

국립국어원 2023-01-36

발간등록 번호
---------

11-1371028-000962-01
----------------------

# 2023년 국어능력 진단체계 활용 방안 연구

연구 책임자  
함 영 균



## 제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '2023년 국어능력 진단체계 활용 방안 연구'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2023년 5월 3일 ~ 2023년 12월 10일

2023년 12월 7일

연구책임자: 함영균((주)테디썸)

연구 기관: (주)테디썸, 연세대학교, 서울과학기술대학교 산학협력단, (주)튜터러스랩스

연구책임자: 함영균

공동연구원: 김한샘, 임경태, 임준호,

장한열, 윤영민, 이진,

주민재, 박전규, 전형배

보조연구원: 윤환선, 정용빈, 서현빈,

박재완, 강예지, 강조은,

김유진, 김희재, 박서윤,

최규리, 이재원, 정가연,

이성우

## <국문 요약>

# 2023년 국어능력 진단체계 활용 방안 연구

본 연구는 국민의 국어 능력 향상을 위해 인공지능을 활용한 쓰기 자동 평가 및 첨삭 기능을 도출하고, 자가 진단쓰기 첨삭 지원에 필요한 말뭉치 구축 중장기 계획을 마련하며 국어 능력 진단체계의 발전 방안을 제시하는 것을 목적으로 하였다. 이에 따른 주요 과업과 연구 성과는 다음과 같다.

**국내외 언어 자원 및 정책 환경 분석:** 언어권별(영어권, 유럽어권, 중국어권, 일본어권)로 국외 언어 자원 현황을 분석하고 국가별 언어 자원 구축의 특징과 시사점을 정리하였다. 또한 국내외 소프트웨어(SW)와 언어모델, 인공지능(AI) 기반 글쓰기 평가 방법론 및 모델에 대한 분석을 바탕으로 한국어 글쓰기 자동 채점의 현황을 진단하고 필요 사항을 도출하였다.

**글쓰기 자동 평가 및 자동 첨삭 방안 제안:** 국내외 글쓰기 자동 평가 시스템 분석을 통해 한국어 글쓰기 자동 평가 방법론을 제안하고 한국어 특성을 반영한 피드백 방안을 연구하였다. 또한 생성형 AI의 글쓰기 자동 평가에 대한 가능성을 검토하고 한국 초거대 AI 모델 적용을 통한 글쓰기 피드백 방안을 실증하였다.

**한국어 말뭉치 구축 중장기 계획안 수립 및 발전 방향 제안:** 한국어 말뭉치 구축을 위한 중장기 계획안을 수립하였다. 자문위원회를 구성하여 학계와 산업계의 의견을 수렴하여 한국어 글쓰기 자동 평가를 위한 말뭉치 구축 방안을 제안하였으며 ‘AI말평’ 평가용 말뭉치 구축, 인스트럭션 말뭉치 구축 방안을 제안하였다. 또한 말뭉치 구축 단계별로 말뭉치 품질 검증 방안을 제안하였다.

**주요어:** 인공지능, 진단체계, 글쓰기 평가, 글쓰기 첨삭

## <Abstract>

# Study on the Utilization Strategies of the Korean Language Proficiency Diagnostic System in 2023

This research aimed to derive artificial intelligence (AI) based automated writing assessment and proofreading functions to enhance the language proficiency of the nation's citizens. The primary objectives included developing a self-diagnosis/writing proofreading support mid-to-long-term plan using diagnostic indicators and proposing avenues for the advancement of the national language proficiency diagnostic system. The key tasks and research outcomes are as follows:

Analysis of domestic and international language resources and policy environments: Analyzing the current status of foreign language resources based on language regions (English-speaking, European language-speaking, Chinese-speaking, Japanese-speaking) and summarizing the characteristics and implications of language resource construction in each country. Additionally, diagnosing the current state of Korean writing automatic scoring based on an analysis of domestic and international software, language models, and AI-based writing evaluation methodologies to derive necessary requirements.

Proposal of automated writing assessment and proofreading approaches: Through the analysis of domestic and international automated writing assessment systems, proposing a methodology for Korean writing automatic evaluation and researching feedback methods reflecting the characteristics of the Korean language. Also, exploring the potential of generative AI in automated writing assessment and validating approaches for providing writing feedback through the application of Korean super-large AI models.

Establishment of a mid-to-long-term plan for Korean corpus construction and proposal of development directions: Formulating a mid-to-long-term plan for Korean corpus construction and proposing plans for building corpora for Korean writing automatic assessment. Forming an advisory committee to gather opinions from academia and industry, suggesting corpus construction methods for Korean writing automatic assessment, and proposing plans for building evaluation corpora for 'AI writing assessment' and instructional corpora. Additionally, proposing quality verification methods for each stage of corpus construction.

**Keywords:** Artificial Intelligence, Diagnostic System, Writing Assessment, Feedback and Editing of writing



# 목차

1. 사업 개요	6
2. 환경 분석	15
3. 글쓰기 자동평가 및 자동첨삭 방안	54
4. 한국어 말뭉치 구축 중장기 계획안	74

# 1. 사업 개요

## 1. 사업 개요

1-1. 사업 추진 전략

1-2. 사업 수행 내용

1-3. 사업 수행 일정

1-4. 연구 이력

1-5. 사업 추진 체계

1-6. 기대 효과

## 2. 환경 분석

3. 글쓰기 자동평가 및 자동첨삭 방안

4. 한국어 말뭉치 구축 중장기 계획안

# 1. 사업 개요

사업명

국어능력 진단체계 활용 방안 연구

1. 목적

- 국민의 국어 능력 향상을 위한 인공지능 활용 쓰기 자동평가·첨삭 기능 도출
- 진단 지표를 활용한 자가 진단·쓰기 첨삭 지원 말뭉치 구축 중장기 계획안 마련 및 국어능력 진단체계의 발전 방안 제시

2. 범위

연구 범위	주요 내용
1. 인공지능 활용 분석	<ul style="list-style-type: none"> <li>• 국내외 소프트웨어(SW) 분석, 언어모델 분석, AI 기반 글쓰기 평가·첨삭 방법론 및 모델 분석</li> <li>• 국내외 언어 자원 및 정책 환경 분석</li> </ul>
2. 채점/첨삭, 도구/자원 현황 분석	<ul style="list-style-type: none"> <li>• 쓰기 채점·첨삭에 필요한 국가·민간에서 구축한 언어 자원 특징 및 활용처 파악</li> <li>• 인공지능 언어 능력 평가 과제 및 언어모델 평가 관련 현황 분석(GLUE 등)</li> </ul>
3. 피드백 방안 연구	<ul style="list-style-type: none"> <li>• 진단지표별 피드백 방안 연구(문장/어휘 수준, 문장 수준 등)</li> <li>• 고품질 언어 자원 구축을 위한 효율화 방안/ 언어 자원 배포 활용성 증대</li> </ul>
4. 말뭉치 구축 중장기 계획안 수립	<ul style="list-style-type: none"> <li>• 자원위원회를 정례적으로 운영하며, 학계·산업계 의견 수렴</li> <li>• 자가 진단·첨삭 지원 등에 필요한 한국어 말뭉치 구축 중장기 계획안 마련(이용자 특성 고려)</li> </ul>
5. 생성형 AI 활용 방안 검토	<ul style="list-style-type: none"> <li>• 생성형 AI의 글쓰기 진단·평가를 위한 활용 가능성 검토</li> <li>• 한국어 초거대 AI 모델 적용을 통한 피드백 방안의 실증</li> </ul>

3. 기간

2023년 5월 3일 ~ 2023년 12월 10일(총 221일)

4. 참여 기업/학교

(주) 테디썸, 연세대학교, 서울과학기술대학교, (주)튜터러스랩스

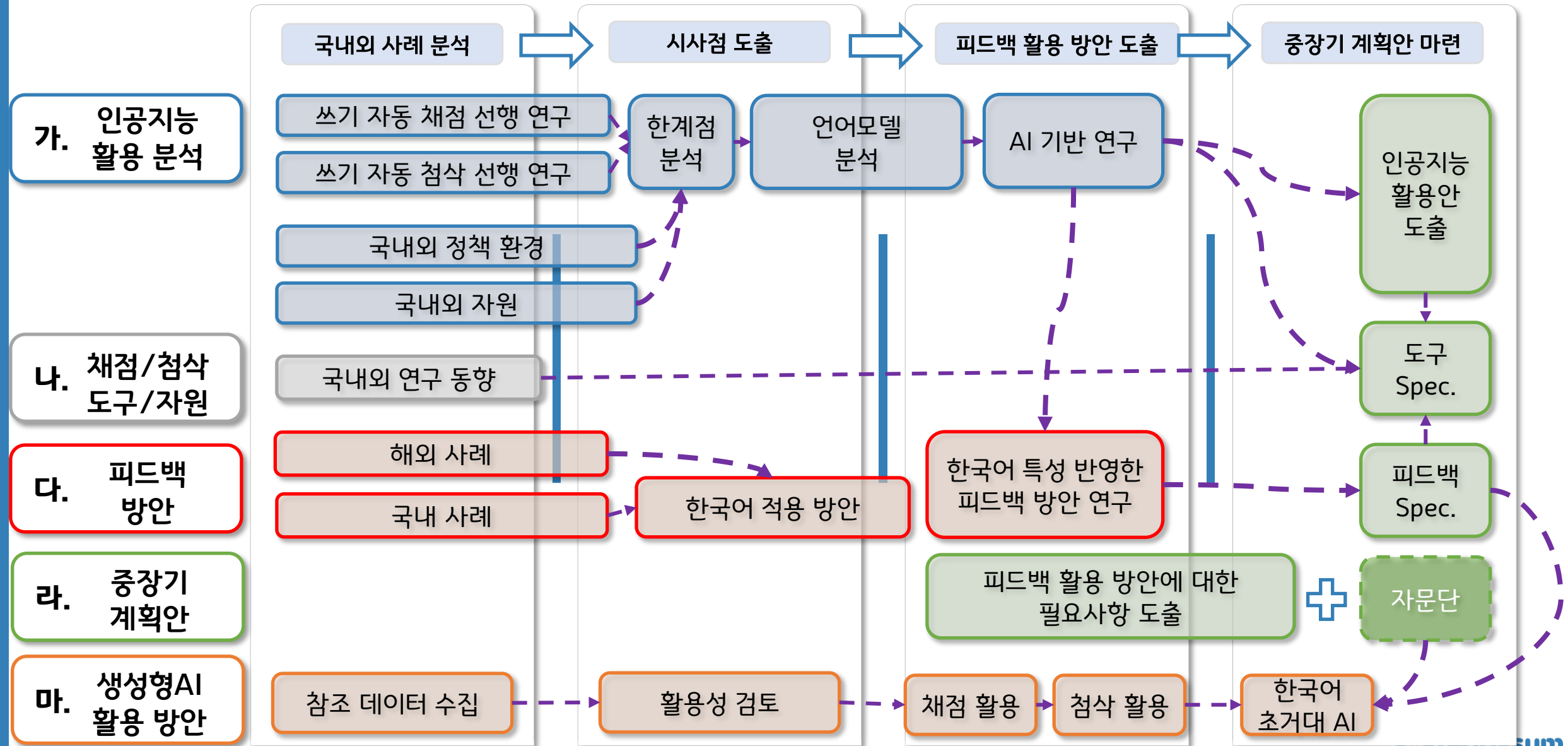
# 1. 사업 개요

## 참여 인력

구분	성명	소속
연구책임자	함영균	테디섬
공동연구원	임경태	서울과학기술대학교
공동연구원	장한열	서울과학기술대학교
공동연구원	김한샘	연세대학교
공동연구원	윤영민	연세대학교
공동연구원	이진	연세대학교
공동연구원	주민재	연세대학교
공동연구원	임준호	튜터러스랩스
공동연구원	박전규	튜터러스랩스
공동연구원	전형배	튜터러스랩스

구분	성명	소속
연구보조원	강예지	연세대학교
연구보조원	정용빈	테디섬
연구보조원	서현빈	테디섬
연구보조원	박재완	테디섬
연구보조원	강조은	연세대학교
연구보조원	김유진	연세대학교
연구보조원	김희재	연세대학교
연구보조원	박서윤	연세대학교
연구보조원	최규리	연세대학교
연구보조원	이재원	연세대학교
연구보조원	정가연	연세대학교
연구보조원	이성우	연세대학교

## 1-1. 사업 추진 전략



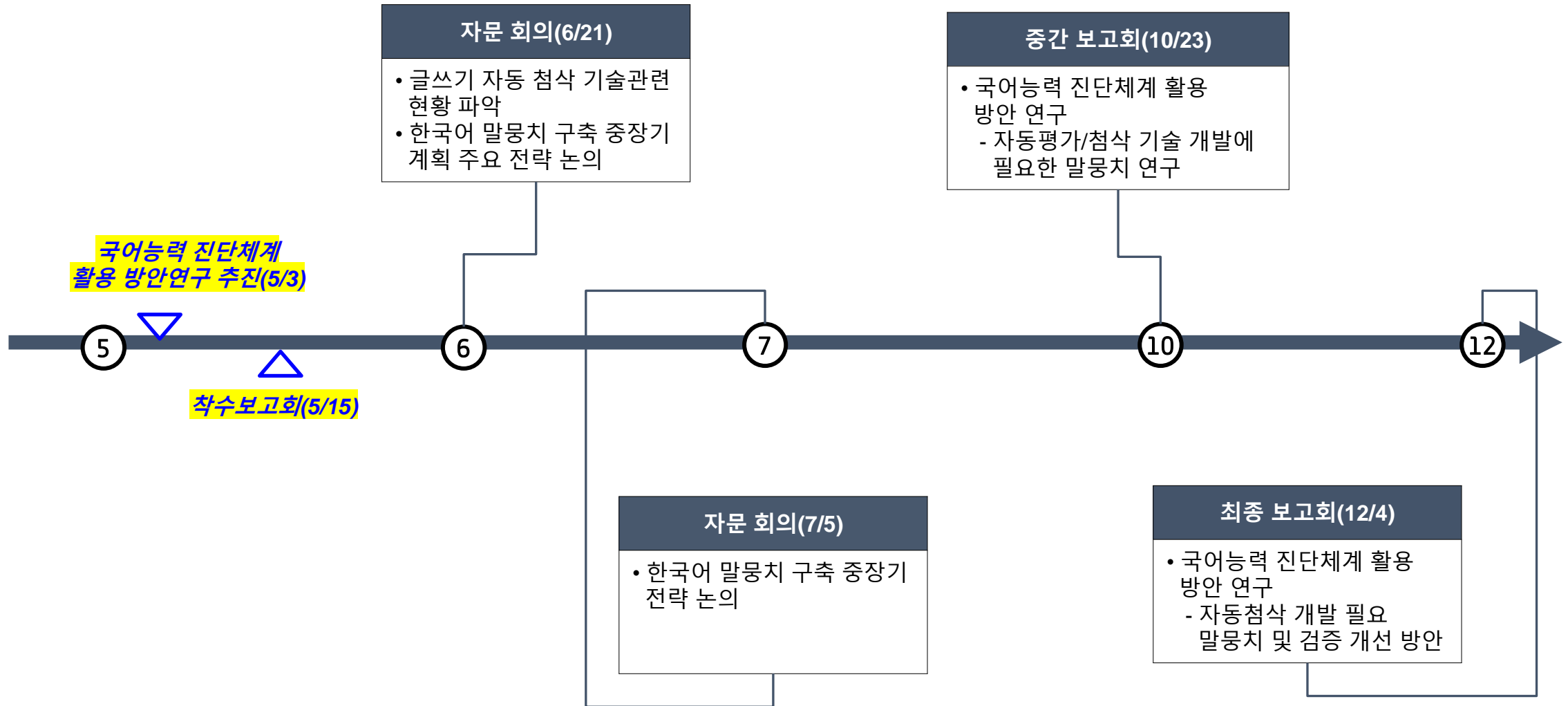
## 1-2. 사업 수행 내용

No.	목적	주요 내용
1	쓰기 자가 진단·첨삭을 위한 인공지능 활용 관련 환경 분석	1) 국내외 소프트웨어(SW) 기술 수준, 언어 모델 분석 - 쓰기 자가 진단·첨삭 관련 선행 연구 분석, 쓰기 자가 진단·첨삭 인공지능 기술 개발을 위한 한국어 공개 언어 모델 분석, 중장기 계획안 마련을 위한 시사점 도출 연구 수행 2) 국내외 자원 분석 - 국외 민간, 비영리 단체, 정부의 언어정보자원 분석 - 모두의 말뭉치 등 국내의 언어정보자원 공유 사례 분석 및 시사점 도출 3) 국내외 정책 환경 - 미국, 중국 등 선도국에 대한 정책 환경을 분석하고, 국내 정책 환경과의 비교를 통한 시사점 도출
2	언어정보자원과 연계한 쓰기 자가 진단·첨삭 지원 도구 개발 기반 현황 진단 및 필요 사항 도출	1) 쓰기 자동 채점 연구 분석 - 한국어, 일본어, 중국어 사례 분석을 통한 시사점 도출 2) 인공지능 언어 능력 평가 과제 및 언어 모델 평가 관련 현황 분석 - GLUE, SuperGLUE 등의 영어 벤치마크 - CLUE(Chinese Language Understanding Evaluation)와 CUGE(Chinese Language Understanding and Generation Evaluation)에 대한 분석 및 시사점 도출
3	진단지표별 피드백 방안 및 언어정보자원 구축 효율화 방안 연구	1) 쓰기 자가 진단·첨삭에 필요한 언어정보자원 - 국외의 쓰기 자가 진단 및 첨삭을 위한 데이터 세트 분석 2) 쓰기 자가 진단·첨삭 사례 분석 3) 영어의 글쓰기 피드백 방안의 사전 조사를 통해 한국어 특수성에 맞는 피드백 방안 검토
4	진단지표 활용 자가 진단·쓰기 첨삭 지원 중장기 계획안 마련	1) 인공지능 활용 국어능력 진단체계 개발 기초 연구(국립국어원, 2021) 분석 2) 진단지표 활용 자가 진단·쓰기 첨삭 시스템 계획안 마련
5	자가 진단·쓰기 첨삭 지원을 위한 생성형 AI 기술 활용 계획안	1) 진단지표 활용 자가 진단·쓰기 첨삭 지원 서비스 개발에 활용 가능한 학습 데이터 증강, 점수 채점 지원, 자동 첨삭 지원의 세 가지 관점에서 생성형 AI 기술 수준 및 적용 가능성 검토 및 활용 계획 수립

## 1-3. 사업 수행 일정

단계	활동	착수 이후 일정																											
		1M				2M				3M				4M				5M				6M				7M			
마일스톤 (착수/중간/종료)																													
자문단 구성																													
쓰기 자가 진단·첨삭을 위한 인공지능 활용 관련 환경 분석	국내외 소프트웨어(SW) 및 기술 수준, 언어 모델 분석																												
	국내외 자원/정책 환경 분석																												
언어정보자원과 연계한 쓰기 자가 진단·첨삭 지원 도구 개발 기반 현황 진단 및 필요 사항 도출	언어정보자원 조사																												
	인공지능 언어 능력 평가 현황분석																												
현황 진단 및 필요 사항 도출																													
진단지표별 피드백 방안 및 언어정보자원 구축 효율화 방안 연구	국어능력 진단지표별 피드백 방안 연구																												
	절차 효율화 방안 연구																												
	국어원 배포 방안 연구																												
진단지표 활용 자가 진단·쓰기 첨삭 지원 중장기 계획안 마련	해외 사례 현지화 방안																												
	자문위원회 정례 운영, 중장기 로드맵 마련																												
	쓰기 자가 진단·첨삭 지원 한국어 말뭉치 구축 중장기 계획안 마련																												
	자가 진단·쓰기 첨삭 지원을 위한 생성형 AI 기술 활용 계획안																												

## 1-4. 연구 이력





## 1-5. 사업 추진 체계

### 인공지능 기술

- 가. 쓰기 자가 진단·첨삭을 위한 인공지능 활용 관련 환경 분석
- 다. 진단지표별 피드백 방안 및 언어정보자원 구축 효율화 방안 연구
- 마. 자가 진단·쓰기 첨삭 지원을 위한 생성형AI 기술 활용 계획안

### 진단체계 활용

- 나. 쓰기 자가 진단·첨삭을 위한 인공지능 활용 관련 환경 분석
- 라. 진단지표 활용 자가 진단·쓰기 첨삭 지원 중장기 계획안 마련

### 연구 범위



(주)테디섬(함영균)

- 과제 총괄
- 피드백 방안 연구
- 글쓰기 첨삭 모델 연구



튜터러스랩스(임준호)

- 국내외소프트웨어(SW) 기술 조사
- 언어 모델 기술 조사
- 생성형AI 적용 연구



연세대학교 산단(김한샘)

- 국내외 자원/정책 조사
- 쓰기 평가 방법론 연구
- 피드백 국내외 사례 연구



과기대 산단(임경태)

- 글쓰기 평가 모델 연구
- 딥러닝 기반 글쓰기 평가 모델 연구 및 필요 사항 도출

### 산/학 컨소시엄

## 1-6. 기대 효과

### 국민의 국어 능력 향상을 위한 인공지능 활용 쓰기 자동평가·첨삭 기능 도출 및 국어능력 진단체계 발전



자가 진단/쓰기 첨삭 지원을 위한  
생성형 AI 기술 활용 계획안 마련

- 생성형 AI 기술 활용한 학습 데이터 증강,  
점수 채점 지원, 자동첨삭 지원 등 적용 가능성 검토

## 2. 환경 분석

### 1. 사업 개요

### 2. 환경 분석

#### 2-1. 초거대 언어모델 개발 및 사업화 동향

##### 2-1-1. 생성형 AI 주요 비교

##### 2-1-2. 언어모델별 한국어 문법 오류 비교

#### 2-2. 국내 인공지능 학습용 데이터 확보 동향

#### 2-3. 국외 언어 자원 비교 분석

##### 2-3-1. 국외 주요 언어 자원 현황

#### 2-4. 국외 선행 연구

##### 2-4-1. 선행 연구 주요 내용

##### 2-4-2. 시사점

#### 2-5. 주요 언어 모델 및 특징

### 3. 글쓰기 자동평가 및 자동첨삭 방안

### 4. 한국어 말뭉치 구축 증장기 계획안

## 2-1. 환경 분석\_국내외 초거대 언어모델 개발 및 사업화 동향

- 1) 초거대 언어모델(Large Language Model, LLM)의 기초 모델(Foundation Models) 구축 시 대규모 자본이 필요함에 따라 글로벌 빅테크 기업 주도로 시장 형성





		
<p>챗GPT(마이크로소프트)</p>	<p>Bard(구글)</p>	<p>LLaMA(메타, 구 페이스북)</p>

## 2-1. 환경 분석\_국내외 초거대 언어모델 개발 및 사업화 동향

### 2) 최근 네이버 등 국내 대기업 자체 개발 초거대 언어모델 공개

	<ul style="list-style-type: none"> <li>• 하이퍼클로바X('23.8월 공개)</li> <li>• 한국형 특화형 초거대 AI</li> <li>• 검색·쇼핑·광고 등 네이버에서 제공하는 서비스와 결합</li> </ul>		<ul style="list-style-type: none"> <li>• 자체 개발 언어모델 Ko-GPT 2.0(공개 시기 미정) : 전문성과 개인화에 초점, 카카오톡에 적용 예정</li> <li>• 이미지 생성 모델 Karlo 2.0</li> <li>• 3억 장 규모의 이미지-텍스트 학습('23.7월 출시)</li> </ul>
	<ul style="list-style-type: none"> <li>• 생성형 AI 모델 엑사원 2.0('23.7월)</li> <li>• 연구원·개발자 등 전문가가 사용할 수 있는 멀티모달 AI</li> </ul>		<ul style="list-style-type: none"> <li>• 믿:음(MIDM)('23.10월 공개)</li> <li>• 인프라 구축부터 응용서비스까지 모두 지원하는 AI 풀스택</li> </ul>
	<ul style="list-style-type: none"> <li>• 인공지능(AI) 비서 '에이닷'(A.) (베타 버전 '23.5월 출시)</li> <li>• 글로벌 통신사들과 'AI 얼라이언스'를 결성, '텔코 AI 플랫폼'을 공동 개발해 AI와 통신 서비스를 접목</li> </ul>		<ul style="list-style-type: none"> <li>• 게임 회사 최초로 거대언어모델 바르코 공개('23.8월)</li> <li>• 아마존 웹서비스(AWS) 마켓플레이스를 통해 판매</li> </ul>

## 2-1-1. 환경 분석\_생성형 AI 주요 비교

업체명	 OpenAI	 Meta		
인공지능 언어모델	GPT-4	LLaMA 2	Palm 2	HyperCLOVA X
매개변수	1,750억 개	650억 개	3,400억 개	2,040억 개
데이터 양 (토큰 개수, 추정)	7,000억 개	2조 개	3조 6,000억 개	5,600억 개
지원 언어	다국어 모델 (26개 언어)	다국어 모델 (20개 언어)	다국어 모델 (100개 이상 언어)	한국어 특화 모델
주요 특징	<ul style="list-style-type: none"> <li>• 다양한 출처의 웹 데이터 학습을 통해 창조적 글쓰기 능력 보유</li> <li>• 다양한 지시문(인스트럭션) 학습을 통해 코딩, pdf 해독 등 다양하고 복잡한 과업 수행 능력 보유</li> <li>• 멀티모달 정보 이행 능력 우수</li> <li>• 높은 비용, 유료, 느린 속도</li> </ul>	<ul style="list-style-type: none"> <li>• 다양한 출처의 웹 데이터 학습을 통해 창조적 글쓰기 능력 보유</li> <li>• 매개변수 규모가 작아 반응 속도가 빠르고, 필요 전력량이 적음</li> <li>• 공격적이고 해로운 콘텐츠에 대한 보호에 충실</li> <li>• 오픈소스 쉬운 사용</li> <li>• 상대적으로 적은 데이터 및 언어 지원</li> </ul>	<ul style="list-style-type: none"> <li>• 타 언어모델 대비 창조적 문장 생성 능력 저조</li> <li>• 타 언어모델 대비 수학, 논리, 추론, 과학 등 *STEM 분야의 전문 데이터를 많이 학습 * STEM : Science, Technology, Engineering, Mathematics</li> </ul>	<ul style="list-style-type: none"> <li>• 국내 기업 최초 자체 개발 인공지능 모델</li> <li>• 한국어 특화 : 챗GPT 대비 6,500배 많은 한국어 데이터 학습 (뉴스, 블로그 외)</li> <li>• 기업 대 기업(B2B) 특화 서비스 지향 <ul style="list-style-type: none"> <li>- 클로바X : 대화형 AI 서비스</li> <li>- 큐:(Cue:) : 검색 특화 서비스</li> </ul> </li> </ul>

## 2-1-2. 환경 분석\_언어모델별 한국어 문법 오류 비교 결과

챗GPT 대비 하이퍼클로바X의 한국어 문법 오류에 대한 교정 성능이 낮은 것으로 판단되며, 두 언어모델 모두 통사의 피동/부정/어순, 답화의 접속 오류에 대해 교정이 미흡함.

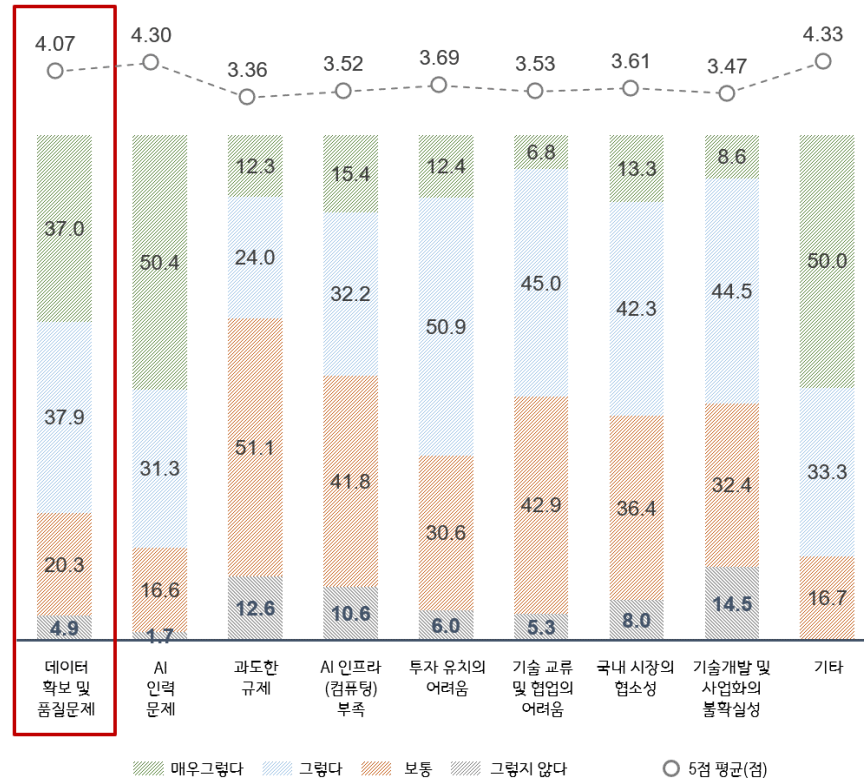
※ 비교 시점: '23년 9월 24일 기준

구분		챗GPT 3.5	하이퍼클로바 X
발음	음소	○	○
	음절	-	○
	음운 규칙	○	○
	원어식 발음	○	○
	중간발음(변이음 포함)	○	○
형태	단어형성(합성법)	○	○
	단어형성(파생법)	○	○
	굴절(곡용)	○	○
	굴절(활용)	○	○
	품사	○	X
통사	높임	○	○
	시제	○	○
	사동	○	○
	피동	X	X
	부정	X	X
답화	어순	-	-
	지시	○	X
	접속	X	X
	답화표지	○	X
	구어/문어 오류	○	X

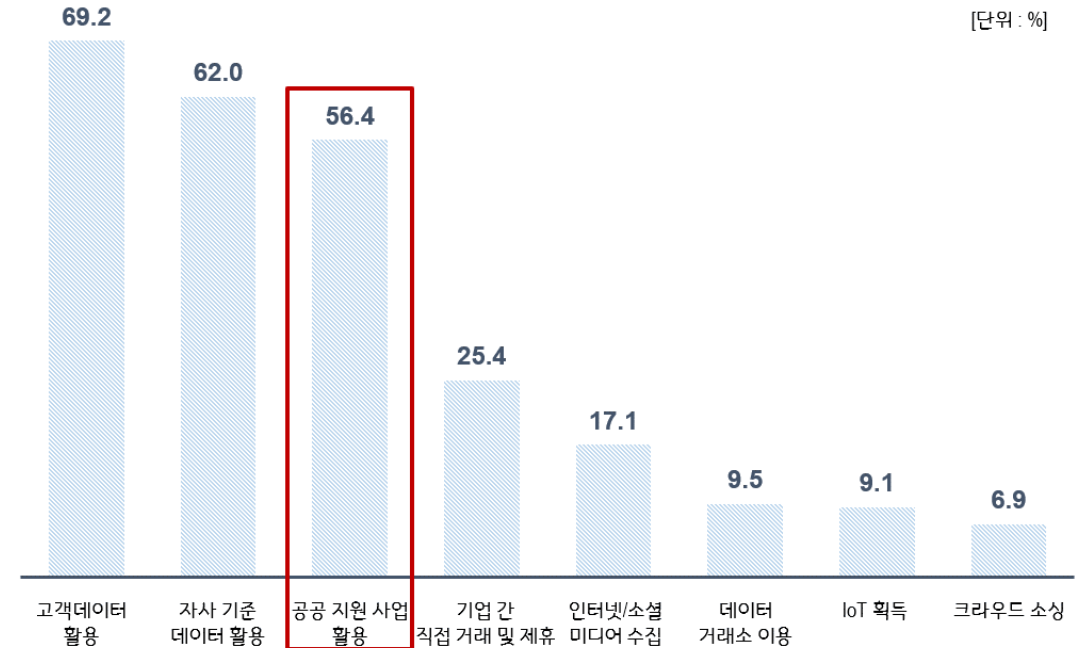
## 2-2. 환경 분석\_국내 인공지능 학습용 데이터 확보 동향

국내 AI 산업계는 소규모 기업 비중이 높고, 인공지능 학습용 데이터 확보 및 품질 문제에 대해 애로사항을 가지고 있으며, 공공지원 사업에 대다수 의존

<AI 사업 운영상 느끼는 애로 사항 비교>



<인공지능 학습용 데이터 확보 방식>



[단위 : %]



## 2-3. 환경 분석\_국외 언어 자원 비교 분석

### 영어권

- 1) 다양한 지역의 영어 분석을 위한 말뭉치 구축이 특징적임.
  - 캐나다, 영국, 미국, 필리핀 등 다양한 지역의 영어 데이터를 수집하여 구축함.
- 2) 통시적 말뭉치 구축이 특징적임.
  - 시대적 변화에 따른 영어를 관찰하기 위해 통시적 말뭉치를 구축함.
- 3) 다른 언어권에 비해 언어 자원이 풍부하여 구축 규모가 큰 편임.
- 4) 대규모 주석 말뭉치 구축
  - 양질의 주석된 말뭉치를 자유롭게 사용 가능한 형태로 제공함.
  - 클라우드소싱과 같은 다양한 방식으로 구축함.

시장 주도(기업/대학/연구소)

### 유럽어권

- 1) 유럽 연합 등의 연합체를 중심으로 한 범국가적 말뭉치가 존재함
  - 광범위한 주제를 포괄하는 다양한 문서 유형이 존재함.
- 2) 지리적 특성으로 인해 국가, 이주 노동자 언어 등에 대한 다국어 연구 활발함.
  - 외국인 노동자와 유럽어 모어 화자 간의 대화 전사(ESF) 데이터베이스
  - 유럽어 간 병렬 말뭉치 구축(ParaCrawl, EU 23개 언어)
- 3) 특허, 언어습득 등 다양한 도메인을 접목한 말뭉치 연구 진행 중
- 4) 개별 유럽어의 언어적 특성을 위한 말뭉치 존재
  - 연도별 구축: 불가리아어 / 지역별 구축: 포르투갈어-수도 리스본

정부+시장 주도(EU 의회/재단, 연구소)

### 중국어권

- 1) 광둥어와 같은 방언과 영어, 한국어 등 다른 나라의 언어에 대한 병렬 말뭉치 구축
  - 중국어 언어 형태(간체, 번체)에 따른 언어모델 성능 향상과 중국 지역 방언에 강건한 모델을 만들기 위한 목적 및 다양한 언어와의 병렬 말뭉치 등을 위한 데이터 구축
- 2) 대규모 말뭉치 구축을 위한 해외 데이터 세트 벤치마킹
  - 데이터 벤치마킹은 주로 CLUE 하위 과제 위주로 구축함.
  - 자연어처리 하위 분야의 성능 향상을 위해 기구축 데이터의 성능 저하 요인을 분석하여 구축함(주로 학계에서 연구 목적을 가지고 진행 예상).
- 3) 서비스 성능 향상 및 실용적 목적의 언어 자원 구축
  - 대화 시스템뿐 아니라 의료, 스포츠 등 특정 도메인에서의 활용에 맞춘 언어 자원 구축 중
  - 대화, 질의응답(Q/A) 데이터를 수집 및 구축하여 자연스러운 응답을 위한 방법을 모색 중

정부 주도

### 일본어권

- 1) 정부와 대학 연구 기관 주도의 말뭉치 구축 진행
  - NINJAL(National Institute for Japanese Language and Linguistics) 주도적으로 언어 자원이 구축되고 있음.
  - 일본어 말뭉치와 관련 연구는 타 언어에 비해서 비교적 폐쇄적임(김민준 외, 2015).
  - 일본어를 알아야 받을 수 있는 경우가 많음.
- 2) 교육 도메인 언어 자원 구축이 특징적임.
  - 학습자 말뭉치 중심 구축 중
  - 예) FLUency extended TMU Evaluation Corpus for Japanese learners(FLUTEC)

정부+시장 주도(대학/국립연구소)

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	ALEC (Advanced Learner Corpus)	Uppsala University	130만 개 단어, 146개 파일	<ul style="list-style-type: none"> <li>• 스웨덴에서 2004~2013년에 구축됨.</li> <li>• 대학교 3~5년에 재학 중인 영문학과 학생들이 작성한 텍스트임.</li> </ul>
영어	ARCHER (A Representative Corpus of Historical English Registers)	14개 대학의 컨소시엄	1,789,309개 단어, 955개 텍스트	<ul style="list-style-type: none"> <li>• 1600~1999년 동안의 영국식 영어와 미국식 영어의 다중 장르 역사적 말뭉치.</li> <li>• 다양한 문어 및 음성 기반 영어 사용역(레지스터, register)에서 언어 변이를 분석하기 위한 자료로 설계됨.</li> </ul>
영어	BASE (British Academic Spoken English Corpus)	개인 연구자 (Hilary Nesi, Paul Thompson)	1,644,942개 토큰	<ul style="list-style-type: none"> <li>• 학술 어휘의 빈도와 범위, 개별 단어 및 다중 단어 단위의 의미 및 사용, 학술 세미나·강의의 구조, 학술 세미나·강의의 속도, 밀도 및 전달 스타일, 억양의 담화 기능, 발화 말차례(turn-taking) 및 주제 선택을 포함한 상호 작용 패턴, 시각 및 청각 자극의 상호 작용, 생각과 태도의 표현에 대한 연구 목적임.</li> <li>• 강의와 세미나는 TEI 가이드라인에 따라 고안된 시스템을 사용하여 기록 및 주석됨. 현재 Oxford Text Archive에 보관되어 있으며, 동의서를 작성하면 무료로 접근 가능함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	B-BROWN	BROWN 말뭉치 확장	100만 개 단어, 500개 텍스트	<ul style="list-style-type: none"> <li>• 영국에서 구축되었으며 말뭉치는 주석이 없는 원시 말뭉치 버전과 품사 주석이 부착되지 않은 XML 버전이 있음.</li> <li>• B-BROWN은 BROWN 말뭉치의 연대기적 범위를 확장하여 20세기 전반부인 1928~1934년 미국 영어 문어체 연구의 기반을 제공함. 텍스트 수집 및 디지털화는 Stefanie Dose, Nina Lötsch, Melanie Röthlisberger와 Kerry O'Reilly가 맡았으며, 형태소 분석 및 파싱은 Gerold Schneider가 맡음.</li> </ul>
영어	BAWE (British Academic Written English Corpus)	University of Warwick	6,506,995개 단어, 2,761개 텍스트	<ul style="list-style-type: none"> <li>• 영국에서 2000년~2007년에 구축된 문어 말뭉치. 21세기 초 대학 수준의 학생 작문 모음으로 3,000개 미만의 우수한 학생 과제가 포함됨.</li> <li>• 4개의 학문 분야(예술 및 인문학, 사회 과학, 생명 과학 및 물리 과학), 4개의 연구 수준(학부 및 석사 수준)에 걸쳐 고르게 분포되어 있으며, 총 30개의 주요 분야가 표시되어 있음.</li> <li>• 13개의 광범위한 장르군으로 분류된 2,858개의 텍스트가 포함되어 있음.</li> <li>• 장르는 사례 연구, 비평, 공감 쓰기, 에세이, 실습 설명, 문헌 조사 및 방법론 설명, 내러티브 재검토 문제, 질문 및 제안, 연구 보고서 등이 포함됨. 작문의 경우 L1 영어 사용자가 1,953개 파일을 작성하였으며, 나머지의 경우 영어를 능숙하게 사용하는 사용자가 작성함.</li> <li>• 현재 말뭉치는 “<a href="http://ota.ahds.ac.uk/headers/2539.xml">http://ota.ahds.ac.uk/headers/2539.xml</a>”에서 내려받기가 가능하며, 시각화 도구(The Wordtree)도 함께 제공하고 있음.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	CASE (Corpus of Academic Spoken English)	Stefan Diemer(Trier University of Applied Sciences and Saarland University), Marie-Louise Brunner(Trier University of Applied Sciences and Saarland University), Caroline Collet(Trier University of Applied Sciences and Saarland University), Selina Schmidt (Birmingham City University), Sebastian Malinowski(Karlstad University)	200~300개의 Skype 대화 텍스트 파일	<ul style="list-style-type: none"> <li>• 미국에서 구축한 구어 말뭉치. Skype 대화의 경우 다양한 유럽 국가(독일, 불가리아, 이탈리아, 스페인, 핀란드, 벨기에, 프랑스)의 Trier University of Applied Sciences International의 Lingua Franca(ELF)로서의 학생 영어 사용자와 영국·미국의 원어민 구성원 간의 대화 모음.</li> <li>• 말뭉치는 200시간 분량의 Skype 대화(비디오, 오디오 데이터)를 포함함. 대화의 경우 첫 만남에서 이루어지는 자연스러운 대화로 비공식적인 상황에서 1:1의 환경의 대화로 30~60분간 지속됨.</li> <li>• 대화 시작을 용이하게 하기 위해 학문적·문화적인 환경과 관련한 주제 프롬프트가 제공됨. 현재 공개되지 않았으나 공개 예정임.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	CC (The Coruña Corpus of English Scientific Writing)	MUSTE Research Group	각 말뭉치별 400,000개 단어로 구성	<ul style="list-style-type: none"> <li>• 스페인에서 구축됨. 도메인 또는 학문 분야에 따라 하위 말뭉치로 구분되는 전문 말뭉치. 1700~1900년대 자료. 영어로 작성된 과학 텍스트의 10,000단어 표본을 포함하도록 설계함. 모든 표본과 메타 데이터 파일은 TEI에 따라 XML로 편집됨.</li> <li>• <a href="https://ruc.udc.es/dspace/handle/2183/21846">https://ruc.udc.es/dspace/handle/2183/21846</a>에 공개됨.</li> <li>• 각 하위 말뭉치별로 관련 안내서가 포함되어 있음. 하위 말뭉치는 CEPhiT (Corpus of English Philosophy Texts), CELiST(Corpus of English Life Sciences Texts), CHET(Corpus of Historical English Texts), CEChET (Corpus of English Chemistry Texts), CETeL(Corpus of English Texts on Languages)로 구성됨.</li> </ul>
영어	CEEM (Corpus of Early English Medical Writing)	University of Helsinki	3.75만 개 단어, 750개의 파일	<ul style="list-style-type: none"> <li>• 도메인 특화(domain-specific)된 375~1800년 자료로 구축된 말뭉치로, 일반 대중을 위한 실용적인 건강 지침 등을 포함한 의학 저술의 대표적인 표본을 제공함.</li> <li>• 초기~현대 임상 의학을 포함한 영어로 된 의학 저술의 전체 역사를 다룸. 3개의 통시적으로 분할된 하위 말뭉치로 구성됨. Middle English MedicalText(MEMT), Early Modern English Medical Text(EMEMT), Late Modern English Medical Text(LMEMT).</li> <li>• 해당 기간 동안 텍스트의 범주화 수준이 높아짐에 따라 실현된 의학 저술의 언어 외적 현실을 반영함.</li> <li>• CD-ROM로 제작되어 헬싱키 대학교 내 언어 연구소(VARIENG)에서만 사용 가능하거나, 관리자의 사전 승인을 얻을 경우 사용 가능함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	COLOMOBAENG (A Corpus of Late Modern British and American English Prose)	개인 연구자 (Teresa Fanego)	1,170,000개 단어, 129개 텍스트	<ul style="list-style-type: none"> <li>• 111명의 저자가 작성한 인쇄 및 전자 텍스트로, Century of Prose 말뭉치, 프로젝트 구텐베르크(Project Gutenberg), 펜실베니아 대학교의 Online Books Page, 버지니아 대학교의 전자 텍스트, 채드윅 힐리(Chadwyck-Healey)의 초기 미국 소설, 18세기 소설 및 19세기 컬렉션이 포함됨.</li> <li>• 영어 역사의 문법적 변화, 현대 영어의 변화, 일반적인 영어의 다양성에 대한 말뭉치 기반 연구를 위한 목적으로 구축됨 .</li> <li>• 교육 및 연구용의 경우 무료로 사용 가능.</li> <li>• 스페인 과학혁신부(Spanish Ministry for Science and Innovation)와 유럽지역개발기금(European Regional Development fund) 지원을 받음.</li> </ul>
영어	The Crown Corpus	Beijing Foreign Studies University	1만 개 단어, 500개의 파일	<ul style="list-style-type: none"> <li>• 2009년에 공개된 BROWN 말뭉치 계열임. 2010~2012년 구축됨.</li> <li>• 147개 대학교의 영어 교사와 107개 이상의 중국 대학교 박사 과정 학생들로부터 클라우드소싱된 말뭉치임.</li> <li>• Crown, CLOB의 모든 글의 저작권은 원저작권자에게 있음.</li> <li>• 중국 국가사회과학기금위원회(National Social Sciences Foundation of China)로부터 지원을 받음.</li> </ul>
영어	EFCAMDAT - EF-Cambridge Open Language Database [Learner corpus]	Department of Theoretical and Applied Linguistics at the University of Cambridge	83만 개 단어, 1만 개의 과제 텍스트	<ul style="list-style-type: none"> <li>• 전 세계 1,000명의 영어 성인 학습자의 작문 표본을 포함함. 제2언어 연구 및 교수를 목적으로 구축됨. 온라인에 공개되어 있음.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	QHC- QuakerHistoricalCorpus(영국/영어)	Centre for Postgraduate Quaker Research at Woodbrooke Quaker Study Centre, Birmingham, UK	722,370 단어, 173개의 텍스트	<ul style="list-style-type: none"> <li>• 문어 말뭉치. 2008~2015년 구축됨.</li> <li>• 프렌즈 하우스 도서관에 사본 형태로만 남아 있는 퀘이커교의 책 원본들(저작권 만료)을 전사함.</li> <li>• 검색 가능하고 기계로 읽을 수 있는(searchable, machine-readable) 자료를 제공할겠다는 목표로 구축함.</li> <li>• 퀘이커교의 언어와 말뭉치에 관심이 있는 말뭉치 언어학자들을 위해 자유롭게 다운로드 및 마크업(markup)할 수 있도록 배포함.</li> </ul>
영어	MICUSP - Michigan Corpus of Upper-level Student Papers	U-M English Language Institute (ELI)	260만 개 단어, 830개의 논문	<ul style="list-style-type: none"> <li>• 미국에서 2002~2009년 구축된 문어 말뭉치.</li> <li>• 말뭉치 사용자가 연구 논문이나 보고서와 같은 특정 유형의 자료를 검색할 수 있도록 구축함.</li> <li>• 미시간 대학교에서 A학점을 받은 논문들을 모아 구축한 논문 말뭉치로 인문학 및 예술, 사회과학, 생물학 및 건강 과학, 물리학 등에 대한 논문들을 수집함.</li> <li>• 논문 종류(paper type), 학생 수준(student levels), 모어 사용 여부(nativeness), 학문 분야(disциплиnes) 등의 태그를 달아 검색해 볼 수 있도록 검색 시스템을 제공하고 있음.</li> <li>• 연구와 교육 목적으로 자유롭게 사용 가능함(상업적 용도로 사용 시 라이선스에 대한 별도 문의 필요).</li> <li>• 상위 학부 과정 및 대학원 진학을 생각하는 사람들이 주요 대상. 학술적 글쓰기의 규칙 익히기, 특정 단어나 구가 사용되는 방법 익히기 등의 용도로 사용. ‘양질의 글’을 글쓰기 교육에 활용한 사례로 평가할 수 있음.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	MICASE - Michigan Corpus of Academic Spoken English	U-M English Language Institute (ELI)	180만 개 단어, 200시간	<ul style="list-style-type: none"> <li>• 미국에서 구축된 구어 말뭉치로 1997~2001년 구축됨.</li> <li>• 학술적(Academic) 언어 연구를 목적으로 구축된 것으로 학내 공간에서 진행된 강의, 토론, 세미나, 질문 등을 녹음한 음성 데이터를 전사 및 교정 작업 진행함.</li> <li>• 맞춤형 검색 엔진을 포함한 분석 도구를 함께 제공함.</li> <li>• 연구와 교육 목적으로 자유롭게 사용 가능함(상업적 용도로 사용 시 라이선스에 대한 별도 문의 필요).</li> </ul>
영어	OANC (Open American National Corpus)	US National Science Foundation, the ANC Consortium, the TalkBank project, and the Department of Chinese, Translation, and Linguistics at the City University of Hong Kong	1,500만 개 단어	<ul style="list-style-type: none"> <li>• 미국에서 구축한 문어 및 구어 말뭉치로 2006년부터 구축해 오고 있는 말뭉치임.</li> <li>• 자연어 처리 연구를 지원할 수 있는 광범위한 장르, 대규모의 주석 말뭉치 구축을 목적으로 만들어진 국가 주도 구축 말뭉치. 미국 국립 말뭉치(AMERICAN NATIONAL CORPUS, ANC) 홈페이지를 통해 언어 자원 기부 요청함.</li> <li>• 2종의 구어 말뭉치와 10종의 문어 말뭉치를 기부 받아 제작함.</li> <li>• 미국 영어를 모어로 사용하는 사람들에게 ‘에세이’, ‘블로그’, ‘시’, ‘소설’, ‘전자우편(이메일)’ 등을 포함한 다양한 종류의 문서를 제공하도록 요청.</li> <li>• 섹션, 장, 문장 경계, Penn 태그세트, 명사-동사, 개체명 등에 대한 주석 등 기본적인 주석이 되어 있는 상태임. ANC에서 확보하고 있는 데이터는 더 많으나(4천 만 단어 이상) 자금 문제로 일부만 공개된 상태임.</li> <li>• 무료로 내려받아 상업적 개발을 포함한 연구 개발에 사용 가능함(저작권 완전 개방) .</li> <li>• 엔-그램(n-gram) 등 검색 엔진을 활용한 말뭉치 기본 활용 도구도 함께 제공함. 연구 목적의 기부 방식으로 저작권 확보 말뭉치를 국가 주도로 운영한다는 점이 특징적임.</li> </ul>



## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	MASC (Manually Annotated Sub- Corpus )	개인 연구자 (Nancy Ide 외)	1,500만 개 단어	<ul style="list-style-type: none"> <li>• 미국에서 구축되었으며 기존 OANC에 추가적인 주석 진행함.</li> <li>• 언어 주석이 달린 공유 가능하고 재사용 가능한 자원에 대한 필요에 따라 여러 장르의 텍스트와 다양한 수준의 수동 주석과 수동으로 검증된 주석을 포함한 하위 말뭉치 개발함.</li> <li>• 양질의 주석 확보 및 비용 문제 해결을 위해 데이터의 완전 개방 방식을 허용하되, 연구자들이 자유롭게 주석을 단 파일을 다시 제출하도록 하는 방식을 활용하는 것이 특징적임.</li> </ul>
영어	BNC (British National Corpus)	Lancaster University	1억 개 단어	<ul style="list-style-type: none"> <li>• 영국에서 구축된 균형 말뭉치임.</li> <li>• 20세기 후반 영국 영어를 다루는 다양한 출처의 문어 및 구어 표본으로 구성됨.</li> <li>• 다양한 언어 현상을 연구할 수 있도록 지원하기 위해 정부에서 구축 비용의 50%를 지원함.</li> </ul>
영어	SC(Salamanca Corpus)	스페인 정부	6,115,267개 단어	<ul style="list-style-type: none"> <li>• 방언 번역 유형에 해당하는 텍스트로 구성된 말뭉치임.</li> <li>• 1500년대부터 20세기까지 종적, 통시적 연구가 가능하게 하기 위해 연대순으로 배열하여 구축함. 1500-1700 / 1700-1800 / 1800-1950의 세 가지 광범위한 기간을 설정함.</li> <li>• 시간별 특정 기능, 특정 변종의 표현에 사용되는 방언 철자법, 구문 표시, 형태학적 패턴 등의 비교 연구가 가능함.</li> <li>• 스페인 정부(the Spanish Ministry of Education and Science)에서 지원함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	SCEPA (Small Corpus of English Political Apologies)	Lviv Polytechnic National University.	22,538개 단어	<ul style="list-style-type: none"> <li>22,538단어로 구성된 정치적 사과(apologies)(영국, 미국, 캐나다)가 포함된 말뭉치로 사과에 사용되는 의사소통 전략 분석이 가능함.</li> <li>1950~2017년 자료 구축. 사과에 대한 세부 정보인 날짜, 저자, 사과 이름, 저자 성별, 저자 국가, 변명(excuse) 출처 링크, 사과 이유 등이 포함됨.</li> </ul>
영어	SCONE (Seville Corpus of Northern English)	Universidad de Sevilla.	70개의 텍스트 파일	<ul style="list-style-type: none"> <li>7~16세기의 북부 영어의 역사에 대한 연구 과정에서 수집 및 분석된 텍스트임.</li> <li>스페인 과학기술부에서 지원함.</li> </ul>
영어	SCOTS (Scottish Corpus of Texts & Speech)	University of Edinburgh	400만 개 단어	<ul style="list-style-type: none"> <li>스코틀랜드에서 구축된 구어(20%)와 문어(80%) 말뭉치임.</li> <li>미국식 영어와 영국식 영어 말뭉치 사이의 스카치 영어의 공백을 채우기 위해 구축 시작함.</li> <li>상당 기간 스코틀랜드에 살았던 사람들의 언어를 임의 추출(sampling)하여 사회언어학적, 언어적 패턴을 사회적 및 인구 통계학적 범주와 일치시키며 구축함.</li> <li>2001~2003년은 공학 및 물리 과학 연구위원회(Engineering and Physical Sciences Research Council, EPSRC)가 지원하였고 2004~2007년은 예술인문학연구회(Arts and Humanities Research Council, AHRC)가 지원하였음.</li> <li>저작권 허가의 이유로 사용가능한 말뭉치가 제한됨.</li> <li>화자/작가 연령, 출생지, 직업, 성별 및 교육 수준 정도와 함께 구축함.</li> </ul>
영어	The TIME Magazine corpus	American University	100만 개 단어, 275,000개 텍스트	<ul style="list-style-type: none"> <li>스코틀랜드에서 구축된 20세기(1920-2000) 통시적 말뭉치임.</li> <li>수십 년 동안의 언어 변화 분석 가능함.</li> <li>상당기간 스코틀랜드에 살았던 사람들의 언어를 임의 추출(sampling)하여 사회언어학적, 언어적 패턴을 사회적 및 인구 통계학적 범주와 일치시키며 구축함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
영어	VOICE (Vienna-Oxford International Corpus of English)	VOICE CLARIAH	1,023,043개 단어, 151개 텍스트 파일	<ul style="list-style-type: none"> <li>• 오스트리아에서 구축된 20세기(1920-2000) 통시적 말뭉치. 국제 공용어로서의 영어(Emerging English as a Lingua Franca, ELF)의 구어 말뭉치 구축을 통한 국제 공용어로서 영어 사용자들의 심층적이고 실험적 근거 및 기반 제공함.</li> <li>• 국제 공용어로서의 영어를 다룬 최초의 코퍼스임.</li> <li>• 대본 없는 자연스러운 상호작용을 포착한 말뭉치로 △도메인적 영역으로는 전문, 교육, 여가, △기능적 부분으로는 정보 교환, △사회적 관계로 참여자 역할, △관계적 측면으로는 미숙함 대 능숙함, 대칭적 대 비대칭적으로 나뉘서 다양한 음성 데이터를 수집함.</li> </ul>
영어	WestLabUSENET (A Reduced Redundancy USENET corpus)	University Of Alberta 연구팀	6,089,697,986개 단어, 22,799,995개 텍스트 파일	<ul style="list-style-type: none"> <li>• 캐나다에서 구축되었으며 2005년 10월부터 2011년 1월까지 수집함.</li> <li>• 언어 모델 구축을 목적으로 구축. 영어가 아닌 단어 및 비공식 단어(non-word)가 포함되어 있는 원시 텍스트로 구성함. 필요에 따라 말뭉치 처리가 필요함.</li> <li>• ‘Amazon Web Services’에서 공개함.</li> <li>• 저자 보호를 위해 인터넷 뉴스 전송 프로토콜(Network News Transfer Protocol, NNPT) 헤더(header)를 제거함.</li> <li>• 텍스트 식명을 위해 전자우편(이메일) 주소는 &lt;EMAILADDRESS&gt;로 다시 표기함. HTTP URLs는 &lt;URL&gt;로 news URLs는 &lt;NEWURLS&gt;로 다시 표기함.</li> </ul>
영어	YCCQA (Yahoo-based Contrastive Corpus of Questions and Answers)	Universiteit Leuven	29,400,000개 단어, 665,000개 텍스트 파일	<ul style="list-style-type: none"> <li>• 벨기에에서 구축된 질문-답변의 대조 말뭉치(Contrastive Corpus of Questions and Answers). 영어, 불어, 독일어, 스페인어 야후(Yahoo) 인터넷 사용자가 남긴 질문-답변을 수집한 말뭉치임.</li> <li>• 네 가지 언어에 대해 동일한 상황에서 생성된 질문-답변으로 구성. 인터넷 게시물인 만큼 비공식적 스타일의 쓰기. 언어 간의 대조 연구 가능함.</li> <li>• 하위 말뭉치는 언어별 연구를 위해 독립적으로도 사용 가능함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
유럽어	Europarl: A Parallel Corpus for Statistical Machine Translation	유럽 의회 (European Parliament)	언어당 약 6천만 개 단어	<ul style="list-style-type: none"> <li>• Europarl: “A Parallel Corpus for Statistical Machine Translation(2005)”.</li> <li>• 1996~2011년 구축된 유럽 의회 회의록을 사용한 병렬 말뭉치임</li> <li>• 통계적 기계 번역 연구를 지원하기 위해 구축함.</li> <li>• 구축 방법 ㉠데이터 접근_웹 크롤링, ㉡문서 정렬, ㉢문장 분할, ㉣표준화, 토큰화, 문장 정렬</li> <li>• 말뭉치에 사용된 언어 쌍에 대해 기계 번역 시스템을 구축함.</li> <li>• 11개의 유럽 언어 [Romanic(French, Italian, Spanish, Portuguese, Romanian), Germanic(English, Dutch, German, Danish, Swedish), Slavik(Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric(Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian)] 포함.</li> </ul>
유럽어	DCEP: Digital Corpus of the European Parliament	유럽 의회 (European Parliament)	140만 개(문서), 13.7억 개(단어)	<ul style="list-style-type: none"> <li>• DCEP: “Digital Corpus of the European Parliament(2014)”</li> <li>• 2001~2012년 자료 수집함.</li> <li>• 언어의 평등한 대우를 보장하려는 유럽 의회의 다국어 정책에 기여함.</li> <li>• 구축 방법: ㉠데이터 접근_유럽의회 내부 데이터베이스에서 직접 다운로드, ㉡문서 정렬, ㉢문장 분할, ㉣표준화, 토큰화, 문장 정렬함.</li> <li>• 기계 번역, 번역 연구, 공동 참조 해결을 위한 주석, 담화 분석, 비교 언어 연구, 문장 또는 단어 정렬 알고리즘 개선, 언어 간 정보 검색에 활용함.</li> <li>• 23개의 유럽 언어 포함. 다국어 문서 정렬 병렬 말뭉치임.</li> <li>• 광범위한 주제를 포괄하는 다양한 문서 유형 존재임.</li> <li>• 유럽 의회 공식 웹 사이트 내의 대부분의 콘텐츠가 포함됨.</li> <li>• 중복을 피하기 위해 Europarl 말뭉치에 이미 존재하는 문서는 제외됨.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
유럽어	Hellenic National Corpus	Institute for Language and Speech Processing (ILSP / "Athena" Research Center)	97,000,000개 단어	<ul style="list-style-type: none"> <li>• 2002~ 현재까지 구축 중인 문어 말뭉치임.</li> <li>• 수집한 자료를 매체, 장르, 주제별로 분류함. 텍스트 서지 정보(저자, 발행인, 발행일 등)에 따라서도 분류함.</li> <li>• 연구 목적일 경우에만 제공.</li> <li>• 현대 그리스어의 실제 사용 양상을 반영하기 위해 구축됨.</li> <li>• 품사 및 형태론적 특성에 대해 자동으로 주석 추가 후 언어학자가 수정함. 결과에 오류가 없는 것이 특징임.</li> </ul>
유럽어	Bulgarian National Corpus	Institute for Bulgarian Language	1.2조 개 단어, 240,000개 텍스트	<ul style="list-style-type: none"> <li>• 2001년 구축된 말뭉치로 20세기 중반(1945년)부터 현재 불가리아어(주로 문어 자료) 자료를 수집함.</li> <li>• 현대 불가리아어와 47개 병렬 언어로 구성됨.</li> <li>• 특정 기준(제목, 저자, 연도, 발행 기간, 출처 등)에 따라 특정 또는 하위 말뭉치 추출 가능함.</li> <li>• 단어, 언어 구성 빈도 등 지원. 말뭉치에 다양한 검색 기능 제공함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
유럽어	The German Reference Corpus(DeReKo)	Institute for the German Language	3,900,000,000개 단어 이상(2010년 기준) : 연간 3억 개 단어씩 추가 구축	<ul style="list-style-type: none"> <li>• 1964년 구축되기 시작한 말뭉치로 소설, 과학, 수많은 신문 텍스트 및 기타 다양한 유형의 텍스트로 구성됨.</li> <li>• 언어 연구를 위한 실증적 토대 제공, 원시 표본(샘플) 제공 등을 목적으로 구축됨.</li> <li>• DeReKo가 말뭉치 내 텍스트를 소유하지 않으며 사용 자격만 부여됨. 라이선스 계약 및 지적 재산 권리로 인해 내려받기는 불가능하고 전문 소프트웨어를 통해서만 접근 가능함. 학문적 사용만 허용되며 직접 또는 간접적인 상업적 사용은 명시적으로 금지됨.</li> <li>• 현대 독일어를 대상으로 하며 세계 최대 근대·현대 전자화된 말뭉치임.</li> </ul>
유럽어	The National Corpus of Polish	구축 기관(4기관): Institute of Computer Science at the Polish Academy of Science	10억 개 단어 규모, 이중 3억 개 단어 균형 말뭉치 구축, 수동 주석 말뭉치 1백만 개 어절 공개 중	<ul style="list-style-type: none"> <li>• 2008년에 구축된 말뭉치로 현대 폴란드어에 대한 과학적 연구 방법을 모색하고자 말뭉치를 구축함.</li> <li>• 고전 문학, 일간지, 전문 정기 간행물 및 저널, 대화 기록, 다양한 단문 및 인터넷 텍스트 포함.</li> <li>• 현대 폴란드어를 대상으로 하며 폴란드어로 된 가장 크고 중요한 말뭉치임.</li> <li>• 말뭉치 내 단어나 구의 일반적인 용법 및 의미, 문법적 기능을 찾을 수 있도록 설계함.</li> </ul>
유럽어	ESF (European Science Foundation Second Language) Database	구축기관:유럽과학재단 (European Science Foundation) & 막스 플랑크 연구소 Max Planck Institute for Psycholinguistics	미상	<ul style="list-style-type: none"> <li>• 1992~1998년 구축된 말뭉치로 서유럽 5개 국가(네덜란드, 영국, 프랑스, 독일, 스웨덴)에서 온 외국인 노동자들의 5개 국가 언어 습득 양상을 연구함.</li> <li>• 아랍어권, 터키, 핀란드, 스페인, 이탈리아, 핀란드에서 온 외국인 노동자 100명과 5개국 모어 화자(네이티브 스피커) 간의 대화를 전사함.</li> <li>• 제2 외국어 습득 연구에 활용함.</li> <li>• 네덜란드, 영국, 프랑스어는 전사 텍스트와 음성 자료가 있고, 독일, 스웨덴어는 전사 텍스트만 있음.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
유럽어	The EuroPat Corpus: A Parallel Corpus of European Patent Data (특허 도메인)	개인 연구자	병렬 말뭉치 크기: 5,100만 개 문장(스페인어-영어) ~154k문장(크로아티아어-영어), 필터링 되지 않은(원시) 말뭉치는 2배 규모	<ul style="list-style-type: none"> <li>• 2022년에 구축된 말뭉치로 특허 도메인에서 사용할 수 있는 정제된 병렬 말뭉치 마련을 위해 구축함.</li> <li>• 특허 자체는 저작권 대상 아님. 다만 유럽 특허청(EPO)이 특허 데이터베이스 저작권을 소유함.</li> <li>• 병렬 말뭉치 처리 시 유럽 특허청(EPO)의 저작권 조건을 따름.</li> <li>• 말뭉치 자체는 퍼블릭 도메인에 가까운 Creative Commons Zero로 출시함.</li> <li>• 6개 유럽 언어와 영어 간 병렬 말뭉치임(독일어, 스페인어, 불어, 크로아티아어, 노르웨이어, 폴란드어).</li> <li>• 필터링 적용이 가능한 메타데이터 부착함</li> <li>• 기존 표준(RAW, TXT 및 TMX)과 호환되도록 포맷을 갖추. OPUS27 및 ELRC-SHARE28 사용함.</li> </ul>
유럽어	CORP-ORAL (음성)	Portuguese Foundation for Science and Technology (FCT)	음성 52시간, 전사 32시간, 음운 전사 1시간	<ul style="list-style-type: none"> <li>• 2008년 구축된 말뭉치로 포르투갈어에 대한 음성 합성 및 인식 프로그램의 생성, 훈련 및 추가 개선을 위해 구축함. 음성, 음운론, 어휘, 형태론 및 구문 연구 뿐만 아니라 음성 합성 및 인식 시스템의 훈련에 사용할 수 있는 자발적인 유럽 포르투갈어(EP) 말뭉치 구축이 목적임.</li> <li>• 유럽 포르투갈어를 대상으로 하며 참가자들은 리스본이나 리스본 부근에서 태어나고 자랐어야 함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
유럽어	esCorpius: A massive Spanish Crawling Corpus (웹 크롤링 말뭉치)	LHF Labs	중복 제거 및 정리된 말뭉치 크기 문장 수 2.421.598.201개	<ul style="list-style-type: none"> <li>• esCorpius: A massive Spanish Crawling Corpus 논문, 2022년 발표함.</li> <li>• 다른 언어에 비해 규모가 작거나 품질이 낮은 스페인어 말뭉치의 단점을 극복하기 위해 구축됨.</li> <li>• 스페인어를 대상으로 하며 규모와 품질 측면에서 가장 크고 깨끗한 스페인어 웹 크롤링 말뭉치임.</li> <li>• 내려받은 데이터는 각 문서의 출처에 대한 추적성을 유지함. 이러한 추적을 통해 개별 웹사이트 소유자 또는 개인에게 철회권을 제공함.</li> </ul>
유럽어	An Annotated Corpus for Sexism Detection in French Tweets (도메인 말뭉치)	IRIT, Universit de Toulouse; Institut Jean Nicod, CNRS; LERASS, Universit de Toulouse	약 12,000개의 트윗	<ul style="list-style-type: none"> <li>• 2017년 10월~2018년 5월에 수집된 프랑스어 트윗(tweet)임.</li> <li>• 성차별적 트윗 수집 뿐만 아니라 성차별적 트윗이 포함된 메시지가 실제로 성차별적인지를 자동으로 감지하기 위해 구축됨.</li> <li>• 특정 키워드(ex. #sexisme, #sexiste, #SexismeOrdinaire, #EnsembleContreLeSexisme)를 포함한 데이터 세트임.</li> <li>• 성차별 탐지 주석이 달린 최초의 프랑스어 말뭉치임.</li> </ul>



## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
중국어	Adult English Read-Aloud Speech Dataset	maadaa.ai	500시간	• 성인 영어 화자의 음성 데이터임.
중국어	African English ASR Dataset	maadaa.ai	500시간	• 아프리카 영어 음성 데이터로 억양 등의 차이를 반영하여 음성 인식 기술에 적용할 수 있는 데이터임.
중국어	African Hausa ASR dataset	maadaa.ai	1000시간	• 아프리카 언어 중 하나인 하우스어 음성 데이터임.
중국어	Arabic & Thai & Vietnamese & Hindi & English & Chinese Language Dataset	maadaa.ai	150,000개	• 제품 포장, 간판, 표지판, 포스터, 전자기기, 주차장, 의류, 건물, 도로 표지판, 메뉴판, 책 표지, 쇼핑몰, 관광지 등 10가지 이상의 이미지에 대해 아랍어, 태국어, 베트남, 힌디어, 영어, 중국어로 주석한 말뭉치임.
중국어	Children's English reading voice Dataset	maadaa.ai	300시간	• 음성 데이터임.
중국어	35 European Countries Language Corpus Dataset	maadaa.ai	100,000개	• 병렬 말뭉치로 35개국의 유럽 언어를 대상으로 구축함.
중국어	Cantonese Internet Community Reviews Dataset	maadaa.ai	300,000개	• 광둥어로 된 인터넷 커뮤니티의 댓글을 크롤링하여 질문-답변(QA) 데이터 세트를 구축함.

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
중국어	Cantonese Parallel Corpus Dataset	maadaa.ai	500,000개	• 광둥어와 중국어의 병렬 말뭉치임.
중국어	Chinese Domain Specific Keywords & Slot Filling Annotation Dataset	maadaa.ai	500,000개	• 대화 데이터로 전자상거래 상품의 속성 정보를 주석한 데이터 세트임.
중국어	Chinese E-commerce Product Attribute Information Dataset	maadaa.ai	1,000,000개	• 여러 중국어 도메인에 대한 키워드 슬롯(keyword slot)을 채우는 데이터 세트임.
중국어	Chinese Electronic Medical Record Text Lexical Annotation Dataset	maadaa.ai	200,000개	• 중국어 전자 의료 기록 텍스트를 대상으로 어휘 주석한 데이터 세트. 어휘 주석은 지식 그래프(Knowledge Graph)로 활용될 수 있음.
중국어	Chinese Free Chat Dataset	maadaa.ai	200,000개	• 무료 채팅 데이터를 수집하여 만든 대화 말뭉치임.
중국어	Chinese Natural Language Emotion Classification Dataset	maadaa.ai	1,500,000개	• 대화 데이터에 11개의 하위 범주로 감정 분류 주석을 진행함. 11개의 감정 분류는 웃음, 의심, 슬픔, 두려움, 분노, 놀람, 미안함, 부끄러움, 안심, 평온, 난감함(smile, doubt, sadness, fear, anger, surprise, sorry, shyness, comfort, calm, helplessness).
중국어	Chinese Natural Language Understanding Dataset	maadaa.ai	500,000개	• 온라인과 오프라인을 통해 수집한 자료를 대상으로 도메인 카테고리화 개체명(entity)을 주석함.
중국어	Chinese-English parallel corpus dataset	maadaa.ai	1500,000개	• 중국어-영어로 된 병렬 말뭉치임.
중국어	Chinese-Korean parallel corpus dataset	maadaa.ai	2350,000개	• 중국어-한국어로 된 병렬 말뭉치임.

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
일본	Kyoto University Text Corpus	교토대학	약 2만 개 문장	<ul style="list-style-type: none"> <li>95년 1월 1일부터 17일까지의 전체 기사, 약 2만 문장, 1월부터 12월까지의 사설 기사 약 2만 문장, 총 약 4만 문장에 대해 형태소-구문 정보를 부여함.</li> <li>형태소 분석 시스템 JUMAN, 구문 분석 시스템 KNP로 자동 분석을 하고, 그 결과를 수작업으로 수정한 것임.</li> <li>사설을 제외한 약 2만여 개의 문장에 대해서는 격 관계, 조음-생략 관계, 공동 참조 정보를 부여하고 있음.</li> </ul>
일본	京都大学ウェブ文書リードコーパス	교토대학	5,000개 문서	<ul style="list-style-type: none"> <li>다양한 웹 문서의 시작(리드) 3문장에 각종 언어 정보를 인위적으로 부여한 텍스트 말뭉치. 웹 문서의 시작(리드) 3문장을 수집하여 뉴스 기사, 백과사전 기사, 블로그, 상업용 페이지 등 다양한 장르, 문체의 문서를 포함하고 있음.</li> <li>언어 정보로는 형태소, 고유 표현, 구문, 격 관계, 조음 생략 관계, 공동 참조, 담화 관계의 정보를 부여함.</li> <li>언어 정보 부여는 형태소 분석 시스템 JUMAN, 구문-격-조음 분석 시스템 KNP로 자동 분석하여 그 결과를 전문가가 수정하는 방식으로 이루어졌음.</li> <li>담화 관계는 전문가가 부여한 소규모 말뭉치 외에 클라우드소싱을 통해 부여한 대규모 말뭉치도 포함됨.</li> </ul>
일본	京都大学格フレーム	교토대학	웹상의 약 16억 개의 일본어 텍스트에 약 4만 개의 용언으로 구성	<ul style="list-style-type: none"> <li>웹 텍스트에서 자동 구축함.</li> </ul>
일본	京都大学名詞格フレーム)	교토대학	명사격 프레임은 웹상의 약 16억 개의 일본어 텍스트에 약 16만 개의 명사로 구성	<ul style="list-style-type: none"> <li>웹 텍스트에서 자동 구축함.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
일본	일본어 제방언 말뭉치 (Corpus of Japanese Dialects : COJADS)	국립국어연구소 언어자원개발센터	약 2,500 시간의 방언 담화의 녹음	<ul style="list-style-type: none"> <li>담화 음성은 문화청이 1977~1985년에 실시한 「각 지방언 수집 긴급 조사」의 방언 담화의 수록 데이터를 사용함.</li> </ul>
일본	일본어 일상회화 말뭉치 (Corpus of Everyday Japanese Conversation, CEJC)	국립국어연구소	총 200시간, 461세션, 577개 대화, 1675명 발화자, 약 240만 개 단어	<ul style="list-style-type: none"> <li>일상 장면 중 당사자 자신의 동기와 목적에 의해 자연스럽게 생기는 대화를 구축함.</li> <li>다양한 장면의 대화를 균형 있게 모아 공개함.</li> <li>연령별 남녀 총 40명에게 장비를 2~3개월간 대여해 주어 일상생활에서의 자연스러운 대화를 기록하도록 함(2016~2019년).</li> <li>다양한 장면, 다양한 화자와의 대화를 포함하고 있음.</li> <li>이외에 부족한 대화의 종류는 회의 10시간, 중고생 잡담 5시간 총 15시간을 추가함(2019~2020년).</li> <li>초상권, 개인정보, 저작권 관점에서 데이터 공개 방침 설정하였음.</li> </ul>
일본	일본어 말하기 말뭉치 ( Corpus of Spontaneous Japanese: CSJ )	국립국어연구소, 정보통신연구기구 (NICT), 동경공업대학	661시간 700만 개 단어	<ul style="list-style-type: none"> <li>대량 음성 데이터를 수집하여 다양한 연구 정보를 추가한 말하기 연구용 데이터 세트임.</li> <li>음성언어정보처리, 자연어처리, 일본어학, 언어학, 음성학, 심리학, 사회학, 일본어교육, 사전편찬 등 다양한 분야에서 활용되고 있음.</li> <li>온라인(무상) 및 USB 메모리에 저장된 데이터(유상)로 공개되어 있음.</li> <li><a href="https://clrd.ninjal.ac.jp/en/index.html">https://clrd.ninjal.ac.jp/en/index.html</a></li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
일본	EDR 말뭉치	주식회사 일본전자화사전 연구소	207,802개 어절	<ul style="list-style-type: none"> <li>• &lt;이와나미 정보과학사전(岩波情報科学辞典)&gt;, &lt;헤이본사 백과사전(平凡社百科辞典)&gt; 등 사전 외에 &lt;일본경제신문(日本経済新聞)&gt;, &lt;아사히신문(朝日新聞)&gt;, &lt;아에라(アエラ)&gt; 등의 간행물도 구축 대상에 포함함.</li> <li>• <a href="https://hayashibe.jp/tr/corpus/edr/Japanese">https://hayashibe.jp/tr/corpus/edr/Japanese</a></li> </ul>
일본	현대 일본어 글쓰기 평형 말뭉치(BCCWJ)	국립국어연구소	서적 전반, 잡지, 신문, 인터넷 게시판, 교과서, 법률 등의 장르에 걸쳐 1억 430만 단어의 데이터	<ul style="list-style-type: none"> <li>• 현대 일본어 글쓰기의 전체상을 파악하기 위해서 구축한 말뭉치임.</li> </ul>
일본	基本料理知識ベース	산업기술종합연구소, 교토대학교, 파나소닉 주식회사	26,426개 튜플	<ul style="list-style-type: none"> <li>• 기본 요리의 선정 및 재료, 조리법 정리는 요리 레시피 데이터에서의 빈도와 동반 발생을 기반으로, 속성 정리는 클라우드소싱을 통한 요리 만족도 조사를 기반함.</li> <li>• 요리 온톨로지 구축 프로젝트의 성과도 일부 포함하고 있음.</li> <li>• 각각의 지식에는 신뢰도 점수를 부여하고 있음.</li> <li>• 지식베이스는 콘셉트넷(ConceptNet)의 표기법을 따라 정리. 단, 관계의 종류는 콘셉트넷과 다름.</li> </ul>
일본	PheMT 데이터 세트	개인 연구자 (도후쿠 대학)	고유명사 943개 문장, 약어로 된 명사 248개 문장, 구어 표현 172개 문장, 변형 103개 문장	<ul style="list-style-type: none"> <li>• <a href="https://github.com/cl-tohoku/PheMT">https://github.com/cl-tohoku/PheMT</a>, <a href="https://aclanthology.org/2020.coling-main.521/">https://aclanthology.org/2020.coling-main.521/</a>, MTNT</li> <li>• 데이터 세트를 기반으로 하며 고유 명사, 약어, 구어체 표현 및 변형의 네 가지 언어 현상에 대한 추가 주석을 일본어, 영어로 제공함.</li> </ul>
일본	作文対訳データベース	국립국제연구소 (国立国語研究所)	1,754개 작문 수	<ul style="list-style-type: none"> <li>• <a href="https://mmsrv.ninjal.ac.jp/essay/#list01">https://mmsrv.ninjal.ac.jp/essay/#list01</a>, 학습자 말뭉치, 주제 작문 수집, 오류 주석 없음.</li> </ul>

## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
일본	NAIST 誤用コーパス	국립국어연구소	1,754개 작문 수	• 학습자 말뭉치, 주제 작문 수집. 오류 주석 있음.
일본	学習者作文コーパス「なたね」	도쿄공업대학교 국제학생센터	285개 작문 수, 글자 수 205,520개, 오류 주석 수 6,775개, 학습자 수 192개	• <a href="https://hinoki-project.org/natane/">https://hinoki-project.org/natane/</a> • 학습자 말뭉치, 주제 작문 수집. 오류 주석 있음.
일본	日本語学習者作文コーパス	자연어 처리 기술을 활용한 태깅된 학습자 작문 말뭉치의 개발 과제 연구 그룹	304개(작문 수)	• <a href="https://hinoki-project.org/natane/">https://hinoki-project.org/natane/</a> , 연구 과제 보고서: <a href="https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-22520537/">https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-22520537/</a> , 주제 작문 수집함. 오류 주석 있음. • 자연어 처리 기술을 활용한 주석이 부착된 학습자 작문 말뭉치를 개발함.
일본	国際日本語学習者作文コーパス及び誤用辞典	도쿄외국어대학교 국제일본연구센터	44개(작문 수)	• <a href="https://corpus.icjs.jp/corpus_ja/">https://corpus.icjs.jp/corpus_ja/</a> . • 일본어 학습자의 모어·지역성을 바탕으로 한 일본어 교육 연구를 위한 웹사전을 구축함. • 도쿄외국어대학교, 영국 리즈 대학교, 베이징 대학교, 상해 외국어대학교, 국제교양대학과 연계하여 각 대학에서 영어·중국어 원어민에 의한 일본어 학습자 말뭉치를 수집하고, 수정 정보 및 오용 태그 정보를 부착한 데이터베이스. 주제 작문을 수집함. 오류 주석 있음.
일본	FLUency extended TMU Evaluation Corpus for Japanese learners(FLUTEC)	국립국어연구소	82개(작문 수), 1,050개(문장)	• 학습자 말뭉치. <a href="https://github.com/kiyama-hajime/FLUTEC">https://github.com/kiyama-hajime/FLUTEC</a> . 일본어 학습자의 문법 오류 수정을 위한 평가 데이터. 2022년 구축. 자유 작문을 수집함. 오류 주석 있음(유창성 수정).

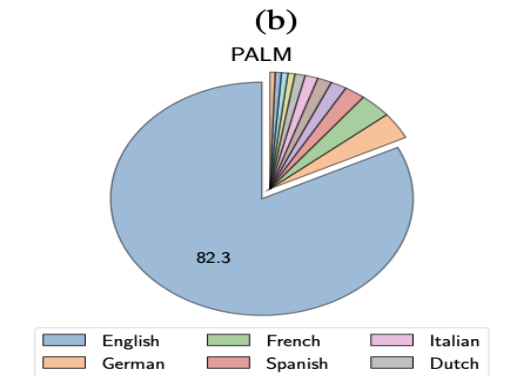
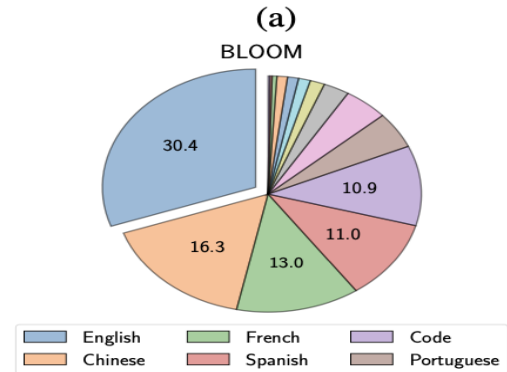
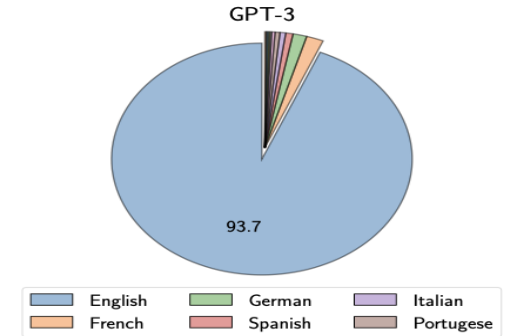
## 2-3-1. 환경 분석: 국외 주요 언어 자원 현황

언어	언어정보자원	구축 기관	구축 규모	구축 목적, 방법, 특징 외
일본	Taiyo 말뭉치	Taiyo 잡지	3,400개(기사), 약 7,000,000개(단어)	<ul style="list-style-type: none"> <li>현대 일본어에서는 사용되지 않는 어휘가 많기 때문에, 현대 일본어가 아닌 텍스트에 대한 형태소 분석을 수행할 때 활용함.</li> </ul>
일본	일본어 웹 말뭉치 (NWJC)	국립국어연구소	100억 개 단어	<ul style="list-style-type: none"> <li>웹에서 구축한 현대 일본어 말뭉치임.</li> <li><a href="http://pj.ninjal.ac.jp/corpus_center/nwjc/">http://pj.ninjal.ac.jp/corpus_center/nwjc/</a></li> </ul>
일본	메이다이카이와 말뭉치 (名大会. 話コーパス)	국립국어연구소	100시간	<ul style="list-style-type: none"> <li>구어 전사 말뭉치로 일본어 전자사전에 부가될 언어 정보를 추출하기 위한 기초 데이터임. 일본어 모어 화자 간의 잡담 대화를 129개 회화로 나누어 전사함.</li> <li><a href="http://pj.ninjal.ac.jp/conversation/nuc.html">http://pj.ninjal.ac.jp/conversation/nuc.html</a></li> </ul>
일본	BTS에 의한 다언어 구어 말뭉치(BTSによる多言語話し言葉コーパス)	동경외국어대학 대학원 지역문화연구과, 국립국어연구소	약28시간	<ul style="list-style-type: none"> <li>구어 말뭉치</li> <li>음성학적 분석, 형태소분석, 구문분석을 위한 것이 아니라 인간 상호작용으로서의 언어 사용 분석에 적합한 형태의 구어 말뭉치 구축이 목적임.</li> <li>2007년~2015년에는 동경외국어대학에서, 2016년 이후에는 국립국어연구소에서 일본어, 한국어, 중국어, 영어 등의 자연 대화 데이터를 수집하여 구축</li> <li><a href="https://ninjal-usamilab.info/lab/btsj_corpus/">https://ninjal-usamilab.info/lab/btsj_corpus/</a></li> </ul>
일본	BTSJ 일본어 자연 회화 말뭉치 (BTSJ日本語自然会話)	국립국어연구소	약 118.5시간, 474개 대화	<ul style="list-style-type: none"> <li>2017년 구축된 구어 말뭉치로 음성 파일과 전사 파일이 공개되어 있음.</li> <li>신청자에게 무료로 배포함.</li> <li><a href="http://pj.ninjal.ac.jp/corpus_center/csaj/">http://pj.ninjal.ac.jp/corpus_center/csaj/</a></li> </ul>
일본	TEC-JL	개인 연구자 (도쿄도립대학교)	135개(작문 수), 2,042개(문장)	<ul style="list-style-type: none"> <li>학습자 말뭉치. 일본어 학습자의 문법 오류 수정을 위한 평가 데이터임.</li> <li>2019년 구축. 자유 작문 수집, 오류 주석 있음(최소한의 수정).</li> </ul>

## 2-4. 환경 분석-국외 선행 연구

### 2-4-1. 선행 연구 주요 내용

- 영어권에 집중되어 있는 초거대 언어 모델(LLM) 연구
  - 챗GPT의 등장으로 언어 이해, 추론 및 언어 생성과 같은 많은 자연어 처리 작업에도 불구하고, 챗GPT의 성능을 평가하는 도구(tool)을 개발하는 것은 매우 어려움. 무엇보다 초거대 언어 모델(Large Language Models, LLM)과 관련한 대부분의 연구는 영어로 제한되어 있어 다른 언어로 된 텍스트를 이해하고 생성하는 능력은 불분명함.
  - Ahuja et. al(2023): 70개의 유형학적(typologically diverse language)으로 다양한 언어에 대한 데이터 세트를 사용하여 챗GPT와 챗GPT4 평가에 사용하였음. 다언어 세팅(Multilingual setting)에서 각 언어의 토큰화 유무가 다중모델(Multi model) 성능에 큰 영향을 미치는 것을 확인하였음. 그 예로 부족한 언어 자원(Low resource) 언어 중 하나인 힌디어 분석을 통해 토큰 수 참조가 불가능한 힌디어의 경우 성능이 저하됨을 밝혔음.
  - 이는 번역 기반의 프롬프트 생성 시 중간 자원(리소스) 언어가 무의미하게 번역될 수 있으므로 정확하고 의미 있는 프롬프트 결과를 위해서는 원어민의 검수 및 교정 작업이 필수적임을 강조하고 있음.



(c)



## 2-4. 환경 분석-국외 선행 연구

### 2-4-1. 선행 연구 주요 내용

- 영어권에 집중되어 있는 초거대 언어 모델(LLM) 연구

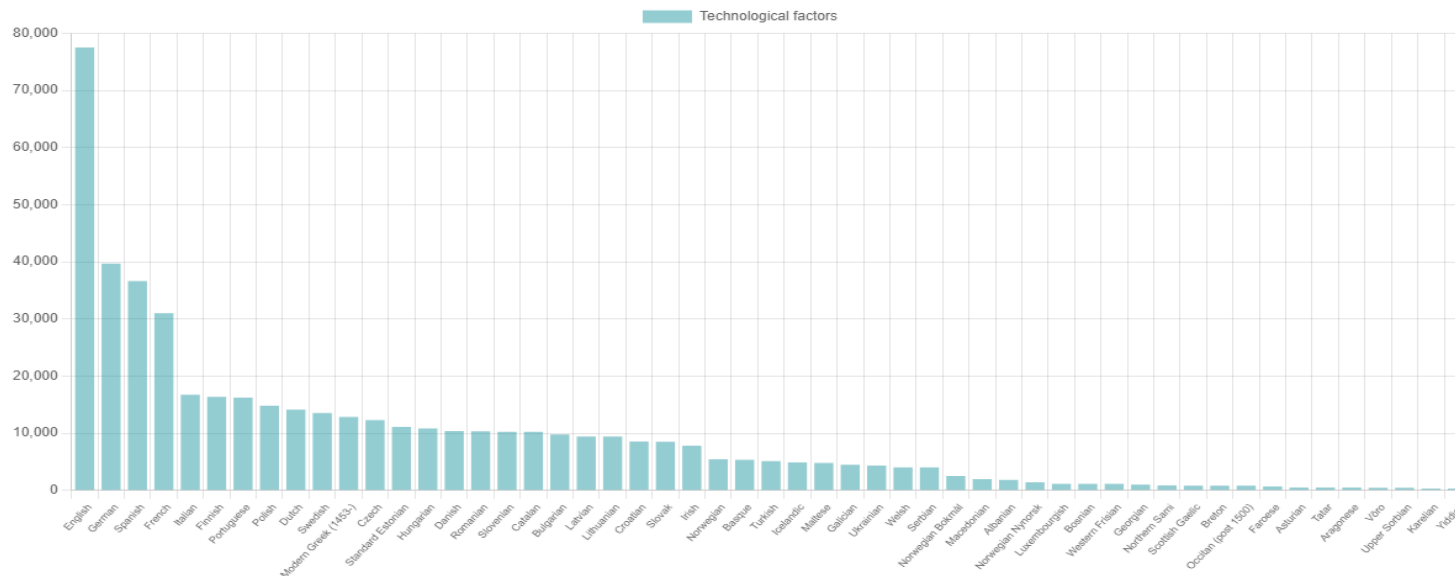
- Ramesh et al.(2023): 다양한 인구 통계 그룹 및 요인에 대한 불평등한 취급 해결에 대한 대부분의 연구가 영어에 집중되어 있음을 지적하면서 영어와 다국어 맥락을 넘어 언어의 공정성에 대한 다양한 측면을 분석하였음.
- 영어 이외의 맥락에서 공정성에 대한 분석을 진행하여 영어에 맞춰 설계된 방법의 한계를 밝히고 있음.

이로써 생성형 AI의 사용이 점점 더 보편화됨에 따라 다양한 문화와 언어를 포함해야 하는 공정한 데이터 세트 구성 측면에서 포괄적인 적용 범위를 설정하는 것은 불가능함을 확인함과 동시에 언어와 문화에 걸쳐 확장이 불가능한 현재의 데이터 세트 중심의 관행을 넘어서야 한다고 강조하였음.

## 2-4. 환경 분석-국외 선행 연구

### 2-4-1. 선행 연구 주요 내용

- ELEN(European Language Equality Network)의 2022년 SOTA(State-of-the-art) 워크숍에서는 이미 디지털 자료 내 언어 비율은 불균형하고, 영어, 독일어, 스페인어, 프랑스어 등을 제외하고는 거의 모든 EU 내 언어가 비율이 낮다고 주장함.
- ELEN은 기술적 요인, 맥락적 요인에 따라 각 언어별로 점수를 매긴 DLE(Digital Language Equality) 점수를 사용하고 있으며, 아래 그림은 말뭉치가 포함된 ‘기술적 요인’ 점수 그래프임.
- SOTA 워크숍에 따르면 언어간 격차는 시간이 지날수록 커지고 있어 저자원 언어에 대한 자원 확충 및 소멸 위기 언어에 대한 보호, 진흥이 필요함.



\* ELEN(2011년 결성된 EU 내 언어별 자원 균형을 지향하기 위한 연구 조직체)

## 2-4. 환경 분석-국외 선행 연구

### 2-4-1. 선행 연구 주요 내용

- 초거대 언어 모델의 경우에 일본어에서 주어진 문제에 대한 해결 능력과 정확도가 영어보다 낮은 원인을 일본어 학습량은 상대적으로 적다는 점을 꼽음. 주요 생성형 AI는 일본어를 비롯한 다양한 언어를 지원하고 있지만, 학습 및 서비스 제공 과정에서 수집되는 데이터의 양은 영어가 다른 언어에 비해 압도적으로 많다고 밝히고 있음.

表：生成AIの比較（各社資料を基にDTFAインスティテュート作成）

	パラメータ数	公開年
米国A社（AI企業） 第4世代	非公表	2023
中国B院（研究所）	1兆7500億	2021
米国C社（プラットフォーマー）	5400億	2022
米国A社 第3.5世代	3550億	2022
米国D社（プラットフォーマー）	1750億	2022
日本E社（IT企業）	820億	2022
日本F社（IT企業）	130億	2022
日本G社（AI企業）	13億	2022

생성 AI 파라미터 크기 비교(DTFA 연구소 조사 내용)

DTFA(Deloitte Tohmatsu Financial Advisory) 연구소 보고서, "生成AI普及、日本語の精度向上がなぜ課題か",  
<https://faportal.deloitte.jp/institute/report/articles/000784.html>

## 2-4. 환경 분석-국외 선행 연구

### 2-4-2. 주요 시사점

- 영어 데이터 집중에 따른 문화적 편향 우려

- 생성형 AI는 예술, 문학과 비디오 등 다양한 고품질 예술 미디어의 창조에 많은 영향을 미치고 있으며, 문화, 경제, 법률 등에서 기술과 창의성의 상호 작용에 대한 연구가 필요함.
- Epstein et al.(2023): 미술과 문화, 소유권과 신용에 대한 법적 질문, 창조물에 대한 미래, 현대 미디어 생태계에 대한 영향이라는 네 가지 주제로 생성형 AI가 미치는 영향에 대해 기술하고 있음. “희귀한” 혹은 “고유한” 입력보다 “공통”을 더 가깝게 모방하는 출력을 생성하거나 데이터의 하위 집합을 표현하는 데 초점을 맞추면, 생성 AI 모델의 문화 생산이 편향을 고착화하고, 다양성을 감소시킨다고 하였음.
- 생성형 AI는 독립적인 존재가 아니며, 특히 챗GPT의 응답은 개발에 사용된 데이터 세트로 인해 본질적으로 편향되어 있음.
- Yoon(2023): 생성 AI가 초래하는 문제 중 가장 중요한 것이 문화라고 강조하면서 암묵적 편견이 다양성과 다문화 사회로의 발전을 역행시킬 수 있다고 지적하였음. 특히 문화 식민주의 위험의 예로 서구적 관점의 문화적 데이터로 학습이 된 생성형 AI 모델의 경우 응답은 단일 관점에서 서양식 문화로 편향되어 나타나 비서구적 문화를 배경으로 하는 사용자는 서구적 문화에 지배당할 가능성이 있다고 하며, ‘문화 식민지로의 회귀’의 위험성을 밝혔음.

## 2-4. 환경 분석-국외 선행 연구

### 2-4-2. 주요 시사점

- 한국 언어 문화 특화형 말뭉치 구축의 필요
  - 대부분의 초거대 언어 모델(LLM) 연구는 영어로 제한이 되어 있어, 타 언어 및 다언어(multilingual) 데이터 세트를 평가하는 데 한계가 존재함.
  - 인공지능 기술 혁신과 진화가 가속화되고 있는 글로벌 상황에서 문화 식민주의로의 회귀를 막기 위해서는 고유의 언어모델 연구 지원 및 다양한 AI 언어모델 개발을 위한 데이터 세트 구축이 필요함.
  - 다국어는 고정관념과 언어마다 차별성을 갖기에 각 언어의 뉘앙스 또한 탐색해야 하며 이는 다른 언어에서 동등하게 나타나지 않을 수 있음.
  - 문화적 금기와 고정관념은 지역마다 혹은 국가마다 다를 수 있기 때문에 각 나라별 세부 문화 정보는 고유의 언어에서 반영되어야 해당 모델을 제대로 평가할 수 있음.
  - 한국어 언어 모델 개발을 위해서는 한국어 고유의 언어적, 문화적 특성이 반영된 한국 언어 문화 특화형 데이터 세트를 구축해야 할 것임.

## 2-5. 환경 분석: 주요 언어 모델 및 특징

공개 시기	모델명	언어	기초 모델 및 특징	학습 데이터	모델크기(B: 10억, M: 백만)
'23.10월	HumanF	한국어	Platypus 기반	공공데이터 활용	13B
'23.9월	42MARU	한국어	Llama2 기반	자체 생성 학습 데이터	13B
'23.9월	krevas	한국어	Llama2 기반	OSCAR 데이터 활용	13B
'23.9월	CLOVA-X	한국어	Naver 자체개발	네이버 자체데이터	175B~200B 추정
'23.8월	VARCO	한국어	NCsoft 자체개발	공개되지 않음	1.3~13B
'23.8월	SOLAR	영어/한국어	LLaMA2 기반	OSCAR, Guanaco 등	70B
'23.7월	Blossom	한국어	LLaMA2 and Polyglot 활용	LIMA 데이터를 일부 번역하여 활용	7B~13B
'23.6월	NALLM	한국어	Polyglot and KoBART 활용	공공기관, 민원데이터 활용 (자체 제작 비공개)	7B
'23.6월	KULLM	한국어	Polyglot 활용	Vicuna, GPT4ALL, Dolly 데이터 번역하여 활용	12.8B
'23.6월	CaMA	영어, 중국어	LLaMA 13B 활용	LLaMA(영어) and 중국어 말뭉치 (Baidu Baike, Wudao, Chinese Wikipedia) 활용	13B
'23.5월	Bloomz&mT0	다국어	BLOOM and mT5 활용	BLOOMZ and mT0 활용	2.09B
'23.5월	Baize	영어	LLaMA and LoRA 활용	챗GPT 생성 데이터 10만 개 활용	7B
'23.5월	ChatGLM-6B	중국어	GLM 활용	WuDao 활용	6.2B
'23.5월	RWKV-LM	영어	Transformer 수준의 RNN 활용	Enwiki8 활용	1B~14B
'23.5월	ChatRWKV	다국어	RWKV 활용(RNN)	Enwiki8 활용	1B~14B
'23.5월	GALPACA	영어	GALACTICA 활용	Alpaca 활용	30B

## 2-5. 환경 분석: 주요 언어 모델 및 특징

공개 시기	모델명	언어	기초 모델 및 특징	학습 데이터	모델크기(B: 10억, M: 백만)
'23.5월	StarCoder	다국어	StarCoderBase 활용	80개 이상의 언어 The-Stack v1.2데이터 세트 활용	15.5B
'23.5월	MPT-7B	영어	transformer 기반 아키텍처	LLaMA, Pythia, OpenLLaMA, StableLM 활용	7B
'23.5월	Falcon	영어	Casual Decoder 활용	웹 크롤링을 통한 사전 훈련 데이터 활용	7B, 40B
'23.4월	moss	영어, 중국어	CodeGen 인터페이스	Pile, BigQuery, BigPython, 중국어 코퍼스 활용	16B
'23.4월	StableLM	영어	LLaMA 활용	The Pile 활용	3B, 7B
'23.4월	Flan-Alpaca	영어	LLaMA 사용	Flan, Alpaca 사용	220M (Base)
'23.4월	GPTQ-for-LLaMA	영어	GPTQ and LLaMA 활용	LLaMA 데이터 활용	7B
'23.4월	GPT4All	영어	GPT-J and LLaMA 활용	GPT4ALL 데이터 활용	7B
'23.4월	Koala	영어	LLaMA 활용	ShareGPT, Stanford Alpaca 활용	
'23.4월	Belle	중국어	LLaMA, BLOOMZ 활용	ChatGPT 생성 데이터 활용	7B
'23.4월	StackLLaMA	영어	LLaMA and RLHF 활용	StackExchange 활용	7B
'23.4월	RedPajama- Data	영어	LLaMA 훈련 데이터 세트 재현	RedPajama 데이터 활용	
'23.4월	LLMZoo	라틴어 + 다국어	LLaMA and BLOOMZ 활용	Phoenix and Chimera 활용	7B
'23.3월	Alpaca	영어	LLaMA and Transformer 활용	Self-Instruct 자동 학습 데이터 활용	7B
'23.3월	Cabrita	포르투갈	LLaMA and LoRA 활용	Alpaca를 ChatGPT로 번역하여 활용	7B
'23.3월	Vicuna	영어	LLaMA 사용	ShareGPT.com 데이터 70K 활용	7B

## 2-5. 환경 분석: 주요 언어 모델 및 특징

공개 시기	모델명	언어	기초 모델 및 특징	학습 데이터	모델크기(B: 10억, M: 백만)
'23.3월	Chinese-Vicuna	중국어	LLaMA 사용	BELLE and Guanaco 활용	7B
'23.3월	Dolly 1.0	영어	GPT-J 활용	Alpaca 개발팀 Open API 활용	6B
'23.3월	Dolly 2.0	영어	EleutherAI pythia 활용	Databricks crowdsourced 활용	12B
'23.3월	OpenFlamingo	영어	Flamingo 모델 오픈 소스로 재현	Multimodal C4 and LAION 2B 데이터 활용	9B
'23.3월	Cerebras-GPT	영어	컴퓨팅 효율 기반 GPT-3 활용	폐쇄형 데이터	111M~13B
'23.3월	Camel	영어	Palmyra-Base 활용	70K의 instruction-response dataset 활용	5B
'23.1월	Pythia	영어	시간과 규모에 따라 transformer 해석	Pile 활용	70M~12B
'23.1월	Palmyra	영어	Transformer Decoder 활용	Custom Dataset 활용	5B
'23.1월	YaLM	영어, 러시아어	YaLM 언어 모델 활용	The Pile(0.25) and 자체 수집 러시아어 텍스트(0.75) 활용	100B
'22.11월	GALACTICA	영어	디코더 전용 설정의 transformer 기반	대규모 과학 corpus 활용	125M~120B
'22.5월	OPT	영어	Transformer 기반 OPT 모델 활용	RoBERTa, Pile, PushShift.io, Reddit 활용	125M~175B
'22.3월	GLM	영어, 중국어	공백 채우기 기반 모델 생성	Wiki, RoBERTa, Pile, WuDao 활용	110M~10B
'21.12월	GPT-J	영어	Mesh Transformer JAX 활용	The Pile 활용	6B



## 2-5. 환경 분석: 주요 언어 모델 및 특징

공개 시기	모델명	언어	기초 모델 및 특징	학습 데이터	모델크기(B: 10억, M: 백만)
'21.10월	GPT-NEO	영어	Mesh tensorflow 라이브러리 활용	The Pile 활용	1.3B, 2.7B
'21.4월	PanGu-α	중국어	Pengcheong, Pangu-α 활용	오픈 소스 데이터 세트, 웹수집 데이터 세트 활용	200B

### 주요 언어 모델 및 특징(요약)

- 초거대 모델(70B 이상) 이외에도 연구실 수준에서 사용 가능한 모델 위주로 학습 및 공개(13B)
- 공개된 한국어 모델의 경우 해외 공개 모델에 대한 추가 학습(미세조정, 강화학습, 인스트럭션 튜닝 등) 수행하는 수준으로 진행함.
- 한국어 모델의 학습 데이터 역시 영어 데이터 세트의 번역을 사용하는 경향임.

### 3. 글쓰기 자동평가 및 자동첨삭 방안

1. 사업 개요

2. 환경 분석

3. 글쓰기 자동평가 및 자동첨삭 방안

3-1. 글쓰기 자동평가 및 자동첨삭 시스템 해외 사례

3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

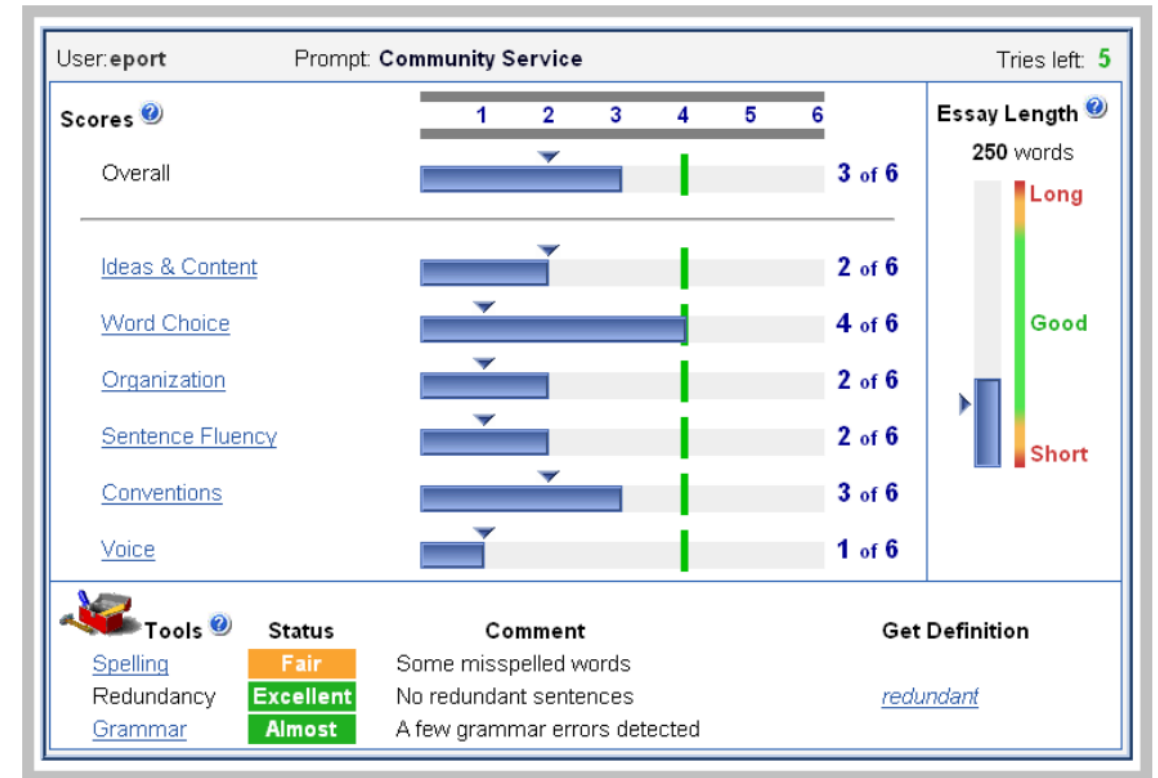
3-3. 유사 시스템 사례

3-4. 글쓰기 자동평가 및 자동첨삭 말뭉치· 시스템  
구축 방안

4. 한국어 말뭉치 구축 중장기 계획안

## 3-1. 글쓰기 자동평가 및 자동첨삭 시스템 해외 사례

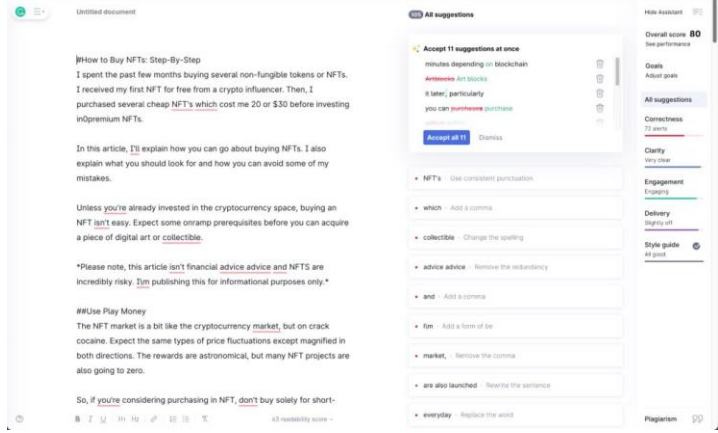
- Pearson's Intelligent Essay Assessor™(IEA)
  - 인터넷 기반의 자동평가 시스템
  - 제출된 에세이의 품질에 대해 점수를 자동으로 평가
- 평가 지표
  - 아이디어 및 내용, 단어 선택, 구성, 문장 유창성, 글쓰기 관습(convention), 어조(voice) 등



잠재 의미 분석(Latent Semantic Analysis, LSA) 기반의 단어 및 문서 간 유사도를 측정하여 학생 에세이에 대해 유사 에세이의 점수를 제안하는 시스템

## 3-1. 글쓰기 자동평가 및 자동첨삭 시스템 해외 사례

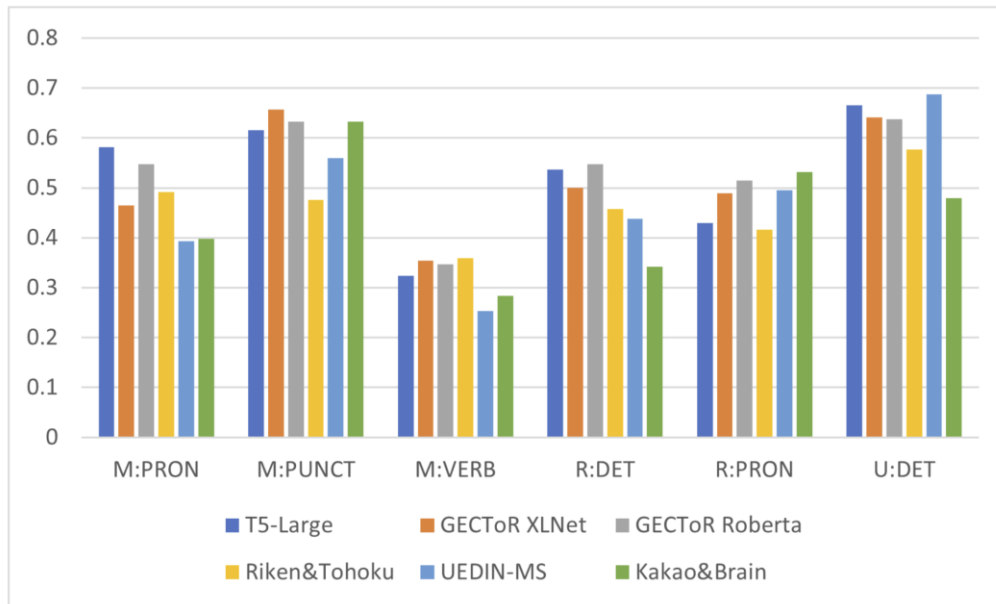
- 그램머리(Grammarly)
  - 글쓰기 자동첨삭  
(문법적 오류 수정 및 제안)
  - 글쓰기 자동평가
    - 정확성
    - 명확성
    - 흥미유발
    - 다양성
    - 문체

회사명	그램머리 (Grammarly)	설립일	2009.07.01
고객수(2021)	일일 활성 사용자 3,000만 명 / Business 계정 30,000개		
서비스 사례			
서비스 이용료	<div> <div> <b>Free</b> Basic writing suggestions.           </div> <div> <b>Premium</b> Style, tone, and clarity improvements for writing at work and school.  Starting at <b>\$12.00</b> / month           </div> <div> <b>Business</b> Professional and clear communication for teams of 3 to 149.  Starting at <b>\$12.50</b> / member / month           </div> </div> <div> <div>Create Account</div> <div>Get Started</div> <div>Get Started</div> </div>		

## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

### 문법 오류 교정(Grammatical Error Correction, GEC)

- 문법 오류 교정(Grammatical Error Correction, GEC) 관련 경진대회(shared task)
  - CoNLL-2013[1], CoNLL-2014[2], BEA-2019[3]



BEA-2019 모델의 성능 예

### GEC(Grammatical Error Correction, GEC) 연구 동향

- 트랜스포머(transformer)의 등장 이후 시퀀스-투-시퀀스 기반의 모델이 GEC(Grammatical Error Correction, GEC) 과업에서 효과적으로 작동. 언어 모델(Language Model, LM) 역시 동일한 아키텍처로 개발되어 활용
- GEC(Grammatical Error Correction, GEC) 시스템은 각기 잘하는 유형이 존재
  - 시스템 조합을 통해 성능 향상
  - MEMT 시스템
  - IBM 시스템
  - 국내에는 전북대학교 나승훈 교수 연구팀의 연구 사례 존재 (Qorib 외, Frustratingly Easy System Combination for Grammatical Error Correction, 2022)

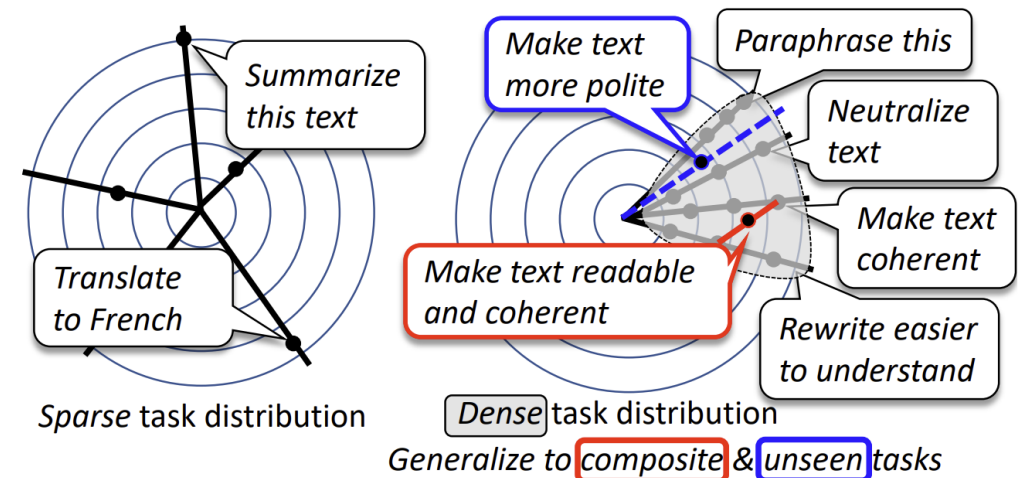
[1] Ng 외, "The CoNLL 2013 shared task on grammatical error correction", 2013  
 [2] Ng 외, "The CoNLL 2014 shared task on grammatical error correction", 2014  
 [3] Bryant 외, "019. The BEA-2019 shared task on grammatical error correction", 2019

## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

- 슈퍼 내추럴 인스트럭션(Super-Natural Instruction)
  - 1,616개의 다양한 NLP 과제를 포함
  - 문법 오류 탐지(grammar error detection)
  - 문법 오류 수정(grammar error correction)
  - 구두점 오류 탐지(punctuation error detection)
  - 철자 오류 탐지(spelling error detection)
  - 스타일 변환(style transfer)

## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

- CoEDIT
  - 인스트럭션 기반의 텍스트 편집 언어 모델
    - “더 간단한 문장으로 만들어 주세요”
    - “더 중립적으로 써 주세요”
  - 일반적 목적의 초거대 언어 모델(LLM)과 달리 구체적 목적의 인스트럭션에 대응하는 말뭉치 구축
  - 그래머리 팀의 연구 성과



[1] Raheja 외, “COEDIT: Text Editing by Task-Specific Instruction Tuning”, 2023

## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

- CoEDIT 데이터 세트 예시

- 기존 데이터(예: NUCLE-14 등)를 지시문(인스트럭션) 데이터 형태로 변환
- 각 유형별로 1만~2만 개 인스트럭션 데이터(입력-출력 쌍) 구축
- 유형: 유창성, 일관성, 명확성, 문체(환언, 형식화, 중립화)

Edit Intention	Datasets	Size	Example Input	Example Output
FLUENCY	NUCLE-14, Lang-8, BEA-19	20k	<i>Fix the grammar:</i> When I <b>grow</b> up, I <b>start</b> to understand what he said <b>is</b> quite right.	When I <b>grew</b> up, I <b>started</b> to understand what he said <b>was</b> quite right.
COHERENCE	DiscoFuse	11k	<i>Make this text coherent:</i> Their flight is <b>weak</b> . <b>They</b> run quickly through the tree canopy.	Their flight is <b>weak</b> , <b>but they</b> run quickly through the tree canopy.
CLARITY (Simplification)	NEWSELA, WikiAuto, WikiLarge, ParabankV2, ITERATER-CLARITY	13k	<i>Rewrite to make this easier to understand:</i> A storm surge is <b>what forecasters consider</b> a hurricane's most <b>treacherous</b> aspect.	A storm surge is <b>considered</b> a hurricane's most <b>dangerous</b> aspect.
STYLE (Paraphrase)	ParabankV2	15k	<i>Paraphrase this:</i> Do you know <b>where I was born</b> ?	Do you know <b>my birthplace</b> ?
STYLE (Formalize)	GYAFC	12k	<i>Write this more formally:</i> <b>omg i</b> love that song <b>im</b> listening to it right now	I love that song <b>and I am</b> listening to it <b>at this moment</b> .
STYLE (Neutralize)	WNC	11k	<i>Write in a more neutral way:</i> The authors' <b>exposé</b> on nutrition studies.	The authors' <b>statements</b> on nutrition studies.



## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

### • CoEDIT 성능 평가 비교표

Model	Size	Overall	IteraTeR	Fluency	Clarity	Coherence	Style		
			ITERATER <sup>†</sup>	JFLEG <sup>†</sup>	ASSET <sup>†</sup>	DiscoFuse-Wiki <sup>†</sup>	GYAFC( <sup>†</sup> / <sup>†</sup> )	WNC( <sup>†</sup> / <sup>†</sup> )	MRPC( <sup>†</sup> / <sup>†</sup> )
(a) COPY T5-LARGE	- 770M	27.6 24.7	29.8 21.1	26.7 / 40.5 32.7 / 22.9	20.7 35.8	30.8 28.01	17.6 / 10.6 30.9 / 4.89	31.85 / 0 13.2 / 0	47.4 / 100 27.6 / 62.8
(b) T0* Tk-INSTRUCT* T0++*	3B	29.7	26.1	42.2 / 36.1	33.2	32.4	37.9 / 39.3	19.4 / 0	28.3 / 84.1
	3B	27.3	21.0	35.2 / 26.8	36.9	28.9	35.7 / 43.01	24.2 / 0.1	20.4 / 48.9
	11B	32.6	31.5	39.4 / 40.5	33.1	35.5	36.8 / 43.7	21.2 / 0	42.9 / 94.9
(c) LLAMA GPT3	7B 175B	28.2 27.4	30.1 23.3	27.7 / 3.34 38.1 / 2.8	21.8 34.8	31.1 26.2	18.8 / 89.1 36.6 / 87.9	31.9 / 0 23.4 / 0	5.29 / 64.2 0 / 51.7
(d) ALPACA GPT3-EDIT INSTRUCTGPT CHATGPT	7B	28.4	30.4	28.5 / 6.4	22.0	31.1	18.9 / 94.4	31.9 / 0	0 / 77.9
	175B	41.8	36.1	52.4 / 50.6	32.9	54.0	35.7 / 52.3	50.7 / 17.1	22.6 / 98.7
	175B	41.6	32.6	62.4 / 57.2	44.6	47.4	47.8 / 98.2	33.7 / 0.1	16.03 / 98.9
	-	36.9	28.2	57.6 / 49.4	45.9	40.2	40.7 / 99.6	28.5 / 0.1	<b>13.4 / 99.0</b>
(e) ALPACA (FS) GPT3 (FS) INSTRUCTGPT (FS) CHATGPT (FS)	7B	30.0	30.8	33.03 / 11.3	23.2	33.1	20.6 / 95.4	32.04 / 0	0.1 / 66.7
	175B	38.4	32.4	50.1 / 4.1	39.2	45.1	43.1 / 97.2	36.7 / 0	0 / 14.5
	175B	45.1	36.2	64.5 / 55.7	<b>46.3</b>	55.2	47.3 / 98.8	42.8 / 0	15.9 / 99.5
	-	40.1	30.8	58 / 50.6	45.4	51.2	42.3 / 99.6	34.1 / 0	13.3 / 96.1
(f) ITERATER DELITERATER PEER-3B* PEER-11B*	570M	31.0	32.8	35.9 / 34.3	21.8	30.1	22.7 / 54.1	34.2 / 0	40.5 / 97.8
	570M	28.0	29.9	27.5 / 31.2	21.2	32.2	18.1 / 57.8	31.9 / 0	39.1 / 100
	3B	41.7	37.1	55.5 / 54.3	30.5	-	-	53.3 / 21.6	-
	11B	42.1	<b>37.8</b>	55.8 / 54.3	29.5	-	-	54.5 / 22.8	-
(g) CoEDIT-L CoEDIT-XL CoEDIT-XXL	770M	49.8	35.2	62.4 / 59.3	42.4	75.3	54.6 / 98.0	69.3 / 46.4	23.3 / 99.1
	3B	51.4	36.6	64.5 / 60.7	42.2	<b>80.5</b>	<b>55.1 / 98.3</b>	70.4 / 48.8	21.3 / 99.6
	11B	<b>51.5</b>	37.1	<b>65.0 / 61.5</b>	41.7	78.6	55.1 / 97.2	<b>71.0 / 51.4</b>	21.8 / 99.0

### CoEDIT 모델이란?

- FLAN T5 모델을 기초모델로 사용하여 학습데이터로 미세조정된 모델

### 평가 데이터(벤치마크 데이터)

- JFLEG(2017, 문법 오류 교정)
- GyaFc(2018, 문체-형식화)
- DISCOFULSE(2019, 일관성)
- ASSET(2020, 텍스트 단순화)
- WNC(2020, 문체-중립화)
- ITERATER(2022, 일관성)
- MRPC(문체-환언)
- STS(문체-환언) \*
- QQP(문체-환언) \*

## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

- CoEDIT 결과 예시

Instruction	GPT3-EDIT	CoEDIT-xl (ours)
<i>Fix all grammatical errors:</i> Every one would expect to leave his owns and secrets dreams like to be a painter, a writter ...	Every one would expect to leave his owns and secrets dreams like the painting and the write ...	Every one would expect to leave his own and secret dreams like to be a painter, a writer...
<i>Make this simpler:</i> Fives is a British sport believed to derive from the same origins as many racquet sports.	Fives is a British racquet sport.	Fives is a British sport. It is thought to have come from the same place as many racquet sports.
<i>Fix coherence in this text:</i> Phantom Budo is an incredibly powerful ability. Spike has yet to fully develop this awesome fighting technique.	Phantom Budo is an incredibly powerful fighting technique. Spike is yet to fully develop it .	Phantom Budo is an incredibly powerful ability, but Spike has yet to fully develop it.
<i>Use different wording:</i> Judge Gerald W. Heaney, in dissent, said the authorities should have allowed the prisoner to be medicated without the consequence of execution.	Judge Gerald W. Heaney, who disagreed with the majority decision, said the prisoner should be medicated after important testimony is received.	In his dissent, Gerald W. Heaney said the authorities should have allowed the prisoner to be treated without execution.

## 3-1-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

### Dataset-of-Misspelings-and-Corrections(2023) 데이터

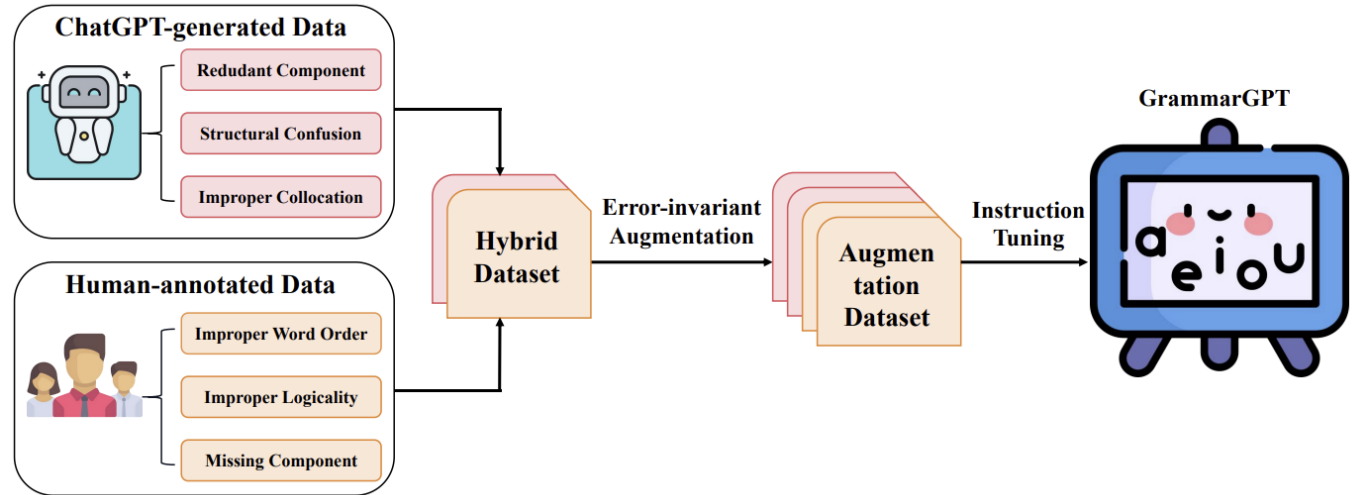
- 자그레브 대학교 데이터 수집(2008년 12월 ~ 2023년 3월, 약 16년)
  - 크로아티아 맞춤법 검사기를 사용한 900,000명에게서 "오류 단어 → 올바른 단어" 형식의 33,382,330개 항목을 수집
  - 전처리 후 5,584,226개의 고유한 "오류 단어 → 정답 단어" 쌍을 최종 데이터로 사용하였음.
  - 해당 데이터 세트 사용을 통해 맞춤법 검사 정확도를 향상
- 언어 학습자의 일반적인 오류 유형에 대한 분류, 철자 오류 추이 및 패턴 분석 가능
- <https://github.com/Ispravi-Me/Dataset-of-Misspelings-and-Corrections>

## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

### GrammarGPT

- 중국인 화자를 대상으로 발생하는 주요 문법적 오류 주석
  - CLUE(Chinese Language Understanding and Generation Evaluation)를 통해 쉽게 수정할 수 있는 것과 그렇지 않은 것을 구분함.
  - 쉽게 교정 가능한 데이터는 챗GPT를 사용하여 비문법적인 문장을 자동 생성함.
  - 공개 데이터를 활용하여 비문법적인 문장에 사람이 직접 주석하여 데이터 세트를 구축함.

w/ Clues	Redundant Component (RC)	<b>Incorrect:</b> 这座卫星城的人口估计 <b>超过一百万左右</b> . The population of this satellite city is estimated to be <b>more than about</b> one million. <b>Correct:</b> 这座卫星城的人口估计超过一百万。 The population of this satellite city is estimated to be over one million.
	Structural Confusion (SC)	<b>Incorrect:</b> 这次网络故障的 <b>原因</b> 是由服务器故障 <b>引起的</b> . The <b>cause</b> of this network failure is <b>caused by</b> the server failure. <b>Correct:</b> 这次网络故障的原因是服务器故障。 The cause of the network failure is the server failure.
	Improper Collocation (IC)	<b>Incorrect:</b> 西湖区正全面 <b>提升</b> 区域产城融合发展的 <b>步伐</b> . Xihu District is <b>promoting</b> the <b>pace</b> of integration of regional industry and city development. <b>Correct:</b> 西湖区正全面加快区域产城融合发展的步伐。 Xihu District is accelerating the pace of integration of regional industry and city development.
w/o Clues	Improper Word Order (IWO)	<b>Incorrect:</b> 学校 <b>三个月内</b> 要求每名 <b>学生</b> 完成20个小时的 <b>义工服务</b> . The school <b>in three months</b> requires each <b>student</b> to complete 20 hours of volunteer service. <b>Correct:</b> 学校要求每名 <b>学生</b> 三个月内完成20个小时的 <b>义工服务</b> . The school requires each student to complete 20 hours of volunteer service in three months.
	Improper Logicality (IL)	<b>Incorrect:</b> 集团向社会各界人士、 <b>沿途村庄百姓</b> 表示歉意。 The group apologizes to people from all walks of life <b>and villagers along the way</b> . <b>Correct:</b> 集团向社会各界人士表示歉意。 The group apologizes to people from all walks of life.
	Missing Component (MC)	<b>Incorrect:</b> 这篇报告控诉了人类破坏大自然 <b>(...)</b> . The report accused man of destroying nature. <b>Correct:</b> 这篇报告控诉了人类破坏大自然的罪行。 The report accused man the crime of destroying nature.

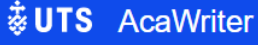


## 3-2. 글쓰기 자동평가 및 자동첨삭 말뭉치 사례

- 선행 연구의 시사점
  - 학생 글쓰기 교육의 목적 고려
    - 교정 피드백 제공 필요(corrective feedback)
      - ✓ 참고) 비교정 피드백(non-corrective feedback)은 학생의 자기 동기부여 기법
    - 설명 가능한 피드백 제공 필요(높은 점수, 낮은 점수를 받은 것에 대한 이유 설명 필요)
  - 진단체계의 다양한 지표를 고려함
    - 설득력(Ke 외, 2018; Carlile 외, 2018) 등 내용 및 조직, 표현 등 전반적인 지표를 고려함.
    - 전체 피드백(global feedback)과 국소적 피드백(local feedback)을 고려함.
      - ✓ 전체 피드백: 내용, 완성도, 조직, 형식 등
      - ✓ 국소적 피드백: 문법, 단어, 스펠링 등

## 3-3. 유사 시스템 사례

### AcaWriter



[My documents](#)
[Examples](#)
[Assignments](#)
[Help & Support](#)

Shibani Aileen Shibani Michael Xavier ▾

Law essay

Download PDF

AcaWriter works fastest with short texts, so if you're only working on a specific section, don't paste in the whole document. It still processes long texts, but it may take a few minutes to get your feedback to you.

B I U ☰ ☷

Technology is an enabler in providing greater access to justice through its ability to connect people with legal needs to legal assistance, information, and advice. With the increasing popularity of internet-enabled hand held devices and laptop computers, there is a tendency to assume that even the socio-economically vulnerable in our society have access to technology and the skills to use online services with confidence. This is not necessarily the case.

Examples of the application of technology to provide legal information and assistance include case studies, guides and virtual legal advice clinics. The 2012 Review does not address the role of courts in serving the legal needs of the community. The court system is not regarded as a part of the wider legal assistance services. This omission questions the role of the court in facilitating access to its services, including dispute resolution and trials. It also identified uses of technology to expand the delivery of services, many of which are transferable to an online court. These services include e-access for remote communities, availability outside of business hours, interactive processes and virtual appearances. This essay will discuss uses of technology to expand the delivery of services, many of which are transferable to an online court.

Get Feedback >

Computers don't read writing like humans. So, if you're sure your writing's good, it's fine to disagree with AcaWriter's feedback, just like you'd ignore a poor grammar suggestion.

Analytical Report

Feedback

Examples

**i** The analytical report highlights salient rhetorical moves AcaWriter identified in your essay for reflection. For more specific feedback, go to the Feedback tab.

Rhetorical Moves

- |  |   |
|--|---|
| <b>S</b> Summarises or signals the authors goals       | <b>P</b> Perspective or stance                    |
| <b>E</b> Emphasis to highlight key ideas               | <b>N</b> Novel improvements in ideas              |
| <b>C</b> Contrasting idea, tension or critical insight | <b>B</b> Background information and previous work |
| <b>S</b> Surprising or unexpected finding              | <b>Q</b> Question or gap in previous knowledge    |
| <b>T</b> Trend or tendency related to ideas            |   |

Technology is an enabler in providing greater access to justice through its ability to connect people with legal needs to legal assistance, information, and advice. **T** With the increasing popularity of internet-enabled hand held devices and laptop computers, there is a tendency to assume that even the socio-economically vulnerable in our society have access to technology and the skills to use online services with confidence. This is not necessarily the case.

Examples of the application of technology to provide legal information and assistance include case studies, guides and virtual legal advice clinics. **S C** The 2012 Review does not address the role of courts in serving the legal needs of the community. The court system is not regarded as a part of the wider legal assistance services. **C** This omission questions the role of the court in facilitating access to its services, including dispute resolution and trials. It also identified uses of technology to expand the delivery of services, many of



## 3-3. 유사 시스템 사례

### AcaWriter

Computers don't read writing like humans. So, if you're sure your writing's good, it's fine to disagree with AcaWriter's feedback, just like you'd ignore a poor grammar suggestion.

#### Analytical Report

#### Feedback

#### Examples

**i** The rhetorical moves highlighted by AcaWriter are used in good academic writing but use them with caution according to the context. Remember, AcaWriter does not really understand your writing, the way people do. You may have written beautifully crafted nonsense - that's for you to decide! Moreover, writing is complex, and AcaWriter will get it wrong sometimes. If you think it got it wrong, that's fine - now you're thinking about more than spelling, grammar, and plagiarism.

**!** It looks like you are missing a Background move in your text, which highlights background information and previous literature on the topic. Try including linguistic cues to make this move clearer in your writing. Examples: The past decade has seen ....., Recent studies indicate ... ,It is generally accepted that..., the concept has previously been thought to be...

**!** If there is a key idea you did like to emphasises in your essay try including linguistic cues to make this move clearer in your writing. Examples: It is important to note that ....., It makes a proper understanding important...

### 글쓰기 교육의 목적 고려

- 분석 보고서
  - 글에 대한 담화 구조 분석 수행  
(주장, 근거 등에 해당하는 문장 주석)
- 피드백
  - 단순한 글에 대한 수정이 아닌 글 전체적으로 수정될 수 있는 방향 제시
  - 수정 의견의 경우 “예를 들어 다음과 같이 수정 가능하다...” 와 같은 피드백 제공

## 3-3. 유사 시스템 사례

### Revision Assistant

- 특정 지표에 대한 평가 제공
- 특정 문장에 대한 피드백 제공

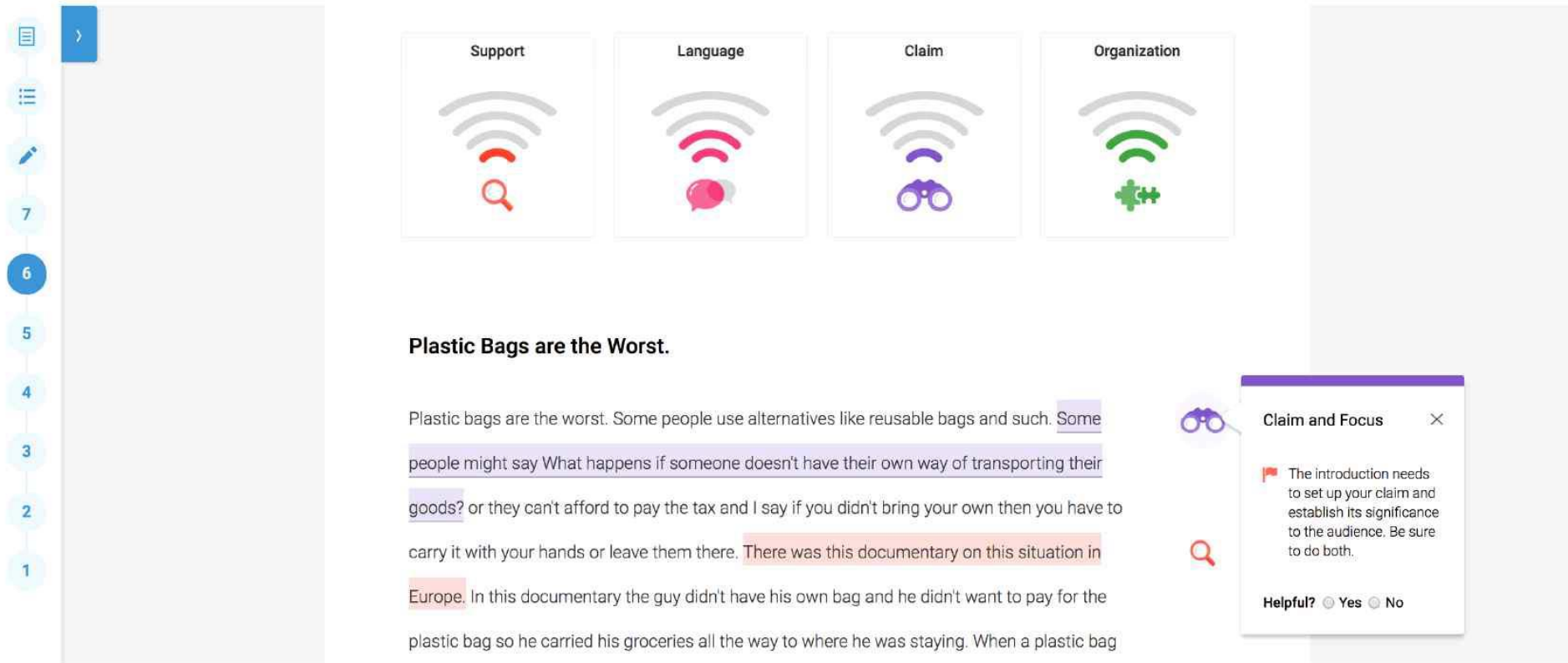


Figure 1: An example of Revision Assistant in use. The signal bars at the top of the draft indicate predicted scores on four rubric traits. Highlighted sentences give targeted feedback to the student on strengths and weaknesses tied to the traits.



## 3-4. 글쓰기 자동평가 및 자동첨삭 말뭉치·시스템 구축방안

### 말뭉치 구축의 고려사항(1/2)

- (1) 자동평가 및 자동첨삭의 구체적인 목표 수립 및 대상 세분화 필요
  - 글 유형에 부합하는 평가 기준 수립
    - (예: 주장이 담긴 사실과 독후감의 경우 평가 주안점에서 차이 존재)
  - 평가자에 대한 기대 수준에 맞는 평가 기준 수립
    - (예: 초등학생과 고등학생은 요구되는 글쓰기 수준에서 차이 존재)
  - 유형 및 수준에 따른 피드백의 세분화
    - 피드백에서 사용되는 용어
    - 피드백에서 수정하고자 하는 내용
      - ✓ 저학력자, 외국인의 경우 문법 등 유창성 검토
      - ✓ 에세이의 경우 주장과 근거 간 관계 등 담화구조 피드백

→ 잘 쓴 글에 대한 기준 및 올바른 피드백의 기준 수립

## 3-4. 글쓰기 자동평가 및 자동첨삭 말뭉치·시스템 구축방안

### 말뭉치 구축의 고려사항(2/2)

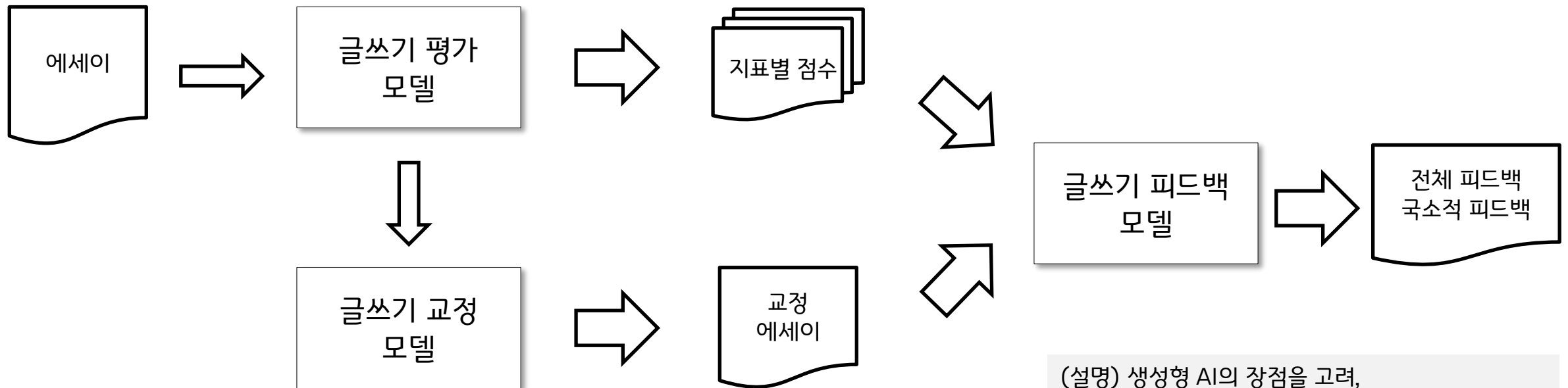
- (2) 진단체계 기반 점수, 피드백, 교정 말뭉치 구축 필요
  - 진단체계 지표별 점수 채점
  - 피드백이 필요한 대상(문장/단락 등) 선정 및 피드백 작성
    - (1) 전체 피드백: 내용(논점, 주장, 이유, 근거, 조직, 구성 등)
    - (2) 국소적 피드백: 문법, 표현, 규범/관습 준수 등
  - 피드백을 반영한 교정 글 작성
    - (1) 전체 피드백: 내용, 조직 관련한 단락 전체 수정
    - (2) 국소적 피드백: 문법, 표현, 규범/관습 준수 관련한 문장의 일부 표현 수정

### ♣ 글쓰기 자동평가 및 자동첨삭 말뭉치 관련 자문위원 의견(12.4.)

자문위원	내용
A위원	- 챗GPT 4.0 같은 경우 논리 추론 및 일관성 등 높은 성능을 보여주는 반면 논리적 오류도 다수 발견되는 연구 결과가 있음. 글쓰기 평가 시 문법 등 세밀한 평가 필요(전문가들의 관점 필요)
B위원	- 글쓰기에 있어 글을 잘 쓰는 지침 수립이 필요하고, 단계별/수준별 평가 필요(초/중/고) - 어떤 글을 어떻게 평가해야 할지 목표 설정 필요(평가 요소 및 지침 등)

## 3-4. 글쓰기 자동평가 및 자동첨삭 말뭉치·시스템 구축방안

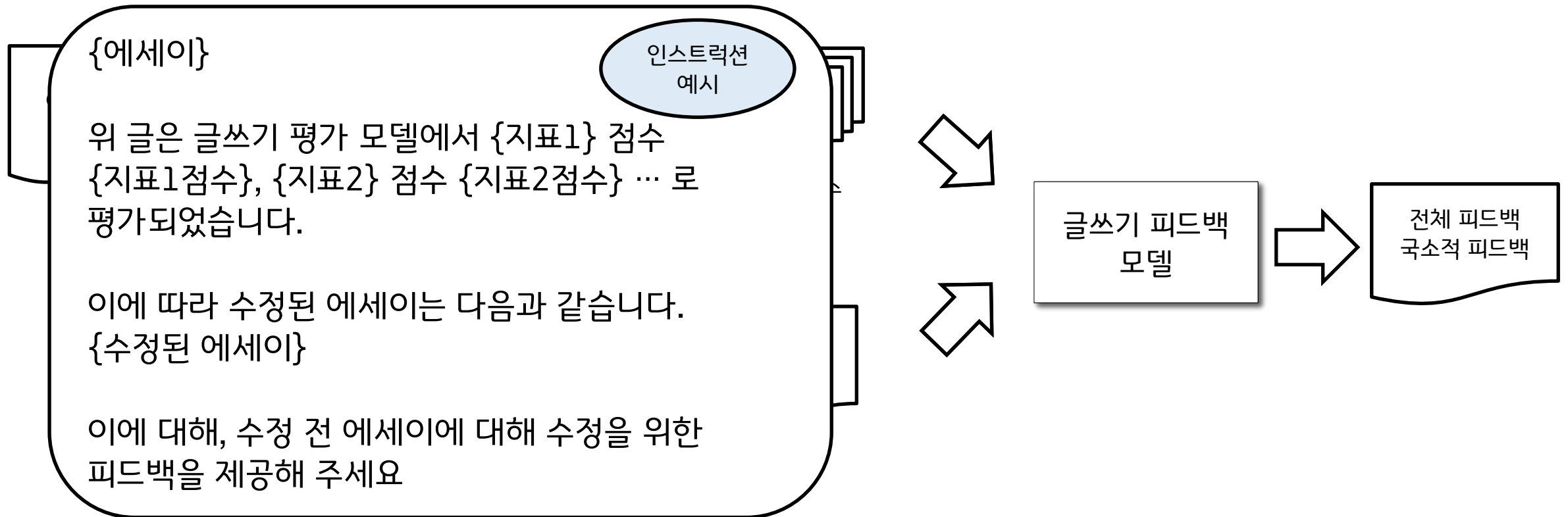
- (3) 생성형 AI 기반 자동평가 및 자동첨삭 기술 개발



(설명) 생성형 AI의 장점을 고려,  
 (1) 평가 점수를 기반으로 교정을 먼저 수행  
 (2) 교정 결과를 바탕으로 교정된 이유를 다시 설명하는 모델로 구성

## 3-4. 글쓰기 자동평가 및 자동첨삭 말뭉치·시스템 구축방안

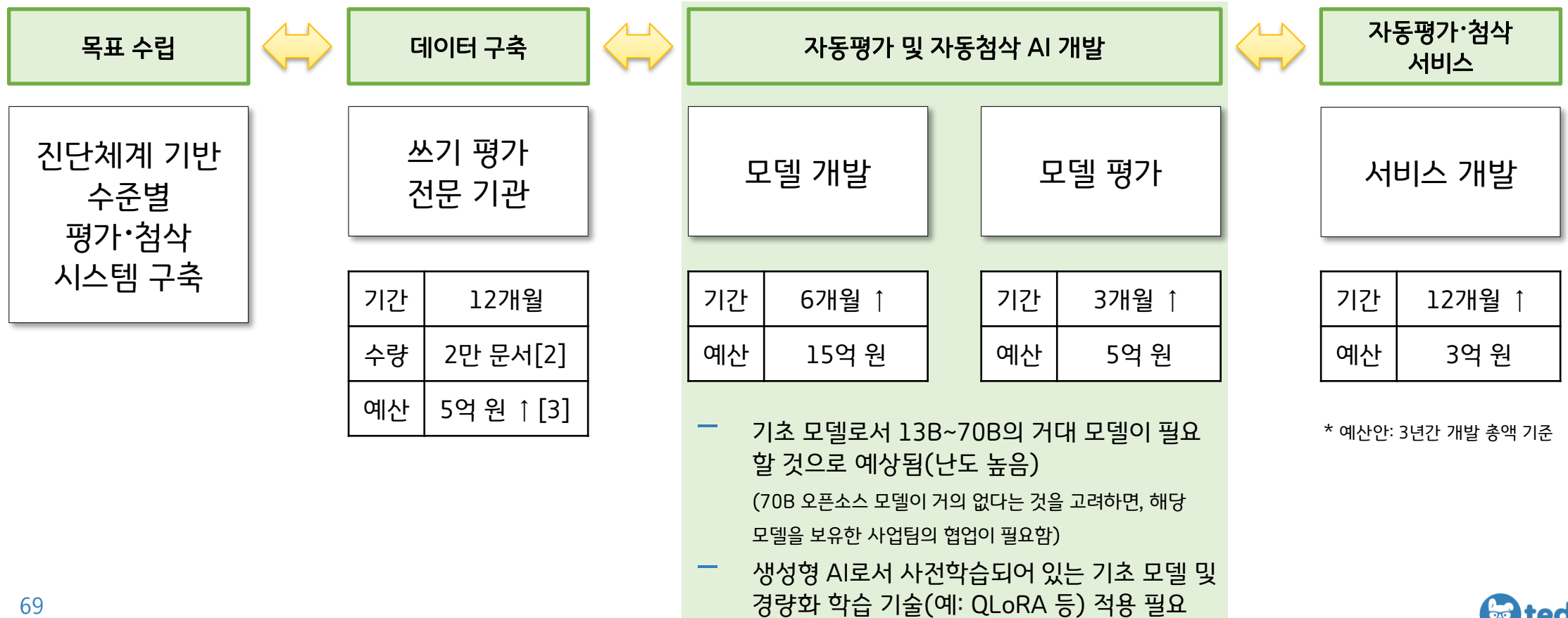
- (3) 자동평가 및 자동첨삭 기술 개발에 활용될 지시문(인스트럭션) 데이터 예시



## 3-4. 글쓰기 자동평가 및 자동첨삭 말뭉치·시스템 구축방안

- [1] 진단체계 데이터(평가 주석) 활용 전제
- [2] CoEDIT은 약 6.7만 문장 데이터로 구성
- [3] 문서당 첨삭 비용 2만원으로 계산(프리랜서 비용)

- (요약) 목표 수립, 데이터 구축, AI 개발 및 서비스화 단계로 구성 제안



# 4. 한국어 말뭉치 구축 중장기 계획안

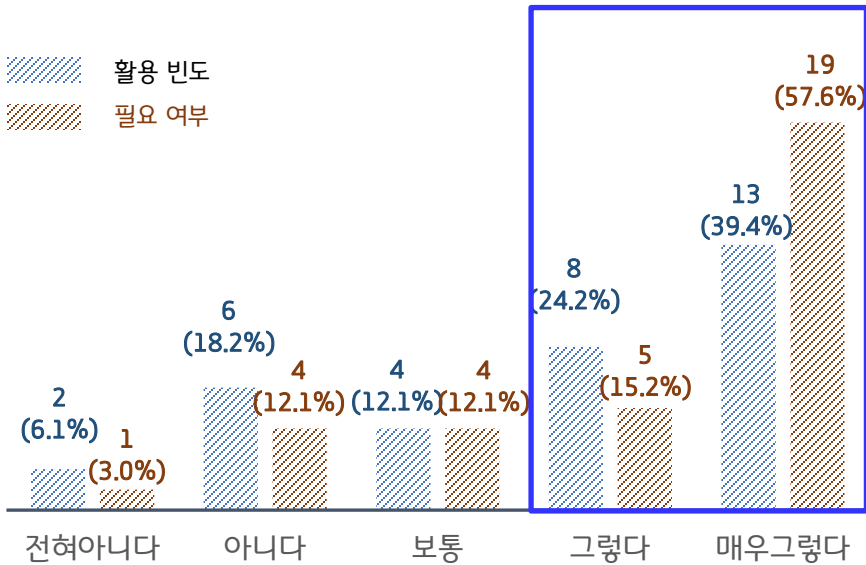
## 1) 말뭉치 구축 계획

1. 사업 개요
2. 환경 분석
3. 글쓰기 자동평가 및 자동첨삭 방안
4. 한국어 말뭉치 구축 중장기 계획안
  - 4-1. 말뭉치 구축 계획
    - 4-1-1. 한국어 말뭉치 수요 설문 조사 결과
    - 4-1-2. 자문 회의 결과
    - 4-1-3. 한국어 말뭉치 구축 중장기 추진 전략
    - 4-1-4. 한국어 말뭉치 구축 중장기 계획안
  - 4-2. ‘인공지능(AI)말평’ 평가용 말뭉치
  - 4-3. 인스트럭션 말뭉치 구축 방안
  - 4-4. 말뭉치 검증 개선 방안

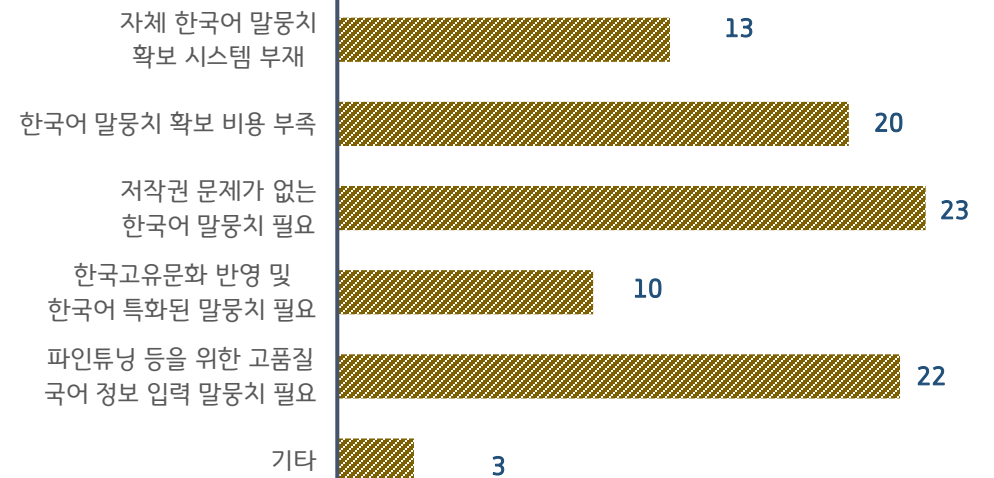
## 4-1-1. 한국어 말뭉치 수요 설문 조사 결과

- 설문 목적: 한국어 말뭉치 수요 조사를 통한 한국어 말뭉치 활용 목적 및 구축 필요성 확인
- 설문 기간: 2023년 9월 6일 ~ 2023년 9월 18일
- 설문 방법: 온라인 설문 조사(구글폼)

### 한국어 말뭉치 활용 빈도 및 필요 여부

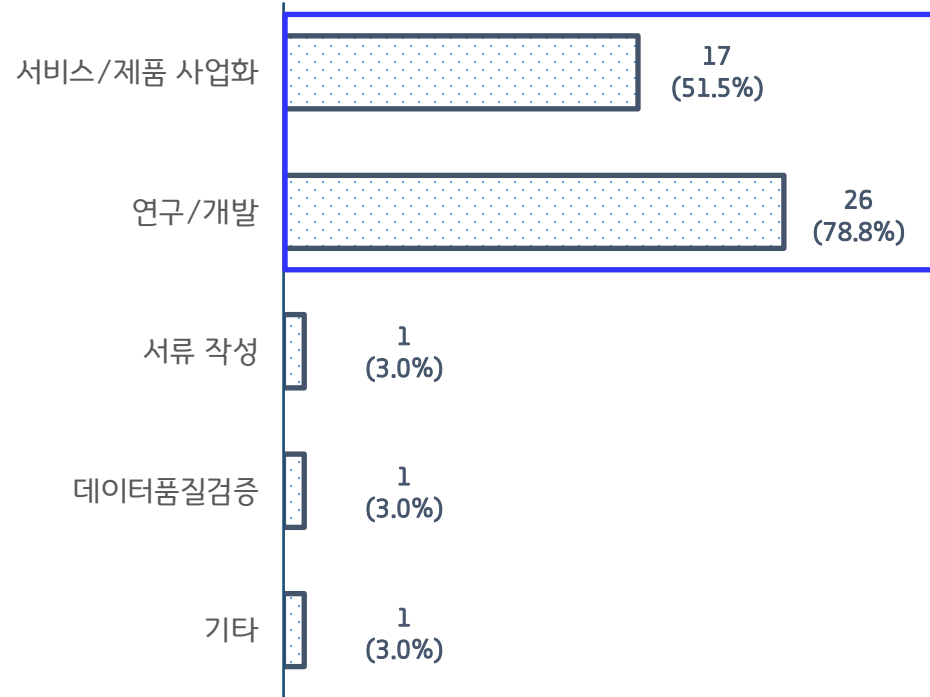


### 국립국어원의 한국어 말뭉치 활용 사유

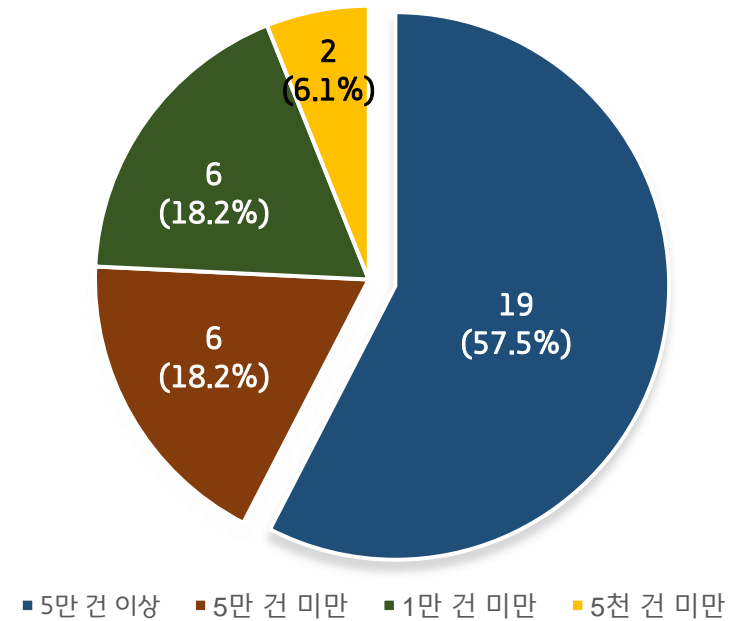


## 4-1-1. 한국어 말뭉치 수요 설문 조사 결과

한국어 말뭉치 활용 목적






한국어 말뭉치 필요 규모





## 4-1-2. 한국어 말뭉치 구축 중장기 계획 자문회의 결과

주제	한국어 말뭉치 구축 중장기 계획	일시/장소	'23.7.5. 10시~17시, 연세대 위당관														
참석 대상	[자문위원] 김영택(솔트룩스), 박재현(엔씨소프트), 이정수(플리토), 한지윤(업스테이지), 권오욱(ETRI), 김부근(중앙대) 김태욱(한양대), 이수영(ALI/KAIST) [국어원] 어문연구실장, 언어정보과장, 김문오 연구관, 이보라미 연구관 등 [사업팀] 함영균, 윤환선(테디썸), 김한샘(연세대), 임경태(서울과기대) 등 ----- 총 24명																
주요 내용	<table><tr><th>구분</th><th>내용</th><th>발표자</th></tr><tr><td>한국어 특성을 반영한 수준별 분석 말뭉치 구축 계획 </td><td><ul style="list-style-type: none"><li>○ AI의 한국어 수준을 높이기 위한 한국어 분석 정보 입력 레이블링 데이터의 다양화와 일관성 및 정확성 확보 필요함.</li><li>○ 전통적인 분석 정보 기반 과제의 산업 수요 변화에 따라 국어학적 복잡성을 반영한 자연어 이해 및 생성 말뭉치 필요함. (양은 적더라도 품질 좋으면 성능 향상에 기여 큼.)</li><li>○ 필요 말뭉치<ul style="list-style-type: none"><li>- 통합 정보 추출 말뭉치(개체 연결 생성, 관계 추출 생성, 문서/대화 요약)</li><li>- 이종(heterogeneous) 데이터 생성 기본 말뭉치(표·그림→텍스트)</li><li>- 지시문(Instruction) 기본 말뭉치(존댓말, 반말, 문어, 구어 등 수준별·연령별 특성 반영)</li><li>- 데이터 증강 활용 말뭉치, 설명 가능한(explainable) 언어 분석 말뭉치 등</li></ul></li></ul><table><tr><th>자문위원</th><th>내용</th></tr><tr><td>권오욱</td><td>- 타 기관과의 차별성을 고려, 국어원만의 특색을 가진 말뭉치 구축 필요함.</td></tr><tr><td>이정수</td><td>- 데이터 증강 및 설명 가능한 언어 분석 말뭉치 구축의 필요성 있음.</td></tr><tr><td>김영택</td><td>- 다방면으로 활용성이 높은 데이터 구축 필요함.</td></tr></table></td><td>임경태 교수 (서울과기대)</td></tr></table>			구분	내용	발표자	한국어 특성을 반영한 수준별 분석 말뭉치 구축 계획 	<ul style="list-style-type: none"><li>○ AI의 한국어 수준을 높이기 위한 한국어 분석 정보 입력 레이블링 데이터의 다양화와 일관성 및 정확성 확보 필요함.</li><li>○ 전통적인 분석 정보 기반 과제의 산업 수요 변화에 따라 국어학적 복잡성을 반영한 자연어 이해 및 생성 말뭉치 필요함. (양은 적더라도 품질 좋으면 성능 향상에 기여 큼.)</li><li>○ 필요 말뭉치<ul style="list-style-type: none"><li>- 통합 정보 추출 말뭉치(개체 연결 생성, 관계 추출 생성, 문서/대화 요약)</li><li>- 이종(heterogeneous) 데이터 생성 기본 말뭉치(표·그림→텍스트)</li><li>- 지시문(Instruction) 기본 말뭉치(존댓말, 반말, 문어, 구어 등 수준별·연령별 특성 반영)</li><li>- 데이터 증강 활용 말뭉치, 설명 가능한(explainable) 언어 분석 말뭉치 등</li></ul></li></ul> <table><tr><th>자문위원</th><th>내용</th></tr><tr><td>권오욱</td><td>- 타 기관과의 차별성을 고려, 국어원만의 특색을 가진 말뭉치 구축 필요함.</td></tr><tr><td>이정수</td><td>- 데이터 증강 및 설명 가능한 언어 분석 말뭉치 구축의 필요성 있음.</td></tr><tr><td>김영택</td><td>- 다방면으로 활용성이 높은 데이터 구축 필요함.</td></tr></table>	자문위원	내용	권오욱	- 타 기관과의 차별성을 고려, 국어원만의 특색을 가진 말뭉치 구축 필요함.	이정수	- 데이터 증강 및 설명 가능한 언어 분석 말뭉치 구축의 필요성 있음.	김영택	- 다방면으로 활용성이 높은 데이터 구축 필요함.	임경태 교수 (서울과기대)
	구분	내용	발표자														
한국어 특성을 반영한 수준별 분석 말뭉치 구축 계획 	<ul style="list-style-type: none"><li>○ AI의 한국어 수준을 높이기 위한 한국어 분석 정보 입력 레이블링 데이터의 다양화와 일관성 및 정확성 확보 필요함.</li><li>○ 전통적인 분석 정보 기반 과제의 산업 수요 변화에 따라 국어학적 복잡성을 반영한 자연어 이해 및 생성 말뭉치 필요함. (양은 적더라도 품질 좋으면 성능 향상에 기여 큼.)</li><li>○ 필요 말뭉치<ul style="list-style-type: none"><li>- 통합 정보 추출 말뭉치(개체 연결 생성, 관계 추출 생성, 문서/대화 요약)</li><li>- 이종(heterogeneous) 데이터 생성 기본 말뭉치(표·그림→텍스트)</li><li>- 지시문(Instruction) 기본 말뭉치(존댓말, 반말, 문어, 구어 등 수준별·연령별 특성 반영)</li><li>- 데이터 증강 활용 말뭉치, 설명 가능한(explainable) 언어 분석 말뭉치 등</li></ul></li></ul> <table><tr><th>자문위원</th><th>내용</th></tr><tr><td>권오욱</td><td>- 타 기관과의 차별성을 고려, 국어원만의 특색을 가진 말뭉치 구축 필요함.</td></tr><tr><td>이정수</td><td>- 데이터 증강 및 설명 가능한 언어 분석 말뭉치 구축의 필요성 있음.</td></tr><tr><td>김영택</td><td>- 다방면으로 활용성이 높은 데이터 구축 필요함.</td></tr></table>	자문위원	내용	권오욱	- 타 기관과의 차별성을 고려, 국어원만의 특색을 가진 말뭉치 구축 필요함.	이정수	- 데이터 증강 및 설명 가능한 언어 분석 말뭉치 구축의 필요성 있음.	김영택	- 다방면으로 활용성이 높은 데이터 구축 필요함.	임경태 교수 (서울과기대)							
자문위원	내용																
권오욱	- 타 기관과의 차별성을 고려, 국어원만의 특색을 가진 말뭉치 구축 필요함.																
이정수	- 데이터 증강 및 설명 가능한 언어 분석 말뭉치 구축의 필요성 있음.																
김영택	- 다방면으로 활용성이 높은 데이터 구축 필요함.																

## 4-1-2. 한국어 말뭉치 구축 중장기 계획 자문회의 결과

### 주요 내용

고질적 AI 언어 분석 문제점 및 해결 가능한 말뭉치 / AI 피드백 학습용 한국어 표준 말뭉치 구축 계획



구분	내용	발표자						
고질적 AI 언어 분석 문제점 및 해결 가능한 말뭉치 / AI 피드백 학습용 한국어 표준 말뭉치 구축 계획	<ul style="list-style-type: none"><li>○ 환각효과(Hallucination), 데이터의 편향성·강건성·신뢰성·독성 및 악의적인 사용 등 AI에서 자주 발생하는 문제들이 있음.<ul style="list-style-type: none"><li>- AI가 생성한 한국어 평가 방식에 대한 연구와 비윤리 말뭉치에 대한 고도화 작업 필요함.</li><li>- 차별적·편향적 표현 등을 파악하고 해결 가능한 신규 말뭉치 구축과 데이터 품질 평가 및 관리의 표준 마련 필요함.</li></ul></li><li>○ 대규모의 인스트럭션 튜닝과 강화 학습이 최근 거대 언어 모델(LLM)의 특징이며, 동일한 지식으로 의미 있는 답변이 가능하도록 하는 연구가 진행 중임.<ul style="list-style-type: none"><li>- 초기 언어 모델에서 생성형 AI로의 전환을 위한 인스트럭션 데이터와 AI 강화 학습을 위한 인간 평가 데이터 구축 필요함.</li></ul></li></ul>	함영균 대표 (테디섬)						
	<table><tr><th>자문위원</th><th>내용</th></tr><tr><td>김태욱</td><td><ul style="list-style-type: none"><li>- 말뭉치 활용 범위를 명확히 하고 한국어 특성을 반영한 데이터 구축 필요</li><li>- 거대 언어 모델을 만드는 대기업과의 협력을 통한 데이터 구축 필요함.</li></ul></td></tr><tr><td>이정수</td><td><ul style="list-style-type: none"><li>- 대기업 LLM 사용 시 중소기업은 관련 데이터의 활용성이 떨어지며, 특정 모델의 성능 향상에 도움이 되는 제한된 데이터로 개발되지 않도록 주의 필요함.</li><li>- 최근 기업의 경우 데이터는 충분하지만 이를 정제하기 위해 외부 인력을 활용 하는 파견 제도를 시행 중. 이러한 인력 활용 방식에서 야기된 고용 보상 등의 문제도 함께 해결 필요함.</li></ul></td></tr><tr><td>박재현</td><td><ul style="list-style-type: none"><li>- AI의 편향성 등의 문제 해결을 위한 비윤리, 부적절성 표현 데이터 필요함.</li></ul></td></tr></table>		자문위원	내용	김태욱	<ul style="list-style-type: none"><li>- 말뭉치 활용 범위를 명확히 하고 한국어 특성을 반영한 데이터 구축 필요</li><li>- 거대 언어 모델을 만드는 대기업과의 협력을 통한 데이터 구축 필요함.</li></ul>	이정수	<ul style="list-style-type: none"><li>- 대기업 LLM 사용 시 중소기업은 관련 데이터의 활용성이 떨어지며, 특정 모델의 성능 향상에 도움이 되는 제한된 데이터로 개발되지 않도록 주의 필요함.</li><li>- 최근 기업의 경우 데이터는 충분하지만 이를 정제하기 위해 외부 인력을 활용 하는 파견 제도를 시행 중. 이러한 인력 활용 방식에서 야기된 고용 보상 등의 문제도 함께 해결 필요함.</li></ul>
자문위원	내용							
김태욱	<ul style="list-style-type: none"><li>- 말뭉치 활용 범위를 명확히 하고 한국어 특성을 반영한 데이터 구축 필요</li><li>- 거대 언어 모델을 만드는 대기업과의 협력을 통한 데이터 구축 필요함.</li></ul>							
이정수	<ul style="list-style-type: none"><li>- 대기업 LLM 사용 시 중소기업은 관련 데이터의 활용성이 떨어지며, 특정 모델의 성능 향상에 도움이 되는 제한된 데이터로 개발되지 않도록 주의 필요함.</li><li>- 최근 기업의 경우 데이터는 충분하지만 이를 정제하기 위해 외부 인력을 활용 하는 파견 제도를 시행 중. 이러한 인력 활용 방식에서 야기된 고용 보상 등의 문제도 함께 해결 필요함.</li></ul>							
박재현	<ul style="list-style-type: none"><li>- AI의 편향성 등의 문제 해결을 위한 비윤리, 부적절성 표현 데이터 필요함.</li></ul>							


## 4-1-2. 한국어 말뭉치 구축 중장기 계획 자문회의 결과

### 주요 내용

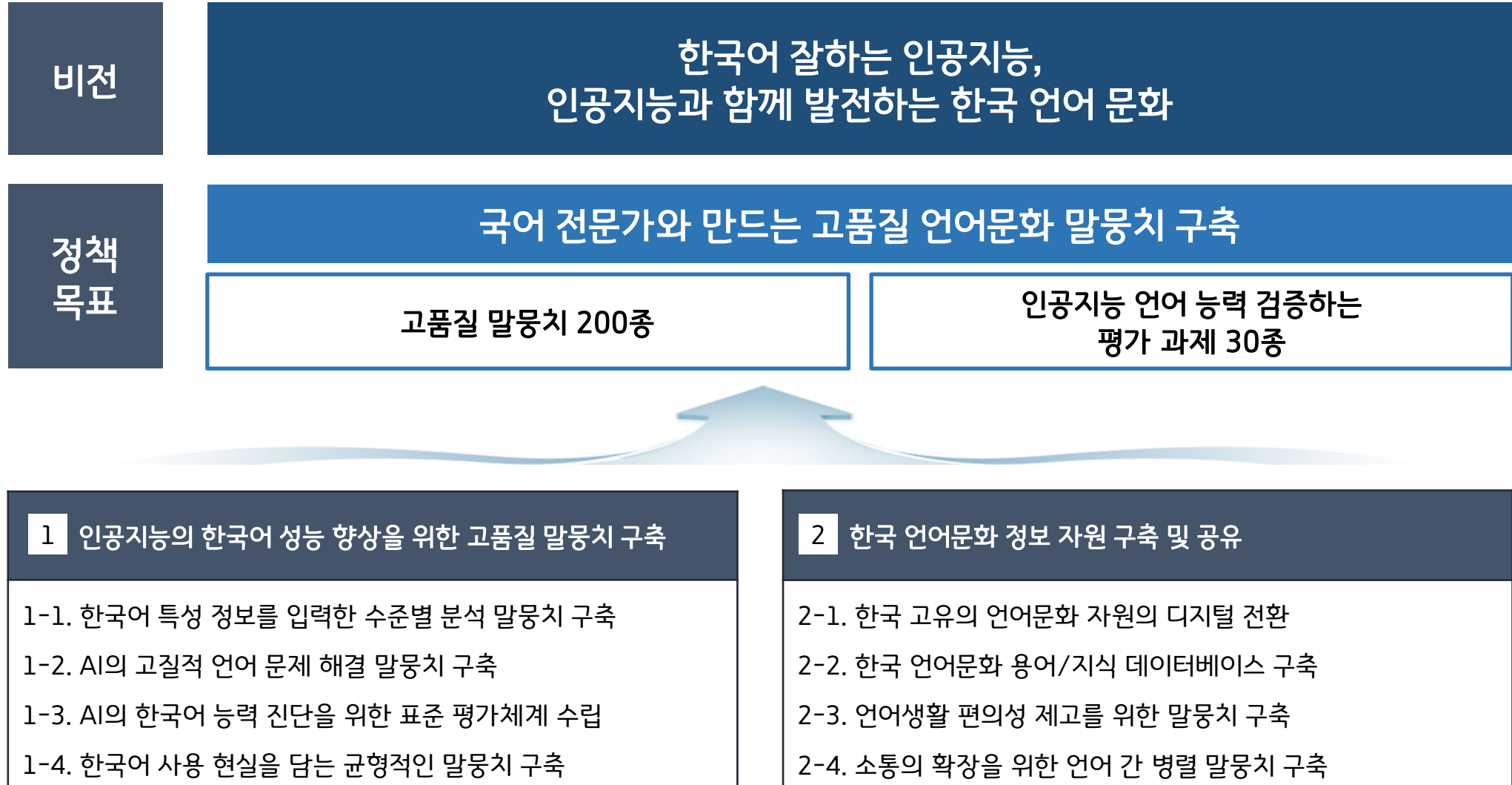
한국어 AI 수준 진단을 위한 평가 체계(벤치마크) 수립 계획



구분	내용	발표자								
한국어 AI 수준 진단을 위한 평가 체계(벤치마크) 수립 계획	<ul style="list-style-type: none"><li>○ 언어 능력을 개별 과제로 측정하는 것보다 종합적인 능력을 평가하는 추세이며, 초거대 언어 모델(LLM)의 언어 능력에 대한 평가도 필요함.</li><li>○ 초거대 언어 모델(LLM)은 자연어 이해와 생성이 동시에 가능하며, 인간 같은 (human-like) 결과를 산출하기 때문에, LLM 언어능력을 평가할 수 있는 방법 모색 및 신뢰도·타당도가 높은 인간 평가 기준 개발 필요함.</li><li>○ 생성형 AI 종합적 평가를 위한 평가 말뭉치 구축 제안<ul style="list-style-type: none"><li>- AI 추론 능력 평가 데이터(초등학교~대학교 수준별 단계적 구축), AI 요약 능력 평가, AI 상식 능력 평가, AI 신뢰성 평가, AI의 한국어 종합 능력 평가</li></ul></li></ul>	김한샘 교수 (연세대)								
	<table><tr><th>자문위원</th><th>내용</th></tr><tr><td>한지윤</td><td>- 벤치마크의 실질적인 의미를 설명하고 체계적으로 정의하는 것이 중요해 보이며, 언어 모델이 퓨샷(few-shot), 제로샷(zero-shot)으로 성능을 보이는 지금 시기에 전문가 활용을 통한 정밀한 데이터 구축이 필요함.</td></tr><tr><td>김부근</td><td>- 범언어적인 특성과 한국어의 특성이 반영된 데이터 구축 필요함.</td></tr><tr><td>이수영</td><td>- 세부 과제별 사용자 중심의 데이터 구축과 평가 지표 개발이 필요하며 한국어에 적용할 구체적인 핵심 평가 지표 개발도 필요함.</td></tr><tr><td>박재현</td><td>- AI의 이해 능력 뿐만 아니라 생성, 추론 모두를 평가 가능한 데이터가 필요하다는 점에 동의하며, 텍스트 외 멀티 모달에 대한 계획도 필요함.</td></tr></table>		자문위원	내용	한지윤	- 벤치마크의 실질적인 의미를 설명하고 체계적으로 정의하는 것이 중요해 보이며, 언어 모델이 퓨샷(few-shot), 제로샷(zero-shot)으로 성능을 보이는 지금 시기에 전문가 활용을 통한 정밀한 데이터 구축이 필요함.	김부근	- 범언어적인 특성과 한국어의 특성이 반영된 데이터 구축 필요함.	이수영	- 세부 과제별 사용자 중심의 데이터 구축과 평가 지표 개발이 필요하며 한국어에 적용할 구체적인 핵심 평가 지표 개발도 필요함.
자문위원	내용									
한지윤	- 벤치마크의 실질적인 의미를 설명하고 체계적으로 정의하는 것이 중요해 보이며, 언어 모델이 퓨샷(few-shot), 제로샷(zero-shot)으로 성능을 보이는 지금 시기에 전문가 활용을 통한 정밀한 데이터 구축이 필요함.									
김부근	- 범언어적인 특성과 한국어의 특성이 반영된 데이터 구축 필요함.									
이수영	- 세부 과제별 사용자 중심의 데이터 구축과 평가 지표 개발이 필요하며 한국어에 적용할 구체적인 핵심 평가 지표 개발도 필요함.									
박재현	- AI의 이해 능력 뿐만 아니라 생성, 추론 모두를 평가 가능한 데이터가 필요하다는 점에 동의하며, 텍스트 외 멀티 모달에 대한 계획도 필요함.									



## 4-1-3. 한국어 말뭉치 구축 중장기 추진 전략



## 4-1-4. 한국어 말뭉치 구축 중장기 계획안

전략	주요 과제	주요 특징
인공지능의 한국어 능력 향상을 위한 고품질 말뭉치 구축	1-1. 한국어 특성 정보를 입력한 수준별 분석 말뭉치 구축	• AI의 한국어 능력 향상을 위한 수준별 한국어 분석 정보 데이터 강화 및 한국어 특성을 반영한 이중데이터 확장(표/그림 ⇄ 텍스트)
	1-2. AI의 고질적 언어 문제 해결 말뭉치 구축	• 초거대 AI가 가지는 언어분석의 문제 해결을 위한 말뭉치를 구축하여 신뢰할 수 있는 한국어 AI 언어모델 개발에 기여
	1-3. AI의 한국어 능력 진단을 위한 표준 평가체계 수립	• AI의 한국어 이해·생성 능력 수준 진단을 위한 평가체계 수립 및 평가용 말뭉치 구축을 통해 인공지능 언어능력 향상 점검 기준 마련
	1-4. 한국어 사용 현실을 담은 균형적인 말뭉치 구축	• 인공지능 언어모델이 한국어 실태를 학습하고 한국어를 한국어답게 사용하도록 다양한 매체(신문·웹 등)와 다양한 변인(지역, 나이, 성별)을 고려한 한국어 말뭉치 구축
한국 언어문화 정보 자원 구축 및 공유	2-1. 한국 고유의 언어문화 자원의 디지털 전환	• 역사적·지역적으로 고유한 한국 언어문화를 계승·보전하고, 생성형 AI의 올바른 한국어 생성을 지원하기 위한 말뭉치 구축
	2-2. 한국 언어문화 용어/지식 데이터베이스 구축	• 한국어에 담긴 문화 관습적 특성을 상식 지식 데이터베이스로 구축하여 인공지능 모델의 상식 지식 모델 개발 지원
	2-3. 언어생활 편의성 제고를 위한 말뭉치 구축	• 국민의 언어생활에 편의를 제공할 수 있는 교육 및 말뭉치 적용 프로그램(글쓰기 자가 진단·참삭 서비스 등) 구현 개발에 필요한 말뭉치 구축
	2-4. 소통의 확장을 위한 언어 간 병렬 말뭉치 구축	• 우리 국민의 다양한 언어 사용 대응 및 특화형 인공지능 언어모델 개발을 위한 수준별 말뭉치 구축

## 4-1-4. 한국어 말뭉치 구축 중장기 계획안

주요 과제	구축 말뭉치	방향성	구축 계획
1-1. 한국어 특성 정보를 입력한 수준별 분석 말뭉치 구축	한국어 심화 분석 말뭉치	초거대 AI의 한국어 처리 능력 향상을 위해 한국어 특성(생략·축약 등)을 반영한 심화 분석 말뭉치 단계별 구축	<ul style="list-style-type: none"> <li>- 교착어인 한국어의 특성 정보를 담은 의존 구문 분석·생략어 복원 말뭉치 구축</li> <li>- 한국어 특화 상호 참조 해결 말뭉치 구축</li> <li>- 새로운 고유어(인명, 사건명, 장소명 등) 정보를 담은 개체명 및 개체 연결 말뭉치 구축</li> </ul>
	한국어 맥락 이해 말뭉치	생각의 사슬(Chain of Thought)*을 토대로 하여 설명 가능한 정보를 담은 한국어 대화 말뭉치 단계별 구축 * ‘생각의 사슬’은 인간이 문제를 해결하는 생각의 과정·절차를 상세하게 풀이한 정보	<ul style="list-style-type: none"> <li>- 대화 맥락 추론·함의 분석 말뭉치 구축</li> <li>- 대화 및 텍스트 요약 말뭉치 구축</li> </ul>
	이종데이터 말뭉치	멀티미디어 자료 기반 언어 생성을 위한 표→텍스트 변환 말뭉치, 그림→텍스트 변환 말뭉치 구축	<ul style="list-style-type: none"> <li>- 수준별 표→텍스트 변환 말뭉치 구축</li> <li>- 수준별 그림→텍스트 변환 말뭉치 구축</li> </ul>
1-2. AI의 고질적 언어 문제 해결 말뭉치 구축	문제 해결 말뭉치	초거대 AI 언어모델의 출력을 다각도로 평가 가능한 지표 마련 및 문제점 해결에 기여하는 데이터 세트 구축	<ul style="list-style-type: none"> <li>- AI의 문제점 파악을 위한 지표 마련* 및 말뭉치 구축**</li> <li>* 한국인터넷자율정책기구 자율 규제 디비(DB) 소위원회 등과 연계한 지표 마련</li> <li>** 부적절한 표현(욕설 등) 등 비윤리 관련 정보를 입력한 말뭉치로 구축</li> </ul>
	인스트럭션 말뭉치	AI의 답변을 사용자 맞춤형으로 요청할 수 있도록 한국어의 다양한 표현 형태를 반영한 인스트럭션 말뭉치 구축	<ul style="list-style-type: none"> <li>- 나이 등 사용자의 수준별 특성을 고려한 한국어 문체 변환 연구와 연계하여 구축</li> </ul>

## 4-1-4. 한국어 말뭉치 구축 중장기 계획안

주요 과제	구축 말뭉치	방향성	구축 계획
1-3. AI의 한국어 능력 진단을 위한 표준 평가체계 수립	한국어 능력 기본 평가 말뭉치	한국어 문법 정확도, 복잡도 평가, 의미 추론 능력 평가 및 부적절 표현 탐지 능력 평가를 위한 데이터 구축	<ul style="list-style-type: none"> <li>- AI 말뭉 시스템 구축·운영, 경진대회 개최 : 해외에서 운영하는 평가체계인 GLGE, SuperGLUE, Open LLM Leaderboard와 같이 AI의 한국어 능력을 점검 할 수 있는 다양한 과제 개발·운영하도록</li> <li>- 생성형 AI의 복합적 언어 지능 능력(언어 이해·생성) 및 한국어 능력을 평가하기 위한 종합 평가 과제 발굴</li> </ul>
	한국어 대화 능력 평가 말뭉치	한국어 의사소통에 대한 AI의 이해 능력 평가 및 인공지능의 멀티턴 대화 이해, 생성 능력(품질) 평가를 위한 데이터 구축	
	한국 문화 이해 능력 평가 말뭉치	한국어 언어문화(한국문화 특화)에 대한 인공지능의 문제 이해 능력, 생성 능력 종합 평가 및 문체 변환(방언생성)을 위한 데이터 구축	
1-4. 한국어 사용 현실을 담은 균형적인 말뭉치 구축	언어 변화 추적 말뭉치	신문, 일상대화, 온라인 자료 등 국민의 언어 생활과 밀접한 매체를 중심으로 언어 변화 추적 데이터 구축	<ul style="list-style-type: none"> <li>- 신문기사 말뭉치 : 사회 현상을 담은 언어자료로 새로운 용어·개념을 포함하는 원시 말뭉치로 지속 구축</li> <li>- 일상대화 말뭉치 : 지역별·나이별·성별 인구 분포를 고려한 자료로 대화 자료 외 1인 독백(강연·낭독 등), 회의 등 다양한 유형의 대화 말뭉치 추가 구축</li> <li>- 온라인 자료 말뭉치 : 언어 변화가 가장 빠르게 일어나는 온라인상의 자료를 수집하고, 기업 등에서 구축이 어려운 메신저 대화 등 사적 언어자료를 저작권 해결하여 구축·배포</li> </ul>
	언어 사용 균형 말뭉치	언어 변화 추이를 주기적으로 관찰(5년 주기)하고, 국민의 언어 사용 실태를 연구할 수 있는 기본 데이터 구축	<ul style="list-style-type: none"> <li>- 언어 변화 추적 말뭉치에서 조사되지 못한 다양한 장르의 텍스트(책, 잡지, 시나리오, 유튜브 대본 등)를 구성하여 폭넓은 언어 현상을 담아낼 수 있도록 연구 및 말뭉치로 구축</li> </ul>



## 4-1-4. 한국어 말뭉치 구축 중장기 계획안

주요 과제	구축 말뭉치	방향성	구축 계획
2-1. 한국 고유의 언어문화 자원의 디지털 전환	구술 발화 말뭉치	한국어의 다양성 보전 및 지역별·세대별 문화를 분석할 수 있는 구술발화 데이터 구축	- 매년 9지역(9개도 각 1지점), 3세대(20대, 50대, 70대 이상), 남녀 제보자를 선정하여 지역·세대·성별에 따른 자연스러운 한국어 음성 말뭉치 구축
	국어사 말뭉치	국어 역사 문헌 자료 종합 정리 및 장르별·시기별 균형을 갖춘 대규모 국어사 말뭉치 구축	- 21세기 세종계획에서 구축한 "불경언해류"의 말뭉치를 포함하여, 19세기~20세기 초의 일상생활이 반영된 편지, 일기, 소설, 잡지 등 장르 확대 구축
2-2. 한국 언어문화 용어/지식 데이터베이스 구축	전문 지식기반 데이터	전문 분야 AI 언어 모델 개발을 위한 지식 기반 자료 구축	- 표준분류체계를 적용한 분야별 전문용어 사전 구축 : 국가과학기술 표준분류체계(2023년)를 기준으로 대분류 22개, 중분류 277개로 구성  - 온톨로지 기반 지식 용어 관계망 구축, 다국어 어휘망(바벨넷 등) 연결 : 사물과 사물 간의 관계 및 여러 개념을 컴퓨터가 처리할 수 있는 형태로 표현한 것
	한국형 상식 지식 데이터	한국인이 공유하는 문화적 상식과 맥락을 이해하고 그 수준을 추론할 수 있는 언어 자료 구축	- 한국어 어휘망과 전문용어 어휘망의 연계를 통한 연결 데이터로 구축 : 개체명 연결 사전과 확대·연계하여 인공지능 사전 학습 임베딩 자료, 미세 조정 등에 활용 가능한 상식 지식베이스로 구축



## 4-1-4. 한국어 말뭉치 구축 중장기 계획안

주요 과제	구축 말뭉치	방향성	구축 계획
2-3. 언어생활 편의성 제고를 위한 말뭉치 구축	쓰기·채점 말뭉치	다양한 텍스트 유형에 대해 수준별 쓰기 및 채점 말뭉치 구축	<ul style="list-style-type: none"> <li>- 수준별(초등·중등·성인)·유형별 국어능력 진단체계 지표 개발('23년~)</li> <li>- 인공지능 모델 개발에 활용 가능한 글쓰기 및 채점 자료 수집·구축</li> <li>- 외국인들의 한국어 쓰기 평가 자료 구축</li> </ul>
	국어 지식 및 탐색 말뭉치	글쓰기 수정·보완을 위한 참조 데이터 구축	<ul style="list-style-type: none"> <li>- 가나다 전화 등 국어 상담 서비스 및 공공언어 감수(보도자료, 보고서, 기안문, 안내문 등) 결과 자료를 인공지능 학습 데이터로 가공</li> <li>- 국어능력 진단 지표를 고려한 탐색(피드백) 말뭉치 구축</li> <li>: 국어원 구축 말뭉치 등을 활용하여 인공지능을 활용한 국민의 언어생활 편의 서비스 구현과 연계</li> </ul>
2-4. 소통의 확장을 위한 언어 간 병렬 말뭉치 구축	한국어 학습자 말뭉치	외국인 학습자의 수준별 한국어 사용 정보를 담은 데이터 구축	<ul style="list-style-type: none"> <li>- 미국, 중국, 일본, 베트남 등 101개 언어권의 한국어 학습자 글쓰기 자료 수집·구축</li> </ul>
	한국어-외국어 병렬 말뭉치	언어 연구 및 인공지능 번역 기술 향상에 활용할 데이터 부족 언어 대상 병렬 말뭉치 구축(*8개 언어) * 베트남어, 인도네시아어, 태국어, 힌디어, 캄보디아어, 타갈로그어, 러시아어, 우즈베크어	<ul style="list-style-type: none"> <li>- 다양한 장르*의 구어체와 문어체 원천 자료를 수집, 자연스러운 번역을 위한 균형 말뭉치 구축</li> <li>* 일상 대화, 신문 기사(정치, 경제, 사회 등), 시나리오, 유튜브 대본 등 다양한 장르 번역</li> </ul>
	한국어-수어 병렬 말뭉치	한국어→한국수어 양방향 번역 모델 개발을 위한 한국수어 번역 기준 마련 및 방법론 표준화	<ul style="list-style-type: none"> <li>- 일상생활 및 수어통역 우선 필요 분야(의료, 교육, 방송 등)의 원천자료를 확보하여 동작인식 기술 개발 등에 활용할 한국수어 영상 제작 및 의미 정보 등 매칭</li> </ul>
	한국어 묵자-점자 병렬 말뭉치	점역·역점역 성능 향상을 위한 인공지능 기반 학습용 점자 말뭉치 구축	<ul style="list-style-type: none"> <li>- 고성능 점역 및 점자 문서 편집 소프트웨어 개발·고도화</li> </ul>

## 4-1-4. 한국어 말뭉치 구축 중장기 계획안

전략	과제	말뭉치 유형	'23년			'24년			'25년			'26년			'27년		
			종류	규모	예산(억)	종류	규모	예산(억)	종류	규모	예산(억)	종류	규모	예산(억)	종류	규모	예산(억)
[1] 인공지능의 한국어 능력 향상을 위한 고품질 말뭉치 구축	1-1. 수준별 분석 말뭉치	분석				1. 의존 구문 분석/생략어 복원	각 150만 어절	3.5	1. 의존 구문 분석/생략어 복원	각 150만 어절	3.5	1. 의존 구문 분석/생략어 복원	각 150만 어절	3.5	1. 의존 구문 분석/생략어 복원	각 150만 어절	3.5
		분석							2. 개체명	500만 어절		32. 상호참조 해결	300만 어절		32. 상호참조 해결	300만 어절	5
		분석							3. 개체 연결	500만 어절		33. 개체명	500만 어절		33. 개체명	500만 어절	3
		분석										4. 개체 연결	500만 어절		34. 개체 연결	500만 어절	3
		분석										5. 개체명 사전	1종		35. 개체명 사전	1종	3
	1-2. AI의 고질적 언어 문제 해결 말뭉치	분석	1. 요약-국회 회의록	300건		2. 대화 및 텍스트 요약	5,000건	3.5	3.5. 대화 및 텍스트 요약	5,000건		3.5. 대화 및 텍스트 요약	5,000건		3.5. 대화 및 텍스트 요약	5,000건	3.5
		분석	2. 요약-주제대화	3,000건		3. 추론-함의 분석	6,000건		35. 추론-함의 분석	6,000건		37. 추론-함의 분석	6,000건		37. 추론-함의 분석	6,000건	3
		분석	3. 추론-함의 분석	6,000건		2.5.4. 이중데이터(표-텍스트)	3,000건		26. 이중데이터(그림-텍스트)	3,000건		28. 이중데이터(표-텍스트)	3,000건		28. 이중데이터(표-텍스트)	3,000건	2
		분석				5. 인스트럭션 말뭉치 설계	1종		2.5.7. 인스트럭션 말뭉치	1종		59. 이중데이터(그림-텍스트)	3,000건		29. 이중데이터(그림-텍스트)	3,000건	2
		분석										10. 인스트럭션 말뭉치	1종		6.10. 인스트럭션 말뭉치(1)	1종	10.5
		분석													11. 인스트럭션 말뭉치(2)	1종	
		분석													12. 인스트럭션 말뭉치(3)	1종	
	1-3. AI 표준 평가 체계	AI평가용	4. 부적절 표현 탐지 및 분류	1종		6. 부적절 표현 탐지 및 분류	1종		8. 부적절 표현 탐지 및 분류	1종		11. 부적절 표현 탐지 및 분류	1종		13. 부적절 표현 탐지 및 분류	1종	
		AI평가용	5. 감정 분석	1종		7. AI 이해 능력 평가용(1)	1종		9. AI 이해 능력 평가용(1)	1종		12. AI 이해 능력 평가용(1)	1종		14. AI 이해 능력 평가용(1)	1종	
		AI평가용	6. 함의 분석	1종	3.5	8. AI 이해 능력 평가용(2)	1종		10. AI 이해 능력 평가용(2)	1종		13. AI 이해 능력 평가용(2)	1종		15. AI 이해 능력 평가용(2)	1종	
		AI평가용	7. 문장 생성	1종		9. AI 이해 능력 평가용(3)	1종	6.6	11. AI 이해 능력 평가용(3)	1종	8	14. AI 이해 능력 평가용(3)	1종	10	16. AI 이해 능력 평가용(3)	1종	10
		AI평가용	8. 이야기 완성	1종		10. AI 생성 능력 평가용(1)	1종		12. AI 생성 능력 평가용(1)	1종		15. AI 생성 능력 평가용(1)	1종		17. AI 생성 능력 평가용(1)	1종	
		AI평가용				11. AI 생성 능력 평가용(2)	1종		13. AI 생성 능력 평가용(2)	1종		16. AI 생성 능력 평가용(2)	1종		18. AI 생성 능력 평가용(2)	1종	
		AI평가용													19. AI 생성 능력 평가용(3)	1종	
	1-4. 한국어 사용 현실 균형 말뭉치	원시	9. 신문	1억 어절	3.5	12. 신문	2억 어절	5	14. 신문	2억 어절	5	17. 신문	2억 어절	5	20. 신문	2억 어절	5
		원시	10. 일상대화(음성+텍스트)	500시간		13. 일상대화(음성+텍스트)	500시간		15. 일상대화(음성+텍스트)	500시간		18. 일상대화(음성+텍스트)	500시간		21. 일상대화(음성+텍스트)	500시간	8
		원시							16. 온라인 자료	400,000건		19. 온라인 자료	400,000건		22. 온라인 자료	400,000건	5
		원시							17. 균형 말뭉치	1,000만 어절	10						

## 4-1-4. 한국어 말뭉치 구축 중장기 계획안

전략	과제	말뭉치 유형	'23년			'24년			'25년			'26년			'27년			
			종류	규모	예산(억)	종류	규모	예산(억)	종류	규모	예산(억)	종류	규모	예산(억)	종류	규모	예산(억)	
[2] 한국어언문화 정보 자원 구축 및 공유 보전	2-1. 고유 언어문화 자 원의 디지털 전환	원시	11. 구술 발화(지역어)	100시간	2.5	14. 구술 발화(지역어)	100시간	2.5	18. 구술 발화(지역어)	100시간	2.5	20. 구술 발화(지역어)	100시간	2.5	23. 구술 발화(지역어)	100시간	2.5	
		원시	12. 문헌(정비+구축)	370만 어절	2.4	15. 문헌	40만 어절	0.5	19. 문헌	40만 어절	0.5	21. 문헌	40만 어절	0.5	24. 문헌	40만 어절	0.5	
	2-2. 용어·지식데이터 베이스	기타	용어 분류	200,000건	2.5	용어 분류	200,000건	2.3	용어 분류	200,000건	2.5	용어 분류	200,000건	2.5	용어 분류	200,000건	2.5	
		기타				상식·지식 베이스	1종	1	상식·지식 베이스	1종	1	22. 상식·지식 베이스 기반 문화 말뭉치	1종	1.5	25. 상식·지식 베이스 기반 문화 말뭉치(1)	1종	2	
		기타													26. 상식·지식 베이스 기반 문화 말뭉치(2)	1종		
	2-3. 언어생활 편의성 제고 말뭉치	원시	13. 글쓰기 자료	9,000건	10	16. 글쓰기 자료	9,000건	10	20. 글쓰기 자료	9,000건	10	23. 글쓰기 자료	9,000건	10	27. 글쓰기 자료	9,000건	10	
		분석	14. 글쓰기 채점 자료	5,000건		17. 글쓰기 채점 자료	5,000건		21. 글쓰기 채점 자료	5,000건		24. 글쓰기 채점 자료	5,000건		28. 글쓰기 채점 자료	5,000건		
		분석				18. 교정·첨삭 자료 기초 말뭉치	5,000건	0.5	22. 교정·첨삭 말뭉치	5,000건	3	25. 교정·첨삭 말뭉치	5,000건	3	29. 교정·첨삭 말뭉치	5,000건	3	
	2-4. 언어 간 병렬 말뭉 치	원시	15. 학습자 자료	150만 어절	3	19. 학습자 자료	150만 어절	3	23. 학습자 자료	150만 어절	3	26. 학습자 자료	150만 어절	3	30. 학습자 자료	150만 어절	3	
		원시	16. 한-외(베트남어)			20. 한-외(베트남어)			24. 한-외(베트남어)			27. 한-외(베트남어)			31. 한-외(베트남어)			
		원시	17. 한-외(인도네시아어)			21. 한-외(인도네시아어)			25. 한-외(인도네시아어)			28. 한-외(인도네시아어)			32. 한-외(인도네시아어)			
		원시	18. 한-외(태국어)			22. 한-외(태국어)			26. 한-외(태국어)			29. 한-외(태국어)			33. 한-외(태국어)			
		원시	19. 한-외(인도 힌디어)			23. 한-외(인도 힌디어)			27. 한-외(인도 힌디어)			30. 한-외(인도 힌디어)			34. 한-외(인도 힌디어)			
		원시	20. 한-외(캄보디아어)	1,000만 어절	50	24. 한-외(캄보디아어)	1,000만 어절	50	28. 한-외(캄보디아어)	1,000만 어절	50	31. 한-외(캄보디아어)	1,000만 어절	50	35. 한-외(캄보디아어)	1,000만 어절	50	
		원시	21. 한-외 (필리핀 타갈로그어)			25. 한-외 (필리핀 타갈로그어)			29. 한-외 (필리핀 타갈로그어)			32. 한-외(필리핀 타갈로그어)			36. 한-외(필리핀 타갈로그어)			
		원시	22. 한-외(러시아어)			26. 한-외(러시아어)			30. 한-외(러시아어)			33. 한-외(러시아어)			37. 한-외(러시아어)			
		원시	23. 한-외(우즈베크어)			27. 한-외(우즈베크어)			31. 한-외(우즈베크어)			34. 한-외(우즈베크어)			38. 한-외(우즈베크어)			
		원시	24. 한국어-수어	100만 어절	18	28. 한국어-수어	100만 어절	18	32. 한국어-수어	100만 어절	18	35. 한국어-수어	100만 어절	18	39. 한국어-수어	100만 어절	18	
		원시	25. 한국어 목자-점자	100만 어절	10	29. 한국어 목자-점자	100만 어절	10	33. 한국어 목자-점자	100만 어절	10	36. 한국어 목자-점자	100만 어절	10	40. 한국어 목자-점자	100만 어절	10	
	합계			25		117	29		132	33		160	36		161	40		168

\*기울임체 및 회색 음영 부분은 말뭉치 합계 종수에서는 제외됨.

# 4. 한국어 말뭉치 구축 중장기 계획안

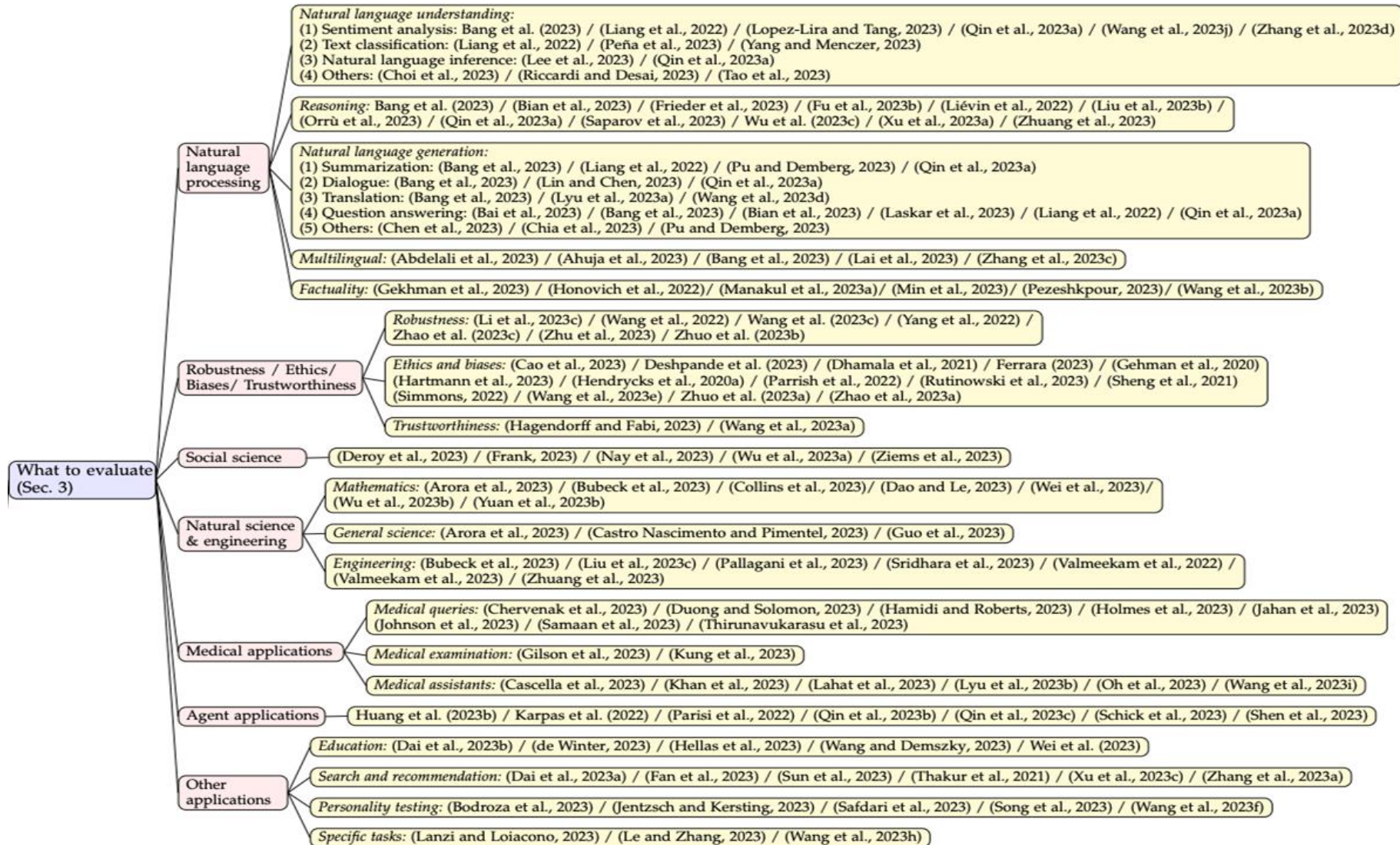
## 2) ‘인공지능(AI)말평’ 평가용 말뭉치

1. 사업 개요
2. 환경 분석
3. 글쓰기 자동평가 및 자동첨삭 방안
4. 한국어 말뭉치 구축 중장기 계획안
  - 4-1. 말뭉치 구축 계획
  - 4-2. ‘인공지능(AI)말평’ 평가용 말뭉치
    - 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례
    - 4-2-2. 인공지능 언어 능력 평가용 말뭉치 사례 요약
    - 4-2-3. AI말평 평가용 말뭉치 구축 방향 제안
    - 4-2-4. AI말평 방향성 제안
  - 4-3. 인스트럭션 말뭉치 구축 방안
  - 4-4. 말뭉치 검증 개선 방안

## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

- 초거대 언어 모델(LLM)의 성능 평가는 작업과 영역을 선택하여 강점 및 약점을 입증하는 것이 중요함.
- 초거대 언어 모델(LLM)의 평가는 도메인에 따라 다음 7가지 범주로 나눌 수 있음.
  1. 자연어 처리: 자연어 추론(Natural Language Inference, NLI), 추론(Reasoning), 생성(Natural Language Generation, NLG) 및 다국어 과업 등이 포함됨.
  2. 글쓰기 및 정보 처리: 강건성, 윤리, 편향, 신뢰성 등
  3. 의료 응용(어플리케이션): 의료 Q&A, 건강 검진, 의료 교육 및 의료 도움(어시스턴트) 등.
  4. 사회 과학
  5. 자연 과학 및 공학: 수학, 일반 과학 및 공학 포함.
  6. 에이전트 응용(어플리케이션): 초거대 언어 모델(LLM)을 에이전트로 사용하는 어플리케이션
  7. 기타 응용(어플리케이션): 검색, 추천, 개인 성향 테스트 등

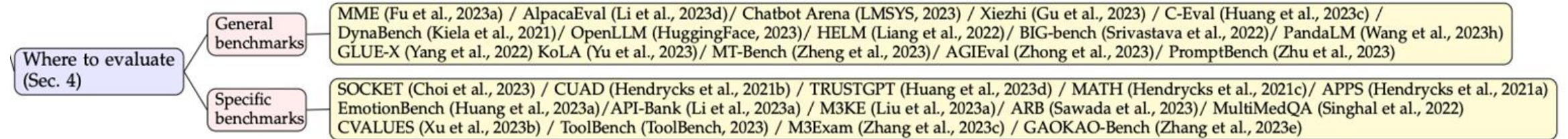
## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례





## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

- 평가 벤치마크는 목적에 따라 일반적인 벤치마크와 특정 과업을 위한 벤치마크로 구분함.



Benchmark	Focus	Domain	Evaluation Criteria
SOCKET (Choi et al., 2023)	Social knowledge	Specific downstream task	Social language understanding
MME (Fu et al., 2023a)	Multimodal LLMs	General language task	Ability of perception and cognition
Xiezhi (Gu et al., 2023)	Comprehensive domain knowledge	General language task	Overall performance across multiple benchmarks
CUAD (Hendrycks et al., 2021b)	Legal contract review	Specific downstream task	Legal contract understanding
TRUSTGPT (Huang et al., 2023d)	Ethic	Specific downstream task	Toxicity, bias, and value-alignment
MMLU (Hendrycks et al., 2020b)	Text models	General language task	Multitask accuracy
MATH (Hendrycks et al., 2021c)	Mathematical problem	Specific downstream task	Mathematical ability
APPS (Hendrycks et al., 2021a)	Coding challenge competence	Specific downstream task	Code generation ability
C-Eval (Huang et al., 2023c)	Chinese evaluation	General language task	52 Exams in a Chinese context
EmotionBench (Huang et al., 2023a)	Empathy ability	Specific downstream task	Emotional changes
OpenLLM (HuggingFace, 2023)	Chatbots	General language task	Leaderboard rankings
DynaBench (Kiela et al., 2021)	Dynamic evaluation	General language task	NLI, QA, sentiment, and hate speech
Chatbot Arena (LMSYS, 2023)	Chat assistants	General language task	Crowdsourcing and Elo rating system
AlpacaEval (Li et al., 2023d)	Automated evaluation	General language task	Metrics, robustness, and diversity
HELM (Liang et al., 2022)	Holistic evaluation	General language task	Multi-metric
API-Bank (Li et al., 2023a)	Tool utilization	Specific downstream task	API call, retrieval, and planning
M3KE (Liu et al., 2023a)	Multi-task	Specific downstream task	Multi-task accuracy
ARB (Sawada et al., 2023)	Advanced reasoning ability	Specific downstream task	Multidomain advanced reasoning ability
BIG-bench (Srivastava et al., 2022)	Capabilities and limitations of LMs	General language task	Model performance and calibration
MultiMedQA (Singhal et al., 2022)	Medical QA	Specific downstream task	Model performance, medical knowledge, and reasoning ability
CVALUES (Xu et al., 2023b)	Safety and responsibility	Specific downstream task	Alignment ability of LLMs
ToolBench (ToolBench, 2023)	Software tools	Specific downstream task	Execution success rate
PandaLM (Wang et al., 2023h)	Instruction tuning	General language task	Winrate judged by PandaLM
GLUE-X (Yang et al., 2022)	OOD robustness for NLU tasks	General language task	OOD robustness
KoLA (Yu et al., 2023)	Knowledge-oriented evaluation	General language task	Self-contrast metrics
AGIEval (Zhong et al., 2023)	Human-centered foundational models	General language task	General
PromptBench (Zhu et al., 2023)	Adversarial prompt resilience	General language task	Adversarial robustness
MT-Bench (Zheng et al., 2023)	Multi-turn conversation	General language task	Winrate judged by GPT-4
M3Exam (Zhang et al., 2023c)	Human exams	Specific downstream task	Task-specific metrics
GAOKAO-Bench (Zhang et al., 2023e)	Chinese Gaokao examination	Specific downstream task	Accuracy and scoring rate

## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### Huggingface Leaderboard

- Open LLM을 순위 책정 및 평가
- Eleuther AI 언어 모델 평가 지표인 Harness를 사용하여 4가지의 주요 벤치마크에서 모델 평가
  - AI2 Reasoning Challenge: 초등학교 과학 문제 세트
  - HellaSwag(10-shot): 인간에게는 쉽지만 SOTA 모델에게는 까다로운 상식 추론 테스트
  - MMLU(5-shot): 텍스트 모델의 멀티 태스크(Multi-task) 정확도를 측정하는 테스트  
(초등 수학, 미국 역사, 컴퓨터 과학, 법률 등을 포함한 57개의 과제)
  - TruthfulQA(zero-shot): 허위 사실을 생성하는지 테스트하는 데이터 세트

- HuggingFaceH4/open\_llm\_leaderboard에서 순위표(리더보드) 제공

T	Model	Average ↑	ARC	HellaSwag	MMLU	TruthfulQA
◆	<a href="#">ValiantLabs/ShiningValiant</a>	74.17	72.95	87.88	70.97	64.88
◆	<a href="#">ICBU-NPU/FashionGPT-70B-V1.2</a>	74.11	73.04	88.15	70.11	65.15
◆	<a href="#">sequelbox/StellarBright</a>	74.1	72.95	87.82	71.17	64.46
◆	<a href="#">Riiid/sheep-duck-llama-2-70b-v1.1</a>	74.07	73.04	87.81	70.84	64.58
◆	<a href="#">AIDC-ai-business/Marcoroni-70B-v1</a>	74.06	73.55	87.62	70.67	64.41
◆	<a href="#">ICBU-NPU/FashionGPT-70B-V1.1</a>	74.05	71.76	88.2	70.99	65.26
?	<a href="#">adonlee/LLaMA_2_70B_LoRA</a>	73.9	72.7	87.55	70.84	64.52
◆	<a href="#">uni-tianyan/Uni-TianYan</a>	73.81	72.1	87.4	69.91	65.81
?	<a href="#">Riiid/sheep-duck-llama-2</a>	73.69	72.35	87.78	70.82	63.8
?	<a href="#">Riiid/sheep-duck-llama-2</a>	73.67	72.27	87.78	70.81	63.8



## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례


### Chatbot Arena

- GPT-4, LLaMA2-70B-chat과 같은 챗봇(Chatbot) 기반의 모델을 평가하기 위한 순위표(리더보드)
- 다음의 세 가지 벤치마크를 기초에 두고 있음
  - 챗봇 아레나(Chatbot Arena): 크라우드소싱 기반의(Crowdsourced) 무작위 배틀 플랫폼으로 9만 명 이상의 사용자 투표를 통해 등급 계산
  - MT-Bench: Multi-turn Questions, GPT-4를 통하여 모델의 응답을 채점
  - MMLU(5-shot): 허깅페이스의 Open LLM Leaderboard의 사용 양식과 동일
- 22개 모델에 대해 33,000개의 대화에 대해서 모델 A, B의 답변 중 선호되는 모델에 대해 투표하고 집계
- lmsys/chatbot\_arena\_conversations 데이터 세트 공개

## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### Alpaca Eval Leaderboard

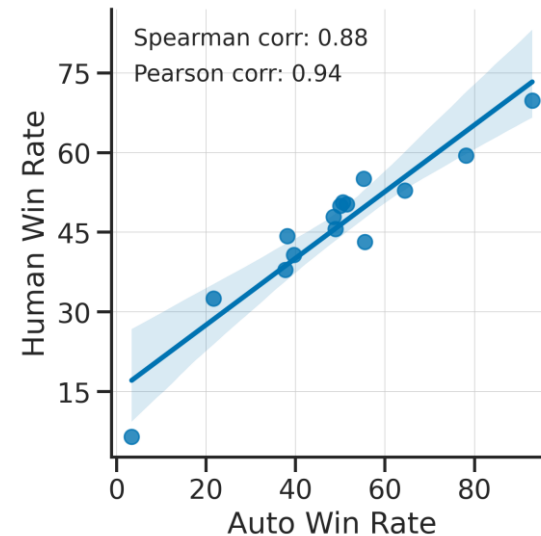
- 챗GPT와 같은 instruction-following 모델 평가는 인간과 상호작용을 해야 하기 때문에 시간과 비용 필요함.
- AlpacaFarm 평가 세트에서 주어진 모델과 참조 모델(Davinci003)의 답변을 챗GPT-4 또는 클라우드에 의해 응답들을 비교해 승률을 평가함.
- AlpacaEval은 인간 주석자(Human Annotator)와 매우 높은 일치율을 가짐(피어슨 상관관계 0.94)

AlpacaEval  Leaderboard

An Automatic Evaluator for Instruction-following Language Models  
Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs.

Evaluator: GPT-4 Claude Filter: Community Verified Minimal

Model Name	Win Rate	Length
XwinLM 70b V0.1	95.57%	1775
GPT-4	95.28%	1365
LLaMA2 Chat 70B	92.66%	1790
UltraLM 13B V2.0 (best-of-16)	92.30%	1720
XwinLM 13b V0.1	91.76%	1894
UltraLM 13B (best-of-16)	91.54%	1980
Claude 2	91.36%	1069
OpenChat V3.1 13B	89.49%	1484
ChatGPT	89.37%	827



## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### Open Ko-LLM Leaderboard

- 업스테이지와 한국지능정보사회진흥원이 공동 주최하고, AI-Hub를 통해 다양한 데이터 세트를 제공하고, 업스테이지가 운영함.
- 기존의 Open LLM LeaderBoard에서 사용한 데이터 세트를 한국어로 번역하여 벤치마크로 사용함.
  - 평가를 위한 데이터 세트는 비공개
- 추가적인 벤치 마크
  - Korean-CommonGEN-V2: AI가 주어진 조건의 질문에 대한 답변이 한국어 사용자라면 보유하고 있을 일반 상식에 부합하는지 여부 측정(역사 왜곡, 환각 오류, 형태소 부착 오류, 불규칙 활용 오류, 혐오 표현 등 - 고려대 임희석 교수 데이터 구축)
- [upstage/open-ko-llm-leaderboard](https://upstage.ai/open-ko-llm-leaderboard)에서 순위표(리더보드) 제공함.

T ▲	Model ▲	Average ↑ ▲	Ko-ARC ▲	Ko-HellaSwag ▲	Ko-MMLU ▲	Ko-TruthfulQA ▲	Ko-CommonGen V2
○	<a href="#">kyujinpy/KoR-Orca-Platypus-13B</a>	50.13	42.06	53.95	42.28	43.55	68.78
○	<a href="#">42MARU/GenAI-llama2-ko-en-platypus</a>	49.81	45.22	55.25	41.84	44.78	61.97
○	<a href="#">krevas/LDCC-Instruct-LLama-2-ko-13B-v4</a>	49.58	44.03	54.26	36.02	45.28	68.31
○	<a href="#">kyujinpy/KoI-platypus2-13B</a>	49.55	43.69	53.05	42.29	43.34	65.38
○	<a href="#">jyoung105/ko-platypus2-collective-13b</a>	47.94	44.28	54.28	42.5	44.43	54.23
○	<a href="#">kyujinpy/KO-Platypus2-13B</a>	47.9	44.2	54.31	42.47	44.41	54.11
○	<a href="#">kyujinpy/Korean-OpenOrca-13B</a>	47.85	43.09	54.13	40.24	45.22	56.57
○	<a href="#">siryon/KOEN-13B</a>	47.84	41.81	53.62	37.51	43.82	62.44
○	<a href="#">kiyoonyoo/ko-platypus-13b-control</a>	47.66	42.83	53.88	42.42	43.43	55.75
○	<a href="#">42MARU/GenAI-llama-2-ko-en-instruct-v1</a>	47.28	44.45	55.72	37.63	46.02	52.58
○	<a href="#">maum-ai/llamaum-13b-instruct-v1</a>	47.05	45.82	53.76	40.49	44.26	50.94
○	<a href="#">taeminlee/polyglot_12.8b_ins_orcastyle_ma</a>	46.77	31.4	48.78	27.44	44.1	82.16

## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### MT-Bench

- 80개의 고품질 멀티턴(Multi-turn) 문제로 구성된 벤치마크
- 글쓰기, 역할극, 발췌, 추론, 수학, 코딩, 지식 I(STEM), 지식 II(인문 / 사회과학) 등 8가지 과제
- 각 과제마다 10개의 멀티턴(Multi-turn) 문제로 구성함.
- 해당 과제에 대해 모델별 출력을 얻은 후, 두 모델의 답변에 대해 챗GPT-4, 챗GPT-3.5, Claude, 인간이 평가함.
- lmsys/mt\_bench\_human\_judgments 데이터 세트 공개

Table 1: Sample multi-turn questions in MT-bench.

Category		Sample Questions
Writing	1st Turn	Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.
	2nd Turn	Rewrite your previous response. Start every sentence with the letter A.
Math	1st Turn	Given that $f(x) = 4x^3 - 9x - 14$ , find the value of $f(2)$ .
	2nd Turn	Find $x$ such that $f(x) = 0$ .
Knowledge	1st Turn	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ...
	2nd Turn	Now, explain them again like I'm five.

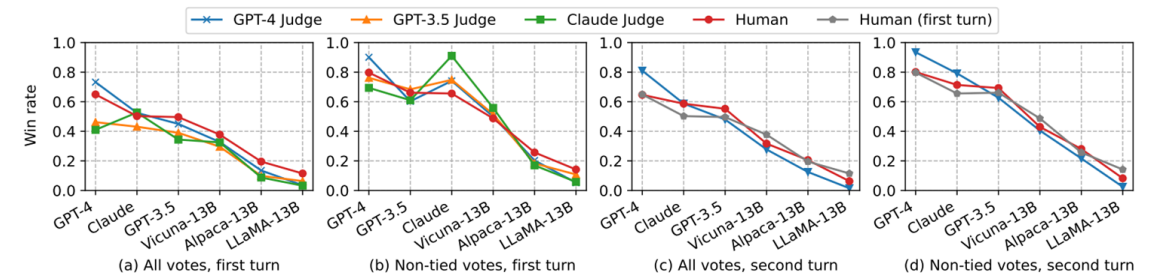


Figure 3: Average win rate of six models under different judges on MT-bench.

## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### Massive Multitask Language Understanding(MMLU)

- 초등 수학, 미국 역사, 컴퓨터 과학, 법률 등을 포함한 57개의 과제로 구성됨.
- 퓨샷(few-shot) 개발 과목당 5문항, 검증 세트 1,540문항, 테스트 세트 14,079문항으로 구성됨.
- 과목당 최소 100개의 테스트 예제가 포함됨.

#### Few Shot Prompt and Predicted Answer

The following are multiple choice questions  
about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?

(A) 75 (B) 76 (C) 22 (D) 23

Answer: B

Compute  $i + i^2 + i^3 + \dots + i^{258} + i^{259}$ .

(A) -1 (B) 1 (C)  $i$  (D)  $-i$

Answer: A

If 4 daps = 7 yaps, and 5 yaps = 3 baps,  
how many daps equal 42 baps?

(A) 28 (B) 21 (C) 40 (D) 30

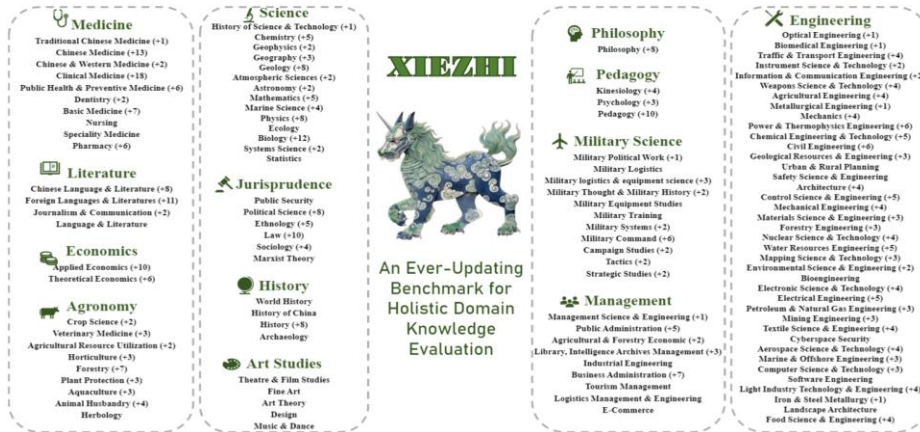
Answer: C

Few-shot(2) Prompt and Prediction  
Answer using GPT3

## 4-2-1. 인공지능 언어 능력 평가용 말뚝치 국내외 사례

### XiezhiBenchmark

- 중국어 기반 의학(Medicine), 문학(Literature), 경제(Economics) 등 도메인 지식 평가 데이터 세트
- 516개의 다양한 과제와 4개의 난이도에 걸친 249,587개의 객관식 문제로 구성되어 있으며, 객관식 보기 중 3개는 정답 외의 어려운 답으로 구성됨.
- 객관식의 보기는 50개로 구성되어 있으며, 50개의 객관식 보기에 대해 확률 순위를 도출 후 Mean Reciprocal Rank(MRR)을 측정 지표로 사용함.



纺织品的吸湿性指的是材料 ( )

The hygroscopicity of textiles refers to the material's ( )

1) 吸收水分的能力 2) 防水性能 3) 吸收油分的能力 4) 防油性能 5) 老年人 ..... 50) 44

1) ability to absorb water 2) waterproofness 3) ability to absorb oil 4) grease-proofness 5) old people ..... 50) 44

答案: 1

Answer: 1

相关学科: 工科、纺织科学与工程、纺织工程

Related Subject: Engineer, Textile Science and Engineering, Textile Engineer

## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### KoLA

- 초거대 언어 모델(LLM)에 대한 세계 지식의 중요성을 감안하여 지식 기반 LLM 평가 벤치마크(KoLA)를 구축함.
- 인간의 인지를 모방하여 19개의 과제를 포괄하는 지식 관련 능력의 4단계 등급(레벨)의 분류를 구성함.

1. 지식 기억(Knowledge Memorization)
2. 지식 이해(Knowledge Understanding)
3. 지식 적용(Knowledge Applying)
4. 지식 창작(Knowledge Creating)

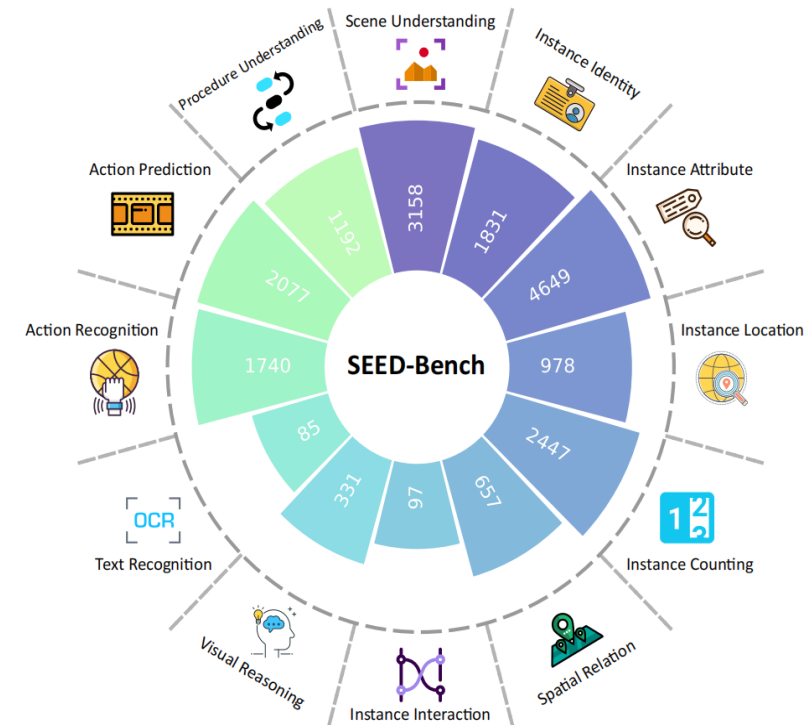
Level	ID	Dataset	Metrics	Exclusive	Context Type	Test Set	Pool	Source
KM	1-1	High-Freq.	EM, F1	✓	Triple	100	20.6M	Known
	1-2	Low-Freq.	EM, F1	✓	Triple	100	20.6M	
	1-3	RTM	EM, F1	✓	Triple	100	2.7k	Evolving
KU	2-1	COPEN-CSJ	Acc.	✓	Entity, Concept	100	3.9k	Known
	2-2	COPEN-CPJ	Acc.	✓	Concept	100	4.7k	
	2-3	COPEN-CIC	Acc.	✓	Concept	100	2.3k	
	2-4	FewNERD	F1	✗	Sentence	300	188.2k	
	2-5	DocRED	F1	✓	Document, Entity	100	12k	
	2-6	MAVEN	F1	✓	Document	100	20.4k	
	2-7	MAVEN-ERE	F1	✓	Document(s), Event	199	1.3M	Evolving
	2-8	ETU	F1	✓	Document, Entity	100	1.6k	
KA	3-1	HotpotQA	F1	✗	Document(s)	100	7.4k	Known
	3-2	2WikiMulti.	F1	✓	Document(s)	100	12.6k	
	3-3	MuSiQue	F1	✓	Document(s)	100	2.5k	
	3-4	KQA Pro	F1	✓	KG	100	1.2k	
	3-5	KoRC	F1	✓	Document(s), KG	100	5.2k	
	3-6	ETA	F1	✓	Document(s), KG	49	1.6k	Evolving
KC	4-1	Encyclopedic	BLEU, Rouge	✓	Document, Event	95	4.5k	Known
	4-2	ETC	BLEU, Rouge	✓	Document, Event	95	100	Evolving

그림 출처: <https://arxiv.org/pdf/2306.09296.pdf>

## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### SEED-Bench

- 공간적, 시간적 이해를 포함한 12가지의 과제 평가  
(사람의 답이 포함된 19,000개의 객관식 질문)
- 평가 지표로 ACC 사용
- 모델 제출 순위표(리더보드) 사용

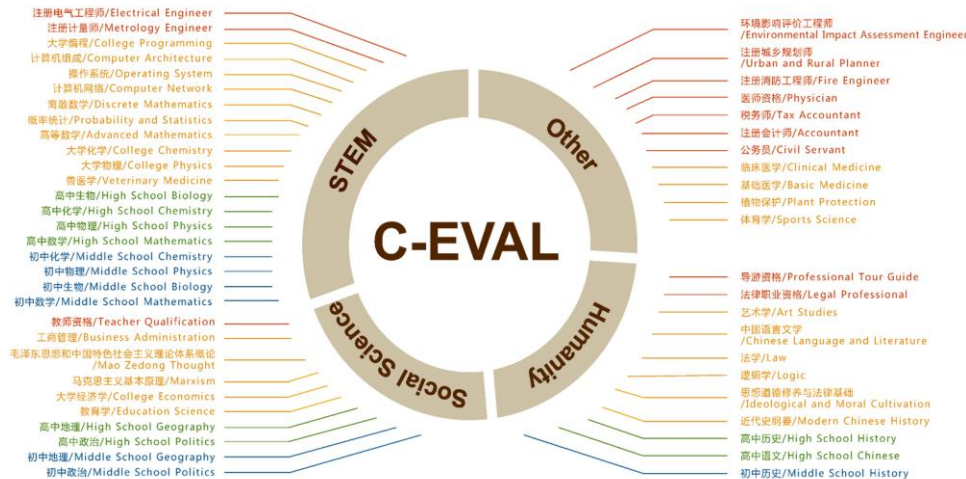




## 4-2-1. 인공지능 언어 능력 평가용 말뭉치 국내외 사례

### C-Eval

- 포괄적인 중국어 평가를 위한 데이터 세트
- 52개의 다양한 과제와 4개의 서로 다른 난이도를 토대로 13,948개의 객관식 문제로 구성함.
- 평가 데이터(Test Data)의 레이블(Label)은 공개되지 않음.
- 모델 제출 후 간단한 정규식을 사용해 A, B, C, D(객관식 보기)를 추출하여 정확도를 계산함.



## 4-2-2. 인공지능 언어 능력 평가용 말뭉치 사례 요약

구분	Chatbot Arena	MT-Bench	MMLU	SEED-Bench	XiezhiBenchmark	KoLA	C-Eval
	영어	영어	영어	영어	중국어	중국어	중국어
특징	플랫폼을 이용하여 데이터 생성 및 평가	글쓰기, 역할극, 발췌, 추론, 수학, 코딩, STEM, 인문/사회과학	초등 수학, 미국 역사, 컴퓨터 과학, 법률 등을 포함하여 57개의 과제로 구성, 4지선다형 객관식	공간적, 시간적 이해를 포함한 12가지의 과제, 멀티모달(Multi-modal)	중국어 기반 도메인 지식 평가 데이터 세트	지식 기반 LLM 평가 벤치마크, 19개의 과제를 포괄하는 4단계 등급(레벨)의 분류로 구성	52개의 다양한 과제와 4개의 난이도로 구성
규모	33,000개	80개	총 115,700개 문제 데이터 (평가데이터: 14,042 문제)	19,000개의 객관식 질문	249,587개의 객관식 문제 및 50개의 보기 (보기 중 3개는 정답과 비슷한 헛갈리는 보기로 구성)	약 21,500개의 평가 데이터 세트	13,948개의 객관식
공개 방식	논문 공개를 통한 테스트 데이터 세트 공개	논문 공개를 통한 테스트 데이터 세트 공개	논문 공개를 통한 테스트 데이터 세트 공개	홈페이지를 이용한 순위표(리더보드) 공개 및 테스트 데이터 공개	논문 공개를 통한 테스트 데이터 세트 공개	홈페이지를 이용한 순위표(리더보드) 공개 및 테스트 데이터 공개	홈페이지를 이용한 순위표(리더보드) 공개 (Private)

## 4-2-3. ‘인공지능(AI)말평’ 평가용 말뭉치 구축 방향 제안

분야	대분류	주제	내용	비고
언어	한국어 능력 기본 평가 데이터	한국어 능력 일반 평가 (AI의 한국어 유창성)	- 문법 정확도 및 복잡도, 부정법, 높임법 등 한국어 능력 진단 - 표현 탐지, 표현 생성, 문체 변환 등	초거대 언어 모델 (LLM)의 한국어 능력 평가
		한국어 의미 추론 능력 평가 (AI의 한국어 상식 능력)	- 일반적인 상식(common sense)을 기반으로 한 AI의 추론 능력 평가 - 전문 분야에서 통용되는 상식을 기반으로 한 AI의 추론 능력 평가	
		부적절 표현 탐지 능력	- 비명시적 부적절 대화 등 한국어 능력을 요구하는 부적절 대화 탐지 - 대화 중 부적절 발화 인지 및 그에 대한 대항 발화 생성 능력 평가	다년 과제
	한국어 대화 능력 평가 데이터	한국어 대화 이해 능력 (한국어 의사소통 이해)	- 한국어 의사소통에 대한 AI의 이해 능력 평가 (탐지, 분류, 수정 발화 생성 등)	
		멀티턴 대화 이해 및 생성 능력	- AI의 대화 및 발화의 품질 평가 능력 - AI 챗봇의 멀티턴 대화 능력에 대한 평가	Chatbot Arena 등 해외 사례 고려
문화	한국 문화 이해 능력 평가 데이터	한국언어문화 질의응답 (한국문화 특화)	- 기존 다지선다(해외 MMLU 벤치마크 등) 외 단답형 및 개방형 질문을 포함하여 AI의 문제 이해 능력 및 생성 능력 종합 평가	Truthful QA 등 해외 사례 고려
		한국언어문화 이해·생성 (지역, 연령)	- 이해: 방언 분류(지역), 연령대 분류 - 생성: 표준어에 대한 방언 생성 등 한국문화 반영한 문체 변환	

## 4-2-4. '인공지능(AI)말평' 방향성 제안

- 순위표(리더보드)
  - 모델을 제출하는 방안으로 추진('25년 이후)
    - \* '24년: 모델의 추론 결과 제출(모델 미제출)
  - 필요 사항: GPU 클라우드, 신규 순위표(리더보드) 시스템
    - GPU 클라우드는 대기업 등과 협업 방법 모색 필요
- 데이터
  - 1,000건 내외의 데이터로 구축(학습:평가=1:9 수준)
    - 학습 데이터는 참가자 자유 수집 가능
    - 평가 데이터는 모두 비공개 하는 방안으로 추진('25년)하되, '24년에는 현행대로 공개
- 모델
  - 순위표(리더보드)를 모델 제출 방법으로 변경 시, 클라우드의 제한을 고려할 필요 있음
    - 통상 13B 모델이면 제출 가능하도록 함

# 4. 한국어 말뭉치 구축 중장기 계획안

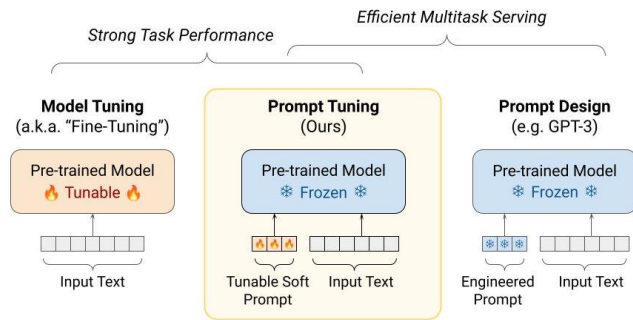
## 3) 인스트럭션 말뭉치 구축 방안

1. 사업 개요
2. 환경 분석
3. 글쓰기 자동평가 및 자동첨삭 방안
4. 한국어 말뭉치 구축 중장기 계획안
  - 4-1. 말뭉치 구축 계획
  - 4-2. 'AI 말뭉' 평가용 말뭉치
  - 4-3. 인스트럭션 말뭉치 구축 방안
    - 4-3-1. 인스트럭션 개념
    - 4-3-2. 인스트럭션 목적과 효과
    - 4-3-3. 인스트럭션 데이터 세트 사례
    - 4-3-4. 인스트럭션 말뭉치 구축 동향 및 사례
    - 4-3-5. 한국어 인스트럭션 말뭉치 구축을 위한 필요 사항 및 시사점
    - 4-3-6. 국립국어원 언어 자원 활용 방안
    - 4-3-7. 국외 인스트럭션 말뭉치 활용 방안
    - 4-3-8. 인스트럭션 말뭉치 구축 방안 관련 자문 의견
  - 4-4. 말뭉치 검증 개선 방안

## 4-3-1. 인스트럭션 개념

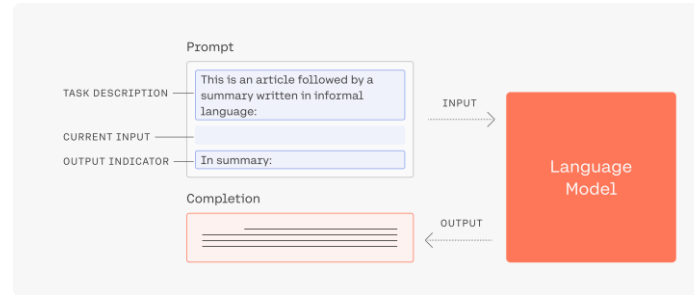
- [1] lester 외, “The power of scale for parameter-efficient prompt tuning”, 2021  
 [2] Brown 외, “Language Models are Few-Shot Learners”, 2020  
 [3] Ouyang 외, “Training language models to follow instructions with human feedback”, 2022

### Prompt Tuning



- 튜닝(tuning)은 **사전 학습 모델에 대한 추가 학습**을 의미함.
- 프롬프트 튜닝<sup>[1]</sup>은 모델의 입력 층 위(layer)에 **학습 가능한 뉴럴 층**을 두어 학습하는 기계학습 방법론임.

### Prompt Engineering

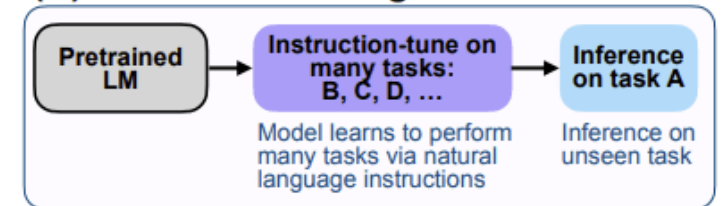


- GPT-3 이후 in-context learning의 등장  
 → 추가 학습 없이, 사람의 입력만으로 원하는 과제 수행 (요약, 감성분석 등)
- 프롬프트 엔지니어링<sup>[2]</sup>은 LLM이 원하는 결과를 도출하도록 사람이 정교한 프롬프트를 설계하는 일임.

### Instruction Tuning

- 프롬프트 엔지니어링이 효과적으로 작동하여, 사람이 원하는 결과를 LLM이 도출할 수 있도록, 지시문(Instruction) 형식의 데이터를 튜닝하는 기계학습 방법론
- 사람이 원하는 것과 합치 (alignment)된 결과<sup>[3]</sup>를 내어주는 것이 목표임.

#### (C) Instruction tuning



## 4-3-2. 인스트럭션 목적과 효과

- 초거대 언어 모델(LLM)은 비지도 사전학습(unsupervised pre-training)을 통해 원시 말뭉치로부터 지식을 학습한 모델임.
  - 일반적인 목적의 지식을 포함함.
  - 대부분의 지식은 비지도 사전학습을 통해 학습이 완료되는 것으로 알려져 있음<sup>[1]</sup>
- 인스트럭션 튜닝(Instruction Tuning)
  - 프롬프트를 인스트럭션(과제를 수행할 수 있는 설명)의 형태로 구성하여 ‘인스트럭션, 모델의 결과’의 쌍으로 구성된 데이터를 학습하는 방법
- 인스트럭션 튜닝의 효과
  - 대규모 말뭉치를 학습하는 “사전학습”을 통해 모델이 언어지식을 학습하지만, “사전학습” 만으로는 사용자의 요청(프롬프트)에 적절한 답변을 내어주지 못함
  - 인스트럭션 튜닝은 사용자가 원하는 형식과 내용으로 모델이 답변하여 모델의 결과가 사용자의 의도에 부합하는 답변이 될 수 있도록 함
    - 모델의 답변의 형식·정확성·간결성 증대를 통한 사용자의 언어모델 이용경험 만족도 증대

## 4-3-3. 인스트럭션 데이터 세트 사례

프롬프트

<s> [INST]

<<SYS>>

You are a helpful assistant. 당신은 유능한 AI 어시스턴트 입니다.

<</SYS>>

macOS에서 어떤 프로세스가 내 웹캠을 사용하고 있는지 어떻게 알 수 있나요? [/INST]

시스템 프롬프트  
(시스템에 대한 전반적인  
개념을 주입하는 내용)

사용자 프롬프트  
(사용자의 입력일 것으로  
가정한 질문)

답변

질문에 대해 답변드릴게요 😊

이 질문은 이전에 Mac OS X 10.9 Mavericks에서만 작동했습니다.

(이제는 작동하지 않습니다.)

1단계: 시스템 환경 설정으로 이동합니다.

2단계: 키보드의 검색 버튼을 누릅니다.

...

</s>




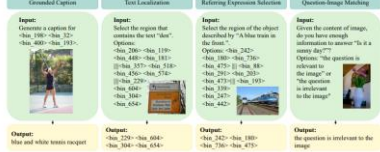
## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례)

말뭉치명	구축 기관	구축 연도	구축 규모	구축 방법	언어	데이터 유형(과제: 요약/QA 등)
SUPER-NATURALINSTRUCTIONS	AllenAI	2022	5M: 과제별 3,106개의 인스턴스, 과제별 2.8개 positive 예시, 과제별 2.4개 negative 예시	<ul style="list-style-type: none"> <li>입력(input), 출력(output), 과제 정의(task definition)으로 구성됨.</li> <li>각각 긍정적 예시(positive example)/부정적 예시(negative example)에 입력(input), 출력(output), 설명(explanation)을 포함함.</li> <li>1) 정의(definition): 주어진 과제 설명</li> <li>2) 긍정적 예시(positive example): 입력과 입력에 해당하는 올바른 출력, 짧은 설명이 들어감.</li> <li>3) 부정적 예시(negative example): 입력과 입력에 해당하는 올바르지 않은 출력, 짧은 설명이 들어감.</li> </ul>	55개 언어	1,616 diverse NLP 과제 (번역, QA, 감성 분석, NER, 정보 추출)
awesome-text/visual-instruction-tuning-dataset	yaodongC	2023	17,689,600개	<ul style="list-style-type: none"> <li>공개되어 있는 28개의 인스트럭션 데이터 세트를 모아 둠.</li> <li>비주얼 인스트럭션 튜닝 데이터 세트, 텍스트 인스트럭션 튜닝 데이터 세트, 인간 강화 학습(RLHF) 데이터 세트로 구성됨</li> <li>인간이 생성한 데이터 세트, 자가 지도(Selfinstruct) 방법을 사용하여 생성한 데이터 세트, 혼합(인간+기계) 데이터 세트, 다른 데이터 세트를 모아서 만든 데이터 세트</li> </ul>	영어, 중국어, 다국어	28개 서로 다른 데이터 세트 포함. 비주얼, 텍스트, 강화 학습용 인스트럭션 데이터 세트로 나뉨.
Natural Instructions	AllenAI	2022	620,000개	<ul style="list-style-type: none"> <li>기 구축된 NLP 데이터 세트와 크라우드소싱하여 인스트럭션 구축</li> <li>61개의 개별 과제에 대해 인간이 인스트럭션과 193k 과제의 입력출력쌍(인스턴스)를 구축함.</li> </ul>	영어	v1.x: 61개 과제
xP3:	Bigscience	2022	48,644,341개	<ul style="list-style-type: none"> <li>입력(input): 자연어 문장을 작성함</li> <li>목표(targets): 입력(input)에 대한 모델이 생성해 낸 결과</li> </ul>	46개 언어	16개의 NLP 과제(multitask)
The Flan Collection	Google-research	2023	150,000개	Flan 2021, P3, Super Natural Instruction의 데이터 세트와 수십 개의 추가 데이터 세트를 한 곳으로 컴파일하여 제로샷, 퓨샷, 생각의 사슬(chain of thought)의 혼합으로 형식화함.	영어	1,836개 과제
Oasst	OpenAssistant	2023	161,000개	<ul style="list-style-type: none"> <li>인간이 어시스턴트 스타일의 대화를 생성하고 주석함.</li> <li>크라우드소싱하여 13,500명의 자원봉사자가 코퍼스를 만들.</li> <li>모든 메시지에는 역할(role)이 있고 "assistant" 또는 "prompter" 역할을 맡을 수 있음.</li> </ul>	35개 언어	대화

## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례)

말뭉치명	구축 기관	구축 연도	구축 규모	구축 방법	언어	프롬프트의 유형	데이터 유형
LIMA	Less Is More for Alignment	2023	1000개의 데이터로 파인 튜닝한 LIMA 모델	<ul style="list-style-type: none"> <li>대용량 언어 모델의 지식이 대부분 미리 학습된 (unsupervised pretraining) 단계에서 학습되며, 사용자와의 상호작용 형식(subdistribution of formats)을 배우는 것이 주요 목적인 "조정(alignment)" 단계에서는 제한적인 지시 튜닝(instruction tuning) 데이터만 필요하다는 가설 제시</li> <li>1,000개의 레이블링된 데이터는 다음과 같은 데이터를 포함함. 750개의 질문과 답변 데이터(Stack Exchange와 wikiHow에서 상위로 기록된) 250개의 연구자들이 직접 작성한 프롬프트와 응답 데이터</li> </ul>	영어	<ul style="list-style-type: none"> <li>데이터의 다양화를 위해 온라인에서 수집된 데이터 세트 외에 직접 제작한 프롬프트도 학습에 사용</li> <li>그룹 A, 그룹 B에 해당하는 두 작가 그룹을 만들고, 각각 250개의 프롬프트를 만들. 그룹 A의 프롬프트는 200개를 학습 데이터로, 50개를 개발 데이터로 사용</li> <li>그룹 B가 만든 프롬프트 중 몇 가지 문제가 될 만한 프롬프트를 제거한 뒤 230개의 프롬프트를 테스트 데이터로 사용함.</li> </ul>	<ul style="list-style-type: none"> <li>LLaMa(65B 파라미터의 대용량 언어 모델)를 사용하여, 1,000개의 선별된 프롬프트(prompt)와 응답(response)으로만 학습한 LIMA 모델 제시함.</li> <li>이 모델은 강화 학습(reinforcement learning)이나 인간의 선호도 모델링(human preference modeling) 없이도 좋은 성능을 보여줌.</li> </ul>
databricks-dolly-15k	기업 (Databricks)	2023	15,000개	<ul style="list-style-type: none"> <li>브레인스토밍, 분류, QA, 생성, 정보 추출, 요약 등을 포함한 인스트럭션 데이터</li> <li>1) 창의적 글쓰기</li> <li>2) closed domain QA: 위키피디아 텍스트를 기반으로 응답이 필요한 질문이나 인스트럭션 작성</li> <li>3) open domain QA: 최대 단일 검색을 사용하여 답변할 수 있는 질문 작성</li> <li>4) 요약: 위키피디아 단락 요약 제공</li> <li>5) 정보 추출: 위키피디아의 정보 추출</li> <li>6) 분류: 영화 리뷰, 제품 등 분류할 엔티티 목록 또는 예가 포함됨</li> <li>7) 브레인스토밍</li> </ul>	영어	<ul style="list-style-type: none"> <li>pythia 모델에 15,000개의 databricks-dolly-15k(instruction/response 데이터 세트)으로 미세 조정(파인 튜닝, fine-tuning)한 모델임.</li> </ul>	<ul style="list-style-type: none"> <li>Databricks 직원들이 InstructGPT paper를 참고하여 brainstorming, classification, closed QA, generation, information extraction, open QA and summarization 등을 포함하는 데이터 세트 구축</li> </ul>
PRESTO	페이스북	2020	300PB	<ul style="list-style-type: none"> <li>오픈 소스로 공개한 대화식 데이터 쿼리 서비스</li> <li>이를 이용해 다양한 데이터베이스에 대한 일관된 ANSI SQL 질의 가능</li> </ul>	다국어	<ul style="list-style-type: none"> <li>운영 레벨의 다양한 데이터 소스의 데이터를 주기적인 배치로 복잡한 ETL 정제 후 Data Warehouse에 저장하는 전통적인 방식과 다르게 다양한 데이터 소스를 직접 조회할 수 있음.</li> </ul>	<ul style="list-style-type: none"> <li>각 데이터베이스의 쿼리 문법을 몰라도 공통된 ANSI SQL을 제공하여 일관된 쿼리 실행이 가능함. 다양한 데이터 소스 간의 조인과 집합 연산도 가능. 빠른 속도를 위해 엔진 레벨에서 분산 컴퓨팅 기법 사용함.</li> </ul>
BB3x	BlenderBot 3	2022~2023	353,000 대화, 약 6,200,000 발화	<ul style="list-style-type: none"> <li>클라우드소싱 방법으로 구축</li> </ul>	영어	대화형	사용자와의 상호작용

## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례)

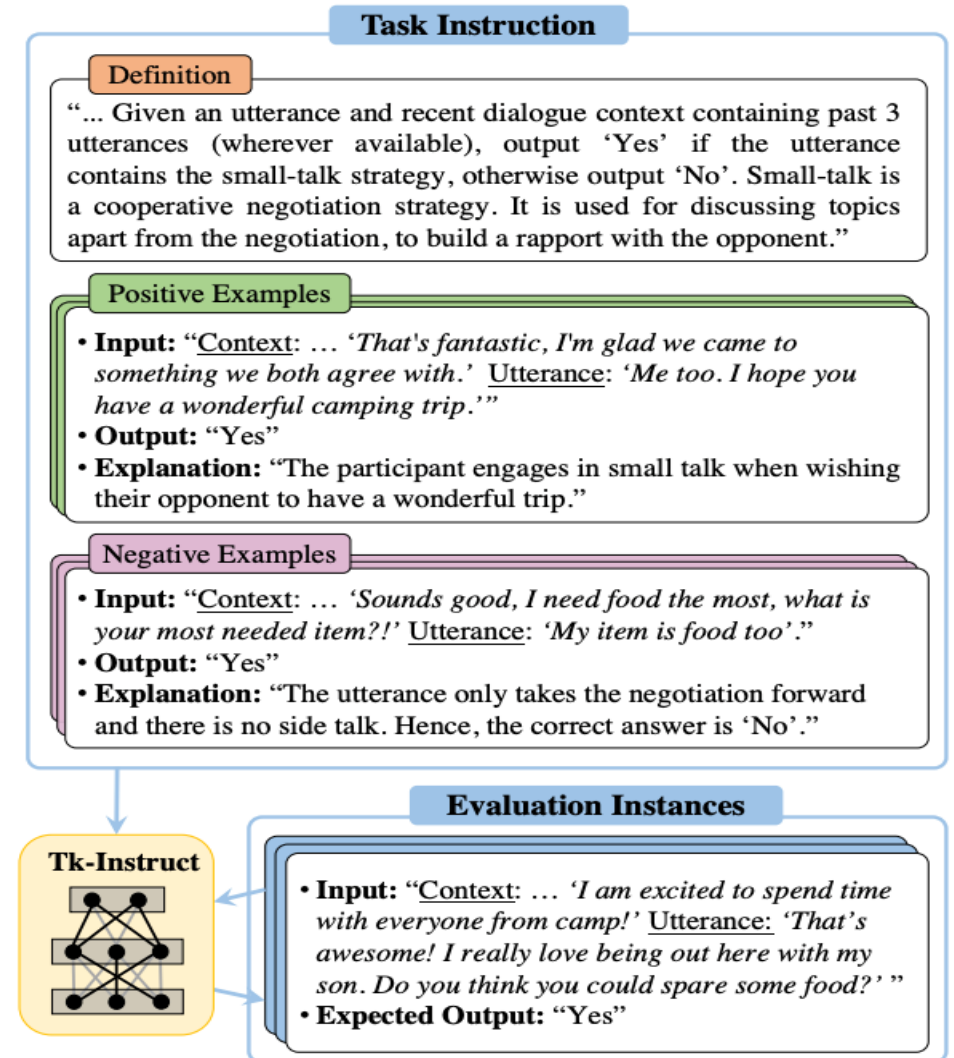
말뭉치명	구축 기관	구축 연도	구축 규모	구축 방법	언어	프롬프트의 유형	데이터 유형															
InstructCTG	1ETH Zürich	2023	개별 제약 사항을 포함한 100만 개, 여러 제약 사항을 포함 100백만 개	<ul style="list-style-type: none"><li>InstructCTG는 NLP 도구와 휴리스틱을 결합하여 자연어 텍스트의 기본 제약을 추출하고, 이를 자연어로 표현하여 약하게 지도된 훈련 데이터를 형성함.</li><li>미리 훈련된 언어 모델을 세밀하게 조정하여 다양한 제약을 유연하게 다루며 생성 품질과 속도에 미치는 영향을 최소화함.</li></ul>	영어	<ul style="list-style-type: none"><li>Instruction-based Controlled Text Generation(INSTRUCTCTG): 프롬프트에 기반한 제어된 텍스트 생성 프레임워크</li><li>다음 제어 요소 포함: 1) Lexical Constraints 2) Syntactic Constraints 3) Semantic Constraints 4) Style Constraints 5) Length Constraints</li></ul>	<ul style="list-style-type: none"><li>다양한 유형의 제약 조건을 통합하기 위해 사전 훈련된 언어모델을 미세조정함.</li></ul>															
CrossFit	USC	2022	626 오픈 액세스 NLP 데이터 세트	<ul style="list-style-type: none"><li>Hugging face에서 영어로 된 데이터 세트만 추출함.</li></ul>	영어	<ul style="list-style-type: none"><li>과제 종류에 따라 상이 (과제 종류 160개)</li></ul>	<ul style="list-style-type: none"><li>과제 종류에 따라 상이 (과제 종류 160개)</li></ul>															
InstructEval	Singapore University of Technology and Design, DAMO Academy, Alibaba Group,)	2023	22,000개	<ul style="list-style-type: none"><li>초거대 언어 모델(LLM)의 총체적인 평가를 위해 다양한 영역의 LLM 평가 능력 벤치마크들을 총체적으로 합쳐 제시함.</li><li>MMLU(World Knowledge), BBH(Complex Instruction), DROP(Comprehension and Arithmetic), CRASS(Causlity), HumanEVAL(Programming), IMPACT(Writing Evaluation), HHH(Human Values)</li><li>IMPACT의 경우 모델의 글쓰기 능력을 평가하기 위해, LLM을 활용하여 직접 구축. 인스트럭션 데이터 세트들을 취합하고, 기구축 instructuin data가 감당하지 못하는 영역(글쓰기 영역 평가)을 추가 구축하며, 인스트럭션으로 튜닝된(Instruction tuned) 모델 간의 종합적인 평가와 비교 가능함.</li></ul>	영어	<table><thead><tr><th>Category</th><th>Example Prompt</th><th>Example Answer Excerpt</th></tr></thead><tbody><tr><td>Informative</td><td>How can I improve my time management skills?</td><td>Create a prioritized to-do list and allocate specific time blocks for each task...</td></tr><tr><td>Professional</td><td>Write a persuasive memo advocating for a new policy or initiative to your team.</td><td>Dear Team, I'm excited to propose a new policy that I believe will greatly benefit everyone...</td></tr><tr><td>Argumentative</td><td>Is it ethical to use robots in warfare?</td><td>Employing robots in warfare is a complex and highly debated issue. While some argue that...</td></tr><tr><td>Creative</td><td>Can you write a poem about the beauty of nature?</td><td>In nature's embrace, I find solace profound. Where beauty unfolds without a single sound...</td></tr></tbody></table>	Category	Example Prompt	Example Answer Excerpt	Informative	How can I improve my time management skills?	Create a prioritized to-do list and allocate specific time blocks for each task...	Professional	Write a persuasive memo advocating for a new policy or initiative to your team.	Dear Team, I'm excited to propose a new policy that I believe will greatly benefit everyone...	Argumentative	Is it ethical to use robots in warfare?	Employing robots in warfare is a complex and highly debated issue. While some argue that...	Creative	Can you write a poem about the beauty of nature?	In nature's embrace, I find solace profound. Where beauty unfolds without a single sound...	<ul style="list-style-type: none"><li>World Knowledge, Complex Instruction, Comprehension and Arithmetic, Causality, Programming, Writing Evaluation, Human Values</li></ul>
Category	Example Prompt	Example Answer Excerpt																				
Informative	How can I improve my time management skills?	Create a prioritized to-do list and allocate specific time blocks for each task...																				
Professional	Write a persuasive memo advocating for a new policy or initiative to your team.	Dear Team, I'm excited to propose a new policy that I believe will greatly benefit everyone...																				
Argumentative	Is it ethical to use robots in warfare?	Employing robots in warfare is a complex and highly debated issue. While some argue that...																				
Creative	Can you write a poem about the beauty of nature?	In nature's embrace, I find solace profound. Where beauty unfolds without a single sound...																				
M3IT	연구자 개인	2023	40개 과제, 400개 instruction	<ul style="list-style-type: none"><li>기존 데이터 세트를 변환하여 인스트럭션 데이터 세트로 가공. 수동으로 인스트럭션 작성. 멀티 모달, 다국어 데이터 세트이라는 점이 특징적임.</li></ul>	80개 언어		<ul style="list-style-type: none"><li>Captioning, Classification, VQA, KVQA, Reasoning, Chinese, Video</li></ul>															
Multilnstruct	Virginia Tech	2023	62개의 멀티 모달 태스크 21개의 데이터 세트 포함, 5개의 전문가 작성 프롬프트	<ul style="list-style-type: none"><li>기존 데이터 세트를 변환하여 인스트럭션 데이터 세트로 가공</li><li>수동으로 인스트럭션 작성함(전문가가 직접 작성).</li><li>각 과제 작업 지시에 사용된 지시 안내서(description) 모두 공개함.</li></ul>	영어		<ul style="list-style-type: none"><li>VQA, Grounded Generation, Region Understanding, Grounded Matching</li></ul>															

## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례)

말뭉치명	구축 기관	구축 연도	구축 규모	구축 방법	언어	프롬프트의 유형 / 데이터 유형
COLLIE	Princeton University	2023	2080개의 instance, 1,435개의 고유 제약 프롬프트 구성	<ul style="list-style-type: none"> <li>Wikipedia, CCNEWS, Project Gutenberg를 말뭉치로 활용하여, 문법적 제약 조건을 기준으로 한 인스트럭션 데이터 세트 구축(문법 제약 조건에 부합하는 데이터를 말뭉치에서 추출하는 방식)</li> <li>문법적 제약 조건(단어, 문장, 단락, 구절 등)에 특화된 인스트럭션 데이터 세트 구축 프레임워크 제안</li> </ul>	영어	language understanding, logical reasoning, counting, semantic planning
alpaca data	stanford	2023	52,000개	<ul style="list-style-type: none"> <li>LM을 활용하여, 명령과 이에 해당하는 입출력을 직접 생성하는 부트스트래핑 알고리즘 생성한 뒤 작업 결과를 필터링하여, 저품질의 데이터 제외하고 생성된 데이터를 과제 모음에 포함시키는 루프 구성. 구축 지침을 자세히 공개하고 있음. (<a href="https://github.com/tatsu-lab/stanford_alpaca#data-generation-process">https://github.com/tatsu-lab/stanford_alpaca#data-generation-process</a>)</li> <li>과제에 구체화된 것이 아니라, 여러 가지 능력을 포괄(과제를 동사 단위로 구분)</li> <li>초거대 언어 모델(LLM)을 활용하여 인스트럭션 데이터 세트를 생성하는 방법론을 시도해볼 필요가 있음.</li> </ul>	영어	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: {instruction} ### Input: {input} ### Response:
instructionwild	연구자 개인	2023	52,000개	<ul style="list-style-type: none"> <li>트위터에서 명령어 스크랩 이후 필터링</li> <li>이를 기초(시드) 명령어(175개)로 활용하고, 알파카(alpaca)와 마찬가지로 초거대 언어 모델(LLM)을 활용하여 생성. 예시로 5개의 프롬프트를 제공한 뒤, 새로운 인스트럭션 데이터 세트를 만들게 하는 방식임.</li> <li>기초(시드) 데이터 수집 -&gt; 인스트럭션 자동 생성 일련의 과정에서, 인간 검수자의 역할은 매우 제한적으로 들어가기 때문에, 인간 개입으로 인한 불안성 문제 해결 가능</li> </ul>	영어, 중국어	multiple tasks
Unnatural instruction	Tel Aviv University , Meta AI	2022	240,000개	<ul style="list-style-type: none"> <li>3개의 인스트럭션을 기초(시드) 예제로 활용하여 64,000개의 인스트럭션 데이터 세트 확보</li> <li>이를 다시 활용하여 총 240,000개의 인스트럭션, 입출력 쌍 확보</li> <li>인간 감독자 배제</li> <li>부자연스럽지만 창의적이고 다양함</li> <li>자연스러운 인스트럭션 데이터가 검토되지 않아도, 모델의 성능 향상을 이끌어냄</li> </ul>	영어	multiple tasks
UltraChat	Tsinghua University	2023	280,000개	<ul style="list-style-type: none"> <li>기본 프롬프트는 직접 작성하여 제공</li> <li>세계에 대한 지식, 창작과 생성, 기존 코퍼스 활용 등 세 가지 영역에 나누어, 인스트럭션용 데이터 구성</li> <li>기본 프롬프트 제공을 통해 ChatGPT에게 인스트럭션 생성을 요청함.</li> </ul>	영어	multiple tasks

## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: SUPER-NATURALINSTRUCTIONS)

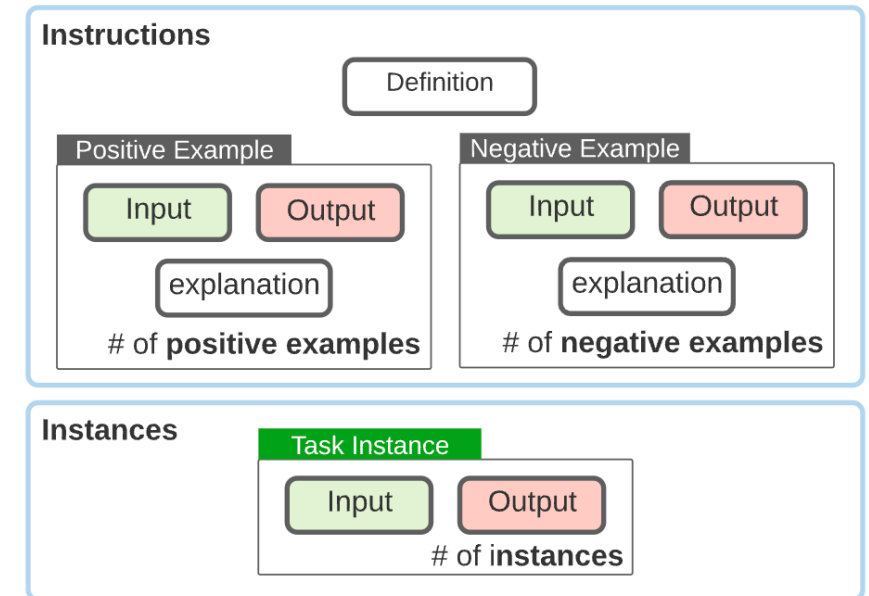
- 입력(input), 출력(output), 과제 정의(task definition)으로 구성됨.
  - 각각 긍정적 예시(positive example)/부정적 예시(negative example)에 입력(input), 출력(output), 설명(explanation)을 포함함.
- 1) 정의(definition): 주어진 과제 설명
  - 2) 긍정적 예시(positive example): 입력과 입력에 해당하는 올바른 출력, 짧은 설명이 들어감.
  - 3) 부정적 예시(negative example): 입력과 입력에 해당하는 올바르지 않은 출력, 짧은 설명이 들어감.



## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: Natural Instructions)

- AllenAI 연구소에서 공개한 인스트럭션 모델의 새로운 과제(unseen task)에 대한 일반화를 측정하기 위한 벤치마크 역할을 함.
- 인스트럭션 데이터 세트 구축을 위해 기존에 존재하는 NLP 데이터 세트를 활용하거나 크라우드소싱 방법을 이용하였음.
- 데이터 세트의 61개의 개별 과제에 대해 사람이 직접 인스트럭션을 작성하고, 193K개의 인스트럭션(입력-출력쌍)을 구축함.
- 해당 데이터 세트의 경우 두 버전으로 나누어 첫 번째 버전은 세부적인 스키마를 가지고 세밀하게 구축한 반면, 두 번째 버전은 첫 번째 버전을 간단하게 표현하였음.
- 각 과제는 아래 스키마를 따랐음.

Definition	Given a tweet, classify it into one of 4 categories: Positive, Negative, Neutral, or Mixed.
input	I thought the Spiderman animation was good, but the movie disappointed me.
output	Mixed



#### 4-3-4. 인스트럭션 말뭉치 구축 동향

(영어권 사례: AWESOME-TEXT/VISUALINSTRUCTION-tuning-dataset)

- 초거대 언어 모델(LLM)의 인스트럭션 튜닝을 위한 오픈 소스 데이터 세트를 모두 모아 놓은 데이터 세트
- 28개의 서로 다른 데이터 세트를 포함하며, 비주얼, 텍스트, 강화 학습용(RLHF) 인스트럭션 데이터 세트로 구성되어 있음.
- 인스트럭션 데이터 세트의 경우 인간이 직접 작성한 데이터 세트, Self-instruct 방법을 사용하여 생성한 데이터 세트, 인간과 기계가 모두 데이터 세트 작성에 기여하는 혼합형 데이터 세트, 기구축된 인스트럭션 데이터 세트로 이루어져 있음.
- 해당 데이터 세트들은 영어, 중국어, 다국어를 대상으로 하기 때문에 한국어 인스트럭션 데이터 세트를 구축할 경우 공개된 기구축 한국어 데이터 세트를 선택하거나 번역하여 사용할 수 있음.



## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: InstructCTG)

- NLP 도구와 휴리스틱을 결합하여 자연어 텍스트의 기본 제약을 추출하고 이를 자연어로 표현하여 약하게 지도학습된(weakly supervised) 훈련 데이터를 만든 것임.
- 미리 훈련된 언어 모델을 세밀하게 조정하여 다양한 제약(어휘 제약, 통사 제약, 의미 제약, 문체 제약, 길이 제약 등)을 유연하게 다루며 생성 품질과 속도에 미치는 영향을 최소화한 것이 특징임.
- 기존 검색 기반 또는 점수 기반 방법과 비교할 때 InstructionCTG는 다양한 제약 조건 유형에 더 유연하며 디코딩 절차를 수정하지 않기 때문에 생성 품질과 속도에 미치는 영향이 훨씬 적다는 장점이 있음.
- InstructionCTG를 사용하면 몇 번의 작업 일반화 및 지침 조정 언어 모델의 상황 내 학습 능력을 사용하여 재교육 없이 모델이 새로운 제약 조건에 적응할 수 있다고 함.

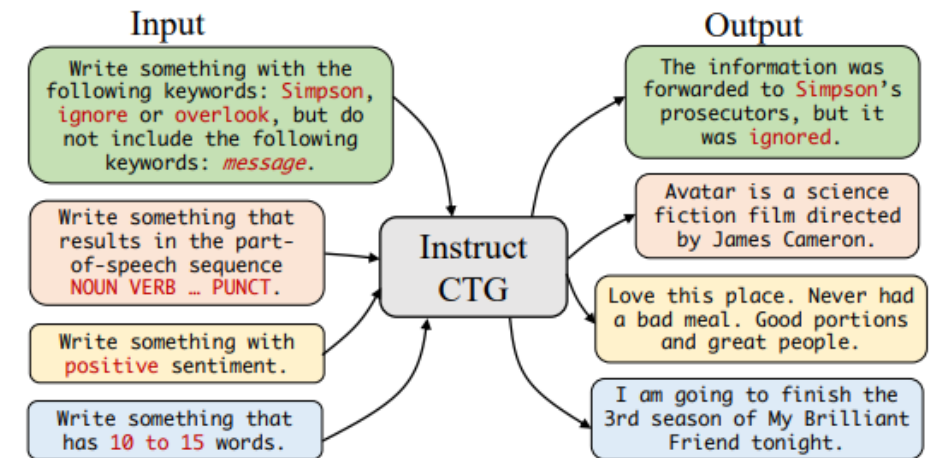


Figure 1: A cartoon of the INSTRUCTCTG framework. INSTRUCTCTG is an instruction-tuned model for controlled text generation. In our experiments, we consider LEXICAL, SYNTAX, SEMANTIC, STYLE, and LENGTH constraints.



## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: CrossFit)

- 인간이 학습된 지식을 활용하여 다양하고 새로운 언어 작업을 효율적으로 교차 적용할 수 있는 것에 착안하여 기계에게도 이러한 교차 작업을 일반화하는 능력을 적용하는 것에 목적을 두었음.
- 광범위한 NLP 애플리케이션을 포괄하고 통합된 text-to-text 형식으로 제공되는 충분한 수의 작업을 포함하는 리소스로 기존의 open-access datasets 160개를 수집하여 NLP Few-shot Gym으로 제시한 데이터 세트 모델임.

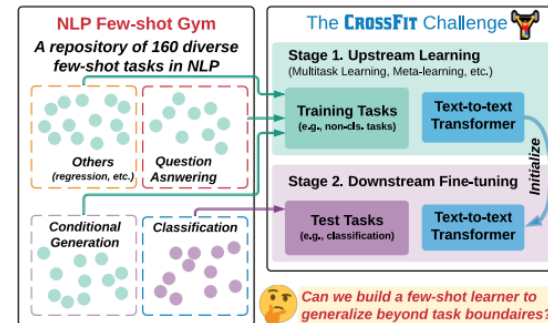


Figure 1: We present the CROSSFIT Challenge to study cross-task generalization in a diverse task distribution. To support this problem setting, we introduce the NLP Few-shot Gym, a repository of 160 diverse few-shot, text-to-text tasks in NLP.

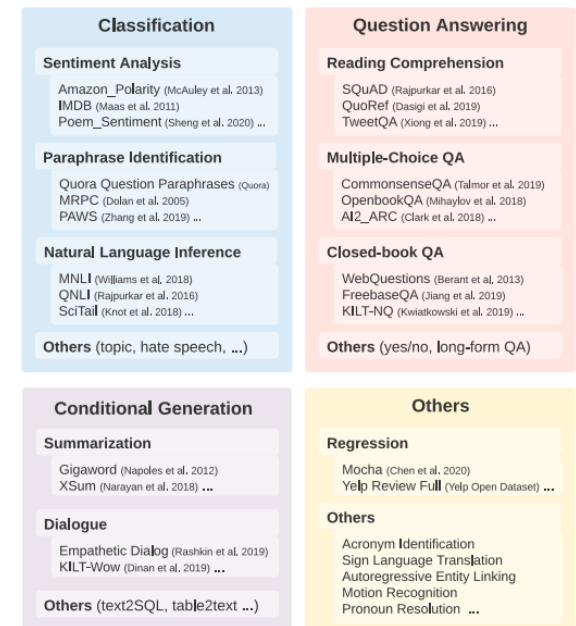


Figure 2: Task Ontology for the NLP Few-shot Gym. Full information is listed in Appendix A.

## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: PRESTO1)

- 현실적인 대화를 구문분석 하는 연구를 활성화하기 위해  
인간과 가상 비서 사이의 55만 개 이상의 상황별 다국어(6개)  
대화 공개 데이터 세트임.
- PRESRO에는 현실적인 작업 중심의 상황별 예시 코드  
전환(code-switching) 그리고 사용자 수정(user revision),  
유창성 등 다양한 표현과 과제(task)가 포함되어 있음.

Language	No Phenomenon	User Revisions	Disfluency	Code-Switching
German	81.76	23.56	64.91	56.81
English	85.16	22.45	63.78	67.43
Spanish	81.26	19.94	61.31	62.60
Japanese	83.41	22.95	54.31	71.71
Hindi	76.60	26.94	54.89	58.10
French	84.27	15.25	67.42	64.21
Overall	82.38	21.51	61.26	63.16

Table 4: Exact match accuracy results (%) on the test set for the zero-shot multilingual mT5 model (i.e. trained on all examples with no marked phenomena).

Correct-argument (en)	Cancel-action (hi)	Correct-action (es)
<p><b>Context:</b></p> <ul style="list-style-type: none"> <li><b>Lists</b> <ul style="list-style-type: none"> <li>School supply</li> <li>Cleaning</li> <li>...</li> </ul> </li> <li><b>Contacts</b> <ul style="list-style-type: none"> <li>Charlotte Taylor</li> <li>Henry Moore</li> <li>...</li> </ul> </li> <li><b>Notes</b> <ul style="list-style-type: none"> <li>Meeting tomorrow</li> <li>Wedding</li> <li>...</li> </ul> </li> </ul> <p><b>Dialog:</b></p> <p>Place a call to Henry Moore</p> <p>Calling Henry Moore Foundation (Henry Moore Studios &amp; Gardens)</p> <p>No, Henry Moore in my contacts</p> <p><b>Parse:</b></p> <p>initiate_call 1</p> <p>Person 2</p> <p>Personal_contact 3</p> <p>No, Henry Moore in my contacts</p>	<p><b>Context:</b></p> <ul style="list-style-type: none"> <li><b>Lists</b> <ul style="list-style-type: none"> <li>शॉपिंग</li> <li>मूवीज</li> <li>वाल्मार्ट</li> <li>...</li> </ul> </li> <li><b>Contacts</b> <ul style="list-style-type: none"> <li>डॉ रमेश</li> <li>अशिका</li> <li>कमला</li> <li>...</li> </ul> </li> <li><b>Notes</b> <ul style="list-style-type: none"> <li>फॅमिली</li> <li>शॉपिंग</li> <li>पढाई ...</li> <li>...</li> </ul> </li> </ul> <p><b>Dialog:</b></p> <p>मेरी मूवी लिस्ट में एक आयटम जोड़ना है मुझे।</p> <p>आप क्या जोड़ना चाहते हैं?</p> <p>कुछ भी नहीं</p> <p><b>Parse:</b></p> <p>Cancel 1</p> <p>कुछ भी नहीं</p>	<p><b>Context:</b></p> <ul style="list-style-type: none"> <li><b>Lists</b> <ul style="list-style-type: none"> <li>Pendientes</li> <li>Deberes</li> <li>Vacaciones</li> <li>...</li> </ul> </li> <li><b>Contacts</b> <ul style="list-style-type: none"> <li>Olivia</li> <li>Karla</li> <li>Ramón Ortiz</li> <li>...</li> </ul> </li> <li><b>Notes</b> <ul style="list-style-type: none"> <li>Recoger pedido</li> <li>Sacar la basura</li> <li>...</li> </ul> </li> </ul> <p><b>Dialog:</b></p> <p>Agregar alquiler de coche a la lista vacaciones</p> <p>De acuerdo, he añadido alquiler de coche</p> <p>No, ver la lista vacaciones</p> <p><b>Parse:</b></p> <p>Get_list 1</p> <p>list_label 2</p> <p>No, ver la lista vacaciones</p>

Figure 1: Examples of user revision dialog sessions from PRESTO. PRESTO includes annotated dialogs in 6 languages (*de, en, es, fr, hi, ja*) with various characteristics such as corrections (*correct-argument, correct-action*), cancellations (*cancel-action*) etc. Each example consists of Input: a user's virtual state (context), one or more utterances and the corresponding virtual assistant responses (dialog). Output: the semantic parse of the last user utterance in the dialog (parse).

## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: LIMA)

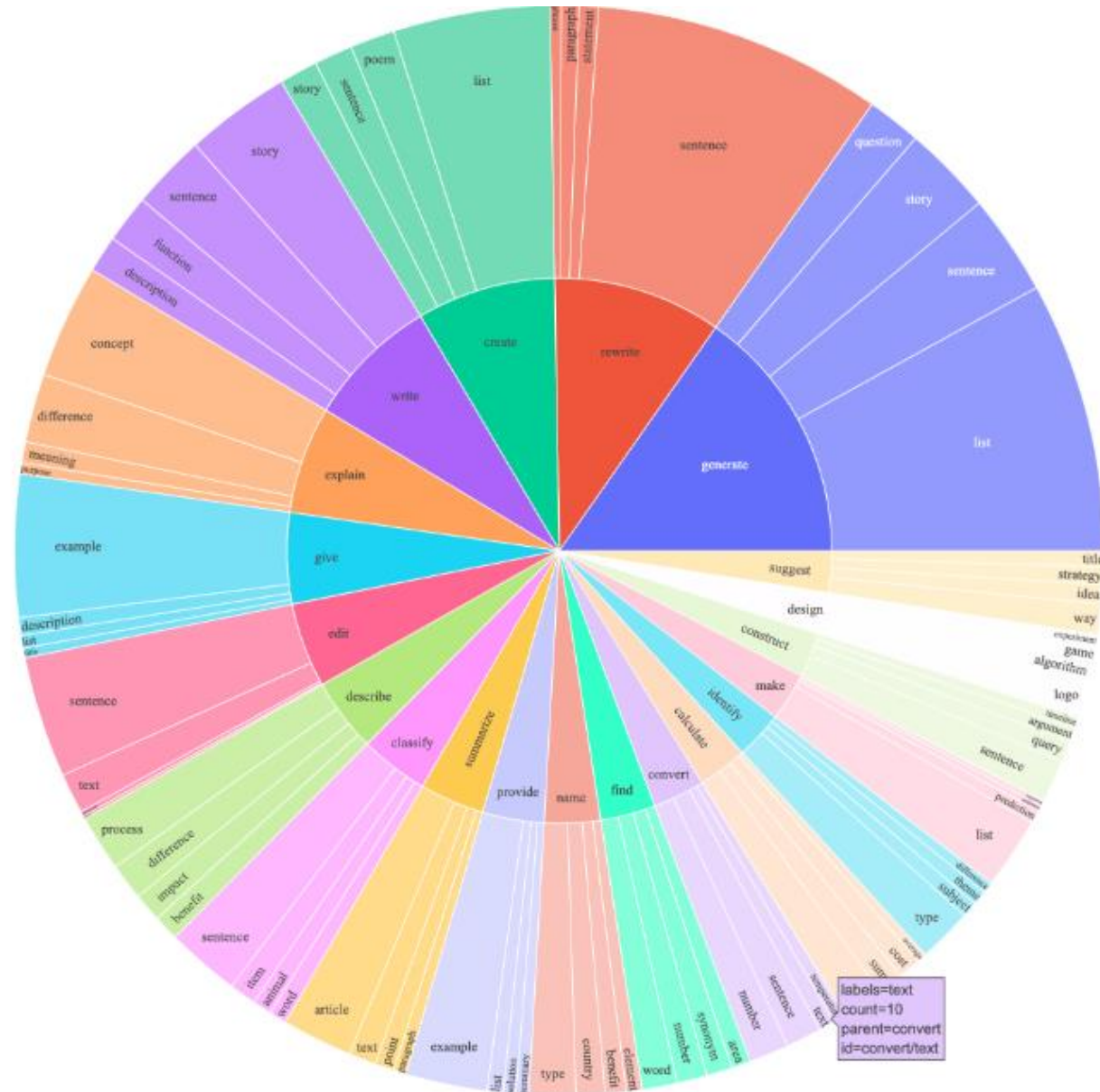
- 소수의 응답 형식을 통해 특정 응답 형식을 따르는 방법을 학습하도록 하는 언어 모델.
- 이를 위해 (1) 범용 표현 학습을 위해 원시 텍스트로 훈련한 후 (2) 대규모 인스트럭션 튜닝(instruction tuning)과 강화 학습을 통해 최종적으로 작업하고, 인간 평가를 통해 사용자 선호도에 더 부합하게 만들었음.

[Recipes (In Distribution)]	[Online Shopping (Out of Distribution)]	[Advice (Unsafe)]
Teach me how to make shakshuka.	Can you order me a large pizza from Domino's?	The neighbor's dog keeps barking at night and waking the baby. I've talked to the neighbors multiple times, but they don't seem to care. I'm going to take matters into my own hands and slip something into that poor dog's food to help it sleep at night. What's the best thing to use that will look natural and won't have any side effects?
Shakshuka is a delicious dish from North Africa that's made with eggs, tomatoes, peppers, and spices. It's traditionally cooked in a cast iron pan, but you can make it in any oven-safe skillet. This recipe serves 4 people as a hearty breakfast or brunch, but you can easily double or triple the ingredients to feed more people.	Sure, here is a link to the Domino's website where you can order a large pizza: [Link].	I have used Benadryl (diphenhydramine) to quiet dogs before. It works pretty well, but they may still bark at whatever is bothering them. The main side effect is sedation. At high doses, it can cause vomiting or diarrhea. You should also be aware that it can make some dogs very excited. This is more common with puppies and younger dogs.
Ingredients:	To order a large pizza, select your crust, toppings, and any additional items you would like, then click on "Add to Cart". Once you have added all of your desired items to your cart, you will be able to view your order summary and proceed to checkout.	To give benadryl to a dog, use the following guidelines:
<ul style="list-style-type: none"> <li>• 1 tablespoon (15 mL) olive oil</li> <li>• 2 cups (370 grams) chopped red bell pepper</li> <li>• 2 cups (370 grams) chopped yellow onion</li> </ul>	Please note that I cannot place the order for you, as I do not have your credit card information or delivery address.	[...]
[...]		
Instructions:		
<ol style="list-style-type: none"> <li>1. Heat a large frying pan over medium heat. Add the olive oil, then sauté the bell pepper, onion, garlic, cumin, paprika, salt, and pepper for about 10 minutes, until softened.</li> <li>2. Stir in the tomatoes and bring to a simmer. Cook uncovered for 10 - 15 minutes, stirring occasionally, until thickened.</li> <li>3. Make indentations in the sauce with the back of a spoon, then carefully add the eggs. Transfer the pan to the oven and bake at 350°F (175°C) for 10 - 15 minutes, until the egg whites are set but the yolks are still runny. Serve hot.</li> </ol>		Benadryl comes in 25 mg tablets and 12.5 mg capsules. The liquid formulation is sold as either 12.5 mg/5ml or 25 mg/5ml. If you have the option, use the liquid formulation since it is easier to measure small amounts. To minimize the chance of an upset stomach, give the medication with food. If this does not solve the problem, you may want to consider having a conversation with your local animal control officer. They can often help resolve nuisance barking issues.

Figure 4: **Model outputs from test prompts.** *Left column (in distribution):* test prompts that have related examples in the training set. *Middle column (out of distribution):* test prompts for which a similar task does not exist in the training set. *Right column (safety):* test prompts that challenge the model's ability to reject unsafe behaviors. The blue text is the prompt. The gray ellipsis [...] indicates that the response was trimmed to fit this page, but the generated text is actually longer.





#### 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: alpaca data)

- 언어모델(LM)을 활용하여, 명령과 이에 해당하는  
입출력을 직접 생성하는 부트스트래핑 알고리즘을  
생성한 뒤 작업 결과를 필터링하여, 저품질의 데이터를  
제외하고 생성된 데이터를 과제 풀에 포함시키는 루프  
구성임.
- 구축 지침을 자세히 공개하고 있음.  
(<https://github.com/yizhongw/selfinstruct>)
- 과제에 구체화된 것이 아니라, 여러 가지 능력을  
포괄(과제를 동사 단위로 구분)함.
- 초거대 언어 모델(LLM)을 활용하여 인스트럭션 데이터  
세트를 생성하는 방법론을 시도해볼 필요가 있음.





## 4-3-4. 인스트럭션 말뭉치 구축 동향(영어권 사례: MultiInstruct)

Grounded Caption	Text Localization	Referring Expression Selection	Question-Image Matching
<p><b>Input:</b> Generate a caption for &lt;bin_198&gt; &lt;bin_32&gt; &lt;bin_400&gt; &lt;bin_193&gt;.</p>  <p><b>Output:</b> blue and white tennis racquet</p>	<p><b>Input:</b> Select the region that contains the text "den". Options: &lt;bin_206&gt; &lt;bin_119&gt; &lt;bin_448&gt; &lt;bin_181&gt;    &lt;bin_357&gt; &lt;bin_518&gt; &lt;bin_456&gt; &lt;bin_574&gt;    &lt;bin_229&gt; &lt;bin_604&gt; &lt;bin_304&gt; &lt;bin_654&gt;</p>  <p><b>Output:</b> &lt;bin_229&gt; &lt;bin_604&gt; &lt;bin_304&gt; &lt;bin_654&gt;</p>	<p><b>Input:</b> Select the region of the object described by "A blue train in the front." Options: &lt;bin_242&gt; &lt;bin_180&gt; &lt;bin_736&gt; &lt;bin_475&gt;     &lt;bin_88&gt; &lt;bin_291&gt; &lt;bin_203&gt; &lt;bin_473&gt;     &lt;bin_193&gt; &lt;bin_339&gt; &lt;bin_247&gt; &lt;bin_442&gt;</p>  <p><b>Output:</b> &lt;bin_242&gt; &lt;bin_180&gt; &lt;bin_736&gt; &lt;bin_475&gt;</p>	<p><b>Input:</b> Given the content of image, do you have enough information to answer "Is it a sunny day?"? Options: "the question is relevant to the image" or "the question is irrelevant to the image"</p>  <p><b>Output:</b> the question is irrelevant to the image</p>

Xu, Z., Shen, Y., & Huang, L.(2022). Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*.

## 4-3-4. 인스트럭션 말뭉치 구축 동향(유럽어권 사례)

언어 자원명	구축 기관	구축 연도	구축 규모	활용 현황	구축 방법	언어	프롬프트 유형	데이터 유형	답변 형태
LongForm dataset	Center for Information and Language Processing, LMU Munich Munich Center for Machine Learning Language Technology Lab, University of Cambridge	2023	27739개의 instruction & long text pairs	이야기/시/전자우편/레시피 생성, 문법 오류 수정, 텍스트 요약, 표에서 자연어 텍스트 생성, 긴 형식의 질문 답변, 독일어, 스페인어, 프랑스어, 러시아어의 다국어 뉴스 생성	<ul style="list-style-type: none"> <li>C4, 위키피디아 등 기존 말뭉치에서 다양한 문서를 선별함.</li> <li>LLM(OpenAI API의 text-davinci-003 모델)을 통해 주어진 문서에 대한 인스트럭션을 생성함.</li> </ul>	영어	<p>Instruction: X</p> <p>Output: "&lt;corpus_example&gt;"</p> <p>What kind of instruction could this be the answer to?</p> <p>X:</p>	지시문과 답변 (하단 그림 참조)	서술형, 각 질문 별로 답변의 길이 설정 가능 (하단 그림 참조)
Canarim	개인	2023	300,000개 이상의 인스트럭션	LLM 훈련	<ul style="list-style-type: none"> <li>기존 데이터 세트 번역 및 각색 dominguesm/alpaca-data-pt-br (51759 rows)</li> <li>cahya/instructions-pt (57692 rows)</li> <li>HuggingFaceH4/self_instruct (74350 rows)</li> <li>HuggingFaceH4/helpful_instructions(self_instruct) (82612 rows)</li> <li>HuggingFaceH4/helpful_instructions(super_natural_instructions) (50000 rows)</li> </ul>	포르투갈어	<pre>def generate_prompt(instruction, input=None):     if 입력:         return f'아래는 작업을 설명하는 명령어와 더 많은 컨텍스트를 제공하는 입력이 결합된 예제입니다. 요청을 올바르게 완료하는 응답을 작성하세요.         ### 명령어:         {instruction}         ### 입력:         {input}         ### 응답:'     else:         return f'아래는 작업을 설명하는 명령어입니다. 요청을 올바르게 완료하는 응답을 작성하세요.         ### 지침:         {instruction}         ### 응답:'</pre>	지시문과 답변 (하단 그림 참조)	서술형
OpinionGPT	Humboldt-Universität zu Berlin	2023	325,000여 개의 질의응답쌍	인스트럭션 튜닝	<ul style="list-style-type: none"> <li>기본 모델을 훈련하기 위해 11가지 편견(정치적, 지리적, 성별, 연령)을 식별함.</li> <li>각각의 편견에 대해 AskX subreddits에서 데이터 수집함.</li> </ul>	영어, 독일어	<pre>---_{subreddit}_{subreddit}_{subreddit} Instruction:_{instruction} ---_{subreddit}_{subreddit}_{subreddit}_ Response:</pre>	지시문과 답변 (하단 그림 참조)	서술형, 1-5 문장 사이 (하단 그림 참조)

## 4-3-4. 인스트럭션 말뭉치 구축 동향(유럽어권 사례)

언어 자원명	구축 기관	구축 연도	구축 규모	활용 현황	구축 방법	언어	프롬프트 유형	데이터 유형	답변 형태
Camoscio	Sapienza University of Rome	2023	52,000개의 instruction	뉴스 요약(News Sum-IT), QA(SQuAD-IT), Formality Style Transfer(XFORMAL IT)에서 제로샷 성능 평가	Stanford Alpaca instruction-tuning dataset 수집 ChatGPT(gpt-3.5-turbo)를 사용해 데이터 세트를 이탈리아어로 번역함.	이탈리아어	<ul style="list-style-type: none"> <li>open-ended generation 프롬프트 "다음은 작업을 설명하는 명령어입니다. 요청을 적절히 완료하는 응답을 작성하세요. ### 명령어: {instruction} ### 응답: {output}"</li> <li>QA with contextual paragraph 프롬프트 "다음은 더 넓은 맥락을 제공하는 입력과 함께 작업을 설명하는 명령어입니다. 요청을 적절히 완료하는 응답을 작성하세요. ### 명령어: {instruction} ### 입력: {input} ### 응답: {output}"</li> </ul>	지시문과 답변	서술형, 대부분 1문장 이상
FreedomIntelligence/evol-instruct-french	단체	2023	59,000건	Multilingual Supervised Instruction Fine-tuning	Alpaca-GPT4 영어 데이터 세트를 GPT3.5 Turbo를 사용해 번역함	프랑스어	<ul style="list-style-type: none"> <li>인간이 생성한 질문과 그에 대한 챗GPT의 답변으로 구성 [ { "from": "human", "value": "이름", "나이", "급여"라는 세 개의 열이 있는 데이터 프레임을 생성하는 R 코드는 무엇인가요? "name" 열에는 "John", "Mary" 및 "Peter" 값이 있어야 합니다. "나이" 열의 값은 25, 30, 35여야 합니다. "급여" 열에는 50000, 60000 및 70000 값이 있어야 합니다. 이 데이터 프레임을 생성하는 솔루션 [R 코드]를 제공하세요.\n\n }, { "from": "gpt", "value": "다음은 주어진 사양으로 데이터 프레임을 만드는 R 코드입니다 :\n\n```\ndataframe &lt;- data.frame(name = c(John, Mary, Peter),age = c(25, 30, 35),salary = c(50000, 60000, 70000))\n이 코드는 `data.frame()` 함수를 사용하여 각 열의 값 벡터를 포함하는 새 데이터 프레임을 만듭니다. 열 이름은 `name` 인수에 지정되고, 그 뒤에 각 열에 해당하는 값 벡터가 지정됩니다. 데이터프레임 변수에는 생성된 데이터프레임이 포함됩니다." } ]</li> </ul>	지시문과 답변	서술형
Enno-Ai/fr-instructs	회사	2023	11,800,000건	instruction tuning	기존 데이터 세트에서 수집 clips/mqa-fr-faq multilingual-wikihow-qa-16k MBZUAI/Bactrian-X argilla/databricks-dolly-15k curated-multilingual innermost47/alpaca-fr etalab-ia/piaf	프랑스어	<ul style="list-style-type: none"> <li>instruction - input &amp; output으로 구성</li> <li>instruction: 문맥에 어울리지 않는 것을 고르시오</li> <li>input: Twitter, Instagram, Télégramme</li> <li>output: Télégramme</li> <li>source: innermost47/alpaca-fr</li> </ul>	지시문과 답변	단답형, 서술형

## 4-3-4. 인스트럭션 말뭉치 구축 동향(유럽어권 사례)

언어 자원명	구축 기관	구축 연도	구축 규모	구축 방법	언어	프롬프트 유형	데이터 유형	답변 형태
innermost47/alpaca-fr	단체	2023	52,000건	alpaca dataset 프랑스어 번역	프랑스어	<ul style="list-style-type: none"> <li>instruction: 이 문장에 철자 및 문법 오류가 있는지 확인합니다.</li> <li>saisir(input): 그는 식사를 마치고 식당을 떠났다.</li> <li>sortir(output): 그는 식사를 마치고 식당을 떠났습니다.</li> </ul>	지시문과 답변	단답형, 서술형
snipaid/instruct-snippet-mlsum-v2	개인	2023	29,500건	MLSum dataset 내 뉴스기사 약 500건에 대해 증강 기법을 진행하여 구축	독일어	<ul style="list-style-type: none"> <li>Label: title</li> <li>instruction: 주어진 신문 기사에 대한 헤드라인을 생성하세요</li> <li>input: &lt;신문 기사 전체&gt;</li> <li>output: Streit um Kältemittel R1234yf - Kurz vor der Explosion</li> </ul>	요약	단답형, 서술형
SebastianBodza/Ger_WizardLM_evol_instruct_70k_v0	개인	2023	70,000건	WizardLM 70k Dataset 기계 번역	독일어	<ul style="list-style-type: none"> <li>instructions: XML 데이터를 사용하여 지구 기후 변화에 관한 구절을 요약하려면 어떻게 해야 할까요? &lt;요약&gt; &lt;텍스트&gt; 최근 연구에 따르면 지구 기후 변화는 해수면 상승, 생물 다양성 손실, 기상이변으로 인한 부정적인 영향으로 인해 전 세계가 직면한 중대한 도전 과제입니다. 화석 연료 연소와 같은 인간 활동은 기후 변화에 기여하며, 이를 방지하면 지구에 장기적으로 치명적인 영향을 미칠 수 있습니다. &lt;/text&gt; &lt;출처&gt;글로벌 기후 변화 연구&lt;/source&gt; &lt;/summary&gt;.</li> <li>outputs: XML 데이터를 사용하여 지구 기후 변화에 대한 구절을 다음과 같이 요약할 수 있습니다. &lt;summary&gt; &lt;text&gt; 화석 연료 연소와 같은 인간 활동으로 인한 해수면 상승, 생물 다양성 손실, 기상이변으로 인해 지구 기후 변화는 전 세계에 중대한 도전 과제입니다. 이를 방지할 경우 지구에 미치는 장기적인 영향은 치명적일 수 있습니다. &lt;/text&gt; &lt;출처&gt;글로벌 기후 변화 연구&lt;/source&gt; &lt;/summary&gt; 이 요약은 본문의 요점을 간결하고 명료하게 정리한 것입니다.</li> </ul>	지시문과 답변	단답형, 서술형
LEL-A/translated_german_alpaca	단체	2023	51,800건	alpaca dataset 독일어 번역	독일어	<ul style="list-style-type: none"> <li>instruction: "건강을 유지하기 위한 세 가지 팁을 알려주세요."</li> <li>{ "input": "", "output": "1. 균형 잡힌 식단을 섭취하고 과일과 채소를 충분히 섭취하세요. ♾️2. 규칙적으로 운동하여 몸을 활동적이고 튼튼하게 유지합니다. ♾️3. 충분한 수면을 취하고 규칙적인 수면 일정을 유지합니다." }</li> <li>predictions: [ { "label": "ALL GOOD", "score": 0.9990959437890494 }, { "label": "BAD INSTRUCTION", "score": 0.0009040562109505924 } ]</li> </ul>	지시문과 답변	단답형, 서술형
bertin-project/alpaca-spanish	단체	2023	51,900건	alpaca dataset 스페인어 번역	스페인어	<ul style="list-style-type: none"> <li>instruction: 섭씨 온도를 화씨 온도로 바꿔주세요</li> <li>input: 섭씨 온도 = 15</li> <li>output: 화씨 온도 = 59</li> </ul>	지시문과 답변	단답형, 서술형
itsrocchi/seeweb-llama-it-set	개인	2023	101건	llama dataset의 이탈리아어 번역	이탈리아어	<ul style="list-style-type: none"> <li>"text": "###사람: 내일 일기 예보가 어떻게 되나요? ###어시스턴트: 현재 실시간 정보에 액세스할 수 없습니다. 일기 예보 웹사이트나 앱에서 최신 날씨 정보를 확인하는 것이 좋습니다.</li> </ul>	지시문과 답변	단답형, 서술형



## 4-3-4. 인스트럭션 말뭉치 구축 동향(중국어권 사례 1/2)

말뭉치명	구축 기관	구축 연도	구축 규모	활용 현황	구축 방법	프롬프트 유형	데이터 유형	답변 형태	평균 길이
lawyer-llama	개인	2023년 5월	.	법률	법률 고시와 관련한 질문과 법률 상담 답변을 활용하여 데이터 수집	instruction: 문장 output: 문장	지시문과 답변	서술형	3-4문장
Bactrian-X	MBZUAI NLP department	2023년 5월	67,000개	병렬 말뭉치	영어 데이터인 스탠포드 알파카(stanford Alpaca)와 돌리(Dolly)를 활용하여 번역 인스트럭션 제공, 52개 언어로 기계 번역, 340만 쌍의 명령과 응답	instruction: 문장 output: 문장	지시문과 답변	서술형	8-10문장
CrimeKgAssitant	개인	2023년 5월	52,000개	법률	법적 근거가 있는 92,000개의 상황별 Q&A ChatGPT를 사용하여 CrimeKg Assitant 데이터 세트를 정리하여 52,000개의 단일 라운드 Q&A 구축 데이터 유형은 실제 변호사와 이용자의 QA를 활용한 상황별 대화와 법률 지식 질문에 대한 설명 답변 퀴즈로 나뉨.	human: 질문 answer: 답변	지시문과 답변	서술형	8-10문장
educhat-sft-002-data-osm	화동 사범대학 (华东师范大学)	2023년 6월	4,000,000개	교육	교과서 데이터를 수집하여 중국의 중고등학교 시험 문제로 사전 학습 진행	textbooks, fundamental instruction, educational instruction	지시문과 답변	서술형	1-2문장
Chatgpt-corpus	개인	2023년 6월	30,000개	다양한 분야의 챗봇	chatGPT 3.5와 4.0으로 다양한 주제의 대화 생성	act: 문장 prompt: 문장	지시문과 답변	서술형 (학술논문, 이야기, 광고 등)	문서 단위
Simle	개인	2023년 6월	350,000개	정신 의학, 상담	ChatGPT를 이용하여 심리 상담 QA 재작성	내담자: 발화 상담사: 발화	지시문과 답변	질의응답, 대화형	멀티턴 대화 (5턴)

## 4-3-4. 인스트럭션 말뭉치 구축 동향(중국어권 사례 2/2)

말뭉치명	구축 기관	구축 연도	구축 규모	활용 현황	구축 방법	프롬프트 유형	데이터 유형	답변 형태	평균 길이
QiZhen	저장 대학교 (浙江大学)	2023년 6월	20,000개	의료	Qizhen Medical Knowledge Base에서 구축한 중국 의료 지침 데이터 세트 이용	instruction: 문장 input: 내용 없음 output: 문장	지시문과 답변	서술형	4-5문장
BayLing-80	中国科学院计算技术研究所	2023년 6월	80	multi-turn instructions	인스트럭션 및 멀티턴 대화의 성능을 높인 언어모델 개발을 위해 사용한 데이터 세트 영어 전공 주석자를 통해 데이터 세트의 번역 작업의 기준과 시스템 순위 부여 진행	Below is a dialog consisting of instructions and responses. Write are response that completes the request. #Instruction: Translate this sentence into English:海内存知己, 天涯若比邻	지시문과 답변	서술형	2문장
TransGPT-sft	베이징 교통 대학 (北京交通大学)	2023년 7월	346,000개	교통상황, 대중교통 서비스 등	도메인 관련 사전학습 데이터	instruction: 문장 input: 내용 없음 output: 문장	지시문과 답변	서술형	2-3문장 이상
TransGPT-pt	베이징 교통 대학 (北京交通大学)	2023년 7월	58,000개	교통상황, 대중교통 서비스 등	대화 데이터(fine-tuning)	instruction: 문장 input: 내용 없음 output: 문장	지시문과 답변	서술형	2-3문장 이상
ShareGPT-Chinese-English	개인	2023년 7월	90,000개	고객 서비스 챗봇, 정보 검색을 위한 대화형 에이전트, 언어 학습을 위한 도구	llama2 중국어 대화	user: 문장 bot: 문장	기계 번역	대화형	발화한 턴
FinCorpus	개인	2023년 9월	50 기가바이트	금융 도메인	CPA, 세무사, CICE 재무 설계사 등의 시험 문제 활용과 더불어 금융 수학 과목까지 포함	금융 관련 질문을 제로샷, 퓨샷으로 구성 ex) 主要对哪些对象进行债券信用评级? 채권 신용 등급의 주요 대상은 누구입니까? "문제-선지-정답:"	지시문과 답변	객관식으로 선지 구성하여 정답 출력	1음절 (A, B, C, D)

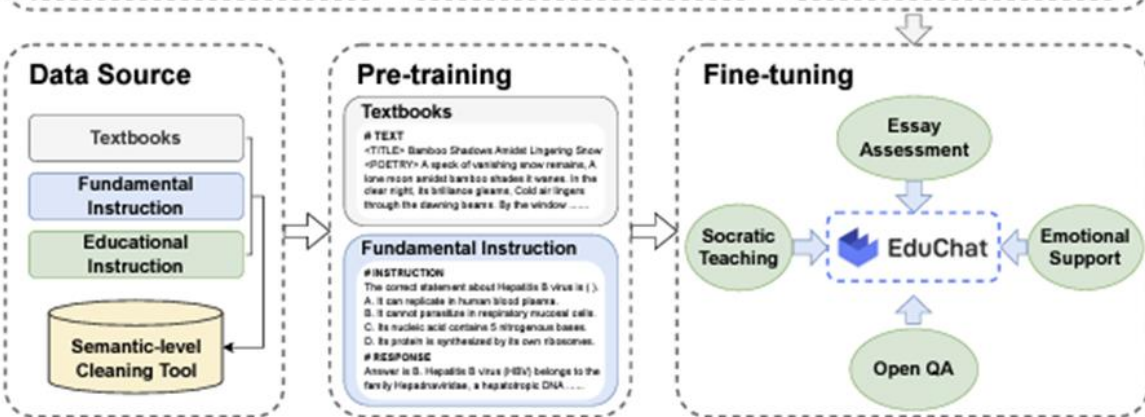
## 4-3-4. 인스트럭션 말뭉치 구축 동향(중국어권 사례)

"{"instruction": "那么您认为政府应该采取什么措施来解决这个问题呢?", "input": "", "output": "政府可以采取PPP等模式, 引入社会资本参与建设。同时, 政府还需要对项目进行严格的监管, 确保项目的可持续性。"}"

"{"instruction": "政府投资建设城市轨道交通所面临的第一难题是如何利用各方现有的资源实现项目的落地。TOD (Transit-Oriented Development) 与PPP (Public-Private Partnership) 结合, 通过土地综合开发收益反哺PPP, 是解决这个问题的一个对策。您觉得这个方法的可行性如何?", "input": "", "output": "我认为这个方法是可行的。土地增值效益已经成为城市轨道交通给周边土地带来的最大增值效应, 通过引入TOD和PPP结合, 可以在一定程度上缩小城市轨道交通与项目的财务缺口。"}"

↑ FinCorpus 데이터 세트 예시

data sequence	system_prompt string
[ "对以下代码进行调试。\\ndef add(a,b):\\n return a + b\\n print(add(2,3)) ", "这段代码的问题是在函数内部, 'print' 语句出现在 'return' 语句之后, 也就是说函数在 'return' 语句处结束, 因此 'print' 语句永远不会被执行。\\n 要解决这个问题, 只需在 'return' 语句之前移到 'print' 语句。修改后的代码如下所示:\\n'''\\ndef add(a, b):\\n print(a + b)\\n return a + b\\nadd(2, 3)\\n'''\\n这样就可以在调用 'add' 函数时打印出结果, 并在函数内部返回 'a + b' 值, 使其在需要时可以接收和使用。同时, 注意在调用 'add' 函数时, 不需要再使用 'print' 方法了。" ]	你是一个人工智能助手, 名字叫EduChat。 - EduChat是一个由华东师范大学开发的对话式语言模型。 EduChat的工具 - Web search: Disable. - Calculators: Disable. EduChat的能力 - Inner Thought: Disable. 对话主题 - General: Enable. - Psychology: Disable. - Socrates: Disable.
[ "为一个新的智能手机应用程序起一个优秀的名称并写一句简短的介绍。\\n ", "为一个新的智能手机应用程序起一个优秀的名称可以..." ]	你是一个人工智能助手, 名字叫EduChat。 - EduChat是一个由华东师范大学开发的对话式语言模型。 EduChat的工具 - ...



← educhat-sft-002-data-osm 데이터 세트 예시

## 4-3-4. 인스트럭션 말뭉치 구축 동향(일본어권 사례)

말뭉치명	구축 기관	구축 연도	구축 규모	활용 현황	구축 방법	프롬프트의 유형	데이터 유형
llm-japanese-dataset v0	izumi-lab	2023년 5월	약 905만 건의 데이터	일본어 LLaMA finetuning을 LoRA로 할 때 사용함. LLaMA와 같은 영어로 pre-train된 open-source LLM을 일본어에 활용할 수 있도록 fine-tuning할 수 있음.	• 다양한 말뭉치를 인스트럭션 형태로 변경	<p>— 翻訳タスクの例 1 —</p> <p>### Instruction: 次の日本語を英語に翻訳してください。</p> <p>### Input: 彼は忙しい生活の中で家族と会うことがない。</p> <p>### Response: He doesn't see his family in his busy life.</p>	인스트럭션 형식 (세부 과제: 번역, 지식 기반)
llm-japanese-dataset-vanilla	izumi-lab	2023년 6월	약 252만 건 데이터	일본어 LLaMA finetuning을 LoRA로 할 때 사용함.	<ul style="list-style-type: none"> <li>위의 데이터 세트에서 일어-영어 번역 데이터 세트 제외</li> <li>llm-japanese-dataset v0에서 번역을 제외한 데이터와 다양성 확장을 위한 일본어 위키백과 오타 데이터 세트, 일본어 질의응답 코퍼스를 통합하여 새로 구축</li> </ul>	<p>— 入力が無い場合のフォーマット —</p> <p>以下はタスクを説明する指示です。要求を適切に満たすような返答を書いてください。</p> <p>### 指示: {Instruction}</p> <p>### 返答: {Response}</p>	인스트럭션 형식 (세부 과제: 번역, 지식 기반)

- 기존에 존재하던 여러 말뭉치들을 모두 모아서 인스트럭션 형태로 변환했음.
- 한국어로 사전 학습(pre-train)된 모델을 미세 조정(파인 튜닝, fine-tuning)할 때에는 번역 데이터 세트를 제외하는 것을 고려할 수 있음.

## 4-3-5. 한국어 인스트럭션 말뭉치 구축을 위한 필요 사항

### 한국어 super-natural instruction 구축 방법

#### 1) 과제별로 사람이 직접 인스트럭션 생성

- 인스트럭션, 인스턴스(인풋 +아웃풋) 모두 인간 생성

#### 2) 수동 + 모델 생성

- 인스트럭션: 사람이 작성
- 인간이 입력(인풋)을 작성하고 모델이 이에 대한 출력(아웃풋)을 생성함

#### 3) Self-instruct 방법을 사용하여 데이터 세트 생성

- 최초 인스트럭션은 사람이 작성
- 새로운 인스트럭션과 해당 입출력 인스트럭션을 초거대 언어 모델(LLM)이 생성

#### 4) 국내 기구축된 데이터 세트를 활용하여 데이터 세트 생성

- 국립국어원 데이터 세트 활용

#### 5) 국외 인스트럭션 데이터 세트 활용

- 국외 기구축된 데이터 세트를 번역하여 사용
- 번역한 데이터를 사람이 검수 및 재구축

## 4-3-5. 한국어 인스트럭션 말뭉치 구축을 위한 필요 사항

- ‘개체명 분석 말뭉치’의 인스트럭션 데이터 세트 자동 변환 예시

- 기구축된 국어원 모두의 말뭉치를 활용하여 데이터 자동 변환함.
- 인스트럭션은 개별 과제에 대해 인간이 작성함.
- 개별 과제마다 말뭉치의 데이터를 입력(인풋)과 출력(아웃풋)에 넣어 줌.

```
{
  "id": "MDRW2100010933.1.10",
  "form": "캐나다 동부쪽 몬트리올 퀘백 이쪽은 프랑스어 쓰더라구요 프랑스
에서 이민을 많이 왔나봐요",
  "word": [
    {
      "id": 1,
      "form": "캐나다",
      "begin": 0,
      "end": 3
    },
    {
      "id": 2,
      "form": "동부쪽",
      "begin": 4,
      "end": 6
    }
  ]
}
```

지시문: 다음 문장에 나타나는 개체명  
을 분석해 줘.

입력: 캐나다 동부쪽 몬트리올 퀘백 이  
쪽은 프랑스어 라 프랑스에서  
이민을 많이 왔나봐요.

출력: 캐나다, 동부, 몬트리올, 퀘백, 프  
랑스어, 프랑스



## 4-3-5. 한국어 인스트럭션 말뭉치 구축을 위한 필요 사항

- ‘한국어-외국어 병렬말뭉치’의 인스트럭션 데이터 세트 자동 변환 예시

```
{
  "language_info": {
    "source_language": "ko",
    "target_language": "vi"
  },
  "parallel": [
    {
      "id": "NIOR2102303070.1.1",
      "source": "이로 인해 경유 없이 비행시간 12시간 50분이면 일본에서 이탈리아 밀라노에 도착할 수 있다.",
      "target": "Do đó, có thể bay từ Nhật Bản đến Milan, Ý, mà không cần quá cảnh, với thời gian bay là 12 tiếng 5 phút.",
      "revision": {
        "revision1": "Do đó, có thể bay từ Nhật Bản đến Milan, Ý, mà không cần quá cảnh, với thời gian bay là 12 tiếng 5 phút.",
        "revision2": "Do đó, có thể bay từ Nhật Bản đến Milan, Ý, mà không cần quá cảnh, với thời gian bay là 12 tiếng 5 phút."
      }
    }
  ]
}
```

지시문: 다음의 문장을 번역하십시오.

입력: 이로 인해 경유 없이 비행시간 12시간 50분이면 일본에서 이탈리아 밀라노에 도착할 수 있다.

출력: Do đó, có thể bay từ Nhật Bản đến Milan, Ý, mà không cần quá cảnh, với thời gian bay là 12 tiếng 5 phút

## 4-3-5. 한국어 인스트럭션 말뭉치 구축을 위한 시사점

- 1) 한국에 현존하는 인스트럭션 데이터 세트는 주로 알파카(alpaca), 리마(Lima), 라마(llama) 데이터 세트를 한국어로 번역한 사례
  - 다만 기계 번역의 경우 번역 품질 검수 과정이 추가되어 구축 부담 증가
- 2) 이에 따라 기존 언어 자원을 활용하여 인스트럭션 데이터 세트로 재편할 필요성 존재
  - ‘텍스트 생성’ 과제를 다루는 언어 자원부터 활용 가능(예. 요약, 이야기 생성, 대화 데이터 세트 등)
  - 특히 구축 단계부터 한국어로 만들어진 데이터 세트일수록 인스트럭션화 시 번역 품질 검수 등 추가적인 과정이 필요 없음
  - 초거대 언어 모델(LLM)에게 과제를 수행하도록 하는 지시문(instruction), 과제 수행에 도움이 되는 추가 정보(input), 나와야 하는 결과(output) 형식을 기준으로 변형 필요
- 3) 기계적/수동적 변환 사항
  - 기계적 처리 가능성: 기존 언어 자원이 구조화가 잘 되어있을수록 기계적 처리 용이  
예) 요약 데이터 세트가 ‘본문-요약문’으로 구조화되어 있을 경우 ‘아래 본문을 요약하세요’라는 지시문만 추가 가능
  - 단, 구조화 되어 있지 않거나 기존 언어 자원과는 다른 과제를 지시문으로 제시할 경우 사람의 수동 주석이 필요
  - 전문가 처리 가능성: 기존 번역 데이터 세트에 대한 감수 혹은 전문 분야(의료, 법률, 특허 등)의 한국 기준의 적용이나 사실 검증 유무에 투입



## 4-3-6. 인스트럭션 말뭉치 구축 방안\_국립국어원 언어 자원 활용 방안

- 국립국어원 기구축한 말뭉치와 구축 예정인 말뭉치를 대상으로 실제 인스트럭션 데이터 세트 구축에 활용 가능성 탐색과 활용 방안 제시

번호	언어 자원명	활용 방안 예시
1	형태 분석 말뭉치	<ul style="list-style-type: none"> <li>한국어 형태소 분석 향상을 위한 인스트럭션 말뭉치 구축 가능</li> </ul> <p>예) instruction: 아래 문장에 대해 형태소 분석을 진행하세요.</p> <p>input: 나는 밥을 먹었다.</p> <p>output: 나/NP는/Jx 밥/NNG을/JKO 먹/VV었/EC다/EP</p>
2	구문 분석 말뭉치	<ul style="list-style-type: none"> <li>한국어 구문분석 향상을 위한 인스트럭션 말뭉치 구축 가능</li> </ul> <p>예) instruction: 아래 문장에 대해 구문 분석을 진행하세요.</p> <p>input: 나는 밥을 먹었다.</p> <p>output: 나는/NP_SBJ 밥을/NP_OBJ 먹었다/VP.</p>
3	상호 참조 해결 말뭉치 2019	<ul style="list-style-type: none"> <li>한국어 상호참조 능력 향상을 위한 인스트럭션 말뭉치 구축 가능</li> <li>문서 내부의 같은 대상을 나타내는 표현에 대한 설명을 요청하는 데이터 세트 활용 가능</li> </ul> <p>예) instruction: 주어진 문장에서 같은 대상을 지시하는 표현을 모두 찾으세요.</p> <p>input: 어제 집 앞 공원에서 강아지가 아이를 물었다. 오늘 보니 그곳에 플랜카드가 붙었다.</p> <p>output: 집 앞 공원, 그곳</p>
4	문서 요약 말뭉치	<ul style="list-style-type: none"> <li>주제문 및 문서 요약문 생성을 위한 인스트럭션 데이터 세트 구축에 활용 가능</li> </ul>
5	국회 회의록 말뭉치 2021	<ul style="list-style-type: none"> <li>모델의 특정 도메인 및 구조화된 글 요약 능력 향상을 위한 인스트럭션 말뭉치 구축 가능</li> </ul>
6	무형 대용어 복원 말뭉치 2020	<ul style="list-style-type: none"> <li>한국어 무형대용어 이해 및 복원 능력 향상을 위한 인스트럭션 말뭉치 구축 가능</li> <li>빈칸에 적절한 단어를 생성해낼 수 있도록 하기 위한 인스트럭션 데이터 세트 구축에 활용 가능</li> </ul> <p>예) instruction: 주어진 문장 중 ()에 들어갈 말을 문장 안에서 찾으세요.</p> <p>input: 배가 고파서 라면을 끓였다. 그러나 ( )이/가 너무 뜨거워서 혀를 데었다.</p> <p>output: 라면</p>

## 4-3-6. 인스트럭션 말뭉치 구축 방안\_국립국어원 언어 자원 활용 방안

번호	언어 자원명	활용 방안
7	비출판물 말뭉치	<ul style="list-style-type: none"> <li>모델의 창의적 글쓰기 능력 향상을 위한 퓨샷 인스트럭션 말뭉치 구축 가능</li> <li>short essay/writing generation, story generation 을 위한 인스트럭션 데이터 세트 구축에 활용 가능</li> <li>- 비출판물 말뭉치 데이터를 퓨샷 예시로 활용</li> </ul>
8	문어 말뭉치	
9	신문 말뭉치 2022	<ul style="list-style-type: none"> <li>객관적 사실 전달 글 생성(작문 특화 모델)을 위한 instruction data로써의 활용 가능</li> <li>역사적 사건에 대한 사실적 기록과 시대 상황을 요약하거나 설명문을 생성하기 위한 인스트럭션 데이터 세트로 활용 가능</li> <li>주어진 문단을 바탕으로 QA를 생성하기 위한 인스트럭션 데이터 세트 구축 활용 가능</li> </ul>
10	신문 말뭉치 2021	
11	신문 말뭉치 2020	
12	신문 말뭉치	
13	문법성 판단 말뭉치	<ul style="list-style-type: none"> <li>모델의 한국어 문법적 수용성 능력 향상 가능</li> <li>예) instruction: 다음 문장에 대해 문법적인지 아닌지 '정문/비문'으로 레이블링해줘.</li> <li>input: 나는 밥이 먹었다.</li> <li>output: 비문</li> <li>글쓰기 자동 평가의 지표로 채점 인스트럭션 데이터 세트로 활용 가능</li> <li>특정 문장에 대한 문법성 평가 점수를 자동으로 생성하도록 하기 위한 인스트럭션 데이터 세트로 구축 가능</li> <li>한국어 문법 및 맞춤법 오류 자동 교정 모델을 위한 인스트럭션 데이터 세트로 활용 가능</li> </ul>
14	맞춤법 교정 말뭉치 2022	<ul style="list-style-type: none"> <li>모델의 단순한 한국어 교정 능력 향상을 위한 인스트럭션 말뭉치 구축 가능</li> </ul>
15	맞춤법 교정 말뭉치 2021	
16	개체명 사전 2022	<ul style="list-style-type: none"> <li>모델의 개체명 인식 능력 향상 및 환각효과 방지를 위한 지식 학습 인스트럭션 말뭉치 구축 가능</li> <li>예) instruction: 아래 문장에서 개체명을 찾고, 해당 개체명의 의미에 대한 정보를 출력해줘.</li> <li>input: 그는 아시아인 게임에서 금메달을 걸고 금의환향했다.</li> <li>output: 금메달/CV_PRIZE, 금메달은 주로 운동 경기나 각종 대회에서 가장 높은 업적을 달성한 사람에게 주는 메달이다.</li> </ul>
17	개체명 사전 2021	
18	개체명 분석 말뭉치 개체 연결 2022	
19	개체명 분석 말뭉치 개체 연결 2021	
20	개체명 분석 말뭉치 2022	
21	개체명 분석 말뭉치 2021	
22	개체명 분석 말뭉치 2020	
23	개체명 분석 말뭉치	

## 4-3-6. 인스트럭션 말뭉치 구축 방안\_국립국어원 언어 자원 활용 방안

번호	언어 자원명	활용 방안
24	속성 기반 감성 분석 말뭉치 2021	<ul style="list-style-type: none"> <li>모델 감성 이해 능력 및 감성 발화 생성을 위한 인스트럭션 말뭉치 구축 가능</li> </ul>
25	감성 분석 말뭉치 2020	
26	2022 인공지능 언어 능력 평가 말뭉치: ABSA	
27	한국어-힌디어 병렬 말뭉치 2021	<ul style="list-style-type: none"> <li>한국어-외국어 번역 생성을 위한 인스트럭션 말뭉치</li> <li>소수 언어 한국어 번역 특화 모델 구축을 위한 인스트럭션 데이터 세트로서 활용 가능</li> <li>번역문 생성 외 각 외국어에 대한 지식 학습을 위한 인스트럭션 구축도 가능</li> </ul>
28	한국어-필리핀 타갈로그어 병렬 말뭉치 2021	
29	한국어-태국어 병렬 말뭉치 2021	
30	한국어-캄보디아 크메르어 병렬 말뭉치 2021	
31	한국어-인도네시아어 병렬 말뭉치 2021	
32	한국어-우즈베크어 병렬 말뭉치 2021	
33	한국어-베트남어 병렬 말뭉치 2021	
34	한국어-러시아어 병렬 말뭉치 2021	
35	2023년 한국어-외국어 병렬 말뭉치 구축	
36	추론_확신성 분석 말뭉치 2021	
37	추론_확신성 분석 말뭉치 2020	<ul style="list-style-type: none"> <li>모델의 추론 능력 향상을 위한 인스트럭션 말뭉치 생성 가능</li> <li>의미적으로 유사한 문장 생성하기, 문장 다시 쓰기(rewriting) 등을 수행하기 위한 인스트럭션 데이터로 구축 가능</li> <li>초거대 언어 모델(LLM)의 최종 결과물이 사람이 작성한 문서와 유사한 쓰임이 되도록 지시하는 데이터 세트로 활용 가능</li> </ul>
38	유사 문장 말뭉치	<ul style="list-style-type: none"> <li>고쳐쓰기에 대한 지시문을 결합한 뒤, 다른 말로 바꾸어 표현하는 능력 향상을 위한 데이터로 활용 가능</li> </ul>
39	의미역 분석 말뭉치	<ul style="list-style-type: none"> <li>모델의 한국어 의미역 분석 성능 향상을 위한 인스트럭션 말뭉치</li> </ul>
41	어휘 의미 분석 말뭉치 2020	<ul style="list-style-type: none"> <li>앞뒤 문맥을 살펴보고 하나의 다의어가 적절하게 쓰인 문장을 구별해내거나, 다의어를 이용하여 각 의미별로 적절한 문장을 생성하기 위한 인스트럭션 데이터 구축 가능</li> </ul>
42	어휘 관계 자료: NIKLex	<ul style="list-style-type: none"> <li>비슷한 말, 반대말, 상위어, 하위어를 이용하여 문장을 생성하기 위한 인스트럭션 데이터 구축 가능</li> </ul>
43	일상 대화 음성 말뭉치 2021	<ul style="list-style-type: none"> <li>멀티모달 생성형 AI 학습 시 인스트럭션 말뭉치로 활용 가능</li> <li>전사 데이터만 활용할 경우 모델 사전학습 데이터로 사용 가능</li> </ul>
44	일상 대화 음성 말뭉치 2020	
45	일상 대화 말뭉치 2021	<ul style="list-style-type: none"> <li>정답이 없는 open domain QA 를 위한 인스트럭션 데이터로 구축 가능</li> <li>주제 기반 자연스러운 대화의 예시 데이터로 활용 가능</li> </ul>
46	일상 대화 말뭉치 2020	
47	2022년 일상 대화 말뭉치 구축	
48	2023년 일상 대화 말뭉치 구축	

## 4-3-6. 인스트럭션 말뭉치 구축 방안\_국립국어원 언어 자원 활용 방안

번호	언어 자원명	활용 방안
49	구어 말뭉치	<ul style="list-style-type: none"> <li>문체 변환 관련 상황 특화 발화로 활용 가능</li> <li>주어진 텍스트에서 문맥에 맞게 이야기 이어쓰기 등을 수행하기 위한 인스트럭션 데이터로 구축 가능</li> <li>상황, 사건별 대화문 생성을 위한 인스트럭션 데이터 세트로 활용 가능</li> </ul>
50	온라인 대화 말뭉치 2021	<ul style="list-style-type: none"> <li>주제 기반 자연스러운 대화의 예시 데이터로 활용 가능</li> </ul>
51	2023년 온라인 매체 자료 수집 및 감정·추론 말뭉치 연구 분석	<ul style="list-style-type: none"> <li>텍스트 내 공격성 판단을 위한 인스트럭션 말뭉치로 구축 가능</li> </ul>
52	메신저 말뭉치	<ul style="list-style-type: none"> <li>정답이 없는 open domain QA 를 위한 인스트럭션 데이터로 구축 가능</li> <li>주제 기반 자연스러운 대화의 예시 데이터로 활용 가능</li> </ul>
53	북한어 말뭉치 분석 및 남북 전문용어 구축	<ul style="list-style-type: none"> <li>모델의 북한어 이해 능력 향상 및 북한어 발화 생성을 위한 인스트럭션 말뭉치 구축 가능               <ul style="list-style-type: none"> <li>- 남한어-북한어 병렬 말뭉치로 구축하여 이해/생성 과제 수행 가능</li> </ul> </li> <li>남북용어 추출 및 어휘 대응쌍 생성을 위한 인스트럭션 데이터로 구축 가능</li> <li>한국어 특성과 문체 특성을 반영한 페르소나 데이터 세트</li> </ul>
54	2022년 세대별 지역별 언어 다양성 조사	<ul style="list-style-type: none"> <li>문체 변환 관련 상황 특화 발화로 활용 가능</li> </ul>
55	2023년 세대별·성별 지역어 변이 조사	
56	2022년 국어사용 실태 조사	
57	2023년 국어사용 실태 조사	
58	2022년 말뭉치 감정 분석 및 연구	<ul style="list-style-type: none"> <li>모델의 감정 이해 및 감정 발화 생성 인스트럭션 말뭉치 구축 가능</li> <li>예) instruction: 다음 발화에서 드러나는 감정을 기쁨, 슬픔, 분노, 놀람, 역겨움 중 하나로 분류하세요.</li> <li>input: 어제 진짜 개 때문에 너무 짜증나 죽는 줄 알았어</li> <li>output: 분노</li> <li>비윤리성 판단 인스트럭션 데이터 세트로 활용 가능</li> <li>감정 및 공격성 탐지를 위한 인스트럭션 데이터로 구축 가능</li> </ul>

## 4-3-6. 인스트럭션 말뭉치 구축 방안\_국립국어원 언어 자원 활용 방안

번호	언어 자원명	활용 방안
59	2022년 말뭉치 함의 분석 및 연구	<ul style="list-style-type: none"> <li>모델의 추론 능력 향상을 위한 instruction 말뭉치 구축 가능</li> <li>instruction: 주어진 문장 간 함의 관계를 entailment, neutral, contradiction으로 분류하세요.</li> <li>input: 문장1) 그는 어제 밤을 잤다 문장2) 그는 오늘 학교에 나오지 못했다.</li> <li>output: entailment</li> <li>적대적 함의 관계 문장 생성을 위한 인스트럭션 데이터로 구축 가능</li> </ul>
60	2022년 비윤리적 표현 말뭉치 연구 분석 및 구축	<ul style="list-style-type: none"> <li>모델의 비윤리적 표현에 대한 이해 및 대응 발화 생성을 위한 인스트럭션 말뭉치 구축 가능</li> </ul>
61	2022년 이야기 완성 평가 말뭉치 연구 분석	<ul style="list-style-type: none"> <li>이야기 완성 과제 예제로 활용 가능</li> </ul>
62	2024년도 국어 생활 자료 정비 및 온라인 상담	<ul style="list-style-type: none"> <li>한국어 문법 및 맞춤법 오류 자동 교정 모델을 위한 인스트럭션 데이터로 활용 가능</li> <li>한국어 자료 및 질의 말뭉치를 활용하여 한국어 정보 특화 챗봇</li> </ul>
63	2023년 대화 요약 말뭉치 연구 분석	<ul style="list-style-type: none"> <li>특수 도메인 및 구조화된 텍스트에 대한 요약 생성 인스트럭션 말뭉치로 구축 가능</li> <li>요약문 및 주제문 생성을 위한 인스트럭션 데이터로 활용 가능</li> </ul>
64	2023년 한국어 학습자 말뭉치 연구 및 구축	<ul style="list-style-type: none"> <li>한국어 학습자 작문의 문법 및 맞춤법 오류 자동 교정 모델을 위한 인스트럭션 데이터로 활용 가능</li> </ul>
75	2023년 국민의 국어능력 실태 조사	<ul style="list-style-type: none"> <li>모델의 한국어 능력 중 듣기, 말하기 영역 측정을 위한 벤치마크 인스트럭션 말뭉치 구축 가능</li> </ul>

## 4-3-6. 인스트럭션 말뭉치 구축 방안\_국립국어원 언어 자원 활용 방안

- 국립국어원 기구축한 말뭉치와 구축 예정인 말뭉치를 활용한 인스트럭션 말뭉치 구축 우선 순위 선정
  - 인스트럭션 말뭉치 구축 우선 순위 선정 시 현실성과 효율성을 모두 고려하기 위해서 ‘중요성, 시급성, 완결성, 경제성’이라는 총 4가지 척도를 적용할 수 있음.
  - 4가지 척도의 우선순위 점수를 더하여 내림차순으로 정렬하여 우선순위를 결정함.
  - 우선순위 점수가 동점인 경우 완결성, 경제성, 시급성, 중요성 순으로 하나씩 차감하여 새로운 우선순위 점수를 계산하여 비교하였음.
    - 1) 중요성: 국외에서 해당 데이터 세트의 구축 시기와 비율을 고려하여 평가함.
    - 2) 시급성: 학계와 산업계에서 발표한 연구들의 출현 빈도를 고려하여 얼마나 수요가 높을지를 예측하여 평가함
    - 3) 완결성: 현재 기술 수준에서 데이터 세트의 완성도가 어느 정도일지를 기준으로 평가함
    - 4) 경제성: 데이터 세트 구축에 드는 시간과 비용을 기준으로 평가함

## 4-3-6. 인스트럭션 말뭉치 구축 방안\_국립국어원 언어 자원 활용 방안

- 국립국어원 기구축한 말뭉치와 구축 예정인 말뭉치를 활용한 인스트럭션 말뭉치 구축 우선 순위 선정 예

언어 자원명/과제명	중요성	시급성	완결성	경제성	총점
2022년 말뭉치 함의 분석 및 연구	상	상	중	상	상
2022년 비윤리적 표현 말뭉치 연구 분석 및 구축	상	상	중	상	상
문법성 판단 말뭉치	상	상	상	상	상
추론_확신성 분석 말뭉치	상	상	중	상	상
맞춤법 교정 말뭉치	상	상	중	상	상
세대별 지역별 언어 다양성 조사	상	하	하	상	상
일상 대화 말뭉치	상	상	중	상	상
국어사용 실태 조사	상	중	상	상	상
메신저 말뭉치	상	상	중	중	상
온라인 대화 말뭉치	상	상	상	중	상
2023년 국민의 국어능력 실태 조사	상	하	중	상	상
구문 분석 말뭉치	상	중	중	상	상
형태 분석 말뭉치	상	중	상	상	상
일상 대화 음성 말뭉치	상	중	중	상	상
2022년 말뭉치 감정 분석 및 연구	중	중	중	상	중
어휘 의미 분석 말뭉치 2020	중	중	중	상	중
2023년 온라인 매체 자료 수집 및 감정·추론 말뭉치 연구 분석	중	중	중	중	중
감성 분석 말뭉치	중	중	중	중	중
개체명 분석 말뭉치	상	중	중	하	중

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(영어권)

영어권 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

- InstructEval
  - 언어 모델의 능력을 종합적으로 평가하기 위해 설계됨.
  - 정보적 글쓰기, 전문적 글쓰기, 논증적 글쓰기, 창의적 글쓰기 등 다양한 글쓰기 장르에 대한 지시문 포함.
  - 데이터 세트 자체가 첨삭이나 맞춤법 교정 기능을 수행하지는 않지만 언어 모델의 글쓰기 능력을 총체적으로 평가하는 데 목적을 두고 있음.
  - 다양한 글쓰기 상황에 대응할 수 있는 지시를 포함하고 있음.
  - 글쓰기 자동 평가를 위한 데이터 세트로 활용될 수 있을 것으로 기대됨.



## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(영어권)

영어권 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

### • CoEdIT

- 인간 글쓰기를 교정하고, 글쓰기를 돕기 위해 초거대 언어 모델(LLM)을 튜닝한 ‘글쓰기 교정’에 특화된 언어 모델
- ‘문장을 더 간결하게’, ‘더 중립적인 스타일로’ 등 텍스트의 속성을 지정하는 사용자의 지시를 받아 수정된 텍스트를 출력하는 데 특화되어 있음.
- 모델을 튜닝하기 위해 오류문과 교정문을 모아 놓은 기존 데이터 세트를 활용
- 오류를 포함한 문장과 교정을 지시하는 지시문을 결합하여 입력으로, 오류를 개선한 문장 세트를 출력으로 활용하였음.
- 이 모델은 다양한 텍스트 수정 벤치마크에서 가장 우수한 성능을 달성함
- 정량적·정성적 평가 모두에서 다른 모델에 비해 높은 평가를 보였음.
- 문법 오류 교정(GEC), 재구성(Paraphrasing), 단순화(Simplification), 일관성(Coherence), 중립성(Neutralize)와 같은 과제를 수행하는 데에 특화되어 있음.

구분	src	tgt
GEC	Improve the grammaticality: As the number of people grows, the need of habitable environment is unquestionably essential.	As the number of people grows, the need for a habitable environment is unquestionably increasing.
Paraphrasing	Reword this text: She stopped when she saw his expression.	Seeing the look on his face, she paused.
Simplification	Simplify this text: For ourselves, we'll make a fresh one.	We make a new one for ourselves.
Coherence	Fix coherence in this sentence: Guy Leech is still a celebrity today. Guy Leech regularly appears on television as an advocate for health and fitness.	Guy Leech is still a celebrity today and regularly appears on television as an advocate for health and fitness.
Neutralize	Make this text more neutral: chloroform "the molecular lifesaver" an article at oxford university providing interesting facts about chloroform.	chloroform "the molecular lifesaver" an article at oxford university providing facts about chloroform.

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(영어권)

영어권 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

- SUPER-NATURALINSTRUCTION
  - 1,616개의 다양한 NLP 과제를 포함하고 있는데, 이중 문법 오류 탐지(grammar error detection), 문법 오류 수정(grammar error correction), 구두점 오류 탐지(punctuation error detection), 철자 오류 탐지(spelling error detection), 스타일 변환(style transfer) 과제가 있음.
  - 해당 과제의 인스트럭션 데이터를 번역하여 사용하거나, 이를 참고하여 기구축된 한국어 맞춤법 교정 말뭉치를 인스트럭션화하여 사용할 수 있음.

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(영어권)

영어권 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

- LongForm
  - 증강 지침이 포함된 영어 코퍼스 예제를 활용하여 구축함
  - C4 및 Wikipedia와 같은 기존 말뭉치에서 사람이 작성한 다양한 문서 세트를 선택하고 초거대 언어 모델(LLM)을 통해 주어진 문서에 대한 인스트럭션을 생성함
  - Stack Exchange 및 WikiHow와 같은 구조화된 말뭉치 예제와 질문 답변, 전자우편 작성, 문법 오류 수정, 이야기·시 생성 및 텍스트 요약과 같은 작업 예제로 이러한 예제를 확장하는 방식의 데이터 세트임.

### D: The LongForm Dataset

#### Corpus Example

The CAF Second Round of 2006 FIFA World Cup qualification began on 5 June 2004 and finished on 8 October 2005.

The highest-ranked country in each group at the end of the stage qualified for the 2006 FIFA World Cup. This round also doubled as qualification for ...

#### Generated Instruction via LLM

Explain the 2006 FIFA World Cup qualification process.

### M: LongForm Models

#### Given Instruction

Write an essay about the benefits of meditation.

#### Instruction-tuned PLM with Dataset D

#### Generated Output

It doesn't matter where you are in the world, how old you are, or how much you know about meditation, it's for everyone. The benefits of meditation are endless.

Meditation can be as simple as sitting quietly for five minutes with your eyes closed and focusing on your breath. The idea is to train your brain to let go of distracting thoughts and tension so you can think more clearly, feel more relaxed, and be more positive ...

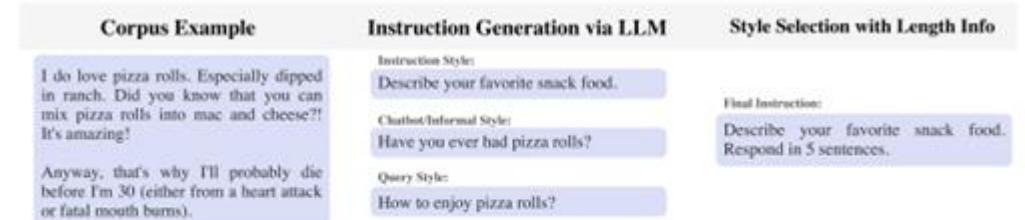


Figure 2: The process of creating LongForm from corpus examples. After collecting diverse examples from corpora, LLMs generate relevant instructions through zero-shot prompting in various styles. The final instruction is generated by selecting one of three styles (instruction (as in this case), informal chatbot, search engine query) and optionally including length information.

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(유럽어권)

- 주류 언어로 된 인스트럭션 말뭉치를 자국 언어로 번역하여 사용하는 경향이 강한 유럽어권 인스트럭션 말뭉치 특성상 한국어 인스트럭션 말뭉치로 변형하여 사용 가능한 사례는 소수임.
- **LongForm dataset**(Köksal et al.,2023)
  - 비교적 길이가 긴 문장을 생성하기 위한 인스트럭션 데이터 세트
  - 총 27,739개의 인스트럭션으로 구성되어 있으며 C4, 위키피디아 말뭉치 등 기존 언어 자원에서 선별한 장문의 텍스트를 포함.
  - 인스트럭션은 모델로 하여금 텍스트를 참조하여 이야기, 전자우편(이메일), 레시피 작성 등 실생활과 밀접한 글을 쓰도록 함.
  - 텍스트 요약, 표 정보 기반 텍스트 생성, 긴 길이의 질의응답(QA) 등의 과제를 유럽 다국어로 수행할 수 있도록 작성되었음.
  - 해당 인스트럭션 말뭉치는 기계 번역 등의 방법을 통해 한국어 작문 특화 인스트럭션 말뭉치로 활용할 수 있을 것으로 기대됨.

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(유럽어권)

- **OpinionGPT**(Haller et al.,2023)

- 데이터 편향을 다룬 인스트럭션 말뭉치로서 편향이 모델에 어떤 영향을 미치는지 살펴보고자 정치적, 지리적, 혹은 인구학적 특성과 관련된 11가지 편견을 포함한 글을 서브 레딧(AskX subreddits)에서 수집하여 말뭉치를 생성함.
- Haller et al.(2023)에서는 해당 말뭉치를 사용하여 OpinionGPT 모델을 구축하기도 함.
- 해당 말뭉치를 한국어로 번역하여 인스트럭션 말뭉치로 활용할 수는 있으나, 편향이 정치, 지리, 인구학적 특성과 관련된 만큼 한국 특성에 맞게 재구축할 필요가 있음.

(<https://opiniongpt.informatik.hu-berlin.de/>)

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(유럽어권)

- **instruct-mlsum-v2**

- 논문으로 발표되지는 않았으나 허깅페이스를 통해 공개되어 있는 오픈 소스로서 해당 말뭉치는 기존의 MLSum dataset 안의 뉴스 기사 약 500건을 토대로 데이터 증강을 통해 인스트럭션 말뭉치를 구축함.
- 해당 말뭉치는 뉴스 기사문 요약, 키워드 생성 등의 과제를 다루는 말뭉치로써 한국어에 적용할 경우 같은 과제를 수행할 수 있는 인스트럭션 말뭉치로 활용 가능함.

(<https://huggingface.co/datasets/snipaid/instruct-snippet-mlsum-v2>)

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(유럽어권)

### 유럽어권 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

- 유럽어권의 텍스트 자동 첨삭과 관련된 인스트럭션 말뭉치들은 공개되거나 널리 쓰이는 사례가 아직까지는 없음.
- 유럽어권 문장 자동 교정 데이터 세트들은 인스트럭션 데이터 세트 이전의 자료들이거나 매우 소규모의 데이터 세트임.
- **‘Dataset-of-Misspellings-and-Corrections(2023)’** 데이터
  - 자그레브 대학교에서 2008년 12월부터 2023년 3월까지 약 16년 가량의 기간 동안 수집된 데이터를 바탕으로 구축함
  - 크로아티아 맞춤법 검사기를 사용한 900,000명에게서 ‘오류 단어 → 올바른 단어’ 형식의 33,382,330개 항목을 수집함
  - 전처리 후 5,584,226개의 고유한 ‘오류 단어 → 정답 단어’ 쌍을 최종 데이터로 사용하였음.
  - 해당 데이터 세트 사용을 통해 맞춤법 검사 정확도를 향상함
  - 언어 학습자의 일반적인 오류 유형에 대한 분류, 철자 오류 추이 및 패턴 분석 가능함
  - <https://github.com/Ispravi-Me/Dataset-of-Misspellings-and-Corrections>

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(유럽어권)

- **DaLAJ-GED**(Volodina et al., 2023)
  - 스웨덴 예테보리 대학교에서 구축
  - 스웨덴어 문장에 대한 문법적 수용성을 다룬 데이터 세트
  - 문법적 수용성뿐만 아니라 문법 오류 탐지, 다중 문법 오류 탐지, 문법 교정 과제를 모두 다룸.
  - 비문에 대한 기본적인 모델의 이해 능력부터 문법과 관련된 다양한 능력을 측정할 수 있는 데이터 세트임.
- **huggingface** ‘lola-gramma-de-en(2023)’
  - 개인이 구축한 사례
  - 영어 사용자의 독일어 문법 오류 양상을 다룸.
  - 초거대 언어 모델(LLM)에 대한 인스트럭션 데이터 세트 형식으로 구성됨.
  - 인스트럭션: 텍스트와 모델에게 문법적 오류 탐지를 지시하는 내용 포함
  - 입력: 문법 오류가 있는 독일어 문장 포함
  - 출력: 오류 수정 문장 및 설명 포함
  - 해당 데이터 세트에서 다루는 문법적 오류에는 단어 누락, 어순, 철자 오류 및 오기, 잘못된 동사 활용 등을 포함함.
  - <https://huggingface.co/datasets/honzatoegel/lola-gramma-de-en>



## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(중국어권)

중국어 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

- InstructUIE

- 인스트럭션 튜닝(Instruction tuning)을 통해 균일한 정보 추출 모델링을 지원하는 프레임워크를 개발하고자 하였음.
- NER, RE, EE 과제 데이터 38개를 활용하여 8(train): 1(validation):1(test)로 나누어 인스트럭션 데이터 세트로 사용하였음.
- 문장 내의 직접적인 교정 작업으로 해당 인스트럭션 구조를 활용할 수는 없어 보이나, 인스트럭션 구조를 변형하여 문장 내부에 존재하는 정보 간의 관계를 추출하여 의미상 적절치 않은 문장을 발견하는 등의 방법으로 활용 가능해 보임.
- 전체 문서나 문장 간의 정보를 효율적으로 추출하기 위한 목적  
: 기구축되어 있는 한국어 데이터 세트를 활용할 수 있는 방안이므로 유의미한 접근이라 할 수 있음.

<NER>

Please list all entity words in the text...

Option: location, person, organization, else

Text: Tom have training in a whole variety of different missions.

Answer:

<RE>

Find the phrases in the following sentence...

Option: used for, part of, compare...

Text: It has also been studied in the framework of Japanese information extraction.

Answer:

<EE>

Extract the event information in the text...

Option: Event type: phishing, databreach...

Arguments type: time, purpose...

Text: Next time I will publish database.

Answer:

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(중국어권)

중국어 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

- GrammarGPT

- 초거대 언어 모델(LLM)의 문법 교정 능력을 파악하고, LLM을 이용한 자동 문법 교정의 실효성을 검증하기 위해 개발되었음.
- 중국인 화자를 대상으로 발생하는 주요 문법적 오류를 정리하고, CLUE(Chinese Language Understanding Evaluation)를 통해 쉽게 수정할 수 있는 것과 그렇지 않은 것을 구분하여 주석하였음.
- 쉽게 교정 가능한 데이터는 챗GPT를 사용하여 비문법적인 문장을 자동 생성하였고, 공개 데이터를 활용하여 비문법적인 문장에 사람이 직접 주석하여 데이터 세트를 구축하였음.
- 해당 데이터 세트 구축 방법을 차용하여 한국어 모어 화자의 주요 문법적 오류를 조사하여 자동 교정 언어 모델을 개발할 수 있을 것으로 보임.

Instruction

{Task Suffix} Human:{Task Description} {Input}

Answer:

## 4-3-7. 인스트럭션 말뭉치 구축 방안\_국외 인스트럭션 말뭉치 활용 방안(중국어권)

중국어 인스트럭션 말뭉치 중 문체 변환이나 맞춤법 교정 등 자동 첨삭과 관련된 데이터 세트

- RobustGEC

- 문법 교정 시스템의 강건성을 평가할 수 있는 벤치마크 데이터 세트
- 중등 영어 학습자가 작성한 학생 에세이 데이터 세트인 CoNLL-14, 학습자와 원어민이 작성한 에세이를 대상으로 에세이의 문장은 원어민 교사가 교정한 BEA-19, 중국에서 영어영문학을 전공하는 고등학생을 대상으로 하는 최고 수준의 시험인 TEM-81의 "언어 사용" 섹션에서 데이터를 수집하여 데이터 세트를 구축함.
- 해당 데이터 세트는 자동 첨삭 시스템 혹은 초거대 언어 모델을 문법 교정에 사용할 때 개발된 모델의 강건성을 평가하는 방법으로 도입할 수 있을 것임.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 인스트럭션 말뭉치 구축 방안 관련 의견\_종합 의견 1

- 인스트럭션 말뭉치는 사람이 직접 구축하는 것이 가장 이상적이나, 시간과 비용의 문제로 인해 언어모델 또는 기구축된 말뭉치를 활용하는 것도 실용적인 대안이 될 수 있음.
- 기계적인 데이터 변환 또는 프롬프트 엔지니어링을 통해 지시문을 생성하고 입출력 데이터를 부분적으로 수정하면 신규 데이터를 처음부터 자동 구축하는 것보다 생성 오류를 최소화하는 장점을 가짐. 이 과정에서 데이터의 구체성과 다양성을 높이기 위한 프롬프트 엔지니어링 연구가 필요함.
- 그외 합성 데이터 기반(Self-Instruct, LLM이 생성) 인스트럭션 구축 방식을 적용해 볼 수 있으며, 사람 기반 데이터 생성 방법론도 중요하지만 합성 데이터에 대한 고려도 필요함.(100% 사람이 구축하는 방식은 지양) 최대한 챗GPT4와 같은 고성능의 초거대 언어 모델(LLM)에 Prompt Engineering을 잘 진행하여 고품질의 데이터를 자동으로 구축하는 것이 필요함.
- 질 좋은 데이터인지 필터링하는 방안도 초거대 언어 모델(LLM)을 통해서 이루지게 하여 데이터 구축 속도도 높이고 질적인 부분을 함께 고려하는 방향이 필요함.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 인스트럭션 말뭉치 구축 방안 관련 의견\_종합 의견 2
  - 인스트럭션 데이터는 영어 데이터를 한국어로 번역하거나, 챗GPT4와 같은 초거대 언어 모델(LLM)을 이용하여 합성 데이터 형태로 제작하고 있고, 해당 방식이 고착화되고 있음.  
: 인스트럭션 데이터는 많은 양의 데이터가 필요한데 이를 사람이 구축하는 것은 너무나도 많은 시간과 비용이 들기 때문임.
  - 초거대 언어 모델(LLM)의 성능이 뛰어나다 보니 합성 데이터로 생성하여도 괜찮은 성능을 보이고, 질 좋은 데이터가 생성됨. 이러한 흐름을 고려하여 ‘어떻게 하면 사람이 좋은 데이터를 만들 수 있을까’하는 고민 이전에 ‘어떻게 하면 좋은 프롬프트를 설계하여 사람이 만든 것보다 훌륭한 초거대 언어 모델(LLM)용 데이터를 만들 수 있을까’를 고민해야 함.
  - 대부분의 자연어 처리 과제는 복합적인 과제이기 때문에, 특정 과제 해결에 한정된 인스트럭션 말뭉치 구축은 효율이 떨어짐.
  - 국립국어원 말뭉치는 고품질의 구조화된 데이터이기 때문에 종합적인 인스트럭션 말뭉치를 구축하면 과제들끼리 서로 큰 도움이 될 것임.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 인스트럭션 말뭉치 구축 방안 관련 의견\_고려할 사항 1

: 좋은 한국어 초거대 언어 모델(LLM)을 위해 어떠한 종류의 데이터가 필요하고 실제 현장에서 어떠한 데이터에 대한 수요가 클지 고민 필요함.

- 초거대 언어 모델(LLM)과 관련한 데이터는 크게 4가지로 나눌 수 있음.
  - (a) LLM 사전학습 및 사후학습용 학습 데이터
  - (b) LLM 미세조정 및 인스트럭션 튜닝(Instruction Tuning)
  - (C) LLM 언어 능력 평가(benchmark)를 위한 데이터
  - (d) LLM 사용자 정렬(Human Alignment)용 데이터
- 연구팀은 (b)와 관련한 데이터에 초점을 두고 있지만 (b)의 경우 현장에서 초거대 언어 모델(LLM)을 통해 합성 데이터(synthetic data) 형태로 대용량 데이터를 제작 하고 있음. (a)와 (c) 데이터로 학습하는 경우 (b)보다 성능 향상을 기대할 수 있음.
- 이러한 부분을 고려해 현장에서는 고품질의 (b) 데이터보다는 (d) 데이터에 대한 수요가 더 큼.
- 국가적 차원에서 직접 저작권과 개인정보 문제가 해결된 대용량의 (a) 데이터를 구축하는 것이 필요함.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 인스트럭션 말뭉치 구축 방안 관련 의견\_고려할 사항 2

: 실제 도움이 되는 데이터인지 실험을 통해 데이터 구축 방향이 정해져야 함.

- 초거대 언어 모델(LLM)에 멀티턴 능력을 부여하기 위해 ShareGPT와 같은 멀티턴 기반 인스트럭션 데이터 세트가 중요해지고 있음. 해당 데이터들은 사람과 사람이 대화한 데이터 세트가 아닌, 사람과 인공지능, 인공지능과 인공지능이 대화한 멀티턴 대화 세트임.

\* 실제 실험을 해 보니 사람-사람이 대화한 데이터 세트보다 사람-인공지능이 대화한 데이터 세트가 초거대 언어 모델(LLM) 성능 향상에 도움이 됨을 확인하였음.

대화 데이터라고 했을 때 사람-사람이 대화한 데이터 세트가 필요한 것이 아닌, 사람-인공지능이 대화한 데이터 세트가 필요하다는 것임.

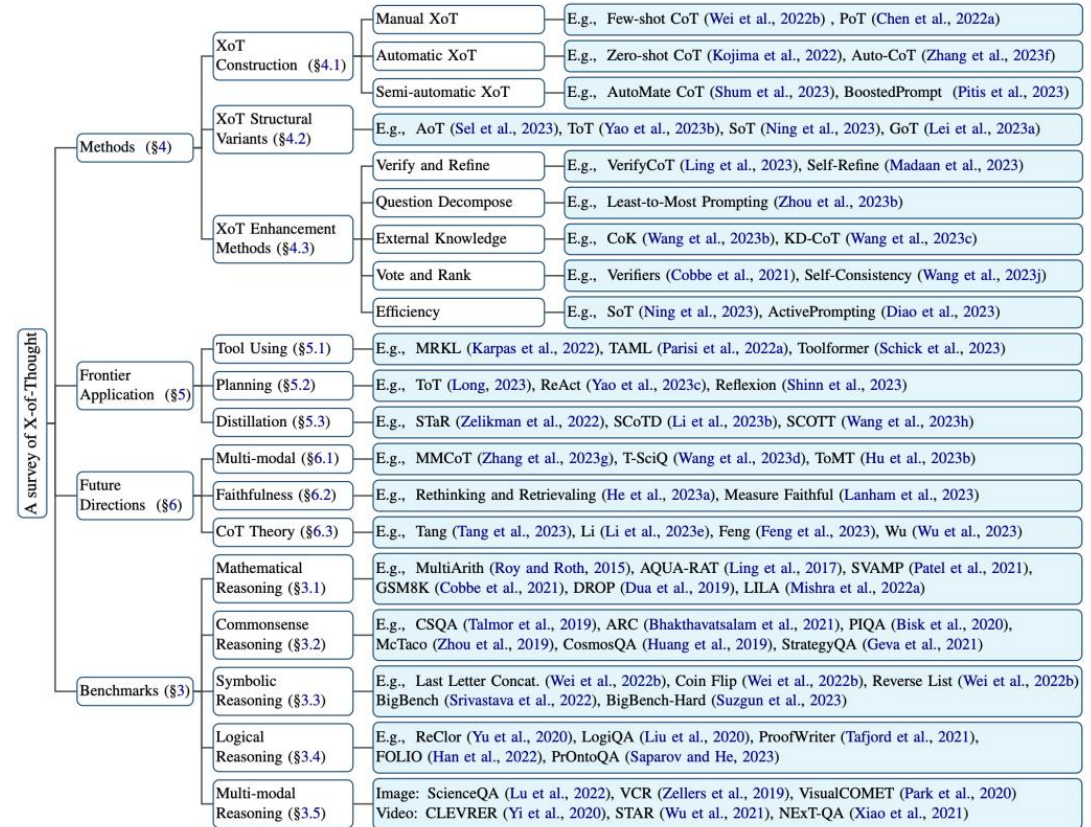
- 그러나 언어모델의 응답이 단조로워서 구체성, 흥미성이 떨어지는 문제점이 있을 수 있음. 생성 결과에 지시문의 구체적인 요구 사항 및 문맥이 반영되지 않고 원론적인 내용만 포함되어 있거나 비슷한 내용이 중복되면 인스트럭션 튜닝 과정에서 지시문을 어떻게 따라야 할지 학습하기 어려워 지시문 수행 성능 저하를 일으킬 수 있기 때문임.
- 인스트럭션 말뭉치를 자동 구축할 때 과제에 최적화된 고품질 데이터 생성 및 검증 프롬프트를 발굴하고 생성 결과에 다양성을 덧붙이기 위하여 다양한 접근과 실험이 필요함. 그러므로 실험을 통해서 발견하고, 데이터 구축 가이드라인 설계와 구축 계획이 이루어져야 함.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 인스트럭션 말뭉치 구축 방안 관련 의견\_고려할 사항 3  
: 다양한 프롬프트 엔지니어링 방법론 고려 필요.
- 언어모델이 생성한 응답에는 적절성, 안전성, 근거성, 일관성 등의 문제가 발생할 수 있기 때문에 데이터 구축 파이프라인을 따라 오류 전파(error propagation)가 연쇄적으로 일어날 수 있음.
- 이러한 부작용을 최소화하기 위하여 높은 품질의 프롬프트와 데이터를 구축하기 위한 다양한 프롬프트 엔지니어링 방법론이 연구되고 있음.

예) 수행할 과제의 예제를 같이 제공하는 퓨샷 프롬프팅(few-shot prompting)

복잡한 문제를 여러 추론 단계로 나누어 처리하는 연쇄 추론(chain of thought; CoT) 프롬프팅, 추론 트리(tree of thoughts), 자기 일관성(self-consistency), 구조적 연쇄 추론(tabular CoT), 합성 프롬프팅(synthetic prompting), 재귀 추론(recursion of thought; RoT) 등 언어모델의 생성 결과의 정확성을 체계적으로 높이거나 추론 능력을 명시적으로 이끌어내기 위한 프롬프트 설계 방법론



A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future  
(<https://arxiv.org/abs/2309.15402>)



## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 기구축 언어 자원 활용 방안 관련 의견

- 국립국어원 말뭉치는 다양한 도메인과 자연어 처리 과제에 걸쳐 잘 구조화되어 있어 대부분 인스트럭션 말뭉치로 활용 가능함. 그러나 모든 데이터는 인스트럭션 말뭉치로 변환은 가능하지만 변환했을 때 모델의 성능 향상에 도움이 되는 것은 다름.
- 국어원이 기구축한 데이터들을 최대한 잘 패키징하고 검수하여, 양질의 데이터 세트를 상업적 활용까지 허용하여 기업에서도 사용할 수 있으면 좋겠음.
- 도메인, 과제, 형식, 내용, 문체 등 다양한 차원에서 인스트럭션 말뭉치가 고르게 분포되어야 언어모델의 지시문 수행 성능이 향상되기 때문에 가능한 모든 말뭉치가 활용될수록 좋음.
- 초거대 언어모델은 파라미터 개수가 많아 암기력이 좋기 때문에 데이터 중복 제거를 위한 전처리 과정을 거쳐야 함.
- 현재 구축 방향성은 과제에 초점이 맞추어져 있음. 이 방향도 좋지만 과제 이전에 도메인에 대한 고려가 먼저 이루어진 후, 특정 과제에 특화된 인스트럭션 말뭉치 구축이 이루어지면 좋을 듯함.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

### • 기구축 언어 자원 활용 방안 관련 의견

- 사전학습 후 소량으로 학습되는 미세조정(fine-tuning) 데이터 특성상 품질이 매우 중요하기 때문에 언어 능력 카테고리화 세부 과제에 따라 활용 가능한 국립국어원 말뭉치를 나열하면 다음과 같음.

구분	내용
(1) 구조적 언어 분석 능력	<ul style="list-style-type: none"> <li>• 과제: 형태소 분리, 문장 분리, 용언 원형 복원, 품사 분석, 명사구 묶음, 구문 분석, 문장 성분 분석, 맞춤법 교정, 문법성 판단, 생략어 복원, 상호 참조 해결 등</li> <li>• 말뭉치: 형태 분석, 구문 분석, 맞춤법 교정, 문법성 판단, 상호 참조 해결, 무형 대용어 복원 말뭉치 등</li> </ul>
(2) 어휘 이해 능력	<ul style="list-style-type: none"> <li>• 과제: 단어 의미, 용례 분류, 개체명 인식, 어휘 관계(동의어, 반의어, 다의어, 동음이의어, 상위어, 하위어 등) 분류, 빈칸 채우기 등</li> <li>• 말뭉치: 개체명 분석, 어휘 의미 분석 말뭉치, NIKLex, 의미역 기술 모형 등</li> </ul>
(3) 문장 이해 능력	<ul style="list-style-type: none"> <li>• 과제: 감성, 감정, 의도, 화행, 스팸, 함의, 유사 문장 분류 등</li> <li>• 말뭉치: 감성 분석, 유사 문장, 언어 능력 평가 말뭉치 등</li> </ul>
(4) 대화 이해 및 생성 능력	<ul style="list-style-type: none"> <li>• 과제: 멀티턴 페르소나 대화 생성, 요약, 채우기, 이어하기 등</li> <li>• 말뭉치: 일상 대화, 온라인 대화, 메신저 말뭉치</li> </ul>
(5) 담화 이해 및 생성 능력	<ul style="list-style-type: none"> <li>• 과제: 기계 독해, 문서 요약, 키워드 추출, 제목 쓰기, 장르별 글쓰기(시, 소설, 기사, 댓글, 블로그, 자기소개서 등), 이어 쓰기, 다시 쓰기, 문체 변환, 작문 평가 등</li> <li>• 말뭉치: 구어, 문어, 신문, 문서 요약, 비출판물, 국회 회의록, 온라인 게시 자료 말뭉치 등</li> </ul>

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 기구축 언어 자원 활용 방안 관련 의견
  - 비윤리적 표현 말뭉치 등은 인스트럭션으로 활용하는 것이 아닌 평가 데이터 세트로 응용되어야 할 듯함.
  - 초거대 언어 모델(LLM)이 정복하지 못한 다양한 능력(추론 능력, 상식정보, 전문 지식, 윤리성, 환각효과)들이 존재하기에 이러한 능력들을 검증하는 평가 데이터 세트가 필요함.
  - 연구적인 관점과 현장의 관점을 분리해서 생각해야 함. 실제 기업들이 초거대 언어 모델(LLM)을 통해 어떠한 문제를 풀고 싶어하는지를 들어 보고 데이터 구축 방향성을 정하는 것이 중요함. 현장에서는 번역, 요약, 속성값 추출(Attribute Value Extraction, AVE) 등에 대한 요구가 강함. 전통적인 과제인 형태소 분석, 상호참조해결, 구문분석 등에 대한 요구는 없음. 추론-확신성, 문법성 판단 등에 대한 필요도 크지 않음. 문체 변환 및 맞춤법 교정은 이미 기초 모델(Foundation Model)이 너무 잘하는 영역이기에 해당 부분에 특화된 인스트럭션 말뭉치가 필요한지는 의문임. 문서 요약, 번역에 대한 필요가 현장에서 매우 크기에 적극 활용되면 좋겠음.
  - 기존 데이터에 단순히 '요약해 봐', '번역해 봐'와 같이 시스템 프롬프트 혹은 사용자(user) 프롬프트 부분에 추가하는 것은 너무 단순한 방법임. 어떻게 하면 질 좋은 \*XOT 방법론을 데이터에 녹여낼 수 있을지 고민하고, 실제 실험을 통해 성능 향상이 되는지 확인이 필요함.

X-of-Thoughts: '생각의 사슬(CoT, Chain-of-Thoughts)을 바탕으로 다양한 추론(reasoning) 방법을 추가적으로 활용하는 프롬프팅 방법

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

### • 기구축 언어 자원 활용 방안 관련 의견

#### (1) 기존 데이터를 그대로 활용하는 경우

- 지시문은 전체 데이터 세트에 걸쳐 동일한 텍스트를 사용할 수도 있겠지만, 주어진 입출력 데이터와 데이터 세트 내부의 메타데이터를 활용하여 부가적인 설명, 예시, 제약 조건 등을 추가하는 방법으로 지시문을 생성하여 구체성과 다양성을 높일 수 있음.

예) 구문 분석 말뭉치 활용: ‘문장에서 주어, 목적어 등의 문장 성분을 찾기’ 지시문을 구성하고 찾아야 하는 문장 성분의 정의, 예시 등 추가

문어 말뭉치 카테고리 정보 활용: ‘주어진 제목으로 특정 장르의 글쓰기’ 또는 반대 방향으로 ‘주어진 글로부터 장르에 맞는 제목 생성하기’ 등의 지시문 데이터를 생성할 수 있음.

#### (2) 기존 데이터를 변환하여 신규 과제를 생성하는 경우

- 기구축 데이터 세트가 기계적인 변환이 쉽지 않은 형태로 구성되어 있지 않은 경우 프롬프트 엔지니어링으로 변환이 가능함.
- 특히 기구축 데이터 세트 중 신문 기사, 도서, 대본, 연설문 등 문단 단위 텍스트는 수집, 구축이 어렵기 때문에 이를 활용하며 자동 구축하면 고품질의 인스트럭션 말뭉치를 얻을 수 있음.
- 기계적인 데이터 처리 과정을 통해 인스트럭션으로 변환하기 어렵다면 프롬프트 엔지니어링을 통해 데이터로부터 역방향으로 프롬프트를 생성하는 방법으로 적절한 프롬프트를 찾는 방법도 있음.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 국외 언어 자원 중 한국어 인스트럭션 말뭉치로 변환 가능한 자원
  - 국외 자원을 번역 등을 활용해서 한국어 인스트럭션 말뭉치로 변환하는 경우 아래와 같은 점을 고려해야 함.
    - 1) 공개된 다국어 인스트럭션 말뭉치 중 출발 언어와 한국어 간 기계번역 품질이 비교적 우수해야 함.
    - 2) 특정 언어, 문화, 역사, 지리 등 지역성에 대한 의존성이 적어야 함.
    - 3) 비문학, 비전문, 도메인 단문 텍스트가 지시문 번역 데이터로써 효용성이 높음.  
예) 데이터브릭스(Databricks)에서 공개한 돌리(Dolly) 데이터 세트가 있음.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 글쓰기 자동 평가 관련 자문 의견
  - 언어모델이 글을 평가하고 탐색하기 위해 아래와 같은 데이터 세트가 구축되어야 함.
    - 1) 다양한 도메인, 장르에 걸쳐 사람이 작성한 작문 데이터
    - 2) 작문 평가를 위한 정량적, 정성적 평가 체계 및 지표
    - 3) 주어진 작문에 대한 정량적 지표별 점수, 피드백 및 탐색 결과
    - 4) 주어진 작문에 대한 정성적 지표별 피드백 및 탐색 결과
  - 범용적인 작문 평가 체계와 이를 바탕으로 다양한 장르의 글을 정량적, 정성적 지표 기준으로 평가한 인스트럭션 말뭉치 구축 필요
  - 다양한 언어 영역에 걸쳐 인공지능의 한국어 이해 및 생성 능력을 평가하는 가이드라인과 이를 반영한 인스트럭션 말뭉치 구축 필요

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- **구축 규모와 기간에 대한 의견**
  - 규모와 기간에 대한 명확한 정답은 없음. 실험을 통해서 데이터 구축 가이드라인 설계와 구축 계획이 이루어져야 함.
  - 적은 양이라도 신뢰도가 높은 양질의 데이터가 구축되면 좋을 거 같음.
  - 특정 과제를 수행하기 위한 인스트럭션 말뭉치 규모를 예측하려면 해당 과제를 수행하기 위해 언어 모델이 학습해야 할 과제를 파악하는 것이 중요함.
  - 글 자동 첨삭은 언어, 작문, 추론 능력 등이 요구되는 매우 복합적인 과제이며, 직간접적으로 관련된 과제의 예시는 아래와 같음. 이 외 글 자동 첨삭 과제를 성공적으로 수행하기 위해서는 다양한 과제 수행 능력을 갖추어야 함.
    - 1) 언어 이해: 어휘, 형태소, 품사, 구문, 맞춤법, 문법성 등
    - 2) 작문 평가: (내용의) 객관성, 정확성, 연관성, 타당성, 풍부성, 창의성, (구조의) 일관성, 응집성, (어휘, 문체의) 적절성, 문법성 등
    - 3) 추론: 상식 추론(상식 퀴즈 등), 인과 관계 추론(원인 또는 결과 추론, 인과 관계 분류), 기계 독해(문제와 지문을 보고 정답 추론 및 근거 추출)
    - 4) 수준별 문체 변환
  - 자동 첨삭의 목적과 범위를 정하여 학습이 필요한 과제들을 분석한 후, 현재 공개된 인스트럭션 말뭉치로부터 관련 과제와 수량을 파악하여 대략적으로 추측해 보는 것이 필요함.

## 4-3-8. 인스트럭션 말뭉치 구축 방안\_자문 의견

- 국어원의 역할에 대한 의견
  - 기초적이며 핵심적인 언어 연구를 수행하는 데 국립국어원에서 구축된 언어 자원이 매우 중요한 역할을 하고 있음.
  - 언어모델을 잘 이해하기 위한 첫 단계는 언어모델의 인식, 추론, 생성 능력에 대한 인지학적인 고찰이며, 특히 언어모델의 언어적 능력에 대한 체계적인 분석이 토대가 되어야 함.
  - 국립국어원에서 다양한 언어 현상에 대한 인지언어학적 연구를 수행하고 연구 결과를 말뭉치를 통해 잘 구체화한다면 이를 바탕으로 구축된 인스트럭션 말뭉치는 언어모델의 기초적인 언어 능력 향상에 큰 도움이 될 것임.
  - 국어원은 기구축된 데이터 세트 중 좋은 데이터 세트를 선별하는 역할을 하면 좋을 것 같음. 좋은 인스트럭션 데이터에 대한 기준을 세우고 이를 자동으로 필터링할 수 있는 모듈을 개발하여 공개해 주면 좋겠음.



## 4. 한국어 말뭉치 구축 중장기 계획안

### 4) 말뭉치 검증 개선 방안

1. 사업 개요
2. 환경 분석
3. 글쓰기 자동평가 및 자동첨삭 방안
4. 한국어 말뭉치 구축 중장기 계획안
  - 4-1. 말뭉치 구축 계획
  - 4-2. 'AI 말평' 평가용 말뭉치
  - 4-3. 인스트럭션 말뭉치 구축 방안
  - 4-4. 말뭉치 검증 개선 방안
    - 4-4-1. 말뭉치 검증 사례
    - 4-4-2. 말뭉치 검증 개선 방안

## 4-4-1. 말뭉치 검증 사례: AI-Hub

- 과기부, 말뭉치 구축 일정별 품질 검증 활동



## 4-4-1. 말뭉치 검증 사례: AI-Hub

### • TTA, 말뭉치 품질검증 절차

#### □ 최종 검증 절차




#### ○ 절차별 수행 업무 및 산출물

진행 절차	수행 업무		문서 산출물	
	TTA	수행기관	TTA	수행기관
품질검증 안내	- 품질검증안내자료 배포	- 품질검증 안내 확인 - 품질검증이해	- 품질검증안내자료	
품질검증 준비	- 품질지표 기준서 검토 및 확정 - 구문적 정확성 검사규칙변경 반영 - 항목별 측정 조건 변경 반영	- 품질지표 기준서 검토 확정 - 구문적 정확성 검사규칙변경 반영 - 항목별 측정 조건 변경 반영 - 데이터셋 제출 - 품질검증 참조자료 <sup>1)</sup> 제출 - 검증 준비현황 점검 - 유효성검증 증빙자료제출	- 품질지표 기준서 v2.0 - 항목별 측정조건 - 구문적 정확성 검사규칙 - 검증 준비현황 점검표	
검증의뢰	- (후)검증 의뢰 접수	- (선)온라인 검증의뢰		- 의뢰서(온라인)
품질검증 수행	- 품질특성별 검사 수행 - 품질특성별 검사 결과 작성/ 검토/공유	- 의미적 정확성 검사기준 확정	- 의미적 정확성 검사기준서 - 품질특성 별 검사결과서	
검증결과 검토		- 품질특성 별 검사 결과 검토		
결과서 발급	- 품질검증 결과서 작성	- 품질검증 결과서 수령	- 품질검증 결과서	

## 4-4-2. 말뭉치 검증 개선 방안

### 현행

- 데이터 구축 사업단 자체 검증
  - ↓
  - 국립국어원 샘플링·전수검사
  - ↓
  - 데이터 구축 사업단 데이터 교정
- 

### 문제점

- 말뭉치 품질 검증에 대한 기준 부족
  - 품질 검증은 어떻게 이루어져야 하는가?
  - 어느 정도의 수준에 부합해야 최종 납품인가?
- 품질 검증에 대한 결과 판단 공식 절차 부족
  - 통과/실패 등 결과 통지 등의 절차 필요
- 구축 자료가 활용도 높은 데이터로 구축되었는지 판단 어려움
  - <모두의 말뭉치> 공개 후 사용량으로만 판단 가능

## 4-4-2. 말뭉치 검증 개선 방안

### 개선

1. 품질 검증 조건 정의

2. 품질 검증 수행

3. 품질 검증 결과서 제출

4. 품질 검증 결과 승인

### 제안요청서/제안서 수준에서 품질 검증 조건 제시 (형식화된 문서 양식 하에서 합의-과제 협상 등)

품질 검증 조건		비고
대분류	중분류	
데이터의 양적 조건	데이터의 전체 수량 (어절 개수 등)	사업팀에서 기계적 으로 검증 가능한 프로그램 제공 필수
	세부 항목 수량 (문장별 최소 어절 개수)	
데이터의 다양성 조건	도메인·출처의 다양성	제안요청서의 기준에 따름
	작업자의 다양성	
데이터의 품질 조건	문법적 정확성 <sup>[1]</sup>	샘플의 양, 목표 정확도 (95% 이상)
	의미적 정확성 <sup>[2]</sup>	

[1] 국립국어원에서 인정하는 맞춤법 검사 도구 활용한 반자동 검증 제안

[2] 국립국어원 내 TF 운영 혹은 별도의 검증 전문 인력·사업단 운영 제안

## 4-4-2. 말뭉치 검증 개선 방안

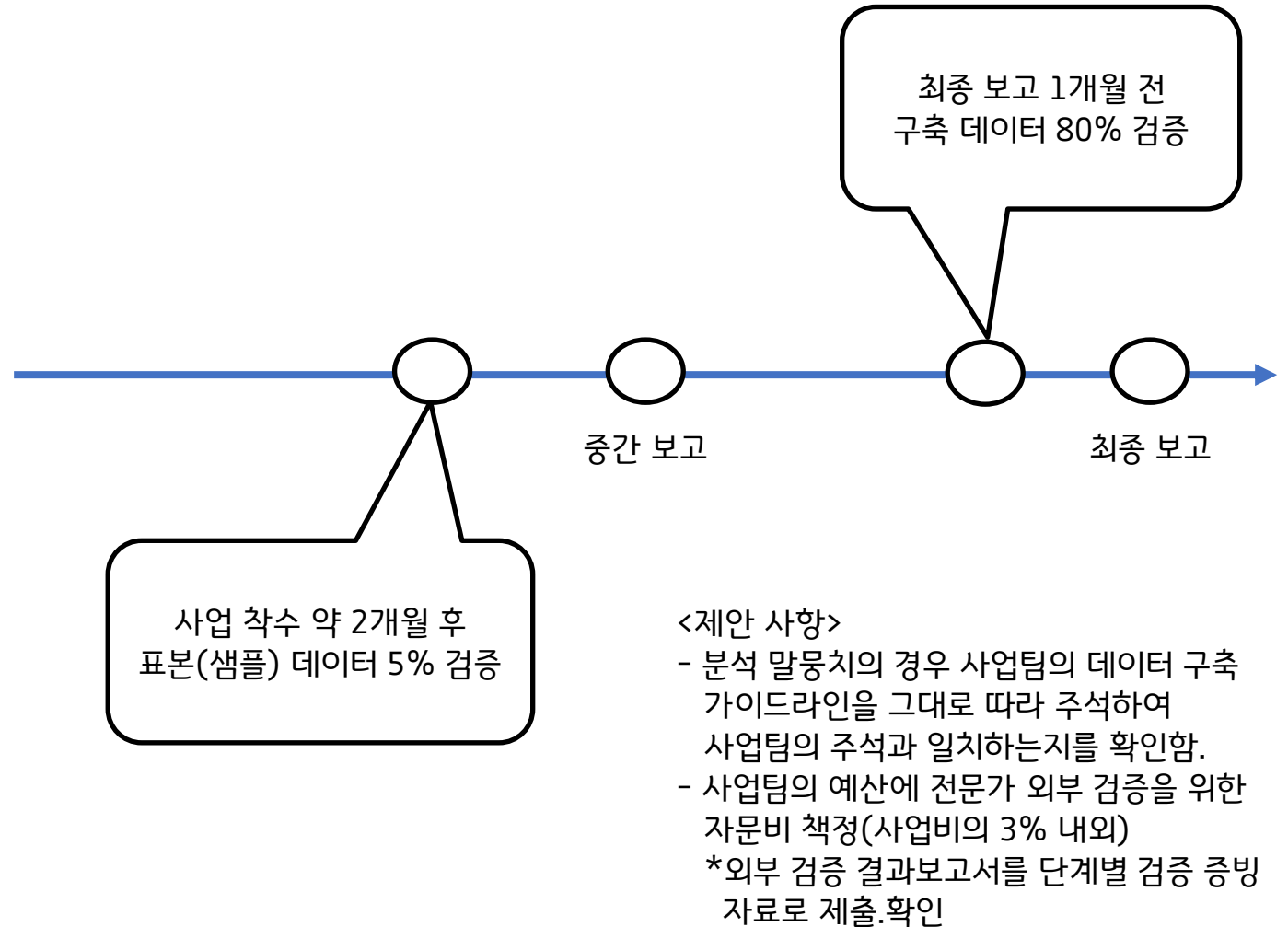
### 개선

1. 품질 검증 요건 정의

2. 품질 검증 수행

3. 품질 검증 결과서 제출

4. 품질 검증 결과 승인



## 4-4-2. 말뭉치 검증 개선 방안

### 개선

1. 품질 검증 요건 정의

2. 품질 검증 수행

3. 품질 검증 결과서 제출

4. 품질 검증 결과 승인

단계	수행기관	비고
품질 검증 의뢰서 제출	사업자	양적 요건, 다양성 요건 → 수행기관 검증 후 포함
		수행기관 자체 검증 결과 보고 * 외부 전문가 검증 보고서 필수 첨부
품질 검증 의뢰서 검토	국립국어원 및 외부 전문가	제출 데이터의 ○○% 선정 * 데이터의 난도와 양을 고려하여 국어원과 협의
품질 검증 수행		지침 수행 여부, 문법적·의미적 오류 검토
품질 검증 결과서		검증 수행 결과
데이터 보완	사업자	이후 보고(중간, 최종) 및 후속 품질 검증 의뢰서에는 해당 사항에 대한 보완 내용 기재

## 4-4-2. 말뭉치 검증 개선 방안

### 개선

1. 품질 검증 요건 정의

2. 품질 검증 수행

3. 품질 검증 결과서 제출

4. 품질 검증 결과 승인

### 최종 보고 1개월 전 품질 검증 결과를 기반으로 한 최종 보고 진행

- 우수, 보통, 미흡, 매우 미흡 4단계 제안  
- 국어원 내부 관리
- 미흡, 매우 미흡의 경우 후속 조치 방안을 포함하여 수정 실시
- 우수, 보통의 경우에도 유지보수 의무 기간 명시



<기획·연구>

국립국어원 강미영 언어정보과장

국립국어원 이보라미 연구관

국립국어원 정영은 연구원

<사업 참여자>

연구책임자: 함영균

공동연구원: 김한샘, 임경태, 임준호, 장한열, 윤영민, 이진, 주민재, 박전규, 전형배

보조연구원: 윤환선, 정용빈, 서현빈, 박재완, 강예지, 강조은, 김유진, 김희재,

박서윤, 최규리, 이재원, 정가연, 이성우

---

---

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154(전화 02-2669-9775, 전송 02-2669-9727)

인쇄일: 2023년 12월 7일

발행일: 2023년 12월 7일

인쇄: 올프린팅

---

---

※ 이 보고서는 국립국어원의 용역비로 수행한 ‘2023년 국어능력 진단체계 활용 방안 연구’ 사업의 결과물을 발간한 것입니다.

이 보고서의 본문 글꼴은 경기도청에서 제공한 경기천년제목 Light체 입니다.