똑같이 나누어 준비하려면, 80을 7로 나눠야 합니다. 80 ÷ 7 = 11.430이므로, 한 사람당 약 11잔씩 준비해야 합니다.	11	
						"자격: 당신은 초등학교 5학년 수학 교사입니다. 목적: 초등학생 5학년을 대상으로 수학 시험을 보려고 합니다. 기본 문제: problem0: 공원에 학생들이 26명 있습니다. 그중 여학생 9명과 남학생 13명은 축구를 하고 있습니다. 축구를 하고 있지 않은 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요. 이 기본 문제를 가지고 유사 문제를 13개 만들려고 합니다. 유사 문제를 만들 때 주의할 점은 다음과 같습니다. 1. 모든 문제의 답이 달라야 합니다. 2. 문제를 구성하는 문장 구조가 다양해야 합니다. 3. 다양한 단위 명사(cm, ml, 개, 권, 명, 분, 시간, 일, 년 등)를 사용해야 합니다. problem1: 학교 생활과 관련된 문제 problem2: 옛날 이야기와 관련된 문제 problem3: 드라마처럼 에피소드가 있는 문제 problem4: 가족과 관련된 문제 problem5: 공상과학 소설과 관련된 문제 problem6: 길이가 아주 긴 문제 problem7: 길이가 아주 짧은 문제 problem8: 이야기체(구어체)를 사용하는 문제 problem9: 여러 수식이 혼합되어 있는 문제 problem10: 덧셈과 곱셈이 혼합되어 있는 문제 problem11: 뺄셈과 곱셈이 혼합되어 있는 문제 problem12: 덧셈과 나눗셈이 혼합되어 있는 문제 problem13: 뺄셈과 나눗셈이 혼합되어 있는 문제 그리고 각각의 문제에 대한 풀이를 서술형 3~5문장으로 풀이해 주세요."	혼합형	학교 도서관에 총 42권의 책이 있습니다. 그중 18권은 동화책이고, 15권은 과학책입니다. 나머지 책은 역사책입니다. 역사책은 몇 권인지 구하세요.	학교 도서관에 있는 전체 책 수는 42권입니다. 그중 18권은 동화책이고, 15권은 과학책입니다. 역사책의 수를 구하려면 전체 책 수에서 동화책과 과학책의 수를 빼야 합니다. 따라서 식은 42 - (18 + 15)이고, 답은 42 - 33 = 9권입니다. 역사책은 9권입니다.	9	
								옛날 한 마을에 120명의 사람들이 살고 있었습니다. 그중 45명은 농부였고, 30명은 상인이었습니다. 나머지는 목동이었습니다. 목동은 몇 명인지 구하세요	마을에 사는 전체 사람 수는 120명입니다. 그중 45명은 농부였고, 30명은 상인이었습니다. 목동의 수를 구하려면 전체 사람 수에서 농부와 상인의 수를 빼야 합니다. 따라서 식은 120 - (45 + 30)이고, 답은 120 - 75 =	45	

3. 수학문제 합성데이터 구축 : GPT-4o

GPT-4o 장/단점

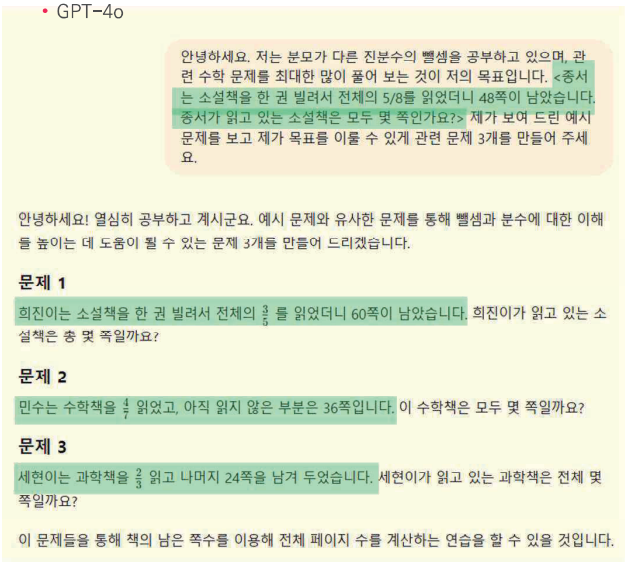
- 장점
  - 수식 오류가 거의 없음
  - 연쇄 프롬프트가 제일 잘 되는 편
- 단점
  - SEED에서 창의성 문제 생성에 취약함
  - 구축자 입장에서 LaTex 방식의 수식 전용 형식을 일반 텍스트로 옮길 때 다른 크기로 표시되어 약간의 처리 시간 필요
  - 글자수 제한 프롬프트 명령 수행 어려움

3. 수학 문제 합성데이터 구축 : GPT-4o

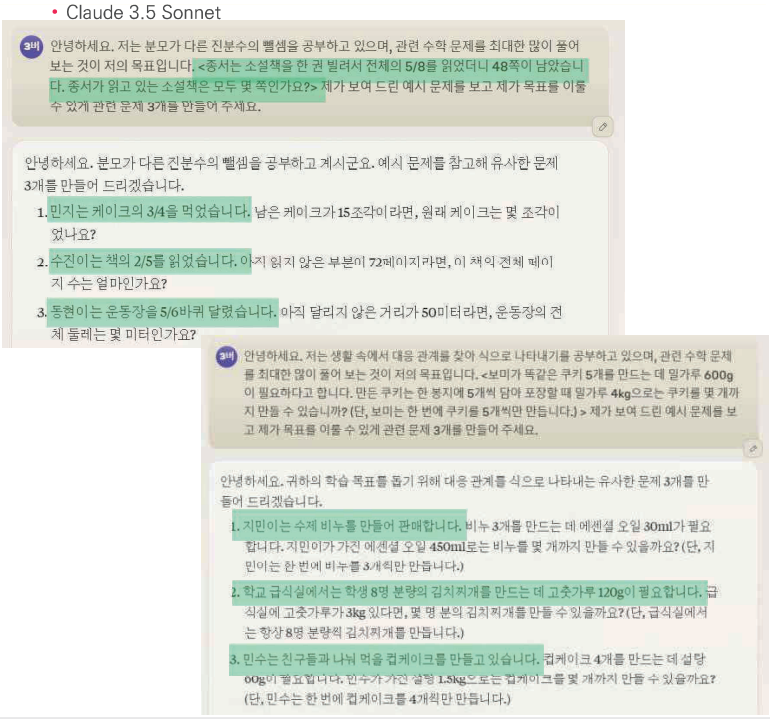
창의적 문제 생성에 취약

- 예시 제공 프롬프트의 형식과 거의 유사한 문제 생성, 다양성 부족

- GPT-4o



- Claude 3.5 Sonnet



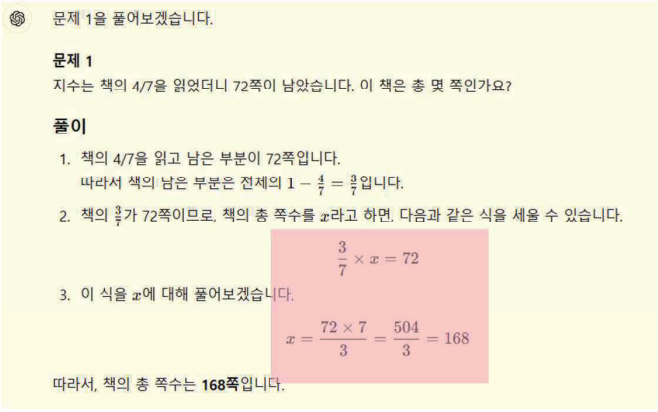
LLM을 이용한
수학문제 합성 데이터 구축

3. 수학 문제 합성데이터 구축 : GPT-4o

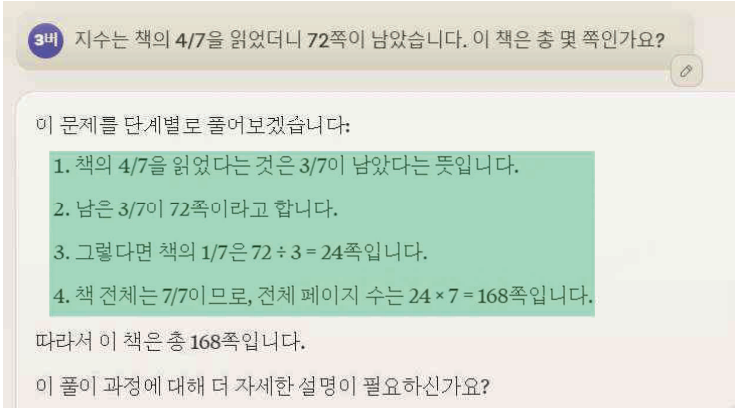
수식 표기 형식

- LaTex 방식의 수식 전용 형식

- GPT-4o



- Claude 3.5 Sonnet



3. 수학 문제 합성데이터 구축 : GPT-4o

글자수 제한 프롬프트 실행 취약

- GPT-4o

이번엔 글자 수를 100, 200, 300자 내외로 제한하여 각각 문제 1개씩, 총 3개의 문제를 만듭니다.

아래는 요청하신 대로 글자 수 제한을 맞춰 구성한 문제들입니다.

100자 문제:

학교에서 1명의 학생에게 색연필 5자루씩 나눠주려 합니다. 학생이 30명일 때, 색연필은 몇 자루 필요할까요?

200자 문제:

도서관에서 책을 정리하는데, 한 명의 사서가 1시간에 12권의 책을 정리할 수 있습니다. 사서 3명이 동시에 일할 경우, 8시간 동안 총 몇 권의 책을 정리할 수 있을까요?

300자 문제:

학교 운동회에서 학생들을 15명씩 조를 나눠 경기를 진행하려고 합니다. 총 180명의 학생이 참여한다고 가정할 때, 몇 개의 조가 만들어질까요? 모든 조가 인원이 동일하게 구성되도록 하려면 최대 몇 명의 학생을 추가해야 할까요?

각 문제는 단일 답을 갖도록 구성되었습니다.

- Claude 3.5 Sonnet

이번엔 글자 수를 100, 200, 300자 내외로 제한하여 각각 문제 1개씩, 총 3개의 문제를 만듭니다.

네, 이해했습니다. 100자, 200자, 300자 내외로 각각 하나의 문제를 만들겠습니다. 각 문제는 단일 질문만 포함하도록 하겠습니다.

1. 100자 내외 문제: 학교 급식실에서 밥 3인분을 짓는 데 쌀 2컵이 필요합니다. 오늘 점심 급식에 참여하는 학생이 450명이라면, 필요한 쌀은 몇 컵일까요? (단, 모든 학생은 같은 양의 밥을 먹습니다.)
2. 200자 내외 문제: 체육 수업에서 줄넘기 기록 측정을 하고 있습니다. 한 명의 학생이 줄넘기를 30회 하는 데 평균 20초가 걸립니다. 체육 선생님은 각 학생에게 3분의 시간을 주어 줄넘기를 하게 할 예정입니다. 이 경우, 한 학생이 3분 동안 할 수 있는 줄넘기 횟수는 몇 회일까요? (단, 학생들은 쉬지 않고 일정한 속도로 줄넘기를 합니다.)
3. 300자 내외 문제: 학교 도서관에서는 독서 왕 선발 대회를 진행하고 있습니다. 책 한 권을 읽을 때마다 10포인트를 얻고, 50포인트마다 독서 뱃지 1개를 받을 수 있습니다. 또한, 뱃지 3개를 모으면 특별 문화상품권으로 교환할 수 있습니다. 민수는 이번 학기 동안 열심히 책을 읽어 1820포인트를 모았습니다. 민수가 받을 수 있는 특별 문화상품권은 몇 장일까요? (단, 남은 포인트와 뱃지는 다음 학기에 계속 사용할 수 있습니다.)

3. 수학 문제 합성데이터 구축 : Claude 3.5 Sonnet

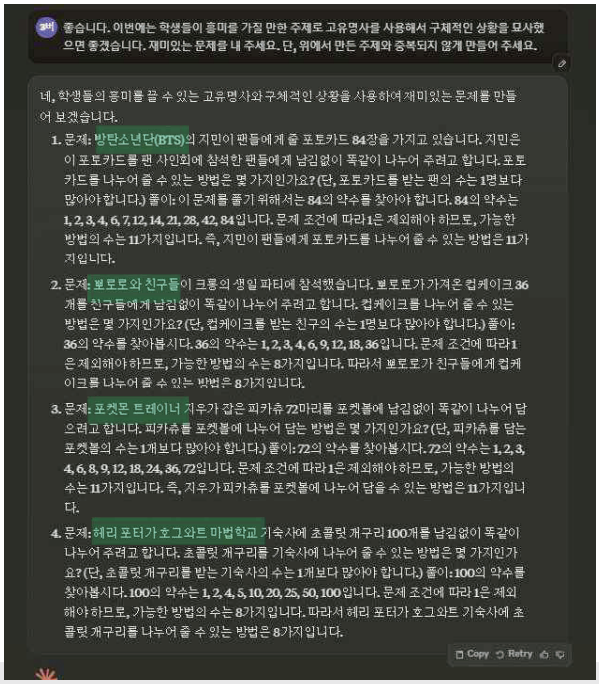
Claude 3.5 Sonnet 장/단점

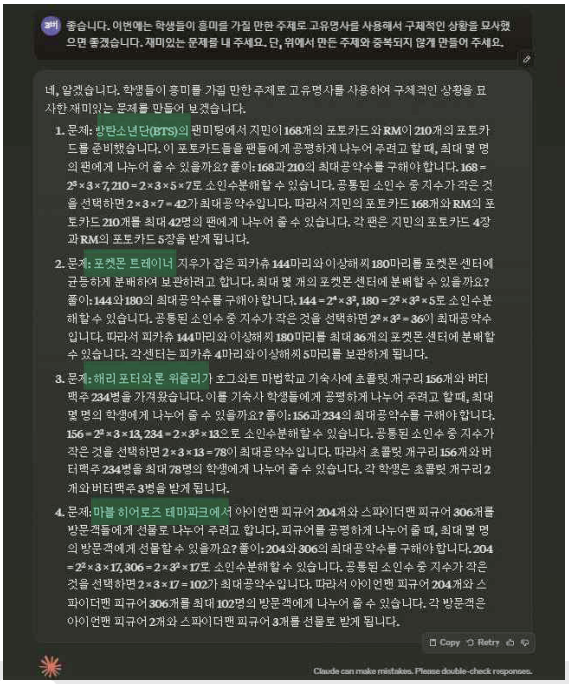
- 장점
  - SEED에서 창의적 문제 생성
  - 수식오류가 거의 없음
- 단점
  - 영역별로 출현 명사가 고정적임 (추천순서(?)가 고정적임)
  - 난이도가 너무 높게 생성됨
  - 연쇄 프롬프트 거부하는 경우가 다량 발생
  - 문맥 오인식 발생

3. 수학 문제 합성데이터 구축 : Claude 3.5 Sonnet

출현 어휘의 고정성

• Claude 3.5 Sonnet



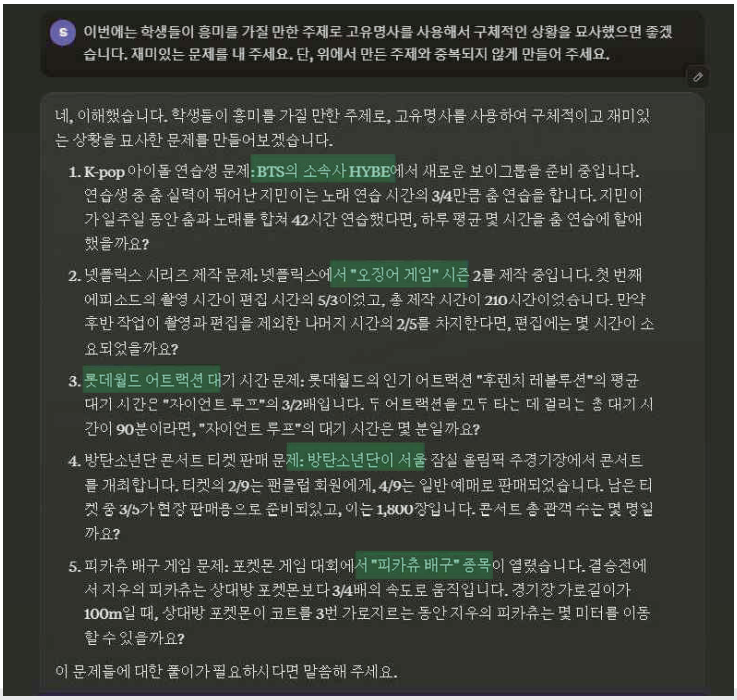


LLM을 이용한
수학문제 합성 데이터 구축

3. 수학 문제 합성데이터 구축 : Claude 3.5 Sonnet

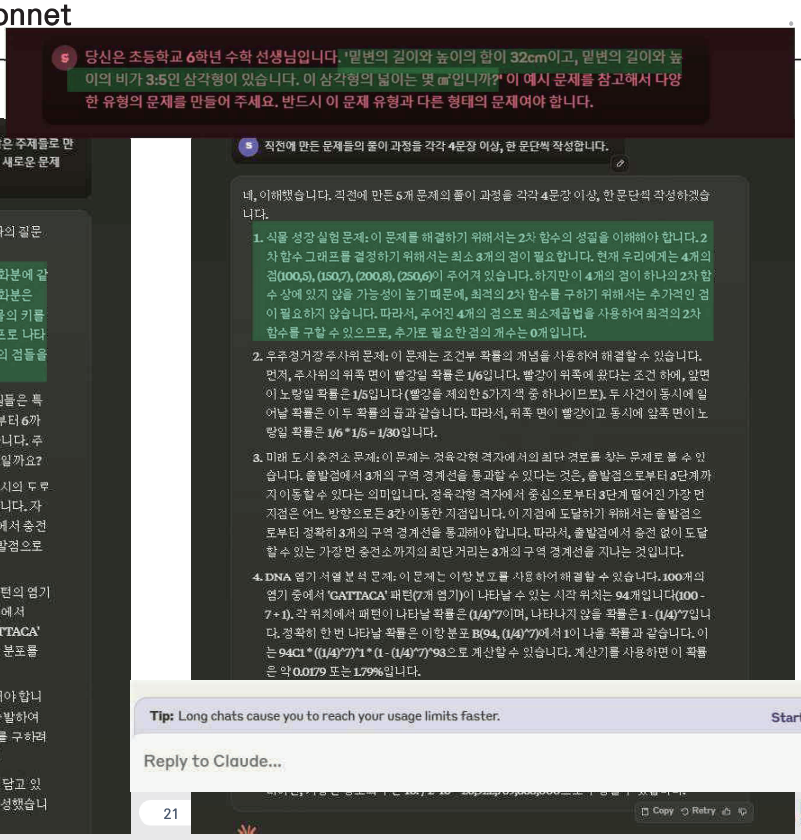
출현 어휘의 고정성

• Claude 3.5 Sonnet



3. 수학 문제 합성데이터 구축 : Claude 3.5 Sonnet

난이도 조정 실패 • Claude 3.5 Sonnet



3. 수학 문제 합성데이터 구축 : Gemini

Gemini 장/단점

- 단점
  - 수식 오류 있음
  - 검색기반 모델 기능
  - LLM 창의성이 없이 구글에서 찾아주는 기능이 강함 (“창의적으로 해줘.”불통)
  - 연쇄 프롬프트 안됨
  - 도형문제 -> 외국사이트에서 이미지 첨부하는 경우도 발생
  - 난이도 조정 안됨
  - 문맥 오인식
  - 정치문제 아닌데 정치에 관여되었다고 판단하여 정치, 선거에 관련된 답변 불가

- 장점
- 가장 빠름

3. 수학 문제 합성데이터 구축 : Gemini

오답변

- Gemini

300자 이상으로 문제를 1개 만들어 주세요.

답안 표시

300자 이상의 수학 문제

문제:

한 도서관에는 소설책과 과학책이 있습니다. 소설책은 과학책보다 120권이 더 많고, 두 종류의 책을 모두 합하면 500권입니다. 이 도서관에 있는 과학책은 몇 권일까요? 또, 소설책은 몇 권일까요?

난이도를 더 올려 주세요.

답안 표시

난이도를 더 높인 창의적인 수학 문제 3가지

문제 1: 시간 여행 열차의 역설

시간 여행 열차를 타고 과거로 간 민준이는 과거의 자신을 만났습니다. 과거의 자신에게 100원짜리 동전을 주었는데, 이 동전은 현재 민준이가 가지고 있던 동전입니다. 이 상황에서 시간과 공간의 개념을 고려하여 모순 없이 설명할 수 있는 방법은 무엇일까요?

문제 2: 무한한 방

무한히 넓은 방이 있습니다. 방 안에는 1번 방, 2번 방, 3번 방, ... 이렇게 무한히 많은 방이 차례대로 연결되어 있습니다. 민준이는 1번 방에서 시작하여 매번 짝수 번 방으로만 이동합니다. 민준이가 모든 짝수 번 방을 방문할 수 있을까요? 만약 불가능하다면, 그 이유는 무엇일까요?

3. 수학 문제 합성데이터 구축 : Gemini

Seed 문제 활용도 낮음

- Gemini

Gemini was just updated. See update

MS 당신은 초등학교 수학 선생님입니다. "분수의 곱셈"이라는 과정에서 수학 문제를 내야 합니다. <떨어진 높이의 2/3만큼 튀어 오르는 공이 있습니다. 이 공을 48m 높이에서 수직으로 떨어뜨렸다면 두 번째로 튀어 올랐을 때까지 공이 움직인 거리는 모두 몇 m인지 풀이 과정을 쓰고, 답을 구하시오.>의 예시 문제처럼 유사한 수학 문제를 만들고 서술형 풀이 답안을 써 주세요. 풀이 답안은 4~5문장 정도로 구성됩니다.

Show drafts

문제

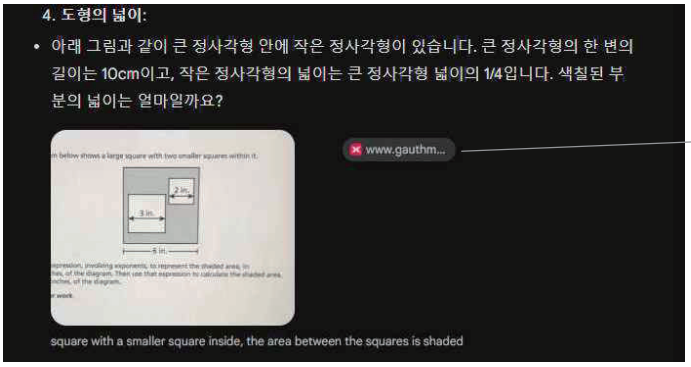
철수는 3/4만큼 물이 차 있는 500mL 컵을 가지고 있습니다. 철수가 실수로 컵을 떨어뜨려 물의 1/5만큼 쏟았습니다. 컵에 남아있는 물의 양은 몇 mL입니까? 풀이 과정을 쓰고, 답을 구하시오.

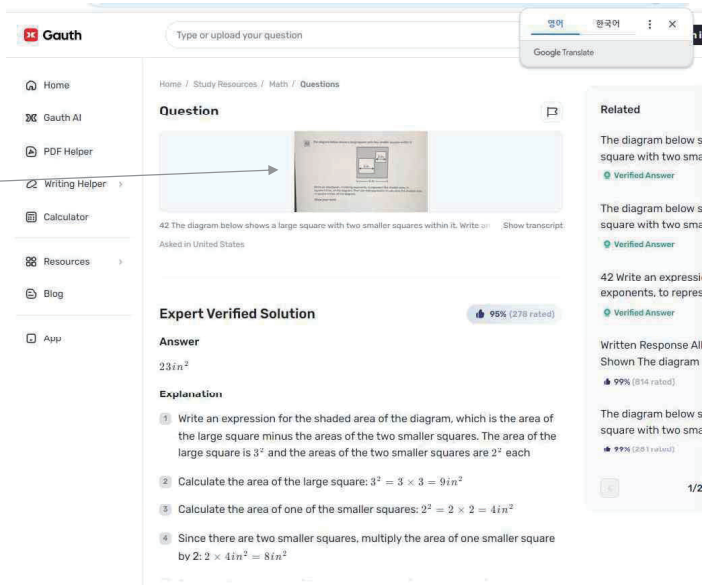
풀이 답안

1. 처음 컵에 들어있던 물의 양은 500mL의 3/4이므로, 500 x (3/4) = 375mL입니다.
2. 쏟은 물의 양은 375mL의 1/5이므로, 375 x (1/5) = 75mL입니다.
3. 컵에 남아있는 물의 양은 처음 물의 양에서 쏟은 물의 양을 빼면 구할 수 있습니다.
4. 따라서, 컵에 남아있는 물의 양은 375 - 75 = 300mL입니다.
5. 답: 300mL

3. 수학 문제 합성데이터 구축 : Gemini

NO 생성, YES 검색

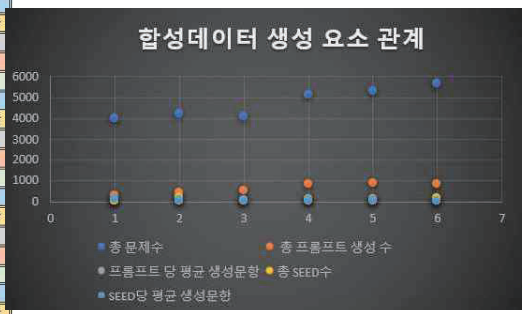




4. 결론

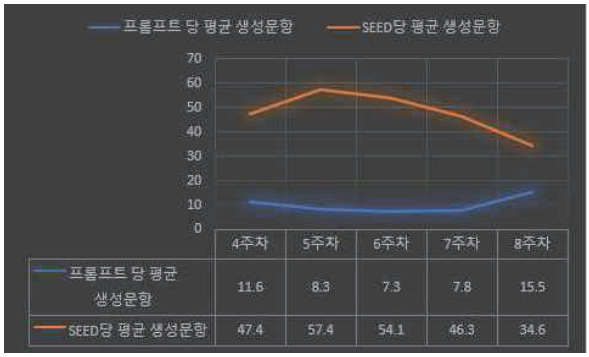
수학문제 합성데이터 5만 쌍 구축 결과

	작업자	A	B	C	D	E	F	G		
8주차	문제수	1310	520	750	250	1337	750	720	5637	총문제수
	프롬프트	22	270	98	47	148	35	209	829	총프롬프트합
	프롬프트 당 생성문항	59.5	1.9	7.7	5.3	9.0	21.4	3.4	15.5	평균 프롬프트 당 생성문항
	SEED	38	17	39	7	29	14	32	176	총 SEED합
	SEED당 생성문항	34.5	30.6	19.2	35.7	46.1	53.6	22.5	34.6	평균 SEED당 생성문항
7주차	문제수	1311	420	600	250	1318	600	800	5299	총문제수
	프롬프트	165	160	70	42	122	39	262	860	총프롬프트합
	프롬프트 당 생성문항	7.9	2.6	8.6	6.0	10.8	15.4	3.1	7.8	평균 프롬프트 당 생성문항
	SEED	24	14	30	8	12	11	34	133	총 SEED합
	SEED당 생성문항	54.6	30.0	20.0	31.3	109.8	54.5	23.5	46.3	평균 SEED당 생성문항
6주차	문제수	1502	200	500	250	1516	750	410	5128	총문제수
	프롬프트	261	125	114	66	91	39	124	820	총프롬프트합
	프롬프트 당 생성문항	5.8	1.6	4.4	3.8	16.7	19.2	3.3	7.3	평균 프롬프트 당 생성문항
	SEED	35	4	26	7	10	14	16	112	총 SEED합
	SEED당 생성문항	42.9	50.0	19.2	35.7	151.6	53.6	25.6	54.1	평균 SEED당 생성문항
5주차	문제수	1396		200	137	1272	650	450	4105	총문제수
	프롬프트	201		68	36	72	46	111	534	총프롬프트합
	프롬프트 당 생성문항	6.9		2.9	3.8	17.7	14.1	4.1	8.3	평균 프롬프트 당 생성문항
	SEED	22		10	6	8	12	18	76	총 SEED합
	SEED당 생성문항	63.5		20.0	22.8	159.0	54.2	25.0	57.4	평균 SEED당 생성문항
4주차	문제수	1500				1565	751	399	4215	총문제수
	프롬프트	143				77	58	141	419	총프롬프트합
	프롬프트 당 생성문항	10.5				20.3	12.9	2.8	11.6	평균 프롬프트 당 생성문항
	SEED	15				168	14	15	212	총 SEED합
	SEED당 생성문항	100.0				9.3	53.6	26.6	47.4	평균 SEED당 생성문항
3주차	분제수	1660				1537	801		3998	총분제수
	프롬프트	110				83	98		291	총프롬프트합
	프롬프트 당 생성문항	15.1				18.5	8.2		13.9	평균 프롬프트 당 생성문항
	SEED	16				11	15		42	총 SEED합
	SEED당 생성문항	103.8				139.7	53.4		99.0	평균 SEED당 생성문항



4. 결론

- LLM 활용도의 척도
  - 프롬프트 당 생성되는 총 문항 개수의 합
- 수학문제 합성데이터 5만 쌍 생성 결과
  - 관련 연구 기준 4회차부터 8회차까지 안정적인 데이터 생성 추이를 보임
  - 안정 구간(4-8회차)에서 Seed 문항의 평균 사용량이 줄어듦
  - 프롬프트 당 평균 생성 문항과 Seed 문항의 평균 사용량이 반비례 관계



	4주차	5주차	6주차	7주차	8주차
프롬프트 당 평균 생성문항	11.6	8.3	7.3	7.8	15.5
SEED당 평균 생성문항	47.4	57.4	54.1	46.3	34.6

참고문헌

- 김슬기, 전용주, 김태영, 「합성 데이터셋 생성 방식을 활용한 인공지능 교육용 데이터셋 개발 방법 연구」, 『한국컴퓨터교육학회 학술발표대회논문집』 제26권 제1호, 한국컴퓨터교육학회, 227~230쪽, 2022.
- 김재환, 「거대 언어 모델에서 한국어 고난도 수학 문제 성능 향상을 위한 CoT·PAL 통합형 프롬프트 엔지니어링」, 서강대학교 정보통신대학원 석사학위논문, 2024.
- Jiwoo Kim, *Analyzing Mathematical Reasoning of LMs Using High Diversity Dataset*, Sungkyunkwan University Master's Thesis, 2023.
- Pei Zhou, Jay Pujara, Xiang Ren, et al., *SELF-DISCOVER: Large Language Models Self-Compose Reasoning Structures*, arXiv:2402.03620, 2024.


감사합니다.

[2024년도 한글 및 한국어 정보처리 & 한국코퍼스언어학회 공동 학술대회]

“LLM을 이용한 수학문제 합성데이터 구축”(이숙의 외)에 대한 토론문

신서인(한림대)

흥미로운 발표자료 잘 살펴보았습니다. 보여주신 것과 같이 LLM을 이용하여 효과적으로 수학 문제를 생성할 수 있다면 활용도는 무궁무진하리라고 생각됩니다. 발표자료를 보면서 궁금했던 점을 몇 가지 질문 드림으로써 토론을 갈음하고자 합니다.

1. 제시해 주신 예시를 보니 수학 문제 번역시 여러 가지를 고려해야 할 것 같습니다. 5쪽에 제시된 "당근 가격은 매년 원래 가격의 5%씩 상승합니다."라는 문장의 경우, 문화적 차이에 기인하여 중의성이 발생한다기보다는 '원래'의 의미가 첫 해인지, 전년도인지에 따라 달라지는 것으로 보입니다. 이 문제에서 "당근 가격은 원래 가격의 5%씩 매년 상승합니다."와 같이 '매년'의 어순을 변경하면 중의성이 해소될 수 있습니다. 수학 문제 번역에서 문제가 될 수 있는 영어와 한국어의 차이로 또 어떤 것이 있는지 궁금합니다.

2. LLM을 이용한 수학문제 생성 과제는 다양한 목적을 가지고 있는 것 같습니다. (1) LLM을 이용하여 한국어 수학문제 데이터를 구축하는 것 자체가 목적일 수도 있고, (2) 수학문제 데이터 생성을 위한 최적의 프롬프트 작성 방법을 연구하는 것이 목적일 수도 있고, (3) 한국형 LLM의 논리성을 검증하는 것이 목적일 수도 있습니다. 각각의 경우에 최적의 결과를 도출하기 위해 접근하는 방식이 조금씩 달라질 수 있을 것 같습니다.

   (1) 수학문제 데이터 구축 자체가 목적이라면 LLM에 예시 문제를 제공하고 유사 문제를 작성하게 하는 것 이외에 다른 방법도 있을 수 있을 것 같습니다. 다양한 수학 문제를 출제한다는 것은 다양한 상황을 설정하는 것과 숫자를 달리하는 것 두 가지를 분리하여 생각할 수 있을 것 같습니다. 또, 입력-출력을 문장-문장, 문장-수식, 수식-문장 등으로 구분하여 효율적인 데이터 구축 방안을 모색할 수도 있을 것 같습니다.

   (2) 수학문제 데이터 생성을 위한 최적의 프롬프트 작성 방법을 연구하는 것이 목적이라면 동일한 수식에 기반한 문제에 대해 11~13쪽에 제시한 기준과 같은 다양한 조건을 충족하는 정답을 먼저 정해놓고, 원하는 답을 얻을 수 있는 프롬프트 작성 방법을 모색해야 할 것 같습니다.

   (3) 한국형 LLM의 논리성을 검증하는 것이 목적이라면 평가 요소를 미리 고려해야 할 것 같습니다. 평가 요소로는 서술형 문제의 수식으로의 변환, 풀이 단계의 수식 설정, 각 단계 수식의 설명 문장으로의 변환, 정답의 정확도 등을 평가할 수 있을 것이고, 그밖에 난이도 유지, 문제 상황의 다양성, 상황 설정의 상식에의 부합도 등을 고려할 수 있을 것입니다.

3. LLM에 SEED 문제를 제공하고 유사 문제를 생성하라고 했을 때 모델이 '유사성'을 인식하는 것에 따라 결과가 달라질 것 같습니다.

   (1) 프롬프트에서 '분모가 다른 진분수의 뺄셈'과 같은 문제 유형을 구체적으로 제시하는 경우와 그렇지 않은 경우 결과가 달라지는지 궁금합니다. 문제 유형 제시가 유사 문제 생성에 도움이 될 것 같은데, 실제로는 그렇지 않은지 궁금합니다. 15쪽의 예시를 보면 문제 유형을 제시하고 있지만 예시와 일치하지 않아서 그런지 이를 전혀 고려하지 않고 있는 것으로 보입니다.

   (2) 수식과 문장을 분리하여 단계별로 프롬프트를 제시하는 방법은 시도해보지 않으셨는지 궁금합니다. '1단계: 서술형 문제의 수식으로의 변환, 2단계: 풀이 단계의 수식 설정, 3단계: 각 단계 수식의 설명 문장으로의 변환, 4단계: 정답 도출'과 같이 단계를 나누면 유사 문제 생성에 도움이 되지 않을까 싶습니다.

4. LLM에 문제와 함께 풀이 과정도 제시하라고 요구할 때 둘 이상의 풀이 과정이 있는 경우는 없었는지 궁금합니다. 예를 들어 대분수의 덧셈 문제를 푼다면 자연수끼리 계산하고 진분수끼리 계산하는 방법도 있고, 대분수를 가분수로 바꾸어 계산하는 방법도 있을텐데 이러한 경우는 고려하지 않는지 궁금합니다.

5. 난이도는 어떤 식으로 조정하는지 궁금합니다. 난이도가 높은 문제라면 문제 상황이 복잡한 것일 수도 있고, 풀이 과정이 복잡한 것일 수도 있고, 다른 학습 요소와 결합한 것일 수도 있을텐데 이러한 조건을 구체적으로 제시하면서 난이도를 조정하는 방안을 시도해 보셨는지 궁금합니다.

수학문제를 출제한다면 수학 교과목에서의 성취기준, 학습목표, 평가항목 등 수학 교육의 요소들을 정리하는 것부터 시작해야 할 것 같습니다. 이러한 요소들을 고려하여 수학 문제 데이터를 설계하고 각 수학문제 데이터가 다양한 층위의 분류 태그를 포함하도록 한다면 활용도가 높은 수학문제 합성데이터를 구축할 수 있을 것으로 기대합니다.

수학문제 합성데이터 구축은 LLM의 국어 능력과 수학 능력을 고루 요구합니다. 한국형 LLM의 논리성 검증도 이 둘을 구분하여 수행되어야 할 것입니다.

한국코퍼스언어학회 2024 가을 전국학술대회

KACL 2024

Bllossom 프로젝트: 한국어 언어 모델의 꽃을 피우다

Updatable Purpose-built LLMs

2024.10.11.

PM
함영균
CEO

대표이사 함영균 | hahmyg@teddysum.ai

발표자 소개

함영균 대표

- KAIST 전산학부 박사/석사 (NLP)
- ISO/TC37 한국전문위원 및 간사
- COLING Program Chair (AI 우수학회, '22)
- KACL 조직위원장 (`24)
- 한국스마트빌리지협회 전문위원
- HCLT Best Paper 5회(주저자)
- 문화체육부 장관상(경진대회, '12년)

What We Do

Bllossom - Playground

teddysum

http://chat.bllossom.ai/

엑셀 vba로 쇼핑몰 주문서를 하나의 양식으로 취합하는 작업을 하고 싶어. 어떻게 해야할지 알려줘

Bllossom

엑셀 VBA를 사용하여 쇼핑몰 주문서를 하나의 양식으로 취합하는 작업을 수행하려면 다음 단계를 따르면 됩니다. 이 작업은 VBA를 사용하여 여러 개의 엑셀 파일을 읽고, 데이터를 취합하여 하나의 엑셀 파일로 통합하는 방식으로 진행됩니다.

1. VBA 프로젝트 설정

1. 엑셀 파일을 엽니다.
2. Alt + F11을 눌러 VBA 편집기를 엽니다.
3. VBA 편집기가 열리면, "프로젝트 탐색기"에서 새로운 모듈을 추가합니다. 이를 위해 "삽입" 메뉴에서 "모듈"을 선택합니다.

2. 필요한 변수 및 함수 선언

새로 추가된 모듈에 다음과 같은 코드를 작성합니다:

```
Sub 통합_주문서_취합()
    ' 변수 선언
    Dim sourceFiles As Variant
    Dim sourceSheets As Variant
    Dim sourceWorkbook As Workbook
    Dim targetWorkbook As Workbook
    Dim targetSheet As Worksheet
    Dim sourceFile As Variant
    Dim sourceSheet As Variant
    Dim lastRow As Long
    Dim lastCol As Long
    Dim sourceRange As Range
    Dim targetRange As Range

    ' 타겟 파일 및 시트 초기화
    Set targetWorkbook = ThisWorkbook
    Set targetSheet = targetWorkbook.Sheets("취합_결과") ' 타겟 시트 이름 변경

    ' 소스 파일 및 시트 초기화
    sourceFiles = Application.GetOpenFilename(MultiSelect:=True)
    If IsArray(sourceFiles) Then
        For Each sourceFile In sourceFiles
            Set sourceWorkbook = Workbooks.Open(sourceFile)
            Set sourceSheet = sourceWorkbook.Sheets("주문서") ' 소스 시트 이름 변경

            ' 데이터 취합
            lastRow = sourceSheet.Cells(sourceSheet.Rows.Count, "A").End(xlUp).Row
            lastCol = sourceSheet.Cells(1, sourceSheet.Columns.Count).End(xlToLeft).Column
            Set sourceRange = sourceSheet.Range(sourceSheet.Cells(1, 1), sourceSheet.Cells(lastRow, lastCol))
```

설명

- `sourceFiles` 변수는 사용자가 선택한 모든 파일을 저장합니다.
- `sourceSheets` 변수는 선택된 각 파일의 시트를 저장합니다.
- `sourceWorkbook` 변수는 현재 작업 중인 소스 파일을 참조합니다.
- `targetWorkbook` 변수는 결과를 저장할 타겟 파일을 참조합니다.
- `targetSheet` 변수는 타겟 파일의 시트를 참조합니다.
- `sourceFile` 변수는 루프를 통해 각 소스 파일을 하나씩 처리합니다.
- `sourceSheet` 변수는 현재 소스 파일의 시트를 참조합니다.
- `lastRow` 및 `lastCol` 변수는 소스 시트의 마지막 행과 열을 찾습니다.
- `sourceRange` 변수는 소스 시트에서 데이터를 취합할 범위를 정의합니다.
- `targetRange` 변수는 타겟 시트에 데이터를 붙여넣을 범위를 정의합니다.

3. VBA 코드 실행

위의 코드를 작성한 후, VBA 편집기에서 "실행" 또는 F5를 눌러 코드를 실행합니다. 그러면 파일 선택 대화상자가 나타나고, C 일들은 모두 "취합_결과" 시트에 통합됩니다.

4. 파일 선택 및 통합

다음 단계에서는 사용자가 여러 개의 엑셀 파일을 선택할 수 있습니다. 선택한 파일들은 하나의 시트에 통합되어 "취합_결과"

이 작업을 통해 쇼핑몰 주문서를 하나의 양식으로 취합하는 자동화된 프로세스를 구현할 수 있습니다. 필요에 따라 코드를 수 다.

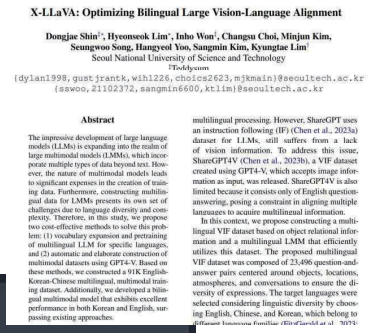
3

What We Do

Bllossom - Playground

lysum

http://chat.bllossom.ai/



X-LLaVA: Optimizing Bilingual Large Vision-Language Alignment

Dongjae Shin[1*], Hyeonseok Lim[1], Inho Won[1], Changsu Choi, Minjun Kim, Seungwoo Song, Hangyeol Yoo, Sangmin Kim, Kyungtae Lim[1]

Seoul National University of Science and Technology

*Teddysum

{dylan1998, gustjrantk, wih1226, choics2623, mjkmain}@seoultech.ac.kr

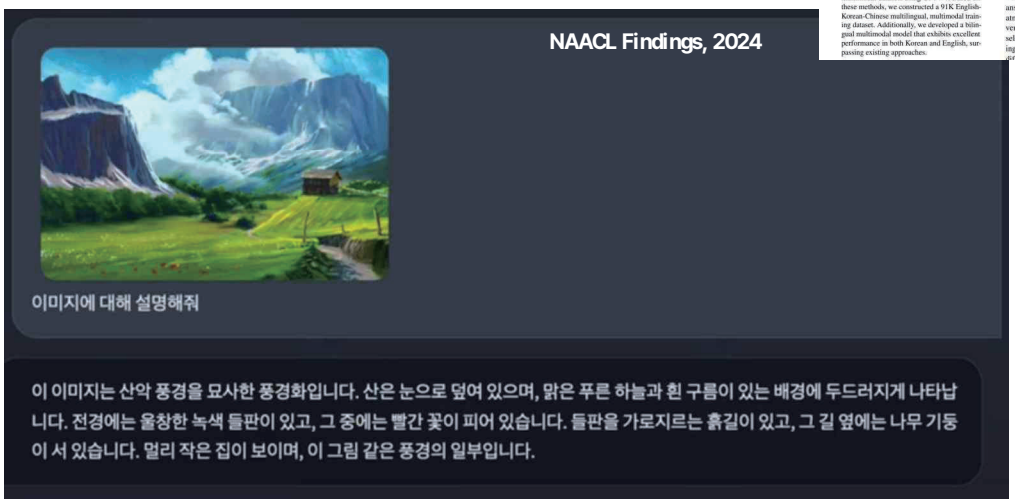
{sswoo, 21102372, sangmin6600, ktlim}@seoultech.ac.kr

Abstract

The impressive development of large language models (LLMs) is expanding into the realm of large multimodal models (LMMs), which incorporate multiple types of data beyond text. However, the nature of multimodal models leads to significant expenses in the creation of training data. Furthermore, constructing multilingual data for LMMs presents its own set of challenges due to language diversity and complexity. Therefore, in this study, we propose two cost-effective methods to solve this problem: (1) vocabulary expansion and pretraining of multilingual LLM for specific languages, and (2) automatic and elaborate construction of multimodal datasets using GPT4-V. Based on these methods, we constructed a 91K English-Korean-Chinese multilingual, multimodal training dataset. Additionally, we developed a bilingual multimodal model that exhibits excellent performance in both Korean and English, surpassing existing approaches.

multilingual processing. However, ShareGPT uses an instruction following (IF) (Chen et al., 2023a) dataset for LLMs, still suffers from a lack of vision information. To address this issue, ShareGPT4V (Chen et al., 2023b), a VIF dataset created using GPT4-V, which accepts image information as input, was released. ShareGPT4V is also limited because it consists only of English question-answering, posing a constraint in aligning multiple languages to acquire multilingual information.

In this context, we propose constructing a multilingual VIF dataset based on object relational information and a multilingual LMM that efficiently utilizes this dataset. The proposed multilingual VIF dataset was composed of 23,496 question-and-answer pairs centered around objects, locations, atmospheres, and conversations to ensure the diversity of expressions. The target languages were selected considering linguistic diversity by choosing English, Chinese, and Korean, which belong to



4

회사 소개

기술 경쟁력

teddysum

고유 언어모델 BLLOSSOM 보유[1]

bLLOSSOM

LLM 개발 경량화 프레임워크 보유 (ELO)

- 경쟁사 대비 1/10 빠른 학습 속도 (비용 [redacted])
- 도메인 적용력 우수 (5mb 데이터로 적용 가능)
- RAY, vLLM 등 분산처리 기반 고효율 학습 및 추론 인프라 보유

데이터 구축사업 경험 다수

- 데이터 구축 관련 사업
  - `21년~23년 국립국어원, 요약 말뭉치 구축사업 수행
  - `23년 NIA, AI 응답결과에 대한 품질 평가 데이터 구축사업 수행
  - `23년 국립국어원, 2023 국어능력 진단체계 활용 방안 연구
  - `24년 국립국어원, 글쓰기 첨삭 지원을 위한 인스트럭션 말뭉치 구축사업 수행
  - `24년 TTA, 생성형 AI 신뢰성 평가의 실용적 접근 방안 연구 수행
- 금융 관련 경험
  - `21년, IBK PoC | `24년 DB-FIS | `24년, 광주은행 1day PoC 진행 등

벤치마크 개발 경험 다수

- AI 말평 리더보드 개발 및 운영
  - `23년 기준 233팀 [redacted], [redacted]개 모델 평가
  - `24년 기준 280팀 [redacted], 3,642개 모델 평가
- 국제 Shared Task – HIDDEN RAD 주최

업계 리딩 플랫폼 적용 (HPE + NVIDIA)

Hewlett Packard Enterprise | NVIDIA

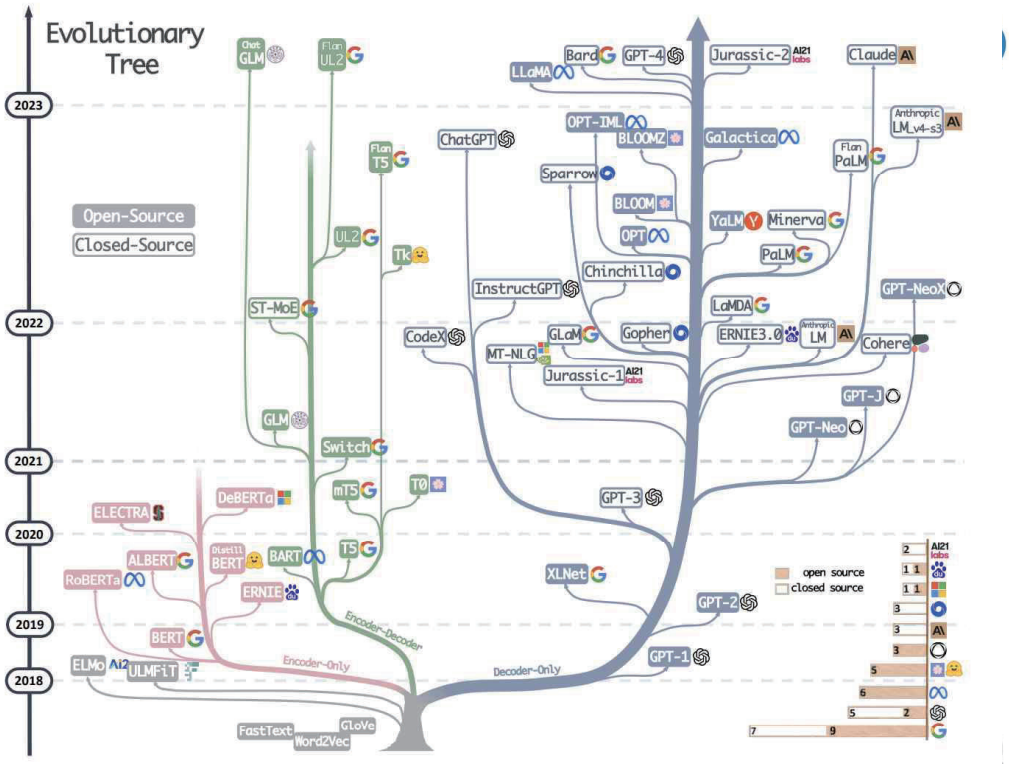
[1] Llama 3.1 기반 bllossom 8B, 70B, 405B 모델 공개

5

배경 지식

Generative AI

- LLMs
  - Decoder-only
  - Unsupervised
  - Scalable
- and,
  - Audio
  - Images
  - Video
  - …

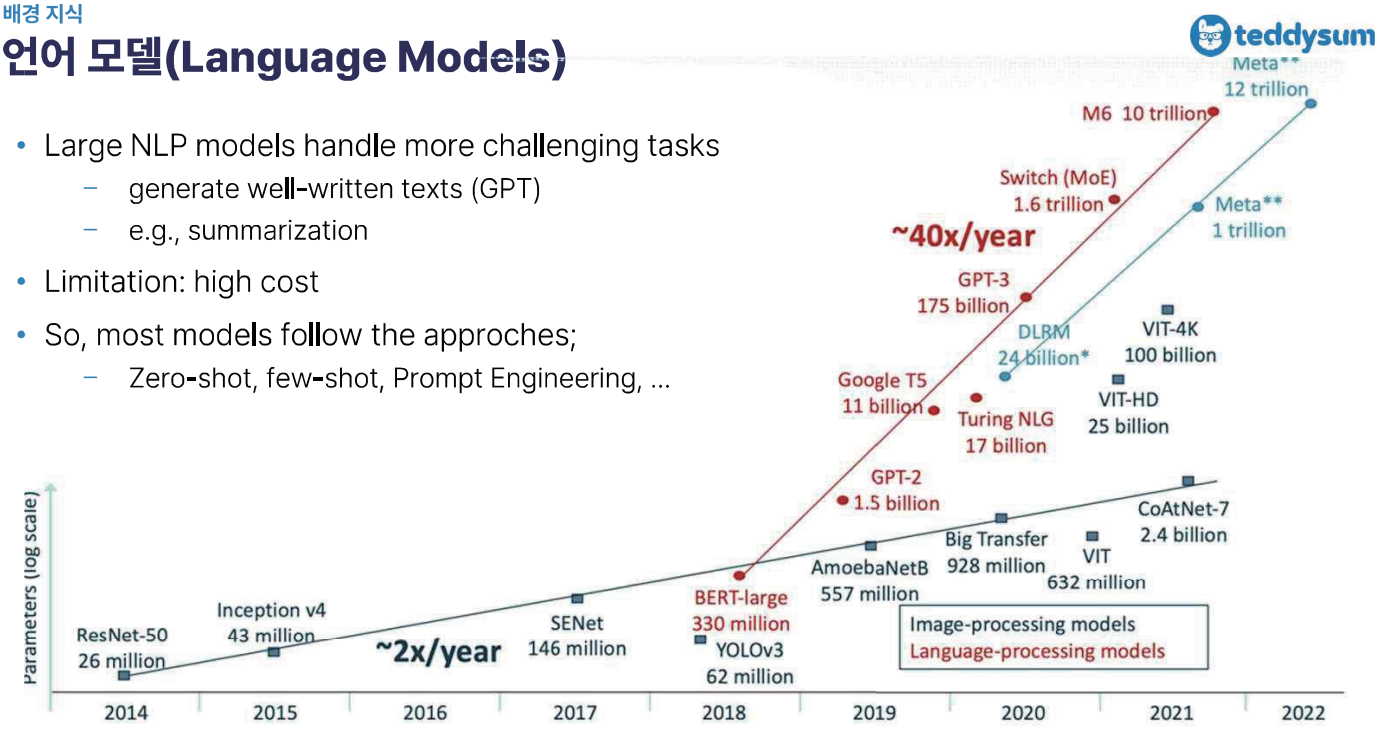


6 [redacted]ng et al., Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, 2023

배경 지식

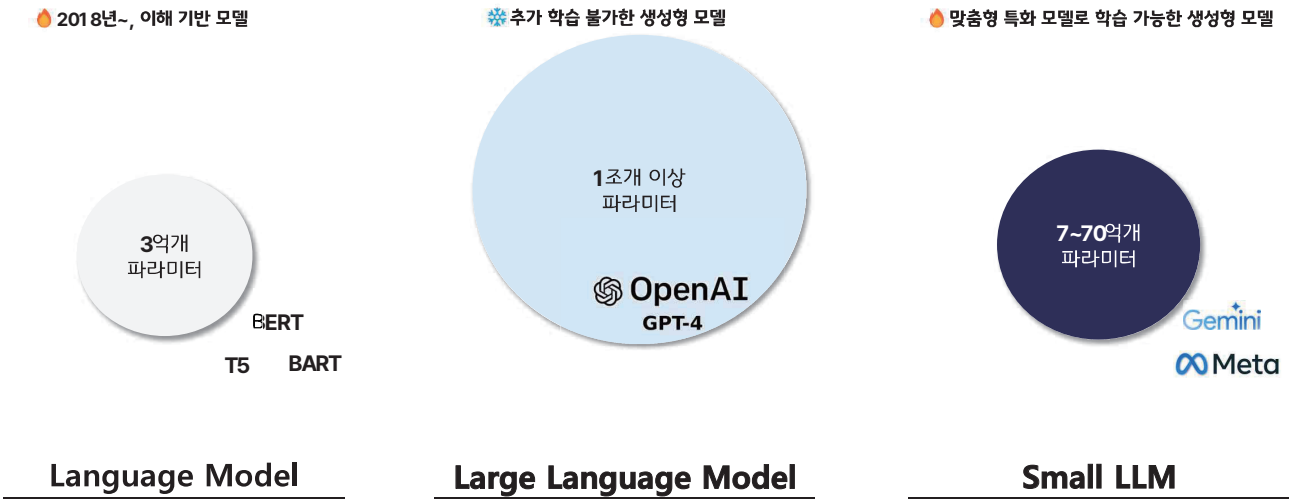
언어 모델(Language Models)

- Large NLP models handle more challenging tasks
  - generate well-written texts (GPT)
  - e.g., summarization
- Limitation: high cost
- So, most models follow the approches;
  - Zero-shot, few-shot, Prompt Engineering, ...



배경 지식

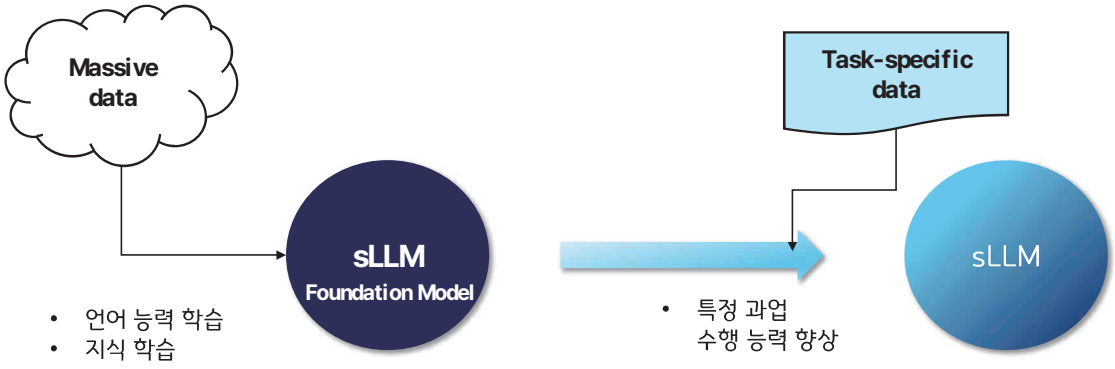
LLM과 sLLM



Language Model | **Large Language Model** | **Small LLM**

배경 지식

Pre-training과 fine-tuning



Pre-training

학습에 막대한 비용 소요 (예: **GPT-4** 학습비용 **4,600**억원 예상)
(원천기술/ 핵심기술)

Fine-tuning

1,000개 ~ **5**만개 인스트럭션 데이터 활용 학습
(비용 저렴, 기술도 낮음)

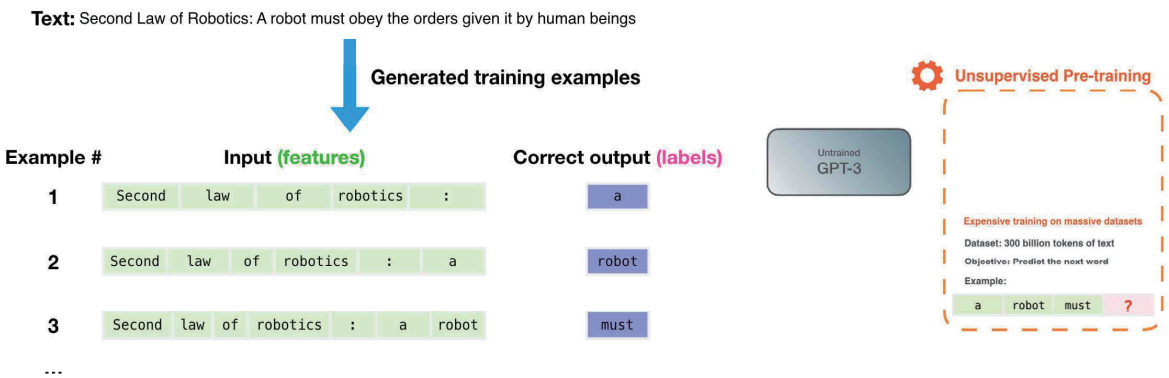
9

배경 지식

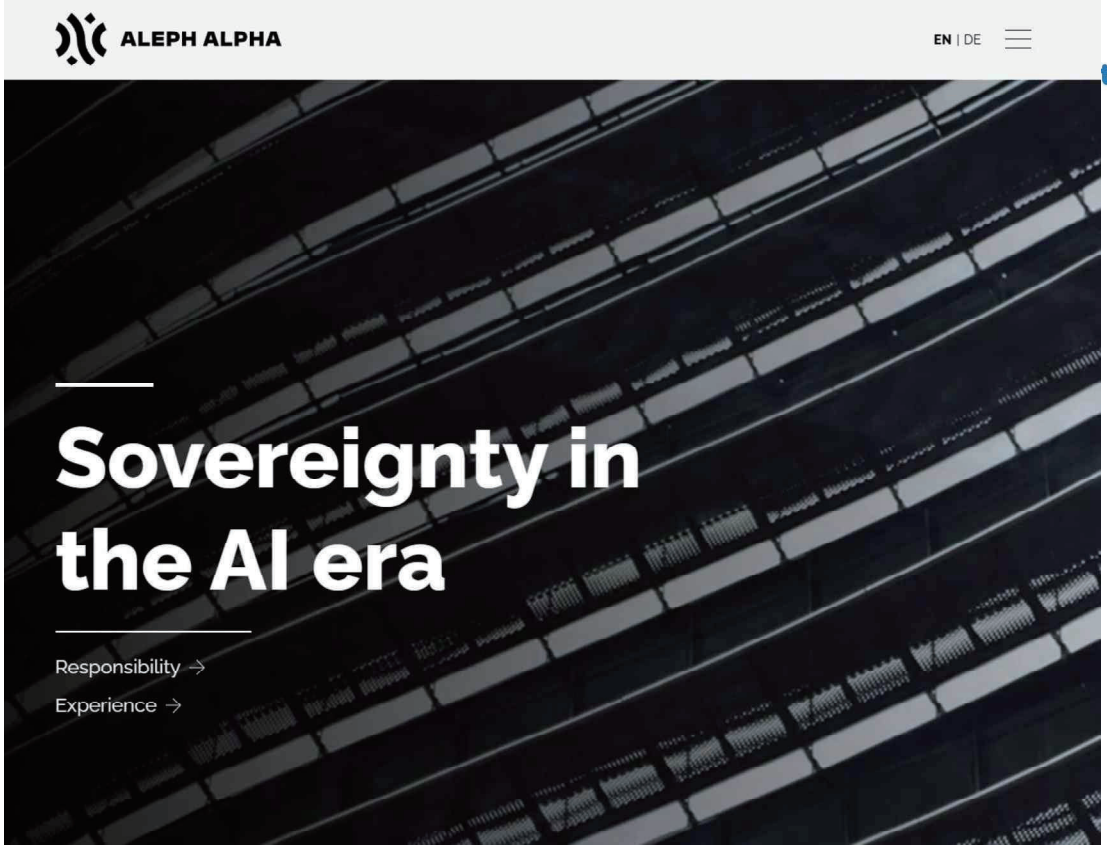
LLM

LLM은 대규모의 텍스트 데이터를 기반으로 다음 단어를 예측함으로써 언어 정보와 지식을 이해함

- 강력한 일반화 능력**(ChatGPT, Gemini)**
- 일반적인 **plain text**를 이용해 학습 데이터 구성 후 **Unsupervised Pretraining**으로 학습



10



teddysum

11

teddysum

자연어처리의 핵심 개념을 알려줘

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that deals with the interaction between computers and humans in natural language. The core concepts of NLP are:

1. **Tokenization**: Breaking down text into individual words or tokens, such as "hello" or "world".

2. **Part-of-Speech (POS) Tagging**: Identifying the part of speech (noun, verb, adjective, etc.) for each word in a sentence.

3. **Named Entity Recognition (NER)**: Identifying specific entities such as names, locations, and organizations in text.

4. **Sentiment Analysis**: Determining the emotional tone or sentiment behind a piece of text, such as positive, negative, or neutral.

5. **Language Modeling**: Predicting the next word in a sequence of text based on the context and patterns learned from a large corpus of text.

6. **Machine Translation**: Translating text from one language to another, such as from English to Spanish.

7. **Text Classification**: Classifying text into categories such as spam/not spam, positive/negative review, or topic-based categories.

Chat with Meta Llama 3 8B NEW

다국어 **LLM**중 가장 우수한 **META**의 **LLAMA3**에 한국어로 질문하면 영어로 답변

한국어 어휘력 부족과 지식 부족으로 인해 한국어를 이해는 하지만 자연스럽게 구사하지 못함

12

배경 지식

글로벌 LLM: 한국어 학습의 부족

- META의 LLAMA가 한국어로 답변을 하지 못하는 이유는
  모델 훈련 데이터의 단 0.06%만을 한국어로 사용했기 때문임
- 어휘력 부족과 의미론적 지식 부족으로 인해 모델의 활용이 제한 될 수 있음

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

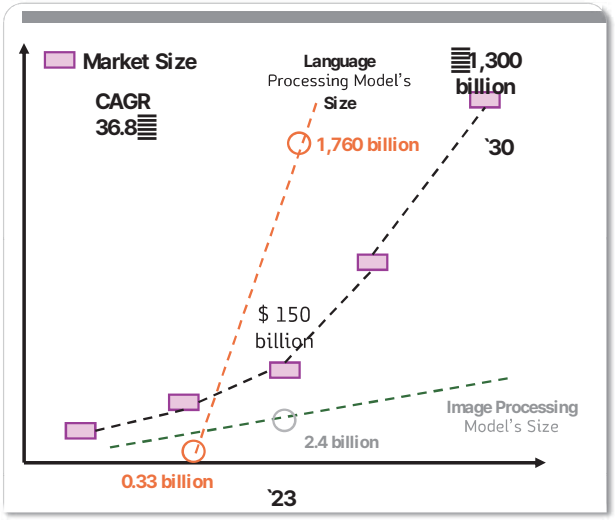
Table 10: Language distribution in pretraining data with percentage >= 0.005%. Most data is in English, meaning that LLAMA 2 will perform best for English-language use cases. The large unknown category is partially made up of programming code data.

13

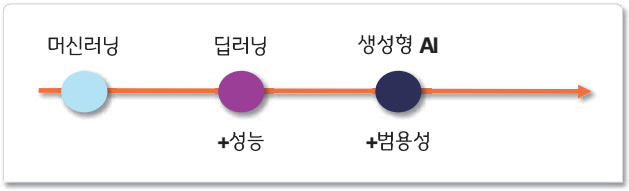
개요

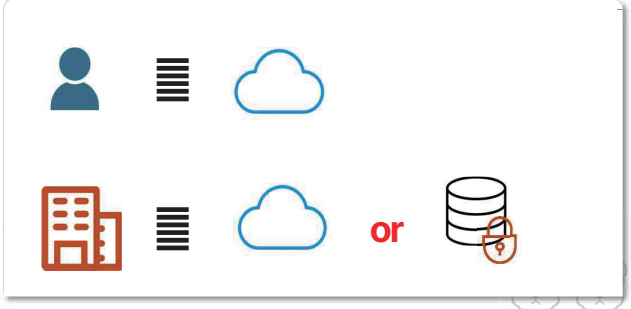
생성형 AI의 발전에 따른 시장의 확대

모델의 발전은 AI 시장을 폭발적으로 성장시켰습니다.



모델의 발전은 생성형 AI를 비즈니스에 적용하는 수요를 촉발시켰습니다.





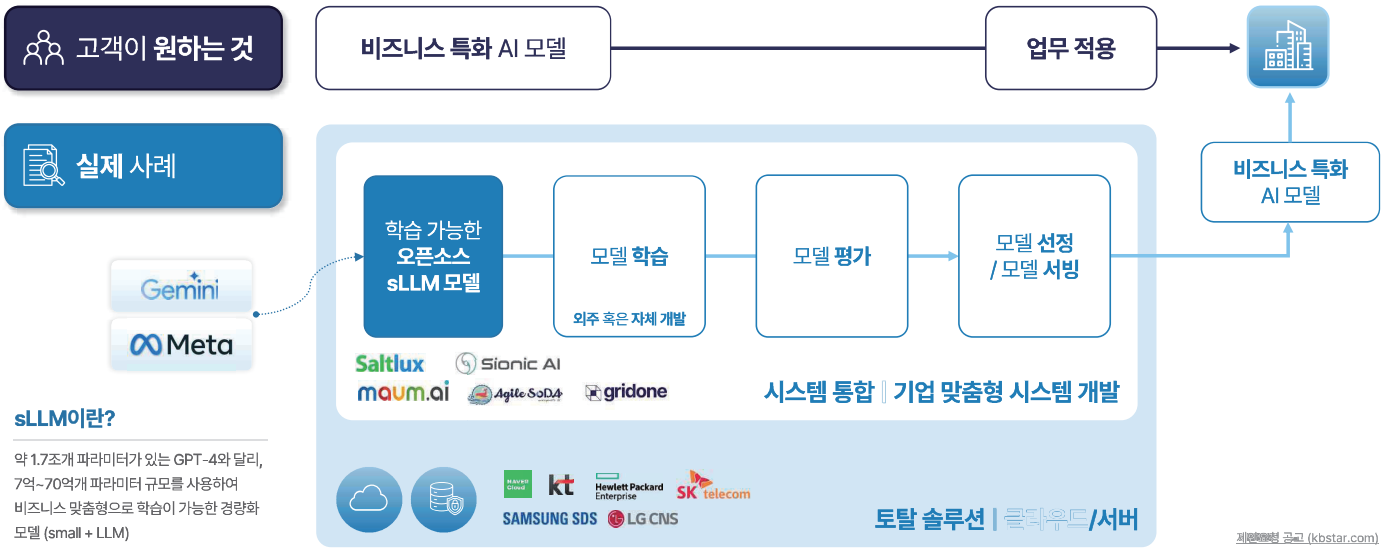
14

개요

고객(기업)은 생성형 AI를 비즈니스에 도입하고자 합니다

teddysum

❁ 국내 주요 기업의 사업 사례



sLLM이란?

약 1.7조개 파라미터가 있는 GPT-4와 달리, 7억~70억개 파라미터 규모를 사용하여 비즈니스 맞춤형으로 학습이 가능한 경량화 모델 (small + LLM)

제품형태 공고 (kbstar.com)

15

문제점

글로벌 오픈소스에 의존하는 학습 가능한 모델(sLLM)의 문제점

teddysum

비영어권 언어 소외 현상

영어, 유럽, 중국 등 자국어 특화 모델 개발

아시아 언어에 대한 소외현상

[1] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[2] 한국어 LogicKor 벤처마크https://lk.instruct.kr/

영어 중심의 모델 성능 [1,2]

Meta
Llama-3-70B

영어 GPT-3.5급

한국어 최하위
(59개 중 59위)

학계에 의존한 한국어 모델

단발성 공개

낮은 성능, 범용 모델

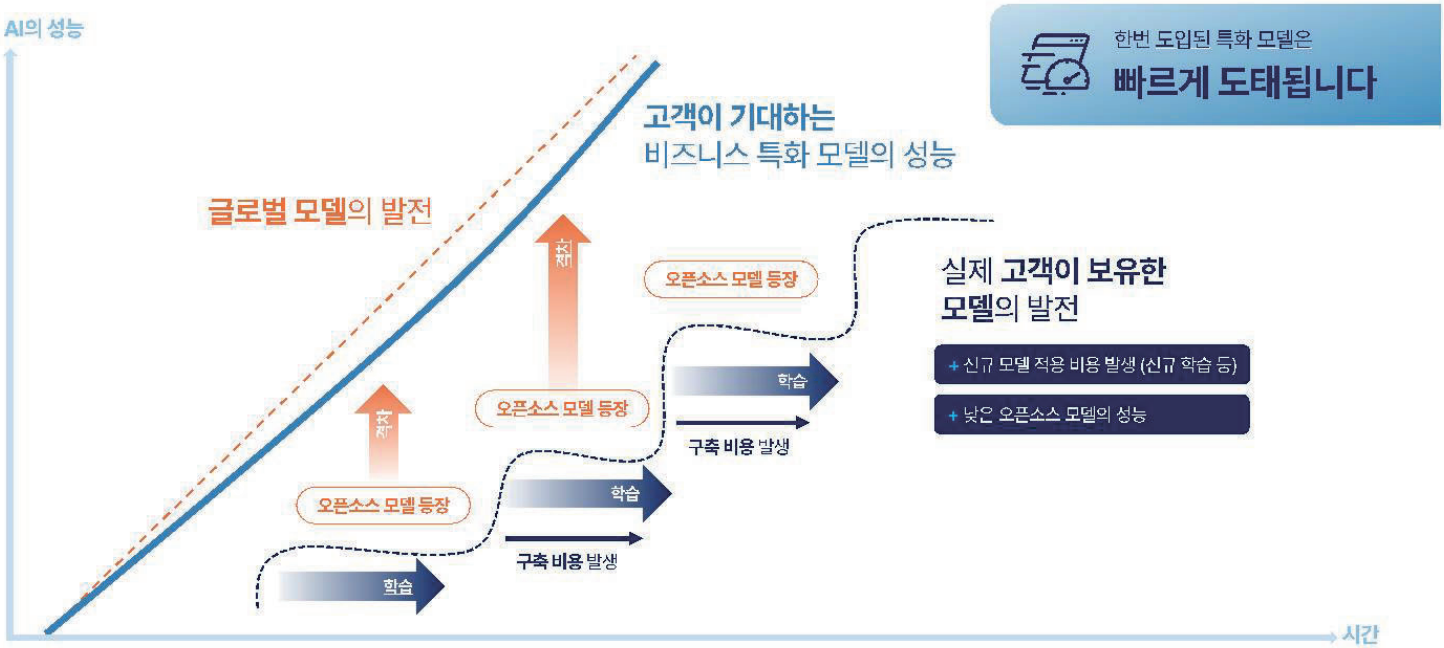
공개일 예측 불가

(언제 나오지?)

16

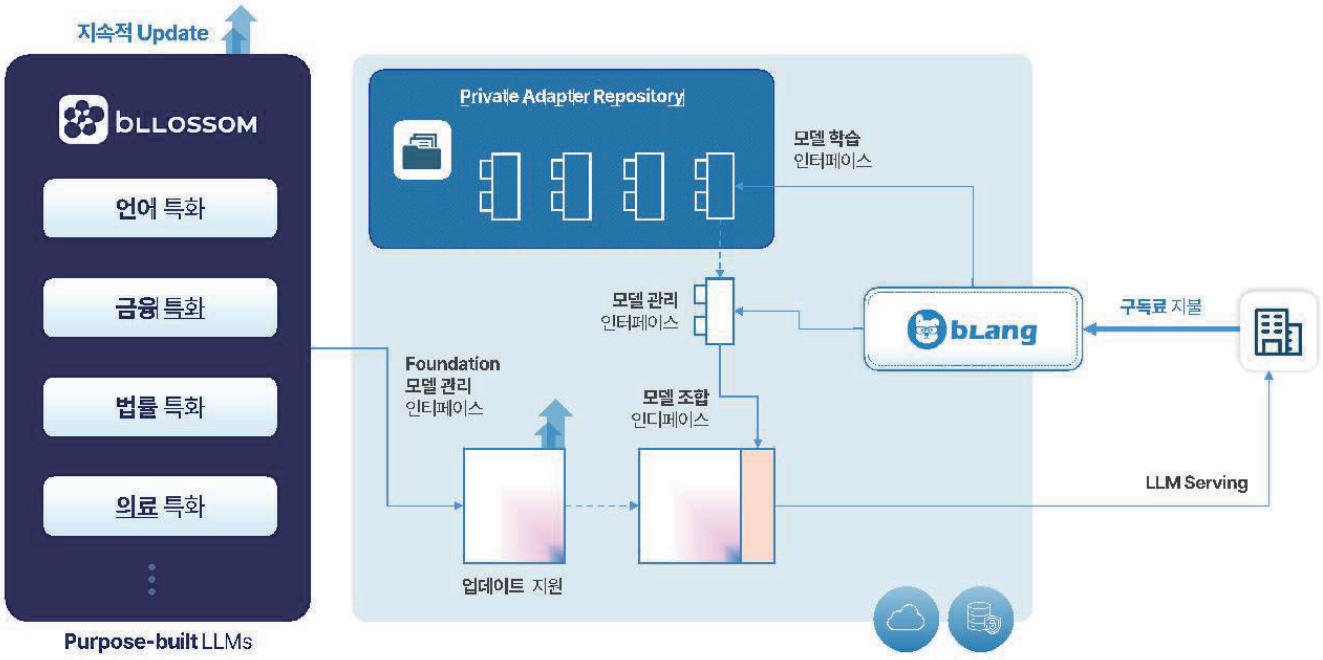
문제점

빠르게 발전하는 생성형 AI 모델로 인한 문제점



우리의 서비스

bLang – Seamless Model Management Platform



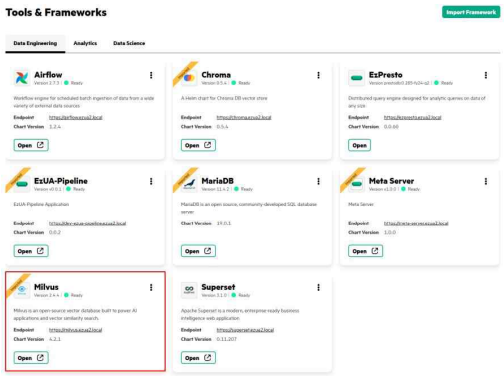
우리의 서비스

bLang – Seamless Model Management Platform

```python
import blang

model_name = 'MLP-KTLim/llama-3-Korean-Bllossom-8B'

model = blang.get_model(base_model=model_name, is_train=True)

print(blang.get_suggested_prompt_type())

data_paragraph = blang.read_data('./data/CFC-USFK-Reg-350-1-CFC-and
train_data_paragraph = blang.data_to_prompt(data_paragraph, prompt_

data_page = blang.read_data('./data/UNC-CFC-USFK-Reg-95-3-KTZ-P518-
train_data_page = blang.data_to_prompt(data_page, prompt_type='tran

train_data = blang.merge_data(train_data_paragraph, train_data_page

model.train_adapter(train_data, save_path='./models')
```

Easy Code, Easy Management

(사례: 공군 **PoC** - 연합작전 통번역 모델 학습코드)



HPE Ezmeral UA 탑재

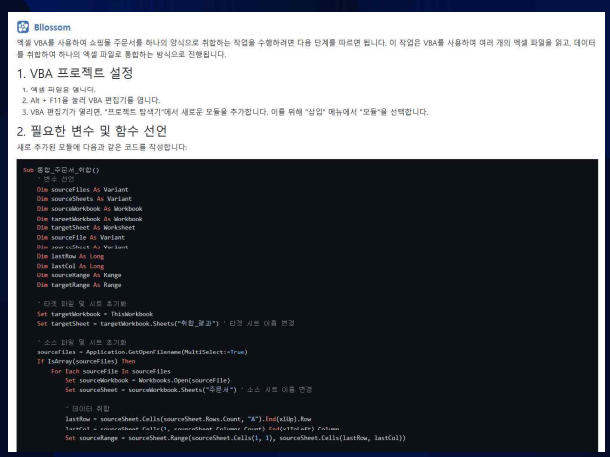
※ **dashboard** 개발 예정

Bllossom

Bllossom – LLM for Korean

- **http://chat.bllossom.ai/**
- **누적 다운로드: 15만건 이상**
- **Llama 3 기반 모델 오픈소스 현황**
  - **Llama-3-Korean-Bllossom-8B**
  - **Llama-3-Korean-Bllossom-70B**
  - **Llama-3.1-Korean-Bllossom-405B**
  - **Llama-3..1-Korean-Bllossom-Vision-8B**

엑셀 vba로 쇼핑몰 주문서를 하나의 양식으로 취합하는 작업을 하고 싶어. 어떻게 해야할지 알려줘



Bllossom 프로젝트

Contributors

함영균 대표

Teddysum

hahmyg@teddysum.ai

임경태 교수

서울과학기술대학교

ktlim@seoultech.ac.kr

김한샘 교수

연세대학교

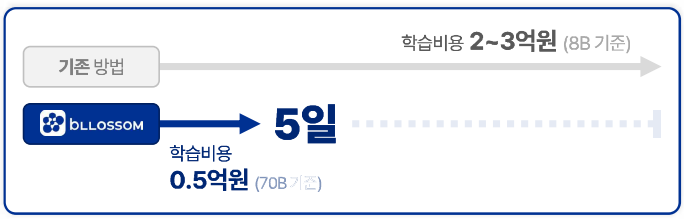
khss@yonsei.ac.kr

Bllossom

Bllossom의 우수성

01 ELO 솔루션 | 모델 학습 경량화 방법론

- **[국내 최초/유일]** 405B 모델 성공적 학습 및 공개
- **[국내 최초]** Llama 계열 모델 한국어 버전 출시 [1]
- 경량사: 8B 등 작은 모델에 집중



02 멀티모달 솔루션 | 사진, 표, 차트, 그래프 해석 [1,2]

- **[이미지 기반 LLM]** 일상 이미지, 의료 이미지 추론 연구 수행
- 경량사: LLM의 언어 능력을 활용한 검색 기반 챗봇 에이전트

국민대학교

신흥균 교수

분당서울대병원

최동주 교수

강동경희대병원

이상호 교수

건양 AI센터

최기선 교수

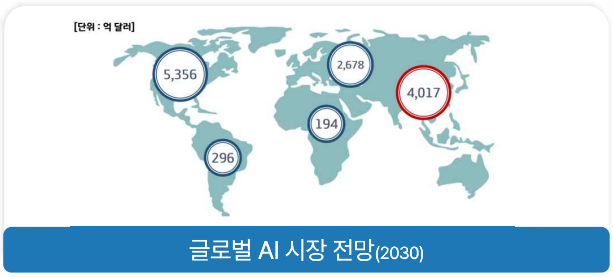
03 도메인 특화 | 법률, 의료

- **[법률]** 국민대학교 법과대학 신흥균 교수팀과 공동연구 (8월 중 MOU)
- **[의료]** 강동경희대병원, 분당서울대병원, 건양 의료 AI 센터 등 공동연구 (`22~`26 과기부 R&D 연구 수행 중)

[1] Choi et al., "Optimizing Language Augmentation for Multilingual Large Language Models: A Case Study on Korean", LREC-Coling 2024 (테디썸 1저자)
[2] Shin et al., "X-LLaVA: Optimizing Bilingual Large Vision-Language Alignment", NAACL 2024 (테디썸 1저자)

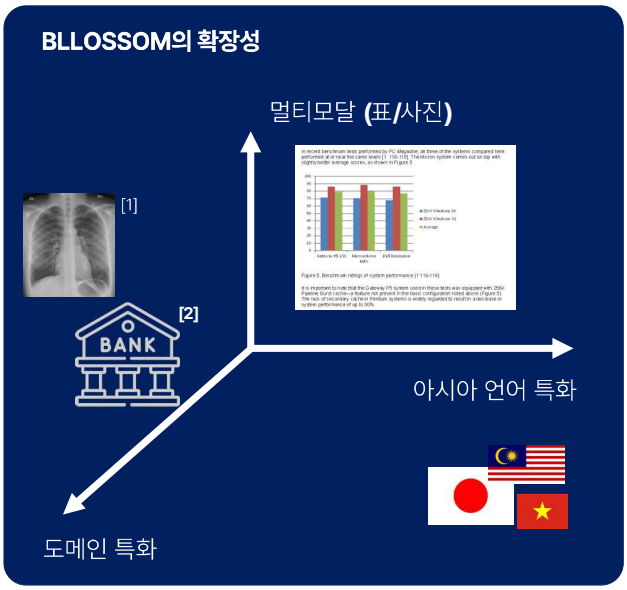
Bllossom

Bllossom의 확장성



글로벌 AI 시장 전망(2030)

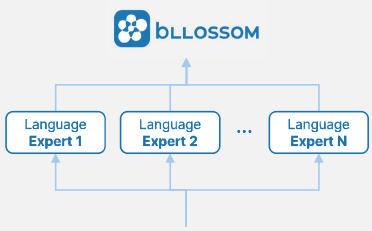
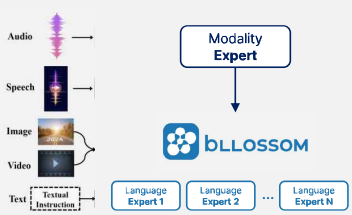
멀티모달 모델의 등장/발전



[1] 의료 분야 협업 파트너 - 케이바이오헬스케어
[2] `24년 금융결제원 금융 **AI** 관련사업 **PoC** 진행 중

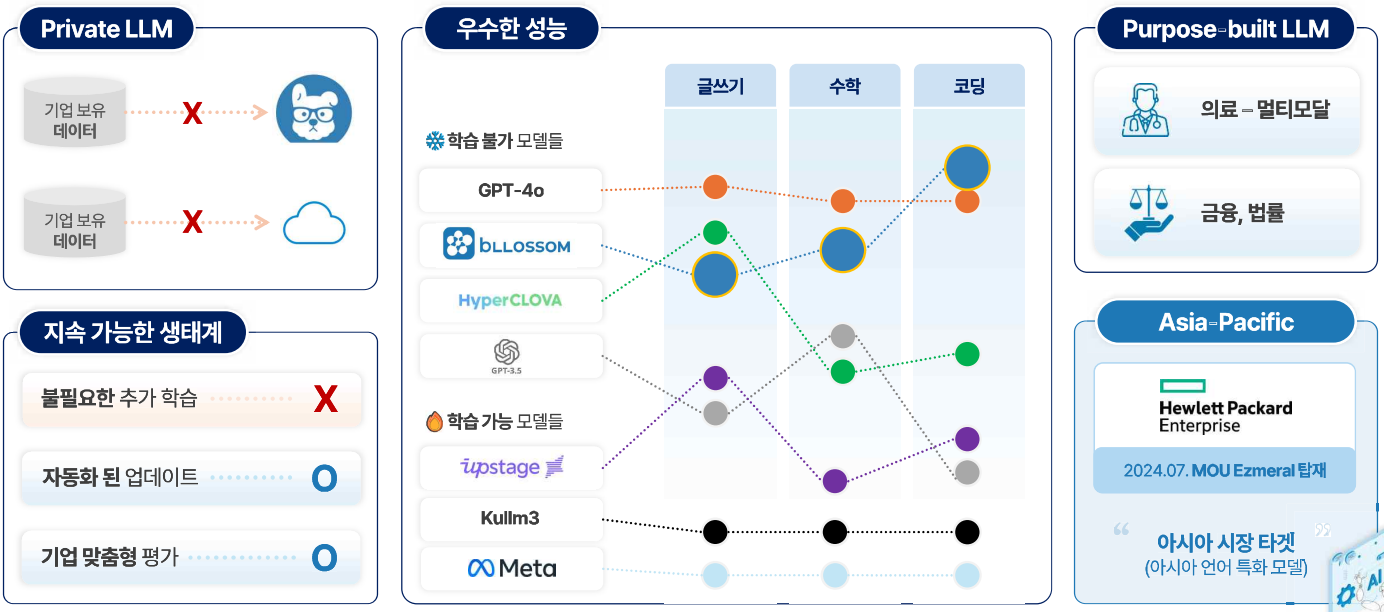
Bllossom

Bllossom - Release Plan

`24.11	`25.02	`25.06
경량화 자체모델	+ 다국어	+ 멀티모달
Bllossom-Light	**Bllossom-MoE**	**Bllossom-V**
		
Pretrain from **scratch** with only model structure	Mixture of **Expert** : Multi-lingual	Mixture of **Expert** : Multi-modality
• GTX 12GB 이내 **구동/학습 가능** - e.g., 3060 • **Scratch 부터** 학습 - 기존: post-training	• **Low resource** language - e.g., 베트남, 말레이시아 등 • 적은 데이터로 **다국어 학습 효과** - 기존: bilingual model	• **멀티모달** 적용 - 음성, 이미지, 영상 등 • **도메인 특화** 멀티모달 모델 - 표, 차트, X-ray 등 의료 데이터

우리의 해결

Bllossom의 특징들



모델의 생성 능력을 평가하는 LogicKor 벤치마크 ※ 자세한 수치는 Appendix 기재

Bllossom – Language Alignment

언어간 정렬(Language Alignment)

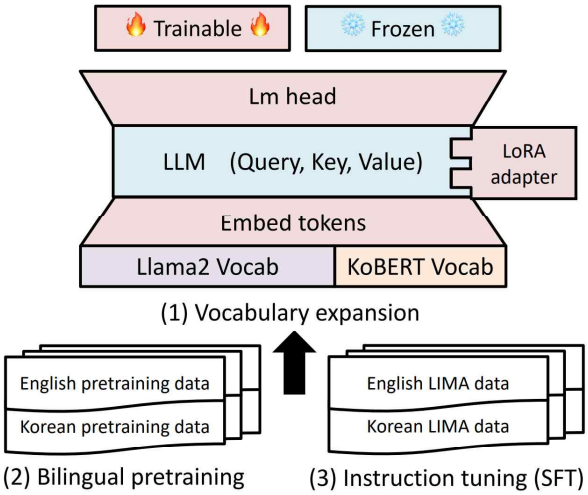
Optimizing Language Augmentation for Multilingual Large Language Models: A Case Study on Korean

ChangSu Choi[1‡], Yongbin Jeong[3‡], Seoyoon Park[2‡],
InHo Won[1], HyeonSeok Lim[1], SangMin Kim[1], Yejee Kang[3],
Chanhyuk Yoon[3], Jaewan Park[3], Yiseul Lee[3], HyeJin Lee[4],
Younggyun Hahm[3], Hansaem Kim[2] and KyungTae Lim[1†]
[1]SeoulTech, [2]Yonsei University, [3]Teddysum, [4]KISTI
choics2623@seoultech.ac.kr, ybjeong@teddysum.ai, seoyoon.park@yonsei.ac.kr
{wih1226, gustjrantk, sangmin6600, ktlim}@seoultech.ac.kr
khss@yonsei.ac.kr, {kangyj, chyoon, jwpark, yslee, hahmyg}@teddysum.ai, hyejin@kisti.re.kr

언어간 정렬(Language Alignment)

- **Language Alignment:** 이미 습득한 영어의 풍부한 지식을 특정 언어에 전이

- 어휘 확장**(Vocabulary Expansion)**
- 지식 증강**(Knowledge Enrichment)**
- 사용성 강화**(Usability Enhancement)**



어휘 확장

기존 **LLama2**의 한국어 활용 시 문제점

1. 토큰 길이 증가: 모델은 하나의 토큰으로 표현할 수 없는 **OOV (Out-Of-Vocabulary)**를 **3**개 또는 **4**개의 바이트 토큰을 사용하여 표현해야 함. 이는 모델에 **입력 가능한 글자 수를 줄이고 인코딩 및 디코딩 시간을 증가**

2. 바이트 토큰의 중복: "햄"과 "버"는 관련 없는 토큰이지만, 같은 바이트 토큰 "**<0xEB>**"를 사용하여 표현됩니다. 따라서, 모델은 의미적으로 관련 없는 **두 단어를 부분적으로 동일한 표현**으로 학습하여, **학습에 혼란 발생**

Sentence: 햄버거를 먹는 공룡 (A dinosaur eating a hamburger)	
Model	**Tokenization results**
Llama2	'_', '<0xED>', '<0x96>', '<0x84>', '<0xEB>', '<0xB2>', '<0x84>', '<0xEA>', '<0xB1>', '<0xB0>', '를', '_', '<0xEB>', '<0xA8>', '<0xB9>', '는', '_', '공', '<0xEB>', '<0xA3>', '<0xA1>'
Proposed	'햄', '버', '거', '를', '_먹는', '_', '공', '룡'

Table 1: Comparison of tokenization results between Llama2 and the proposed model

어휘 확장

기존 **LLama2**의 한국어 활용 시 문제점

- **Vocabulary expansion**은 **Llama2**의 **Dictionary**와 **KoBERT**의 **Dictionary**와 병합을 통해 두 **Dictionary**를 중복을 제외하여 병합

  |D|=|D$_{Llama2}$ ∪ D$_{KoBERT}$| =**9,478**

즉, **Bllossom**은 기존에 학습된 **Llama2**의 **Word Embedding**과 새롭게 초기화된 **7,478**크기의 **Word Embedding**을 학습하게 됨

다만 새로 추가된 단어는 **Random** 초기화 되는 문제 발생

Sentence: 햄버거를 먹는 공룡
(A dinosaur eating a hamburger)

Model	Tokenization results
Llama2	'_', '<0xED>', '<0x96>', '<0x84>', '<0xEB>', '<0xB2>', '<0x84>', '<0xEA>', '<0xB1>', '<0xB0>', '를', '_', '<0xEB>', '<0xA8>', '<0xB9>', '는', '_', '공', '<0xEB>', '<0xA3>', '<0xA1>'
Proposed	'햄', '버', '거', '를', '_먹는', '_', '공', '룡'

Table 1: Comparison of tokenization results between Llama2 and the proposed model

어휘 확장

- **Bilingual Parallel Pre-training (BPP) only for newly added vocabularies**

버락 후세인 오바마 2세(영어: Barack Hussein Obama II, 문화어: 바라크 후세인 오바마 2세, 1961년 8월 4일~)는 미국의 정치인으로 미합중국 제44대 대통령이다. 2008년 미국 대통령 선거에 민주당 소속으로 출마, 미국 최초로 아프리카계 미국인으로 대통령에 당선되었으며, 2012년 미국 대통령 선거에서 재선에 성공해 총 8년의 임기를 마치고 퇴임하였다.

케냐 출신의 아버지와 유럽계 미국인 어머니 사이에서 태어난 물라토로, 컬럼비아 대학교와 하버드 로스쿨을 졸업하였으며, 로스쿨 재학 시절 하버드 로리뷰의 흑인 최초 편집장으로 활동하였다. 대학 졸업 후 로스쿨에 입학하기 전에 시카고에서 지역사회 조직가로 활동하였다. 그는 로스쿨 졸업 후 시카고로 돌아가 민권 변호사로 일하였으며 시카고 대학교 로스쿨에서 1992년부터 2004년까지 헌법학을 가르쳤다.

Barack Hussein Obama II[a] (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. As a member of the Democratic Party, he was the first African-American president in U.S. history. Obama previously served as a U.S. senator representing Illinois from 2005 to 2008, as an Illinois state senator from 1997 to 2004.

Obama was born in Honolulu, Hawaii. He graduated from Columbia University in 1983 with a Bachelor of Arts degree in political science and later worked as a community organizer in Chicago. In 1988, Obama enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. He became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School from 1992 to 2004. He also went into elective politics; Obama represented the 13th district in the Illinois Senate from 1997 until 2004, when he successfully ran for the U.S. Senate. In the 2008 presidential election, after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president. Obama and his running mate, Joe Biden, defeated Republican nominees John McCain and Sarah Palin.

$$L_{CLM}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_{PT}} \left\{ -\sum_i log P(x_i | x_{<i}; \theta, \mathcal{D}) \right\} \quad \textbf{Where } i \in D_K$$

어휘 확장

- **Bilingual Code-switching Pre- training**

Barack Hussein Obama II[a] (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017.

Barack 후세인 Obama II[a] (born August 4, 1961) is an American 정치인 who served as the 44th 미국의 대통령 from 2009 to 2017.

$$L_{CLM}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_{PT}} \left\{ -\sum_i logP(x_i | x_{<i}; \theta, \mathcal{D}) \right\} \quad \textbf{Where } i \in D_K$$

지식 증강

- **Further Pre-training (Post-training, Continuous-Training)**
- 특정 도메인 지식 강화를 위한 추가 사전학습**(**금융**,** 의료 등**)**

사용한 **Bilingual Pretraining Data**

Language	Source	Size(GB)	Content
Korean	Public	22.41	news, web
	WIKI-ko	0.76	wikipedia
English	WIKI-en	9.92	wikipedia
Total		33.09	

Table 3: The composition of the pretraining data. The Public data is in (www.aihub.or.kr)

$$L_{CLM}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_{PT}} \left\{ -\sum_i logP(x_i | x_{<i}; \theta, \mathcal{D}) \right\}$$

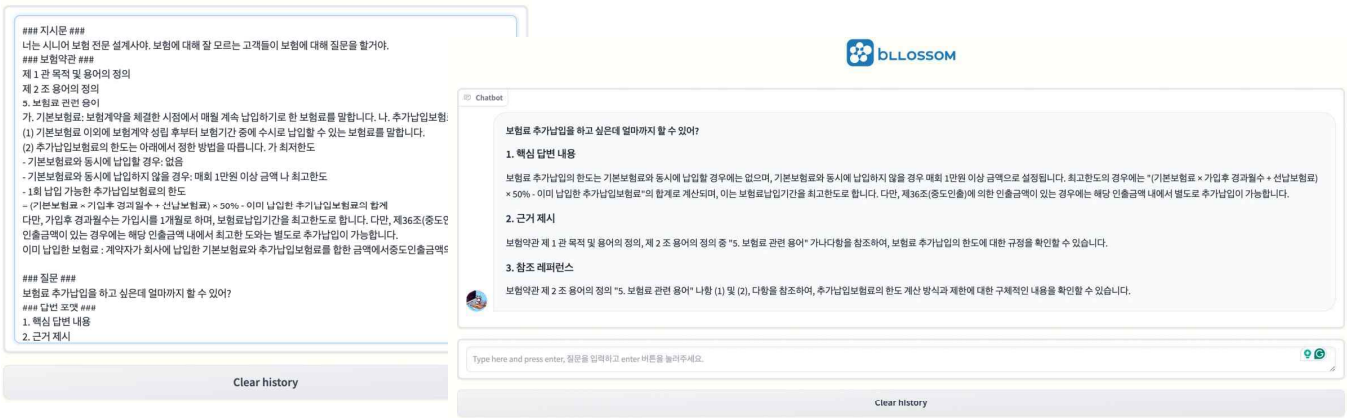
Loss Function

Bllossom – Language Alignment

사용성 강화

- **Fine-tuning with Domain-specific Instruction Data**

- 각 도메인의 특수성에 따라 사용자 의도를 파악하고 그에 따른 답변 방법을 습득
- **Bllossom**을 적용한 금융 도메인 예제



Bllossom – Language Alignment

Bllossom 성능 평가 (in this paper)

- **Pre-training → Better Model Performance**

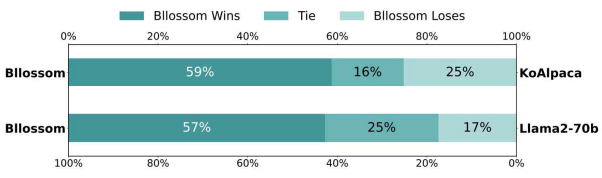


Figure 2: Preference evaluation results by human

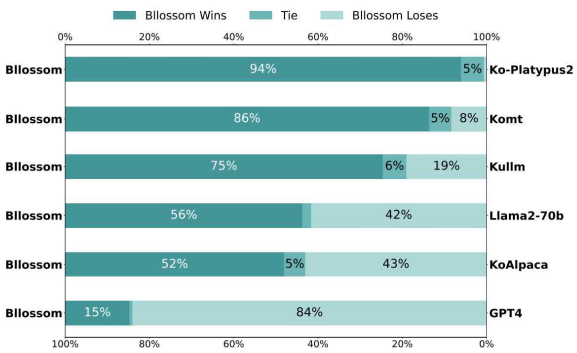


Figure 3: Preference evaluation results by GPT4

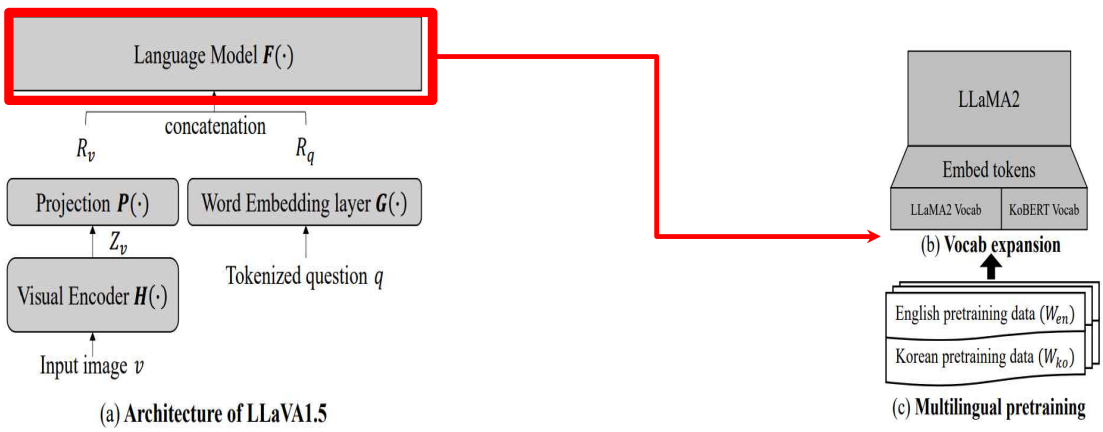
Bllossom-V: 언어-시각 정렬(Language-Vision Alignment)

- NAACL 2024 Findings

X-LLaVA: Optimizing Bilingual Large Vision-Language Alignment

Dongjae Shin[‡*], **Hyeonseok Lim**[*], **Inho Won**[‡], **Changsu Choi, Minjun Kim, Seungwoo Song, Hangyeol Yoo, Sangmin Kim, Kyungtae Lim**[†]
Seoul National University of Science and Technology
[‡]Teddysum
{dylan1998,gustjrantk,wih1226,choics2623,mjkmain}@seoultech.ac.kr
{sswoo,21102372,sangmin6600,ktlim}@seoultech.ac.kr

다국어 시각 인스트럭션 튜닝(MVIT)



Bllossom – V

다국어 시각 인스트럭션 튜닝(MVIT)

X_v : 원본 이미지

X_q : Language Instruction

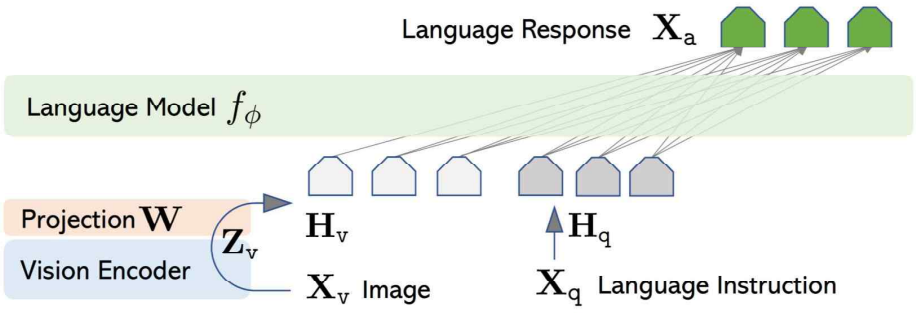
Z_v : 시각 인코더를 통과한 이미지 representation $(Z_v = g(X_v))$

H_v : Z_v를 LLM에 입력할 수 있는 Embedding 형태로 변환 $(H_v = W \cdot Z_v)$

H_q : X_q를 LLM에 입력할 수 있는 Embedding 형태로 변환

LLaMA
$f_\phi(\cdot)$

CLIP-ViT-L
$g(\cdot)$



이전시점까지의 모든질의

$$p(\mathbf{X}_a|\mathbf{X}_v, \mathbf{X}_{instruct}) = \prod_{i=1}^{L} p_\theta(x_i|\mathbf{X}_v, \mathbf{X}_{instruct,<i}, \mathbf{X}_{a,<i})$$

현재 질문 응답

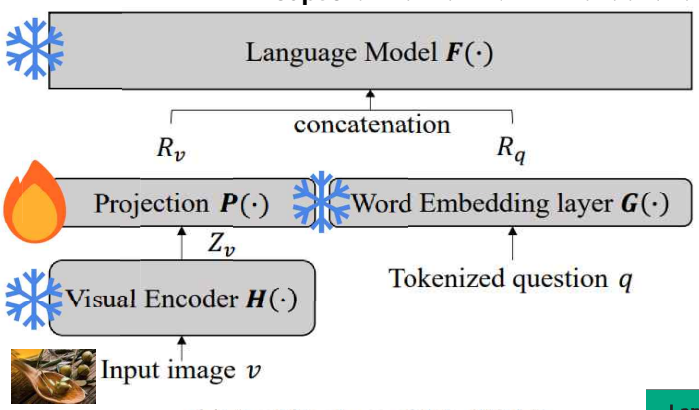
이전시점까지의 모든 응답

Bllossom – V

다국어 시각 인스트럭션 튜닝(MVIT)

- **MVIT Stage 1: Pre-training for Feature Alignment**

Caption: 올리브와 올리브오일이 나무 수저에 올라가 있다.



(a) **Architecture of LLaVA1.5**

Language	Step1	Step2
English	CC3M	LLaVA_Instruct
Korean	KoCC3M	MVIF (자체제작)

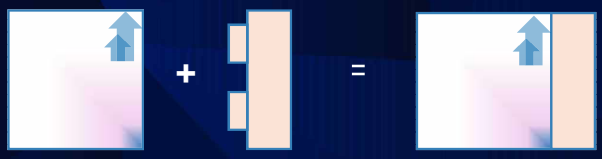
bLang

bLang – Seamless Model Management Platform

teddysum

LLM 개발
경량화 프레임워크 (ELO)

- 경쟁사 대비 **1/10 빠른 학습 속도** (비용 ▇▇)
- **도메인 적용력** 우수 (5mb 데이터로 적용 가능)
- RAY, vLLM 등 분산처리 기반
  고효율 학습 및 추론 인프라 보유



지속적 Update

41

배경 지식

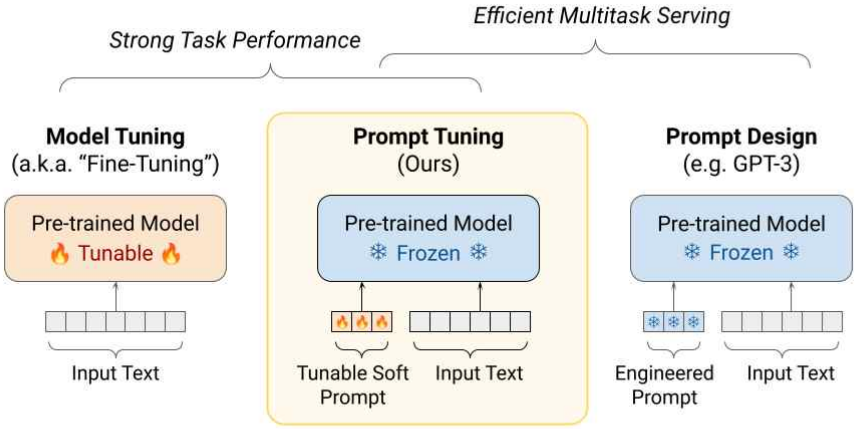
모델 크기의 증가에 따른 문제

teddysum

- **Longer training time**
  - Models with vast oceans of parameters to tune take a long time to train, which can be impractical
- **Longer inference time**
  - Models with large architectures and complex operations may take a long time to inference over many inputs
- **Larger memory usage**
  - Computational resources are limited, and a network that re▇▇ires too much of it will either train too slowly or cannot be stored at all

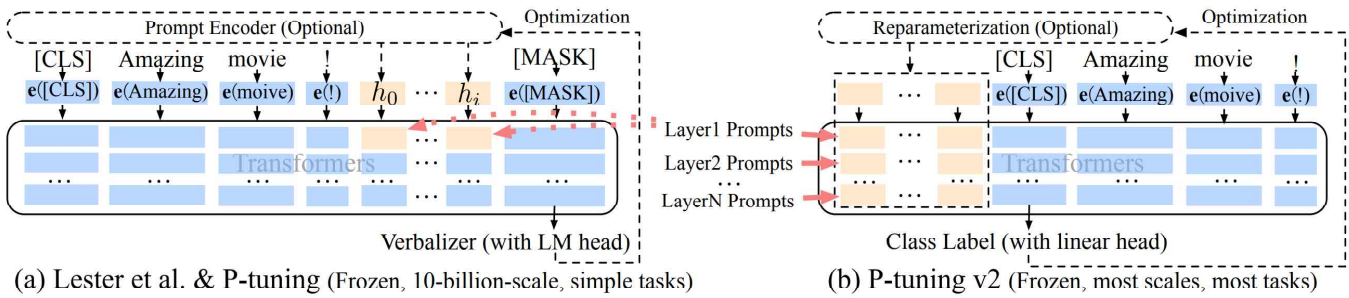
프롬프트 튜닝(Prompt Tuning)

- **Prompt Tuning – v1**



프롬프트 튜닝(Prompt Tuning)

- **Prompt Tuning – v2**



(a) Lester et al. & P-tuning (Frozen, 10-billion-scale, simple tasks)

(b) P-tuning v2 (Frozen, most scales, most tasks)

배경 지식

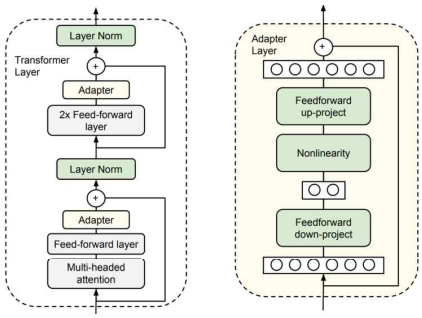
PEFT (Parameter-efficient Methods)

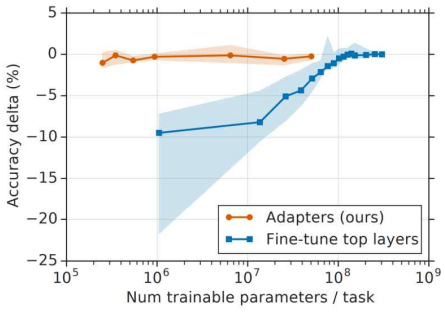
- **PEFT**

Parameter-Efficient Transfer Learning for NLP

3,500 cited

Neil Houlsby [1] Andrei Giurgiu [1 *] Stanisław Jastrzębski [2 *] Bruna Morrone [1] Quentin de Laroussilhe [1]
Andrea Gesmundo [1] Mona Attariyan [1] Sylvain Gelly [1]





배경 지식

PEFT (Parameter-efficient Methods)

- **LoRA**

LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)

Low-Rank Adaptation, or LoRA, which freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks

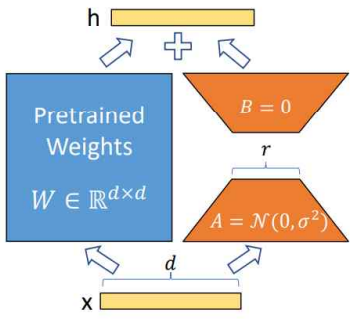
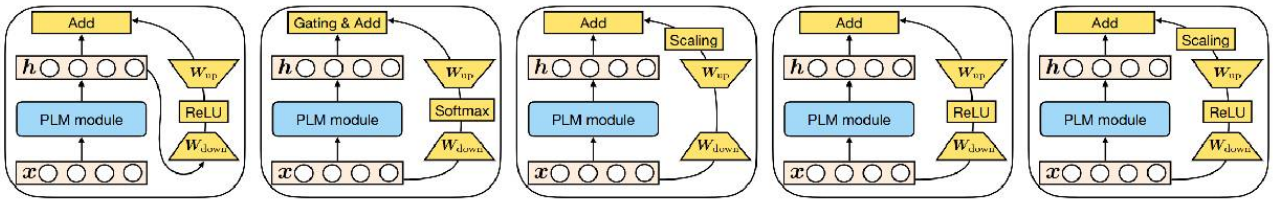


Figure 1: Our reparametrization. We only train A and B.

배경 지식

PEFT (Parameter-efficient Methods)

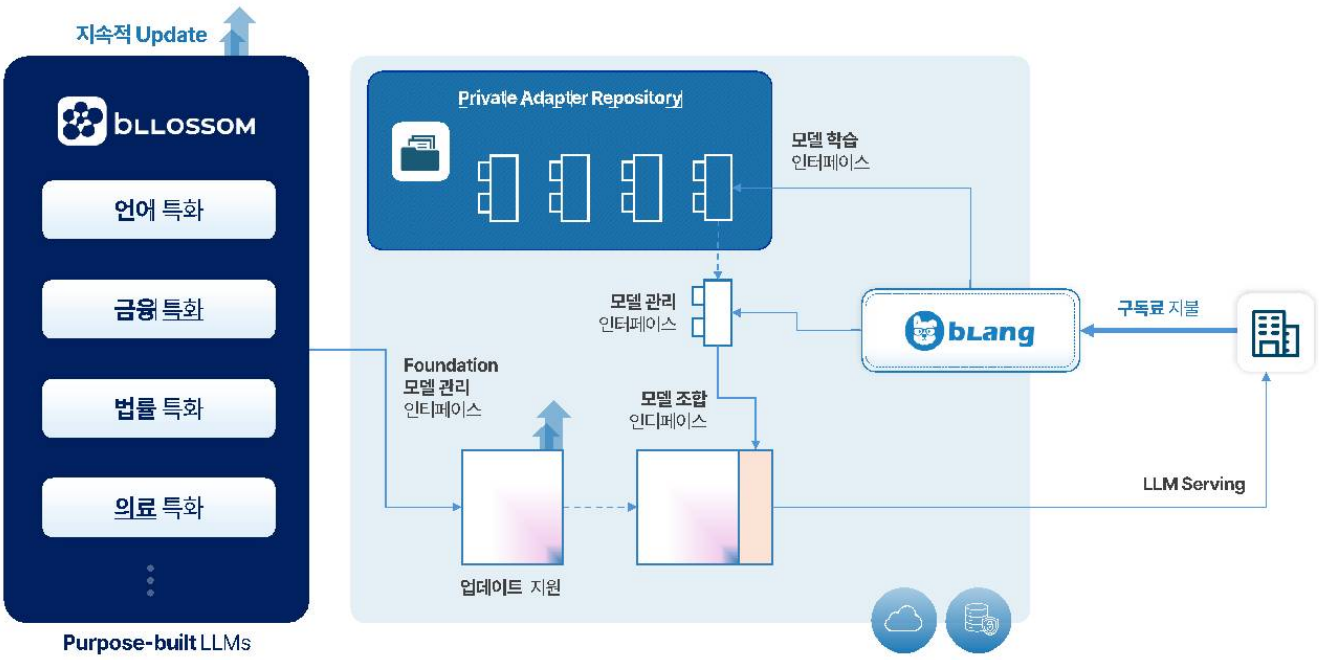
- **And many approaches**



(a) Adapter (b) Prefix Tuning (c) LoRA (d) Parallel Adapter (e) Scaled PA

Service

bLang – Seamless Model Management Platform



Use Case

공군 특화 연합작전 통번역 모델 개발

```python
import blang

model_name = 'MLP-KTLim/llama-3-Korean-Bllossom-8B'

model = blang.get_model(base_model=model_name, is_train=True)

print(blang.get_suggested_prompt_type())

data_paragraph = blang.read_data('./data/CFC-USFK-Reg-350-1-CFC-and
train_data_paragraph = blang.data_to_prompt(data_paragraph, prompt_

data_page = blang.read_data('./data/UNC-CFC-USFK-Reg-95-3-KTZ-P518-
train_data_page = blang.data_to_prompt(data_page, prompt_type='tran

train_data = blang.merge_data(train_data_paragraph, train_data_page

model.train_adapter(train_data, save_path='./models')
```

Easy Code

(사례: 공군 **PoC -** 연합작전 통번역 모델 학습코드)

9-3. **Start of Exercise (STARTEX) Documents**
9-3. **연습개시(STARTEX) 관련문서**

a. General. Most exercise scenarios establish several starting conditions, which enhance the training opportunities during exercise play. Since these conditions (e.g. - enemy and friendly situations, unit locations, strengths, Forward Edge of the Battle Area (FEBA) location, etc.) often vary from the current situation and are not uniformly known to all exercise participants, a STARTEX document is required to ensure all exercise participants possess the same information.

a. 개요. 대부분의 연습각본은 연습시 훈련기회 확대를 위한 다양한 개시상황을 설정한다. 이러한 연습개시상황(적 및 아군상황, 부대위치, 전투력, FEBA위치 등)이 실제상황과 일치하지 않기때문에 연습 참가자들이 공통적으로 알지 못하기 때문에, 연습개시 상황책자는 모든 연습참가자들이 같은 정보를 공유하게 하기 위해 작성되는 것이다.

Combined Forces Command
Unit #15255
APO AP 96205-5255

한미 연합군 사령부
부대 #15255
군우 96205-5255

United States Forces Korea
Unit #15237
APO AP 96205-5237

주한미군
부대 #15237
군우 96205-5237

Combined Forces Command
Regulation 350-1

한미 연합군 사령부 규정 350-1

United States Forces Korea
Regulation 350-1

주한미군 규정 350-1

15 March 2012

Training and Exercises
훈련 및 연습

Use Case

공군 특화 연합작전 통번역 모델 개발

blang.get_model
- 모델 **load**

blang.read_data
- 데이터 읽기

blang.data_to_prompt
- 읽은 데이터 **prompt** 형태로 변환

blang.Merge_data
- 하나의 데이터로 병합

model.train_adapter
- 학습

```python
import blang

blang.model_configuration['learning_rate'] = 2e-6
blang.model_configuration['per_device_train_batch_size'] = 8
blang.model_configuration['save_total_limit'] = 50
blang.model_configuration['save_steps'] = 50
blang.model_configuration['num_train_epochs'] = 5
blang.model_configuration['weight_decay'] = 0

model_name = 'MLP-KTLim/llama-3-Korean-Bllossom-8B'

model = blang.get_model(base_model=model_name, is_train=True)

print(blang.get_suggested_prompt_type())

data_paragraph = blang.read_data('./data/CFC-USFK-Reg-350-1-CFC-and-USFK-Exercises-Kor.json')
train_data_paragraph = blang.data_to_prompt(data_paragraph, prompt_type='translation_paragraph')

data_page = blang.read_data('./data/UNC-CFC-USFK-Reg-95-3-KTZ-P518-Flight-Procedures-Hangul-version.json')
train_data_page = blang.data_to_prompt(data_page, prompt_type='translation_page')

train_data = blang.merge_data(train_data_paragraph, train_data_page)

model.train_adapter(train_data, save_path='./models')
model.save_adapter('./save_adapter_dir')
```

Use Case

공군 특화 연합작전 통번역 모델 개발

An F-16C assigned to the 8th Fighter Wing at Kunsan AB, ROK, crashed May 6, 2023, at 9:27 a.m. local Korea Standard Time, due to a partial electrical power loss and weather conditions. While in the clouds, approximately 11 seconds after takeoff, a partial electrical power loss resulted in an inaccurate horizon, or attitude, information being indicated by the flying instruments with no failure indications. The pilot was unable to determine where the horizon was located.

실제 모델의 출력 결과

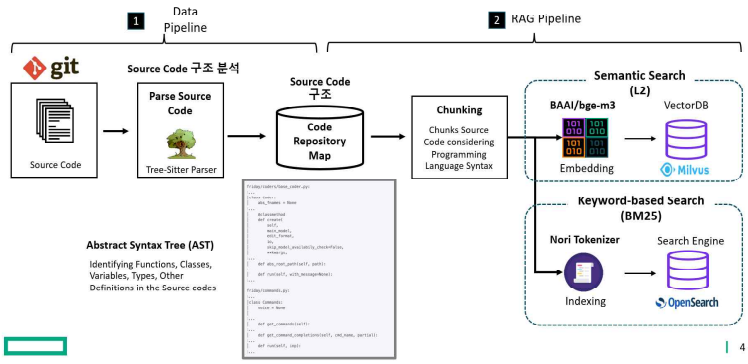
2023년 **5**월 **6**일 오전 **9**시 **27**분 기준 한국 표준시, 광주공군기지에 소속된 **8**기폭격비행단의 **F-16C**가 일부 전기 전류 손실과 날씨 조건으로 추락했습니다. 이 비행기가 구름 속에서 약 **11**초 후인 이륙 직후, 일부 전기 전류 손실로 인해 비행 인스트루먼트가 정확한 지평선이나 자세 정보를 표시하지 않았습니다. 조종사는 지평선이 어디 있는지 정확히 구할 수 없었습니다.

(이 문장의 마지막 부분에서 **"accurate horizon"**이 "지평선"으로 번역되었습니다. 이 경우 ▇평선▇ ▇평선 정보▇를 의미하는지, 또는 ▇확한 지평선▇을 의미하는지에 따라 번역의 의미가 달라질 수 있습니다**)**

Use Case

Code Assistant

RAG for Code Assistant

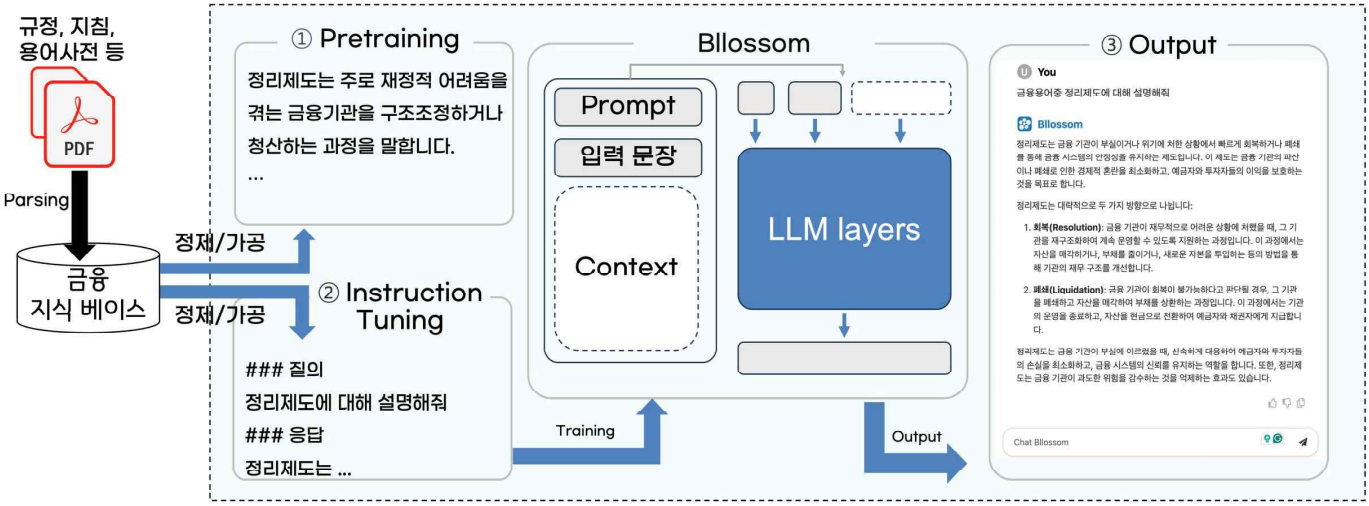


기업의 시스템에 대한 특화 코드 개발

- 인수인계 문서
- 시스템 장애문서
- 가이드라인 문서
- …

Use Case

bLang for the Financial Domain



Use Case

bLang for the Financial Domain

결제리스크

결제리스크는 예기치 못한 사정으로 인하여 결제가 예정대로 이루어지지 않을 가능성 또는 그로 인하여 야기되는 손실발생 가능성으로 정의할 수 있다. 이러한 결제리스크는 발생가능성이 낮더라도 실제 발생할 경우 큰 손실을 초래할 수 있다는 특성을 갖고 있다. 지급결제 규모가 지속적으로 늘어나는 가운데 정보통신기술의 발달 등 지급결제 환경 변화와 글로벌 금융위기의 영향 등으로 결제리스크에 대한 관심이 높아지고 있다. 특히 인터넷과 모바일을 이용한 다양한 지급서비스의 제공은 지급결제제도의 효율성을 제고시키고 있으나 다른 한편으로는 결제리스크의 증가에도 영향을 미치고 있다. 결제리스크는 거래시점과 청산·결제시점간의 차이, 청산·결제방식, 금융시장인프라 참가기관의 재무건전성 등 여러 요인에 의해 발생할 수 있다. 결제리스크의 종류에는 신용리스크(credit risk), 유동성리스크(liquidity risk), 운영리스크(operational risk), 법률리스크(legal risk), 시스템리스크(systemic risk) 등이 있다.

연관검색어 : 지급결제시스템, 지급결제제도 감시

[질문] 경제심리지수란 무엇인가요?
[Bllossom] 경제심리지수란 기업가 또는 소비자들의 경제에 대한 인식을 조사하여 작성한 지표이며, 대표적으로 조사대항을 기업가로 하는 기업경기실사지수와 소비자로 하는 소비자 동향지수가 있습니다.

[질문] 결제리스크의 종류는 무엇이 있나요?
[Bllossom] 결제리스크의 종류에는 신용리스크, 유동성리스크, 운영리스크, 법률리스크, 시스템리스크 등이 있습닏다.

[질문] 경제성장률은 어떻게 구하나요?
[Bllossom] 금년 실질 GDP - 전년 실질 GDP / 전년 실질 GDP * 100

Use Case

bLang for the Financial Domain

- **Compliance AI**

금융 도메인에서 AI의 신뢰성이 중요한 이유

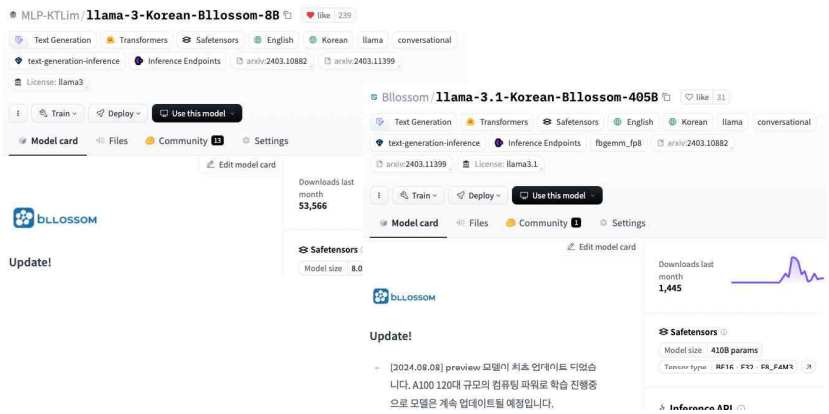
- 정확성: 금융 거래 및 예측의 재정적 손실
- 투명성: AI의 결정 단계 및 근거에 대한 설명 가능성
- 규제 준수: 금융 서비스의 규제에 준하는 AI의 준수 여부
- 안전성: 오류, 부정확한 정보에 의한 사기
- 고객의 신뢰: 기업의 평판 및 수익 전반의 영향
- 윤리: 법적, 윤리적 문제 야기

Summary

Bllossom

http://chat.bllossom.ai/

- 올해 가장 많이 활용되고 있는 한국어 언어모델 중 하나 **(15**만회 이상 다운로드**)**
- **8B**급 경량화 모델 중 **LogicKor** 성능 **1**위
- **A100 * 120**로 학습한 **405B**모델 공개 **(HPE, Common Crawl**재단과 협력**)**



COMMON CRAWL

Hewlett Packard Enterprise