

2024년 한국코퍼스언어학회
여름 전국학술대회

인공지능 시대의
말뭉치 구축과 활용

일시 2024년 7월 17일
장소 성신여자대학교 수정관, 성신관

주최 한국코퍼스언어학회 | 성신여자대학교 인문과학 연구소
후원 고려대학교 4단계 BK21 디러닝 시대의 전문언어인력양
성 사업팀 | 연세대학교 언어정보연구원 | (주) 나라지식정보
| (주) 이르테크 | (주) 테디셈

2024년 한국코퍼스언어학회 여름 전국학술대회

인공지능 시대의 말뭉치 구축과 활용

- 일시_ 2024년 7월 17일 (수) 09:30~18:30
- 장소_ 성신여자대학교 성신관 110호, 수정관 413호

9:30	등록		
9:50	개회사: 장소원(국립국어원장)		사회: 김선혜(연세대)
일반 발표	Session A [성신관 110호]	사회: 윤영민(연세대)	Session B [수정관 413호] 사회: 윤소연(인천대)
10:00~10:30	한국어 감정 형용사 유의어의 말뭉치 기반 변별 정보 연구 • 발표: 봉미경(연세대) • 토론: 최정도(계명대)	Transformers and polysemy of Korean adverbial postposition <i>-(u)lo</i> • 발표: 문성민(아주대) • 토론: 윤수원(서울시립대)	
10:30~11:00	구문 분석 말뭉치를 이용한 주격 중출 구문 연구 • 발표: 신서인(한림대) • 토론: 이준희(동국대)	중국어 의미 주석 말뭉치의 구축과 활용 • 발표: 강병규(서강대) • 토론: 오유정(고려대)	
11:00~11:30	한국어 학습자 오류에 대한 생성형 AI와 모어 화자의 교정 양상 비교 연구 • 발표: 남신혜(경희대) • 토론: 이진(성균관대)	SacreBleu를 이용한 한중 기계번역 평가 • 발표: 황은하(배재대) • 토론: 비립(연세대)	
11:30~12:00	러시아국립말뭉치(Russian National Corpus)의 구조와 기능 • 발표: 김세림, 남혜현(연세대) • 토론: 박카밀라(수원대)	From an interrogative adverb to a discourse marker- way "why" in Korean • 발표: 이해경(아주대) • 토론: 정혜윤(고려대)	
12:00~13:30	점심 시간		
주제 발표1	[성신관 110호]		사회: 박지순(경희대)
13:30~14:10	구어 구문 분석 말뭉치 구축과 구문 분석의 실제	발표: 정연주, 도재학, 이영제, 서셋별(홍익대/경기대/고려대/국립국어원)	
14:10~14:50	일상 대화 말뭉치 구축 현황과 과제	발표: 이용주, 강미영(나라지식정보/국립국어원)	
14:50~15:30	생성형 AI시대와 말뭉치: 글쓰기 첨삭 지원을 위한 인스트럭션 말뭉치	발표: 곽용진, 김소희(이르테크/국립국어원)	
15:30~15:50	휴식		
주제 발표2	[성신관 110호]		사회: 임경태(서울과기대)
15:50~16:30	두 도시 이야기-정치텍스트 분석	발표: 송상헌(고려대)	
16:30~17:10	생성형 AI시대의 한국어 말뭉치 역할	발표: 임수종(ETRI)	
17:10~17:50	Effect of Speed and Breakdown Features on Pronunciation Scoring: Learning from AI-Powered L2 English Speaking Corpus	발표: 윤태진(성신여대)	
17:50~18:20	연구윤리 교육		사회: 김선혜(연세대)
18:20~18:30	폐회사: 김한샘(한국코퍼스언어학회 회장)		사회: 김선혜(연세대)

- 주최: 한국코퍼스언어학회 / 국립국어원 / 성신여자대학교 인문과학연구소
- 후원: 고려대학교 4단계 BK21 딥러닝시대의 전문언어인력양성 사업팀 / 연세대학교 언어정보연구원 (주)나라지식정보 / (주)이르테크 / (주)테디씸

2024년 한국코퍼스언어학회 여름 전국학술대회 인공지능 시대의 말뭉치 구축과 활용 목 차

※ 개회사.....9

§ Session A

- ※ 한국어 감정 형용사 유의어의 말뭉치 기반 변별 정보 연구13
- ※ 구문 분석 말뭉치 기반 변별 정보 연구26
- ※ 한국어 학습자 오류에 대한 생성형 AI와 모어 화자의 교정 양상 비교 연구
.....50
- ※ 러시아국립말뭉치(Russian National Corpus)의 구조와 기능.....63

§ Session B

- ※ Transformers and polysemy of Korean adverbial postposition -
(u)lo.....81
- ※ 중국어 의미 주석 말뭉치의 구축과 활용.....123
- ※ SacreBleu를 이용한 한중 기계번역 평가.....158
- ※ From an interrogative adverb to a discourse maker *-way* “why”
in Korean.....179

§ 주제 발표 1

- ※ 구어 구문 분석 말뭉치 구축과 구문 분석의 실제.....204
- ※ 일상 대화 말뭉치 구축 현황과 과제.....223
- ※ 생성형 AI 시대와 말뭉치: 글쓰기 첨삭 지원을 위한 인스트럭션 말뭉치.....260

§ 주제 발표 2

- ※ 두 도시 이야기-정치텍스트 분석.....290
- ※ 생성형 AI시대의 한국어 말뭉치 역할.....313
- ※ Effect of Speed and Breakdown Features on Pronunciation
Scoring: Learning from AI-Powered L2 English Speaking
Corpus.....365

학술대회 운영

학술대회장: 김한샘(연세대학교)

조직위원장: 김일환(성신여자대학교)

조직위원: 김선혜(연세대학교)

박지순(경희대학교)

윤소연(인천대학교)

윤영민(연세대학교)

임경태(서울과학기술대학교)

학술위원장: 송상헌(고려대학교)

학술위원 : 강병규(서강대학교)

곽용진(이르테크)

남혜현(연세대학교)

문성민(아주대학교)

봉미경(연세대학교)

비립(연세대학교)

신서인(한림대)

이진(성균관대학교)

이혜경(아주대학교)

윤태진(성신여자대학교)

임수종(ETRI)

최정도(계명대학교)

황은하(배재대학교)

후 원

고려대학교 4단계 BK21 딥러닝시대의 전문언어인력양성 사업팀



4단계 BK21 언어학교육연구팀
BK21 FOUR R&E Team for Linguistics

연세대학교 언어정보연구원



연세대학교
YONSEI UNIVERSITY

언어정보연구원

(주)나라지식정보



(주)이르테크



(주)테디섬



개회사

장소원(국립국어원장)

2024년 한국코퍼스언어학회 여름전국학술대회에 참석하신 여러분, 환영합니다.

이번 학술대회 주제는 ‘인공지능 시대의 말뭉치 구축과 활용’입니다. 인공지능 기술이 빠르게 발전하면서 새로운 기회를 맞이하고 있는 코퍼스 언어학 연구자들에게 이 주제는 매우 중요하게 다가오리라 생각합니다.

2022년 11월 챗지피티의 등장을 시작으로 생성형 인공지능이 다양한 방식으로 일상에 들어오면서, 인공지능의 소통 방식이 국어 사용자에게도 영향을 미치게 되었습니다. 이에 국립국어원은 생성형 인공지능 시대를 맞아 영어 중심의 인공지능이 세계 시장을 주도하는 환경 속에서 한국어와 한국문화를 지키고 발전시키기 위한 말뭉치를 연구하고 구축해 나갈 계획입니다. 오늘 학술대회에서는 국어원의 구어 구문 분석, 일상 대화, 글쓰기 첨삭 지원 말뭉치에 관한 발표를 마련하였으니 연구자 여러분들의 귀중한 의견을 부탁드립니다. 이 밖에도 한국의 코퍼스 언어학 분야를 이끌어가시는 분들의 발표가 풍성하게 준비되어 있으니 많은 관심을 부탁드립니다.

이번 학술대회가 학계와 산업계, 정부 기관 간의 소통과 협력을 강화하는 계기가 되기를 희망하며, 그간 국어원이 추진한 많은 정책에 공감하고 참여해 주신 만큼 앞으로 국어원이 나아갈 길에도 많은 기대와 격려를 부탁드립니다.

끝으로, 이번 학술대회를 국립국어원과 공동으로 주최해 주신 한국코퍼스언어학회, 성신여자대학교 인문과학연구소의 노고에 깊이 감사드리며, 고려대학교 4단계 BK21 디퍼닝 시대의 전문언어인력양성 사업팀, 연세대학교 언어정보연구원, 주식회사 나라지식정보, 이르테크, 테디썸 등 여러 기관의 후원에도 깊이 감사드립니다.

감사합니다.

2024년 한국코퍼스언어학회 여름 전국학술대회
인공지능 시대의 말뭉치 구축과 활용



Session A



한국어 감정 형용사 유의어의 말뭉치 기반 변별 정보 연구



구문 분석 말뭉치 기반 변별 정보 연구



한국어 학습자 오류에 대한 생성형 AI와 모어 화자의 교정 양상 비교 연구



러시아국립말뭉치(Russian National Corpus)의 구조와 기능

KACL 한국코퍼스언어학회 2024 여름학술대회

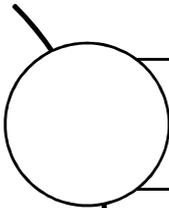
한국어 감정 형용사 유의어의 말뭉치 기반 변별 정보 연구

2024.07.17 성신여대 성신관

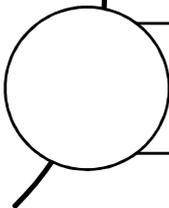
봉미경(연세대 언어정보연구원)

2

연구 목적



감정 형용사 유형 분류와 유의어 선정



말뭉치 기반 변별 정보 분석 - '싫다' 유의어군을 중심으로

감정의 개념

출처	정의
연세한국어사전(1998)	① (사물에 대하여) 마음에 일어나는 느낌. ② 슬픔, 기쁨, 좋음, 싫음 따위의 심리 상태. ③ 외부의 사실에 대하여 자연스럽게 나타내는 반응.
조선말대사전(1992)	주위세계의 사물현상에 대한 매 개인의 태도를 반영하는 심리현상. 주위세계의 사물현상에 대한 주관적인 체험으로서 기쁨, 슬픔, 사랑, 증오, 만족, 불만, 미움 등 다양한 느낌을 통하여 나타낸다.
표준국어대사전(2000)	어떤 현상이나 일에 대하여 일어나는 마음이나 느끼는 기분.
고려대한국어사전(2009)	어떤 일이나 현상, 사물에 대하여 느끼어 나타나는 심정이나 기분.
임석진 외(2009), 철학사전	인간이 주위의 세계, 외계의 여러 현상과 타인의 행동, 또한 자기 자신에 관해 반응하는 심적 상태. 비교적 강하게 단시간 동안 계속되는 감정.
최현석(2011)	외부 자극에 대한 단기적, 인지적 반응.

감정의 개념

출처: *이미지(2012)재인용	감정의 특성
Rohracher (1976)	① 감정은 정신적인 상태이다. ② 감정은 의식적인 의도의 가담 작용 없이 나타난다. 이것은 의식적 의도 없이 원하지 않는데도 나타난다. 따라서 감정은 자발적으로 생성되는 것이다. ③ 감정은 외적 혹은 내적 사상에 대한 정신적 반응이다. 이 정신적 반응의 원인은 대개 의식적이다. 우리는 무엇에 대해 기뻐하며 왜 슬퍼하는지 왜 노여워하는지를 알고 있다. ④ 대개의 감정은 유쾌하거나 불쾌하게 체험한다. 그러나 예외적인 감정도 있는데 후회나 연민의 감정이 불쾌하거나 유쾌하다고 말할 수는 없는 것이다.
김명훈·정영운 (1991)	① 감정은 주관적이고 전체적이다. 개인에 따라 반응 양태가 다르며 신체와 정신 양면에 걸친 전체적 반응 형태로 나타나게 된다. ② 습관적 행동 이외의 모든 지각에 감정이 따른다. 감정은 보통 생활 환경에서의 습관적 행동에는 발생하지 않는다. 이상 상태에 대한 반응에서 나타나는 것이 감정인 것이다. 예컨대, 욕구 불만, 갈등 상태 등 새로운 자극 조건 하에서 발생하는 것이다.
James & Michelle (2005)	① 감정은 관찰되는 것이 아니라 추론된다. ② 통상적으로 감정은 자극에 대한 반응이다. ③ 감정은 인지, 느낌, 행동 경향성을 포함하지만 이 중 하나 또는 두 성분을 포함하는 감정도 일어날 수 있다. ④ 감정은 상황에 대한 실제적 혹은 잠재적 반응을 포함하는데 많은 상황에서 인간이 적응적으로 기능하도록 해준다.

감정의 개념

감정의 의미와 특성 요약

- 외부 자극이나 특정 대상에 대한 정신적 반응
- 인간의 내면에서 일어나는 주관적인 경험
- 비의도적으로 생성
- → 감정이란 외부 자극에 의하여 인간의 내면에서 주관적이고 비의도적으로 발생하는 정신적 반응

감정 형용사의 정의

감정 형용사의 판정

- 감정 유발 자극에 대한 반응: 사람들이 다 쳐다봐서 부끄럽다.
- 마음 내면의 주관적 상태: 몹시 불쾌하고 괴씸한 마음이 들었다.
- 비의도적으로 발생하는 경험: 어둡고 음습한 분위기 때문에 무섭다.
- → 유현경(1998)의 통사적 유형 구분에 따르면, 주관 형용사 중 심리 형용사에 해당됨. 다만, 유의어 연구에서 '감정'이라는 용어가 더 자주 사용되고 '감정 어휘, 감정 명사' 등의 용어와의 연관성을 고려하여 '감정 형용사'라고 부르기로 함.

감정 어휘의 의미 유형 분류

서양의 감정 연구 학자들이 제시한 기본 감정

- Descartes: sadness, joy, desire, hate, love, admiration
- Ekman: happiness, surprise, sadness, fear, disgust, anger
- Oatley: happiness, sadness, fear, disgust, anger
- Fromme: joy, satisfaction, elation, sadness, fear, shock, resignation, anger
- Plutchik: happiness, expectation, surprise, sadness, fear, disgust, anger, acceptance
- Izard: happiness, surprise, sadness, fear, disgust, anger, interest, distress, shame
- Russell: surprise, sleepiness, distress, displeasure, depression, joy, excitement
- Shaver 외(1987): love, joy, anger, sadness, fear, surprise

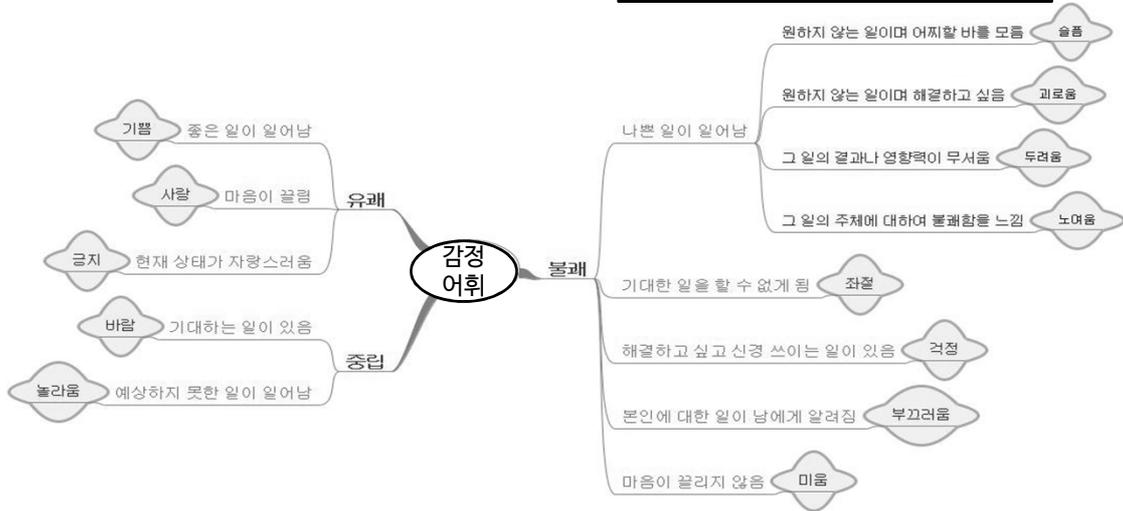
감정 어휘의 의미 유형 분류

서양의 감정 연구 학자들이 제시한 기본 감정

- 이만영, 이홍철(1990): 기쁨, 두려움, 분노, 짜증과 경멸, 슬픔과 괴로움, 각성
- 이준웅 외(2008): 기쁨, 긍지, 사랑, 공포, 분노, 연민, 수치, 좌절, 슬픔
- 손선주 외(2012): 기쁨, 놀람, 슬픔, 공포, 분노, 혐오, 흥미, 지루함, 통증, 중성, 기타
- 장경희, 장세경(1994): 기쁨, 사랑, 슬픔, 두려움, 놀람, 분노, 증오, 부끄러움, 흥미, 괴로움
- 김은영(2004): 기쁨, 좋아함, 동정, 수치심, 노여움, 슬픔, 두려움, 싫어함, 바람
- 홍중선, 정연주(2009): 감동, 감사, 고뇌, 고독, 기쁨, 놀람, 동정, 두려움, 미안함, 분노, 상쾌함, 수용, 수치심, 슬픔, 욕구, 이완, 허무, 혐오, 후회, 복합감정
- 임지룡(2006): 화, 두려움, 슬픔, 부끄러움, 긴장, 미움, 기쁨, 걱정 ('자부심, 경탄, 감동'은 주변 감정으로 제시함)

감정 어휘의 의미 유형 분류

- 기존 연구의 유사한 감정 유형을 통합
- 중복도 높은 감정 유형을 우선 선정
- 뜻풀이 어휘의 연결 관계 고려



감정 의미 유형별 감정 어휘

감정 의미 유형	노여움	부끄러움
감정명사	노여움, 분, 분노, 울분, 화, 성, 울화, 짜증, 신경질	부끄러움, 수치, 수치심, 모욕, 모욕감, 모멸감, 자괴감, 치욕
감정 동사 및 감정 형용사	골나다, 노엽다, 분하다, 성나다, 성질나다, 약오르다, 울분스럽다	굴욕스럽다, 남부끄럽다, 망신스럽다, 머쓱하다, 멧쩍다, 면목없다, 모멸스럽다, 모욕스럽다, 무안스럽다, 무안하다, 민망스럽다, 민망하다, 부끄럽다, 수치스럽다, 쑥스럽다, 창피스럽다, 창피하다, 치욕스럽다
감정 관용구	거품을 물다, 눈에 쌍심지를 쏘우다, 이를 갈다, 뺨이 뒤틀리다, 부아가 솟구치다, 부아가 치밀다, 속에서 천불이 나다, 얼굴을 붉히다, 핏대를 올리다	쥐구멍 속에 들어가 숨고 싶다, 낮이 간지럽다, 주눅이 들다, 풀이 죽다, 가슴이 뜨끔하다

감정 형용사 유의어군 선정

선정 방법

- 유의어 관련 어휘 데이터 활용하여 1차 선정
 - 기존의 유의어 사전, 일반 국어 사전의 유의어 및 관련어 목록, 유의어 연구 논문 데이터, 한국어 어휘망을 통합한 기초 데이터
- 실제 용법 분석 과정에서 추가

감정 유형별 형용사 유의어군

유쾌/불쾌	감정 유형	감정 형용사 유의어
유쾌	금지	자랑스럽다, 대견하다, 기특하다, 강하다
	기쁨	기쁘다, 즐겁다, 좋다, 유쾌하다, 반갑다, 행복하다, 신나다, 흥겹다 / 만족스럽다, 흡족하다, 흐뭇하다, 뿌듯하다 / 활가분하다, 후련하다, 시원하다, 가뽏하다, 가볍다
	사랑	사랑스럽다, 애뜻하다, 반갑다
불쾌	걱정	불쌍하다, 가엾다, 딱하다, 안쓰럽다, 안되다, 애처롭다, 가련하다
	괴로움	귀찮다, 번거롭다, 성가시다 / 불안하다, 초조하다, 조마조마하다, 안절부절못하다 / 지루하다, 따분하다, 지겹다, 심심하다, 무료하다, 재미없다
	노여움	답답하다, 갑갑하다, 거북하다, 불편하다 / 억울하다, 분하다, 원통하다, 노엽다
	두려움	무섭다, 두렵다, 섬뜩하다, 섬찝하다, 공포스럽다
	미움	싫다, 밉다, 알밉다, 못마땅하다, 괘씸하다, 아니꼽다, 눈골사납다, 눈골시다 / 부럽다, 샘나다, 질투하다, 시기하다, 시샘하다
	부끄러움	부끄럽다, 창피하다, 수치스럽다, 민망하다, 무안하다, 쑥스럽다, 수줍다
	슬픔	슬프다, 서글프다, 서럽다, 비통하다, 침통하다, 애절하다, 구슬프다 / 아쉽다, 안타깝다, 섭섭하다, 서운하다, 유감스럽다, 아깝다, 애석하다 / 외롭다, 고독하다, 쓸쓸하다, 적적하다 / 우울하다, 울적하다, 침울하다, 골골하다
	절망	실망스럽다, 좌절스럽다
	중립	놀라움
바람		간절하다, 그림다, 궁금하다

'싫다' 유의어의 의미

유의어	연세한국어사전 뜻풀이
싫다	1. (무엇이) 마음에 들지 않아 만족스럽지 않다. 2. (어떤 행위를) 하고 싶지 않다.
싫다	I (생김새가) 곱지 않다. 볼품이 없다. II 1. (누가) 마음에 들지 않거나 비위에 맞지 않아서 싫다. 2. ['미운'의 꼴로 쓰이어] 마음에 들지 않거나 싫게 느껴지다.
알싫다	다른 사람의 행동이나 말 따위가 싫거나 밉다.
못마땅하다	마음에 맞거나 들지 아니하다.
괘씸하다	(남이 행동이나 말을 제멋대로 함부로 하는 것 같아) 못마땅하고 밉살스럽다.
아니꼽다	말이나 행동이 마음에 몹시 거슬리다.
눈꼴사납다	(하는 것이) 아니꼬워 보기 싫다. *출처: 연세현대한국어사전(미발간)
눈꼴시다	1. (하는 것이) 같잡고 거슬러 보기에 아니꼽다. *출처: 연세현대한국어사전(미발간)



공통 의미: 어떤 사람, 그 사람의 말이나 행동, 어떤 대상이 매우 마음에 들지 않다.

'싫다' 유의어의 빈도와 등급

유의어	절대 빈도	기초어휘 점수(순위)	기초어휘 점수	기초어휘 등급 (국립국어원)	등급의 특성
싫다	41326	623	2.464043	1등급	학령 전
싫다	5391	1872	2.339015	1등급	
알싫다	1409	2608	2.324964	2등급	초등학교
못마땅하다	2385	9229	1.942265	3등급	
괘씸하다	1144	10623	1.938020	3등급	중학교
아니꼽다	609	26957	1.356856	5등급	
눈꼴사납다	22	73781	0.582669	-	
눈꼴시다	20	160493	-0.189610	5등급	

- 분석 말뭉치: 기초어휘 말뭉치(2억 어절, 구어와 문어 1대 1의 균형 말뭉치)
- 기초어휘 점수: <2021년 국어 기초어휘 선정 및 어휘 등급화 연구>(국립국어원) 참조
- 기초어휘 등급: <2023년 국어 기초어휘 선정 및 어휘 등급화 연구>(국립국어원) 참조

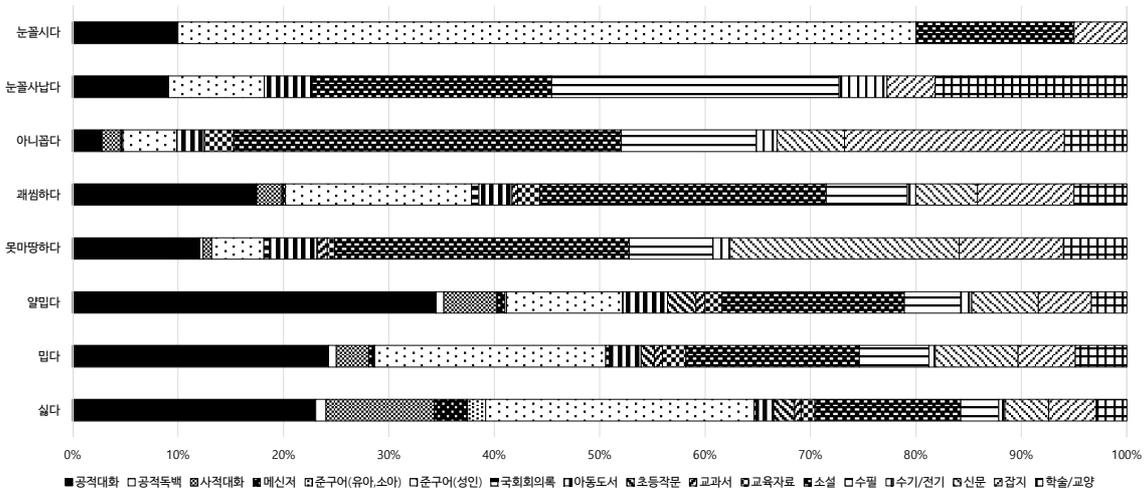
'싫다' 유의어의 장르별 빈도

어형	공적대화	공적독백	사적대화	메신저	준구어 (유아, 소아)	준구어 (성인)	국회회의록	아동도서	초등작문	교과서	교육자료	소설	수필	수기/전기	신문	잡지	학술/교양
싫다	9525	404	4256	1284	725	10520	85	686	820	260	540	5711	1499	237	1713	1870	1191
싫다	1309	39	167	25	5	1182	18	163	67	41	121	886	356	34	423	292	263
알립다	486	10	71	11	2	155	1	59	38	12	23	244	75	15	89	71	47
못마땅하다	288	7	20	0	0	118	12	108	1	22	17	666	190	38	519	236	143
괘씸하다	199	1	27	4	0	202	8	36	0	6	24	311	88	9	67	105	57
아니꼽다	17	0	11	1	0	31	0	16	0	0	17	224	78	12	39	127	36
눈골사납다	2	0	0	0	0	2	0	1	0	0	0	5	6	1	0	1	4
눈골시다	2	0	0	0	0	14	0	0	0	0	0	3	0	0	0	1	0

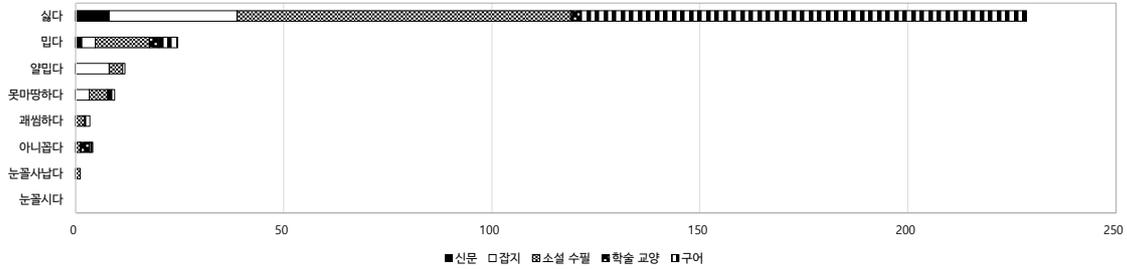


- 분석 말뭉치: 기초어휘 말뭉치(2억 어절, 구어와 문어 1 대 1로 구성된 균형 말뭉치)
- '싫다'가 사용되는 장르, 텍스트 수가 가장 많으며 '눈골시다'로 갈수록 사용 장르, 텍스트 수가 적음.
- 대부분 구어, 소설, 신문에서 자주 쓰임. 학술/교양, 교과서에서 자주 쓰이지 않음.

'싫다' 유의어의 장르별 분포



'싫다' 유의어의 장르별 분포



- 분석 말뭉치: 새연세말뭉치(2백만 어절, 구어와 문어 1 대 1로 구성된 균형 말뭉치)
- 5개 장르로 정규화함.
- '싫다'는 구어, 소설/수필, 잡지에서, '싫다'는 소설/수필, 구어에서, '알싫다'는 잡지에서 자주 쓰임. → 전반적으로 학술 교양에서 자주 쓰이지 않음.

'싫다' 유의어의 공기 관계: 후행어

빈도 순위	후행어	싫은 (2912)			미운 (959)			알미운 (272)			못마땅한 (878)			꽤싫한 (419)			아니꼬운 (291)							
		절대 빈도	상대 빈도	누적 빈도	후행어	절대 빈도	상대 빈도	누적 빈도	후행어	절대 빈도	상대 빈도	누적 빈도	후행어	절대 빈도	상대 빈도	누적 빈도	후행어	절대 빈도	상대 빈도	누적 빈도				
1	것	725	24.9	24.9	것	129	13.5	13.5	생각	33	12.1	12.1	표정	145	16.5	16.5	생각	108	25.8	25.8	생각	39	13.4	13.4
2	소리	284	9.8	34.6	사람	96	10.0	23.5	것	26	9.6	21.7	것	95	10.8	27.3	눈	69	7.9	33.6	것	27	3.1	16.5
3	사람	144	4.9	39.6	생각	85	8.9	32.3	눈	9	3.3	25.0	얼굴	91	10.4	37.7	것	52	5.9	39.6	꼴	25	2.8	19.3
4	일	138	4.7	44.3	오리	58	6.0	38.4	사람	8	2.9	27.9	듯	83	9.5	47.2	일	21	2.4	41.9	눈초리	16	1.8	21.1
5	모양	76	2.6	46.9	눈	48	5.0	43.4	소리	6	2.2	30.1	듯이	60	6.8	54.0	녀석	13	1.5	43.4	소리	14	1.6	22.7
6	말	76	2.6	49.6	정	47	4.9	48.3	표정	6	2.2	32.4	일	31	3.5	57.5	년	12	1.4	44.8	일	12	1.4	24.1
7	기색	47	1.6	51.2	털	31	3.2	51.5	짓	5	1.8	34.2	눈치	30	3.4	60.9	마음	9	1.0	45.8	듯	8	0.9	25.0
8	눈치	47	1.6	52.8	마음	26	2.7	54.2	존재	5	1.8	36.0	모양	29	3.3	64.2	짓	6	0.7	46.5	듯이	5	0.6	25.6
9	생각	42	1.4	54.2	짓	24	2.5	56.7	듯	4	1.5	37.5	눈	24	2.7	67.0	추위	5	0.6	47.1	눈	5	0.6	26.2
10	내색	41	1.4	55.6	감정	21	2.2	58.9	데	4	1.5	39.0	시선	17	1.9	68.9	사람	5	0.6	47.6	표정	4	0.5	26.6
11	눈	39	1.3	57.0	얼굴	20	2.1	61.0	녀석	4	1.5	40.4	점	16	1.8	70.7	듯이	4	0.5	48.1	느낌	4	0.5	27.1
12	얼굴	35	1.2	58.2	소리	17	1.8	62.8	나비	4	1.5	41.9	듯하다	16	1.8	72.6	행동	4	0.5	48.6	노릇	4	0.5	27.5
13	표정	26	0.9	59.1	자식	12	1.3	64.0	자식	3	1.1	43.0	눈초리	14	1.6	74.1	친구	4	0.5	49.0	눈	3	0.3	27.9
14	마음	23	0.8	59.9	얼굴/NR	12	1.3	65.3	새양취	3	1.1	44.1	눈길	13	1.5	75.6	여자	3	0.3	49.4	얼굴	3	0.3	28.2

- 분석 말뭉치: 연세 20세기 한국어 말뭉치(1억 5천만 어절)

'싫다' 유의어의 공기 관계: 후행어

관형사형의 후행어

- 감정이 향하는 대상: 싫은 사람/일/놈/공부/여자/음식, 미운 사람/오리/놈/자식/여자, 알미운 놈/사람/표정/짓/존재/녀석, 못마땅한 일/사람/존재, 괴심한 놈/녀석/년/짓/사람, 아니꼬운 꼴/일/놈/사람, 눈꼴사나운 일/꼴/짓/추파
- 여자는 알미운 표정을 지어 보이고 웃었다.
- 감정이 드러나는 대상: 싫은 말/기색/생각/얼굴/표정, 미운 생각/마음/감정, 알미운 생각/심정, 못마땅한 표정/얼굴/눈치/눈초리/어조, 괴심한 생각/마음, 아니꼬운 생각/눈초리/소리/표정/눈길
- 경아는 내가 외출을 하면 솔직하게 싫은 표정을 하였다.
- '싫다'는 사람 외의 대상에 대한 감정도 표현하나, 그 외 형용사는 주로 사람에 대한 감정을 표현함.
- '못마땅하다, 아니꼽다'는 '눈초리, 눈길, 시선'과 함께 자주 씬.

'싫다' 유의어의 공기 관계: 후행어

빈도 순위	후행어	싫게 (122)			밉게 (221)			알밉게 (76)			못마땅하게 (605)			괴심하게 (133)			아니꼽게 (151)							
		절대빈도	상대빈도	누적빈도	후행어	절대빈도	상대빈도	누적빈도	후행어	절대빈도	상대빈도	누적빈도	후행어	절대빈도	상대빈도	누적빈도	후행어	절대빈도	상대빈도	누적빈도				
1	되다	15	12.3	12.3	보이다	65	29.4	29.4	글다	7	9.2	9.2	여기다	245	40.5	40.5	여기다	58	43.6	43.6	여기다	26	17.2	17.2
2	하다	8	6.6	18.9	보이다	64	29.0	58.4	보이다	7	9.2	18.4	생각하다	220	36.4	76.9	생각하다	35	5.8	49.4	글다	22	3.6	20.9
3	느끼다	7	5.7	24.6	생기다	15	6.8	65.2	생각되다	6	7.9	26.3	생각되다	17	2.8	79.7	생각되다	9	1.5	50.9	생각하다	15	2.5	23.3
4	만들다	4	3.3	27.9	보이다	9	4.1	69.2	생기다	5	6.6	32.9	느끼다	11	1.8	81.5	보이다	3	0.5	51.4	보이다	12	2.0	25.3
5	짱그리다	3	2.5	30.3	글다	9	4.1	73.3	느끼다	5	6.6	39.5	보이다	10	1.7	83.1	노하기/NP	3	0.5	51.9	놀다	6	1.0	26.3
6	생기다	2	1.6	32.0	여기다	8	3.6	76.9	생각하다	4	5.3	44.7	보이다	9	1.5	84.6	바라보이다	3	0.5	52.4	보이다	5	0.8	27.1
7	빼놓여지다	2	1.6	33.6	생각하다	5	2.3	79.2	말하다	2	2.6	47.4	말하다	7	1.2	85.8	생각	3	0.5	52.9	느끼다	4	0.7	27.8
8	소리	2	1.6	35.2	느끼다	4	1.8	81.0	여기다	2	2.6	50.0	바라보이다	6	1.0	86.8	들리다	2	0.3	53.2	노려보이다	4	0.7	28.5
9	싱그럽다	2	1.6	36.9	되다	3	1.4	82.4	보이다	2	2.6	52.6	보이다	5	0.8	87.6	알밉다	2	0.3	53.5	생각되다	3	0.5	29.0
10	변하다	2	1.6	38.5	안/MAG	3	1.4	83.7	얼굴	2	2.6	55.3	노려보이다	4	0.7	88.3	느끼다	2	0.3	53.9	위/NP	3	0.5	29.4
11	일그러지다	2	1.6	40.2	되다	3	1.4	85.1	흘	1	1.3	56.6	쳐다보이다	4	0.7	88.9	치부하다	2	0.3	54.2	네/NP	2	0.3	29.8
12	말하다	2	1.6	41.8	생각되다	2	0.9	86.0	도망	1	1.3	57.9	비치다	4	0.7	89.6	역	1	0.2	54.4	여기/NP	2	0.3	30.1

'싫다' 유의어의 의미 관계

부사형의 후행어

- 주로 동사가 후행함.
- 싫게 생기다/변하다/굴다, 밍게 보다/보이다/생기다/굴다/여기다/생각하다/느끼다, 알밋게 굴다/보이다/생각되다/생기다/느끼다/생각하다/말하다, 못마땅하게 여기다/생각하다/생각되다/느끼다/보다, 꽤심히 여기다/생각하다/생각되다/보다/바라보다, 아니꼽게 여기다/굴다/생각하다/보다, 눈꼴사납게 여기다/보다/벌어지다
- '싫다, 밍다'는 전체 빈도 중 부사형의 빈도가 낮은 편임. 반면, '못마땅하다, 꽤심히, 아니꼽다'는 부사형의 비중의 높음. → 싫게(0.6%), 못마땅하게(20%), 아니꼽게(15%)
- '못마땅하다, 꽤심히, 아니꼽다'의 부사형은 주로 '여기다, 생각하다, 느끼다' 등과 같이 감정의 표출을 나타내는 동사와 함께 쓰임.

'싫다' 유의어의 의미 운용

의미 운용의 분석

- 의미 운용: 결합 관계에서 파악되는 화자의 긍정적, 부정적 태도
- '싫다' 유의어는 어떤 대상이 마음에 들지 않아서 불쾌한 감정이므로 부정적 의미 운용을 가짐.
- 단, '알밋다'는 맥락에 따라 '귀여움'이라는 긍정적 의미 운용을 가지기도 함.

화를 내던 꽃병도 여이가 없다는 듯이 피익_웃어 버렸습니다. "팔자에 없는 요강 노릇을 하다니....." 고놈!	알미운	데도 있지만 귀여운 데도 있던 말야, 허허.....' 삼돌이는_꽃병을 제자리에 갖다 놓고 킁킁 웃었습니다.
경섭이는 깜짝스럽게도 미끄러진 일이 없다고 우기는 을순이가	알밋고도	귀여웠습니다.
그는 징징 우는 소리를 하며 안타까워하는 것이	알밋도록	귀여워 보인다
눈개의 눈웃음은 그림달처럼 가늘고도 처엄(凄艶)했다. "귀엽다마더,	알밋도록	귀엽구나. 그러니까 네 집을 내가 찾아가겠다는 말 아니냐?"
그 모습이 하도 귀여워 진우는 너털웃음을 터뜨렸고 수혜는 더욱	알미운	웃 진우를 깨려보았다. 진우는 한 팔로 수혜의 어깨를 감싸안은 채 공향 청사를 빠져 나왔다.
그리고 경아는 서슴없이 그의 곁에 바짝 붙어 앉더니, 너 참 오래 기다렸지	알미울	정도로 다정하게 그의 볼을 살짝 꼬집어 주는 것이었다.
그리고는 아주		

'싫다' 유의어의 연접 정보

연접 분석

- 연접: 특정한 문법 범주를 선호하는 것
- 전체 용례 중 '않다'의 긴 부정의 비율: 싫다 562개(2.9%), 밉다 393개(9.2%), 알밉다 9개(1%), 못마땅하다 11개(0.4%), 괘씸하다 14개(1.3%), 아니꼽다 6개(0.6%)
 - '싫다, 밉다'는 미움이라는 감정을 부정하는 맥락에서도 사용됨. 반면, '못마땅하다, 아니꼽다'는 미움을 표현하는 맥락에서 사용됨.
- 소수의 부정문 용례에서도 부정문 형식으로 미움의 감정을 더 강조하고 있음.
 - 그녀는 이런 때 남편이 집에 있지 않은 게 여간 못마땅하지가 않았다.
 - 하여튼 창훈이가 그 틈에 끼었다는 것은 한편으로 생각하면 최 참봉보다도 괘씸하지 않은가?
 - 나는 상호의 대답하는 내용이나 태도가 여간 아니꼽지 않았으나 지그시 참았다.

참고문헌

- 국립국어원(연구책임자: 서상규) (2021), 2021년 국어 기초 어휘 선정 및 어휘 등급화 연구
 국립국어원(연구책임자: 서상규) (2022), 2022년 국어 기초 어휘 선정 및 어휘 등급화 연구
 국립국어원(연구책임자: 김한샘) (2023), 2022년 국어 기초 어휘 선정 및 어휘 등급화 연구
 김광해(1987/2000), 비슷한말 반대말 사전, 낱말
 김은영(2004), 국어 감정동사 연구, 전남대학교 박사학위 논문
 봉미경(2016), 한국어 감정명사의 어휘 정보 연구, 연세대학교 박사학위 논문
 손선주·박미숙·박지은·손진훈(2012), 한국어 감정 표현 단어의 추출과 범주화, 감성과학 15-1, 한국감성과학회
 유현경(1998), 국어 형용사 연구, 한국문화사
 이미지(2012), 한국어 교육을 위한 감정표현 어휘 연구, 부경대학교 박사학위 논문
 이만영·이홍철(1990), 형용사 서술 의미의 구조에 관한 연구, 한국심리학회지 2, 한국심리학회
 이준웅·송현주·나은경·김현석(2008), 정서 단어 분류를 통한 정서의 구성 차원 및 위계적 범주에 관한 연구, 한국언론학보 52-1, 한국언론학회
 임지룡(2006), 말하는 몸 - 감정표현의 인지언어학적 탐색, 한국문화사
 장세경·장경희(1994), 국어 관용어에 관한 연구 -정서 표현을 중심으로, 동아시아 문화연구 25, 한양대학교 한국학연구소
 조민정·봉미경·손혜옥·전후민(2013), 학습자를 위한 한국어 유의어 사전, 박이정
 최현석(2011), 인간의 모든 감정, 서해문집「새로 쓰는 비슷한말 꾸러미 사전」, 최종규, 2016, 철수와영희
 홍종선·정연주(2009), 감정 동사의 범주 규정과 유형 분류, 한국어학 45, 한국어학회
 Shaver, P., Schwartz, J., Kirson, D., & O'Connor, G. (1987), Emotion Knowledge: Further Exploration of a Prototype Approach, *Journal of Personality and Social Psychology* 52-6

2024 한국코퍼스언어학회 여름학술대회

구문 분석 말뭉치를 이용한 주격 중출 구문 연구

신서인(한림대)

2024.7.17.(수) 성신여자대학교

연구 목적

- 형태·구문 분석 말뭉치를 이용하여 주격 중출 구문의 실제 사용 양상 조사
- 주격 중출 구문의 분석 방안 모색

주격 중출 구문의 유형

- 주격 중출 구문의 유형
 - "철수가 키가 크다"류
 - "토끼가 피가 많다"류
 - "나는 호랑이가 무섭다"류
 - "국어학은 취직이 어렵다"류
 - "학생이 세 명이 왔다"류 등
- 실제 사용에서 어떤 유형이 많이 쓰이는가?
- 각 유형별로 NP1과 NP2에 어떤 요소가 쓰이는가?
- 기존 주격 중출 구문 논의에서 주목하지 않은 현상이 있는가?

주격 중출 구문 추출 방법

# id: NWRW180000021.234.2.7				
# text: 연 농사가 수익성이 높아 주민들이 쌀 대체작목으로 연꽃을 심은 결과다.				
일련번호	원어절	형태분석어절	구문분석표지	지배소
1	연	연/NNG	NP	2
2	농사가	농사/NNG+가/JKS	NP_SBJ	4
3	수익성이	수익/NNG+성/XSN+이/JKS	NP_SBJ	4
4	높아	높/VA+아/EC	VP	9
5	주민들이	주민/NNG+들/XSN+이/JKS	NP_SBJ	9
6	쌀	쌀NNG	NP	7
7	대체작목으로	대체/NNG+작목/NNG+으로/JKB	NP_AJT	9
8	연꽃을	연꽃/NNG+을/JKO	NP_OBJ	9
9	심은	심/VV+은/ETM	VP_MOD	10
10	결과다.	결과/NNG+이/VCP+다/EF+./SF	VNP	0

- 구문분석표지 열에서 VP 탐색 후 해당 어절의 일련번호 확보
- 그 일련번호를 지배소로 가지는 어절 탐색
- 서술어를 중심으로 의존소 수집
- NP_SBJ NP_SBJ VP 용례 추출

주격 중출 용례

수/NNG+는/JX	2/SN+배/NNG+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000033.392.8.4	# text: 반면 시의원 수는 서울시가 114명으로 뉴욕 시(51명)의 2배가 넘는다.
부가/NNG+가치세/NNG+도/JX	20/SN+% /SW+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000048.347.7.5	# text: 부가가치세도 20%가 넘는다.
넓이/NNG+는/JX	200/SN+평/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000028.122.5.3	# text: 깊이 들어갈수록 쪽방들이 곳곳에 꼬깃꼬깃 들어차 내부 넓이는 200평C
연예인/NNG+만/JX+도/JX	30/SN+명/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000028.135.6.2	# text: 유니폼 외에 아무 대가를 받지 않는데도 시구 순번을 기다리는 연예인만
구역/NNG+은/JX	4/SN+배/NNG+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000052.245.5.3	# text: 태백산국립공원 구역은 △강원 태백시 51.2km² △강원 영월군 0.1km² △강
관객/NNG+은/JX	480/SN+만/NR+명/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000026.84.11.3	# text: 지금까지 '의형제'를 찾은 관객은 480만명이 넘는다.
속도/NNG+는/JX	5/SN+배/NNG+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000052.17.6.2	# text: 매트리스 렌탈·관리가 주축인 코웨이 홈케어 사업의 이런 성장 속도는 자
성공/NNG+를/XSN+은/JX	50/SN+% /SW+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000024.1.3.3	# text: 이런 노력 때문인지, 그의 3점슛 성공률은 50%가 넘는다.
사람/NNG+들/XSN+이/JKS	50/SN+명/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000022.321.3.4	# text: 송화영 살풀이 이수자인 회장 최순희씨를 비롯, 미술치료사이면서 여러
자가/NNG+출판서/NNG+만/JX	50/SN+종/NNG+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000037.117.5.1	# text: 이 서점에서 판매하는 자가출판서만 50종이 넘는다.
사람/NNG+만/JX	500/SN+명/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000052.189.8.4	# text: 현재 120명이 가입 대기 중이고, 상담 후 의향서를 제출한 사람만 500명
회원/NNG+만/JX	5000/SN+명/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000033.219.7.6	# text: 유료 회원만 5000명이 넘는다.
자체/NNG+제작/NNG+를/XSN+은/JX	60/SN+% /SW+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000028.313.3.4	# text: 2009년 기준으로 티비이엔 자체제작률은 60%가 넘는다.
주식/NNG+은/JX	60/SN+% /SW+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000040.31.5.4	# text: 이날 외국인이 판 삼성전자 주식은 전체 외국인 매도 금액의 60%가 넘는
댓글/NNG+이/JKS	600/SN+만/NR+개/NNB+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000046.92.10.3	# text: 이렇게 8년간 자발적으로 선플에 참여한 청소년들이 남긴 댓글이 600만
사람/NNG+은/JX	600/SN+여/XSN+명/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000038.227.10.2	# text: 지난 2008년부터 지금까지 팔짱 프로그램을 거친 사람은 600여명이 넘
간호사/NNG+가/JKS	6200/SN+명/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000038.101.4.1	# text: '간호사=여자'라는 고정관념을 깨고 이제 우리나라엔 남자 간호사가 620
연봉/NNG+이/JKS	8000/SN+만/NR+원/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000048.351.6.2	# text: 세후 515만원이라면 연봉이 8000만원이 넘는다.
총액/NNG+은/JX	87/SN+억/NR+원/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000026.236.8.3	# text: 교수·연구원에게 연구비를 지원하고 대학발전기금도 내놓아 사회에 환원
고용/NNG+를/XSN+은/JX	90/SN+% /SW+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000030.178.4.1	# text: 올 사상 최악의 취업난을 겪는다는 일본의 대졸자 고용률은 90%가 넘는
비율/NNG+은/JX	95/SN+% /SW+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000033.132.11.4	# text: 인터넷 예매사이트 인터파크에 따르면 이들 세 작품의 여성 예매자 비율
수/NNG+/(SS+4330/SN+만/NR+명/NNB+)	배/NNG+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000026.352.11.2	# text: 지난해 말 집계된 종교·인종 차별 등으로 생긴 전 세계 난민 수(4330만명
가격/NNG+은/JX	원/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000033.224.7.1	# text: 김자영이 쓰고 있는 용품 후원사 혼마 클럽의 가격은 7000만 원이 넘는
구매/NNG+가/XSN+는/JX	원/NNB+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000037.199.4.4	# text: 모두 진품일 경우 가방 4개의 총 구매가는 500만 원이 넘는다.
학과/NNG+는/JX	취업/NNG+를/XSN+이/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000022.190.2.3	# text: 2002년 학과를 개설, 3회 졸업생을 배출한 이 학과는 취업률이 80%를 넘
길이/NNG+는/JX	km/SW+가/JKS	넘/VV+는다/EF+./SF	# id: NWRW1800000041.345.8.2	# text: 총 길이는 60만 km가 넘는다.
예산/NNG+만/JX	500/SN+억/NR+원/NNB+이/JKS	넘/VV+는다/EF+./SS+고/JKQ	# id: NWRW1800000036.32.10.4	# text: 김상대 현대차 국내마케팅실장은 "현대차 창립 이후 사전·사후 서비스에
이/NNB+도/JX	3/SN+% /SW+가/JKS	넘/VV+는다/EF+./거나/JX+./SP	# id: NWRW1800000032.405.9.5	# text: 보수층이 결집한 것으로 평가받는 무상급식 주민투표에서 '오세훈 안'에
피해액/NNG+은/JX	4/SN+조/NR+원/NNB+이/JKS	넘/VV+는다/고/EC	# id: NWRW1800000052.21.11.3	# text: 한국금융소비자원은 담합 피해 소비자가 500만명이고 피해액은 4조원

주격 중출 용례

- 주격 중출 구문인 예

id: NWRW1800000021.102.5.1

text: 예를 들어 주가가 50% 하락한다면 초과 하락폭인 5%포인트에 1.818을 곱한 9.09%만큼 원금 손실이 나는 셈이다.

주가가	주가/NNG+가/JKS	NP_SBJ	
50%	50/SN+% /SW	NP_SBJ	
하락한다면	하락/NNG+하/XSV+ㄴ다면/EC	VP	

- 주격 중출 구문이 아닌 예 (오류 포함)

id: NWRW1800000021.103.10.2

text: 서울은 14인승 2대 등 총 3대, 부산이 10인승 2대 등을 보유하고 있다.

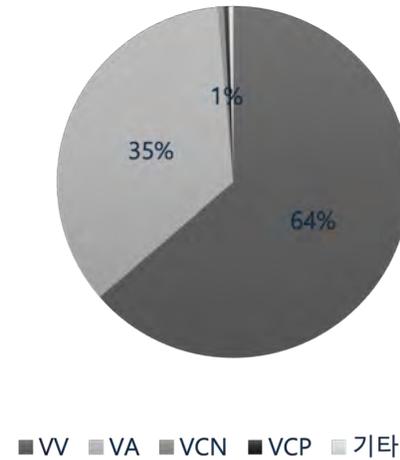
서울은	서울/NNP+은/JX	NP_SBJ	
부산이	부산/NNP+이/JKS	NP_SBJ	
보유하고	보유/NNG+하/XSV+고/EC	VP	

주격 중출 용례의 분포

주격 중출 용례의 서술어 유형

유형	빈도
VV	4,128
VA	2,262
VCN	38
VCP	28
기타	27
계	6,483

주격 중출 용례의 서술어 유형



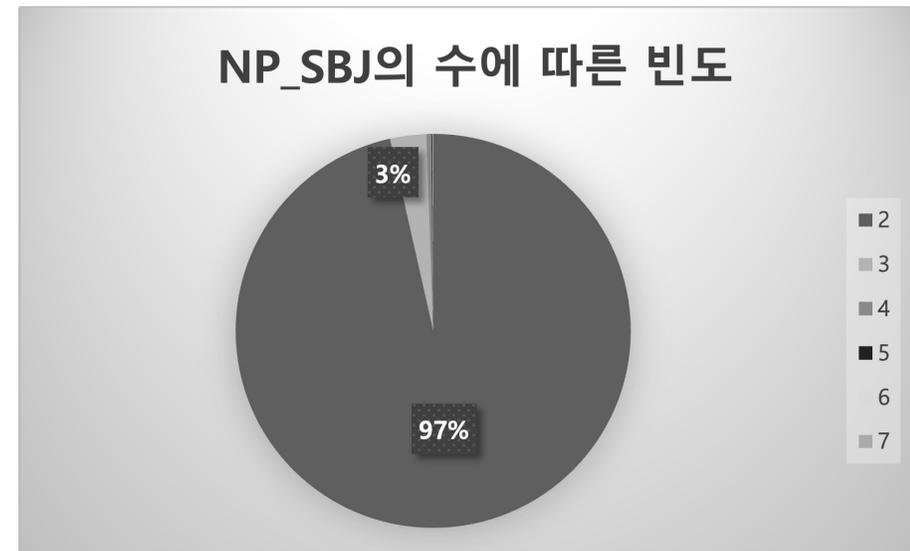
주격 중출 용례 서술어 어간(1,436종)

순위	어간	빈도	누적비율	순위	어간	빈도	누적비율
1	있	488	7.53	26	낮	36	42.79
2	없	473	14.82	27	보이	35	43.33
3	넘	172	17.48	28	강하	34	43.85
4	많	144	19.70	29	받	34	44.38
5	되	139	21.84	30	유일하	34	44.90
6	늘	134	23.91	31	감소하	32	45.40
7	높	130	25.91	32	남	32	45.89
8	크	90	27.30	33	적	32	46.38
9	오르	87	28.64	34	끝나	31	46.86
10	나	81	29.89	35	차지하	31	47.34
11	늘어나	78	31.10	36	걸리	30	47.80
12	좋	72	32.21	37	이르	27	48.22
13	하	70	33.29	38	높아지	26	48.62
14	떨어지	61	34.23	39	부족하	24	48.99
15	지나	60	35.15	40	어렵	24	49.36
16	줄	58	36.05	41	맞	22	49.70
17	가능하	51	36.83	42	이뤄지	21	50.02
18	나오	51	37.62	43	나서	20	50.33
19	줄어들	51	38.41	44	보	20	50.64
20	말	48	39.15	45	내	19	50.93
21	다르	43	39.81	46	바뀌	19	51.23
22	증가하	42	40.46	47	깊	18	51.50
23	들	41	41.09	48	만들	18	51.78
24	아니	37	41.66	49	생기	18	52.06
25	필요하	37	42.23	50	그치	17.00	52.32

주격 중출 구문의 중첩

- 주어가 세 번 이상 나오는 경우가 실제로 얼마나 있을까?
 - 미도파가 양복이 값이 천 원이 싸다.
 - 구두가 바닥이 구멍이 하나가 났다.

NP_SBJ의 수	빈도
2	6,275
3	193
4	27
5	5
6	3
7	1



-> NP_SBJ 4개 이상은 모두 진정한 주격 중출 구문이 아닌 예로 확인됨.

주격 중출 구문의 유형별 특징

수량 표현 구문(1)

- NP1-이 NP2(수량)-이 VV

- '넘다; 늘다, 늘어나다, 줄다, 줄어들다, 오르다, 떨어지다' 등

(1) 가. 이날 외국인이 판 삼성전자 주식은 전체 외국인 매도 금액의 60%가 넘는다.

나. 고리사채 등 불법 사금융 사범은 64% 늘었다.

다. 그 대신 읽기는 28문항으로 5문항이 늘었다.

라. 저장용량은 약 35% 늘어난다.

마. 생산인구의 '허리'로 볼 수 있는 40대 일자리도 6만개가 줄었다.

바. 실제로 우리가 사용해보니 사용 전력량이 30~40% 줄어드는 게 확인됐다"고 주장했다.

사. 인건비도 5년 새 10배 올랐다.

아. 부동산 정보업체 '부동산114'에 따르면 서울 집값은 평균 1.34% 떨어졌고 수도권도 1.8% 넘게 하락했다.

- NP2에 수량사구가 오는 경우 중 위의 예들은 "학생이 세 명이 왔다."류와는 달리 동사의 의미와 긴밀한 관계에 있음.
- 특히, '넘다'는 NP2의 수량사구 뒤에 '이/가'가 결합하는 경향이 강함. '넘게'일 때는 '이/가'가 잘 결합하지 않음.
- 사전에서도 격률 정보에 주격 중출에 대한 정보가 기술되어 있는 경우가 거의 없는데, 추가할 필요가 있음.

수량 표현 구문(2)

- NP1-이 NP2(수량)-이 VA
 - '있다, 많다, 높다, 크다' 등

(2) 가. 치매 전용 데이케어센터는 서울 시내에 10곳이 있다.

나. 예금이 소득 증가분보다 1억9000만원이 더 **많은** 것이다.

다. 교통사고 발생률은 일반 횡단보도보다 5.4배 **높았다.**

라. 새로 배치되는 양륙함은 기존 함정보다 1.5배 **크고**, 수직이착륙 수송기 탑재가 가능한 신형이다.

- 형용사 중 척도의 의미를 가지는 것은 주격 중출 구문으로 실현되는 경우가 많음.
- 특히 비교 구문에서는 'NP1(대상)-이 NP2(비교기준)-보다 NP3(수량)-이 VA'의 문형으로 실현되는 경우가 많음.
- 주격 중출 구문의 유형 중 두 번째 명사구에 수량 표현이 오는 경우는 기타 유형으로 다루어진 경우가 많았으나 실제 사용 양상을 살펴보면 상당한 비중을 차지하고 있음.
 - NP2에 숫자(SN)가 오는 경우: 919건
 - NP2에 숫자(SN)가 오는 경우 중 주격조사(JKS)가 결합한 경우 : 417건

'있다' 구문

- 'NP1-이 NP2-이 있다'는 주격 중출 용례 중 가장 많은 비중을 차지함.
 - 'NP1(처소)-이 NP2(대상)-이 있다': 'NP1-에 NP2-이 있다'와 바꾸어 쓸 수 있음.
 - 'NP1(대상)-이 NP2(수량)-이 있다'
- (3) 가. 그러나 이런 기술은 한계도 있다.
 - 나. 양쪽 모두 고민은 있다.
 - 다. 그러나 나도 감이 있다.
 - 라. 해먼드 소장은 새로운 바이러스가 발견될 때마다 병원성을 규명해 기존에 개발된 백신이 효과가 있는지 지속적으로 추적해야 한다고 강조했다.
- (3') 가. 그러나 이런 기술에는 한계도 있다.
 - 나. ?양쪽 모두에게 고민은 있다.
 - 다. ?그러나 나에게도 감이 있다.
 - 라. *해먼드 소장은 새로운 바이러스가 발견될 때마다 병원성을 규명해 기존에 개발된 백신에 효과가 있는지 지속적으로 추적해야 한다고 강조했다.
- (3)의 예들 중 'NP1-이 NP2-이 있다'를 'NP1-에/에게 NP2-이 있다'로 바꾸었을 때 가능한 것도 있지만, 어색하거나 불가능한 것도 있음.
- 'NP-에'는 'NP-이'에 비해 문법적 지위가 낮는데, 'NP1-에' 대신 'NP1-이'를 사용함으로써 문법적 지위가 'NP2-이'보다 높아지게 하는 효과가 있음.

'되다' 구문

- 'NP1-이 NP2-이 되다'는 'NP1-이 NP2-이다'와 유사한 의미
 - NP1(대상)-이 NP2(결과상태)-이 되다: 상태 변화
 - NP1(대상)-이 NP2(수량)-이 되다: 확인

(4) 가. 이런 차량들의 행렬이 1km는 **족히 돼** 보였다.

나. 올해 일자리 대책을 위한 추경예산이 4조9000억원이 넘지만, 이 가운데 '고용촉진' 관련 예산 비중은 1%도 안 된다.

다. 실제로 해외취업을 원하는 젊은이들이 줄 서 있지만 정부 지원 해외 인턴십 참가자들의 실제 취업률은 10%가 채 안 된다.

- 안명철(2015: 51)에서는 '그 사람이 사장이 아니다.(정체성)', '저 사람이 도둑이 분명하다.(정체성)', '그 사람이 바보가 맞다.(속성)', '이 옷이 10만원이 나간다.(가치속성)'에 대해 대상의 정체성 또는 본질적 속성에 대한 화자의 판단을 나타내는 구문이라고 하였음.
- '되다'가 정태적 의미를 가지는 위의 'NP1-이 NP2-이 되다'의 예들도 이들과 유사함.

'가능하다' 구문

- 'NP1-이 NP2(서술성명사)-이 가능하다'도 비교적 높은 빈도로 사용됨.
 - 주격 중출 구문에 대한 많은 논의에서 NP1과 NP2의 관계에 주목함. NP1과 NP2는 전체-부분, 부류-성원, 유형-개체, 대상-속성 등의 관계를 보이고, 이때의 NP1, NP2는 지시성을 가짐.
 - 한편, NP2에 서술성명사가 오는 경우가 있음. 이때 NP1과 NP2는 논항-술어 관계
 - '가능하다, 불가능하다, 쉽다, 불가피하다, 필요하다' 등
- (5) 가. 바닥도 폴리싱 타일과 온돌마루 중 선택이 가능하다.
 - 나. 청구와 집행에 걸리는 시간을 감안하면 압수수색영장은 사실상 사용이 불가능하다는 게 수사기관들 말이다.
 - 다. 그는 현재 보험 가입이 불가능하다.
 - 라. 어린이집이 자가 소유인 경우는 전환이 쉽지만 건물을 임차하고 있는 경우 절차가 까다롭고 복잡해서 대부분 떨어진다"고 말했다.
 - 마. A 씨도 검찰 소환이 불가피할 것으로 보인다.
 - 바. 노동계도 인식 전환이 필요하다는 목소리가 높다.
- 위의 예들에서 서술어는 NP2의 서술성 명사가 나타내는 사태의 실현 가능성을 나타내는 양태적 의미를 나타냄.

주격 중출 구문의 중첩

- 삼중주어문은 약 3%의 비중을 차지함.

(6) 가. '해커스 토익' 시리즈는 2005년 이래 총 22종 누적 판매량이 1100만부 넘는다.

나. 한나라당 강석호 의원도 주식 평가액이 117억 원에서 62억 원으로 반 토막이 났다.

다. 두 나라는 스킨케어 제품(60%)이 색조 제품(40%)보다 사용 비율이 높지요.

라. 충북과 대전의 대학병원 2곳은 지원자가 한 명도 없었다.

마. 지역별 전세금은 인천이 지난해 평균 1억1671만 원에서 올해 1억3581억 원으로 16.4%(1910만 원) 올라 전국에서 상승률이 가장 높았다.

바. 서울대는 학교장추천 전형인 지역균형선발 전형만 수능 최저학력 기준이 있지만 연세대와 고려대는 실기(특기) 전형을 제외하고 모두 최저학력 기준이 있다.

- 192개 용례 중 48개(25%)가 세 번째 주어의 수량 표현임.

동사의 목적어가 주어로 분석된 경우

- 주제어로 실현된 목적어

(7) 가. 이 제도는 진안군이 전국에서 처음으로 도입했다.

나. 국정기획위 관계자는 “고교 입시 혼란 등 부작용을 줄이기 위해 일괄적인 폐지보다는 단계적 폐지로 가닥을 잡았다”며 “구체적인 시행 방안은 교육부가 마련할 것”이라고 말했다.

다. (DO YOU HEAR?)라는 제목의 이 광고는 서 교수가 기획하고 김 씨가 광고비 전액을 지원했다.

라. 또 다른 선공개곡 ‘종이피아노’는 박용준이 작곡하고 싱어송라이터 조동희가 작사한 곡이다.

마. 이 기체는 방위성 기술연구본부가 개발을 맡고, 기체 제작은 2009년부터 미쓰비시중공업이 담당하고 있다.

바. 따지고 보면 인터넷도 정부가 개발한 것이고, 인간 게놈 지도 연구도 정부가 시작한 것이죠.

- 위의 예들은 서술어의 목적어인데 문장 앞으로 나오면서 주제화된 것으로서 구문 분석 말뭉치에서 주어로 분석된 것임. 이는 주제어를 주어로 인식하는 작업자들의 직관이 반영된 것이라고 할 수 있음.
- (7다)는 두 절의 공통 성분인 목적어가 문장 앞으로 나와 주제화된 경우
- (7라)는 두 내포절의 목적이면서 동시에 주절의 주어인데 주제화된 경우
- (7마)는 두 절에서 대조되는 목적어 성분이 각 절 앞으로 나와 주제화된 경우
- (7가-바)는 모두 주어가 초점화된 문장으로 볼 수도 있음.
- 구문 분석 말뭉치 구축 시, 이러한 예들에서 주제화된 성분을 NP_SBJ로 분석할 것인지 NP_OBJ로 분석할 것인지 결정해야 함.

결론

- 구문 분석 말뭉치를 활용하면 주격 중출 구문의 전체적인 사용 양상을 확인할 수 있음.
- 서술어나 NP1 또는 NP2의 성격에 따라 용례들을 모아 주격 중출 구문의 유형에 따른 특징을 추출할 수 있음.
- 주격 중출 구문에 대한 연구의 결과를 구문 분석에 다시 적용해야 함.

감사합니다.

dryice@hallym.ac.kr

이준희(동국대)

이 논의는 구문 분석 말뭉치를 토대로 실제 주격 중출 구문이 어떠한 양상을 보이는지를 면밀히 살펴보고 있다. 이처럼 말뭉치를 통하여 실제 사용 양상을 살펴보고 있기에, 보다 실제 언어생활에 부합하는 양상 분석 결과를 제시하고 있다. 대표적으로, 수량 표현 구문의 경우 기존 논의에서는 주격 중출 구문을 다룰 때 기타 유형으로 분류하는데, 본 논의에서 언급하고 있는 바와 같이, 실제 양상을 살펴보면 수량 표현 구문이 많은 비중을 차지하고 있다. 논의의 내용에 전반적으로 동의를 하는 바이기에, 논의에 대해 반론을 펼치기 보다는 궁금한 점을 질문하는 것으로 토론을 갈음하고자 한다.

1. 본 논의에서는 3-5쪽에 제시된 방법을 토대로 구문 분석 말뭉치에서 주격 중출 구문을 추출하여 분석 대상으로 삼고 있다. 여기에서는 이들 중 주격 중출 구문으로 추출한 이유가 궁금한 양상들에 대해서 질문하고자 한다. 사실, 주격 중출 구문을 추출한 방법을 보면, 해당 말뭉치에서 NP_SBJ로 분석한 것을 기반으로 하기에 발표자 선생님께 질문할 사항이 아닐 수 있다. 다만, 5쪽을 토대로 보면 발표자 선생님께서도 주격 중출 구문으로의 해당 여부를 판단하고 선별하신 것으로 판단되어, 이들을 주격 중출 구문으로 포함하신 이유에 대해 여쭙고자 한다.¹⁾

1-1. 13쪽에 'NP1이 NP2이 있다' 유형에서 NP1에 보조사가 결합하였을 때 이들을 주격조사가 결합한 것으로 분석하신 이유가 궁금하다. 본 논의에서도 언급하고 있듯이, 해당 유형 중 'NP1(처소)이 NP2(대상)이 있다'의 경우 'NP1에 NP2이 있다'로 교체가 가능하다. 그렇기에 NP1에 주격조사가 실현되어 있는 경우에는 형태를 중시하여 주격 중출 구문으로 해석하는 것이 자연스럽지만, 보조사가 실현되어 있는 경우에는 이들을 '에'가 실현된 것으로 볼 수도 있다. 이는 특히, 예문 (3'가)와 같이 '에' 실현이 자연스러운 경우에 더욱 그러하다.

이는 해당 유형뿐만 아니라 다른 유형에서도 발견된다. 예를 들어, 15쪽 예문인 (5라)의 경우에 '어린이집이 자가 소유인 경우엔 전환이 쉽지만...' 과 같이 '경우엔'로 나타나는 것이 가능하며, 같은 쪽 (5바)도 '노동계에 인식 전환이 필요하다는 목소리가 높다.'처럼 '노동계에'로 실현되는 것이 가능하다.

본 논의에서는 이처럼 보조사가 실현된 경우에도 주격조사가 실현된 것으로 보고 주격 중출 구문으로 분석하고 있는데, 그 이유와 근거에 대해 발표자 선생님께 여쭙고자 한다.

1-2. 본 논의에서 언급하고 있듯이, 비중이 높은 주격 중출 구문은 수량 표현 구문이다. 이들의 경우 11쪽의 예문 (1)과 같이 주격조사가 실현되었거나, 주격조사 외의 다른 조사가 실현되기 어려운 경우에는 무리 없이 주격 중출 구문으로 판단할 수 있다. 다만, 본 논의에서는 주격조사가 실현되어 있지 않아 다른 격조사가 실현된 것으로 볼 수 있는 경우도 주격 중출

1) 물론, 5쪽에서 제시해주신 포함되지 않는 예시의 경우, 주격 중출 구문으로 절대 볼 수 없는 경우가 추출 방법에 의해 포함된 것이기에 제외하신 것이기는 하다.

구문으로 분석하고 있다. 예컨대, 16쪽의 예문 (6가)의 경우 조사가 실현되어 있지 않은데, ‘해커스 토익’ 시리즈는 2005년 이래 총 22종 누적 판매량이 1100만부를 넘는다.’와 같이 대격조사가 실현될 수 있다. 더불어, 11쪽의 (1가)의 경우에 주격조사가 실현되기는 하였지만, ‘이날 외국인이 판 삼성전자 주식은 전체 외국인 매도 금액의 60%를 넘는다.’와 같이 대격조사로 실현이 가능한 것으로 보아, ‘넘다’라는 서술어가 나타난 수량 표현 구문의 경우 수량사구에 조사가 실현되지 않았을 때 주격조사뿐만 아니라 대격조사도 올 수 있을 것으로 추측할 수 있다.

본 논의에서는 이처럼 이들을 주격 중출 구문으로 분석하고 있는데, 그 이유와 근거에 대해서 발표자 선생님께 여쭙고자 한다.

2. 17쪽에서 제시해주신 유형의 경우 본 논의에서 언급하고 있듯이 주어로도 볼 수 있지만 목적어로도 분석할 수 있는 것이고, 그렇기에 이들을 NP_SBJ로 분석할지 NP_OBJ로 분석할지를 정해야 한다. 그렇다면 이들이 각각 NP_SBJ로 분석하였을 때와 NP_OBJ로 분석하였을 때의 장점과 단점이 무엇인지 궁금하다. 얼핏 생각해보면, 《표준》과 《우리말샘》을 기준으로 하였을 때 해당 문장들의 서술어가 목적어를 요구하기에, NP_OBJ로 분석하는 것이 구문 분석의 관점에서만 보면 더 유용할 수도 있을 것으로 보이지만, 이는 너무 단순한 판단일 것 같다. 그렇기에 발표자 선생님께 이들 각각 분석의 장단점에 대해 여쭙고자 한다.

3. 주격 중출 구문의 유형으로 익숙한 것은 2쪽에서 제시해주신, ‘철수가 키가 크다’류, ‘토끼가 꾀가 많다’류, ‘나는 호랑이가 무섭다’류, ‘국어학은 취직이 어렵다’류, ‘학생이 세 명이 왔다’류이다. 그리고 본 논의에 따르면 실제 언어생활에서는 이 중에서 기타 유형으로 분류하는 경우가 잦았던 수량 표현 구문인 ‘학생이 세 명이 왔다’류가 비중이 꽤나 큰 것을 확인할 수 있다. 그렇다면 다른 유형들의 경우 실제 말뭉치에서는 어느 정도의 비중을 지니고 있는지, 그리고 이들 수량 표현 구문이 비중이 생각보다 높은 이유가 무엇일지 궁금하다.

한국어 학습자 오류에 대한 LLM과 모어 화자의 교정 양상 비교 연구

남신혜(경희대)

차 례

1. 들어가는 말
 2. 거대언어모델과 한국어
 3. 연구 방법
 4. 분석 결과
 5. 맺음말
-

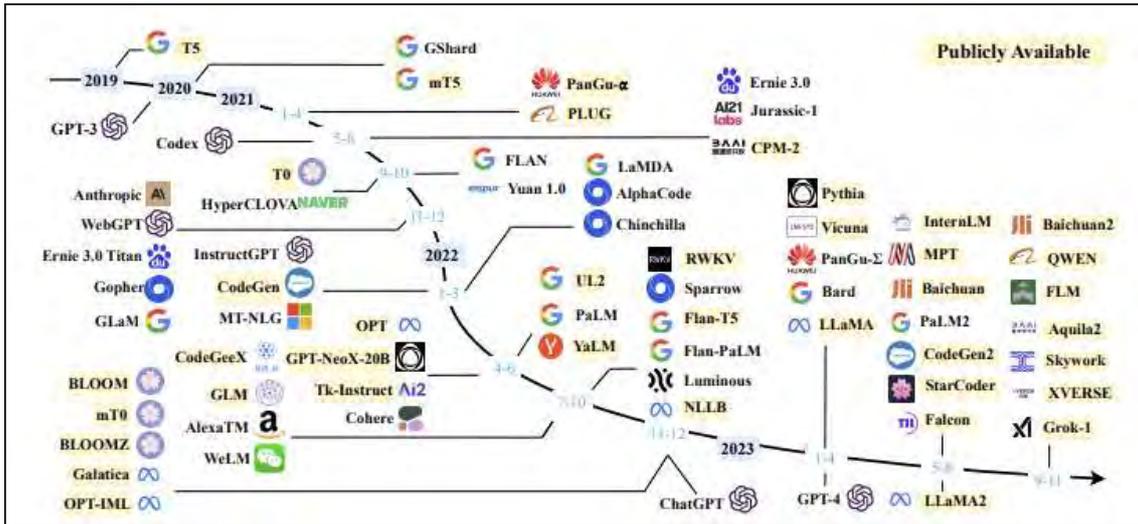
1. 들어가는 말

본 연구는 외국인 한국어 학습자들의 작문 자료를 인간 - 한국인 모어 화자 - 과 거대 언어모델(LLM: Large Language Models)이 어떻게 교정하였는지를 비교·검토해 보고 이를 토대로 현재 시점에서 LLM의 한국어 자동 교정 도구로서의 활용 가능성을 진단해 보는 데 목적이 있다. 이를 위해서 국립국어원에서 구축하여 배포하고 있는 <한국어학습자말뭉치>를 분석의 대상으로 삼았다. 이 말뭉치에는 한국어 학습자들의 학습자 언어(Learner Language)가 담겨 있으며 거기에 나타난 오류들(errors)을 한국인 모어 화자들, 그중에서도 국어학이나 한국어교육학을 전공하고 있는 전문가들이 교정 내용을 표시해 놓은 주석이 부착되어 있다. 본 연구에서는 이 점에 착안하여 <한국어학습자말뭉치>에서 학습자언어와 그에 대한 한국어 모어 화자들의 교정 자료를 수집하고, 동일한 학습자언어를 LLM에게 교정하도록 하여 그 결과를 한국어 모어 화자들의 교정 내용과 비교하였다.

2. 거대언어모델과 한국어

거대언어모델(LLM)은 인공지능 기반의 언어 모델로 대부분 트랜스포머 아키텍처에서

과생된 AI 모델이며 자연어를 이해하고 생성할 수 있도록 설계되었다. Zhao et al. (2023)에 따르면 2023년을 기준으로 발표된 주요 LLM은, 100억 개 이상의 매개변수를 가진 것만으로 제한하더라도 아래 그림과 같이 매우 많다.



<그림 1> 100억 이상의 매개변수를 가진 LLM의 종류 (Zhao, W. X. 2023:9)

산업계에서의 기술 개발과는 별도로 국내 학계에서 이러한 LLM들에 대한 연구들은 국내의 경우 전 분야를 막론하고 이루어지기 시작하고 있는 것으로 보이나, 그중에서도 특히 전산언어학이나 언어정보학 분야에서 주로 다뤄지고 있다. LLM이 결국 자연어로 입력을 받고 그 출력 역시 자연어로 생성해 내는 형태의 인공지능이라는 점에서 컴퓨터 공학과 언어학의 융합 학문의 성격을 지니는 이러한 분야에서 그 논의가 활발히 진행되고 있는 것은 어찌 보면 당연한 일이다.

그러나 위의 그림에 있는 것과 같은 수많은 LLM들 중 상당수는 주로 영어 말뭉치를 기본적인 학습데이터로 사용하였기 때문에 한국어 처리에 대한 성능은 영어에 대한 것과 비교할 때 상대적으로 좋지 못하다는 것이 이제까지의 대체적인 중론이었다. 그러나 최근(2024년 5월 13일) 발표된 GPT-4o 모델은 한글 토큰화 방식을 획기적으로 개선하면서 한글로 이루어지는 한국어 텍스트의 이해(NLU)와 출력 효율이 이전 모델에 비해 월등히 좋아졌다. 물론 한국어 처리의 효율이 개선되었다는 것이 한국어와 관련된 모든 면에 있어서 이 모델의 성능이 향상되었다는 것과 직접적으로 연결되는 것은 아닐 수 있다. 토큰 효율과는 관계없이 한국어 텍스트를 주로 학습한 네이버의 클로바엑스(CLOVA X) 등과 같은 국내 기업의 모델이 특정 국내 상황과 관련된 도메인 영역에서 더 나은 답변을 생성할 수도 있을 것이다. 이에 대해서는 각 세부 분야별로 성능을 평가하는 연구가 앞으로 이루어질 필요가 있다.

3. 연구 방법

3.1. 연구 자료 및 도구

앞서 언급한 바와 같이 본 연구에서는 국립국어원의 <한국어학습자말뭉치>를 활용하였으며, 구체적으로 보면 2024년 5월 14일에 말뭉치 사용을 신청하여 5월 16일에 승인을 받아 다운로드한 자료를 사용하였다.¹⁾ 이 말뭉치에 포함되어 있는 자료 중에서도 본 연구에서는 고급 수준의 학습자가 작성한 작문 자료를 선별하고 그중에서 무작위로 100개를 추출하여 분석하였다.²⁾ 이러한 연구 대상 자료의 규모는 아래 표와 같다.

<표 1> 분석 자료의 규모

항목 구분	규모 (개)
텍스트 수	100
전체 문장 수	1,374
텍스트 당 평균 문장 수	13.74

이후에 현재 상용화되어 있는 LLM 중 가장 널리 활용되고 있는 서비스인 ChatGPT4o를 활용하여 이 말뭉치 자료에 담긴 학습자 작문을 교정하도록 하였다. 한국어 처리에 있어서 클로바엑스(CLOVA X)가 어느 정도 강점을 가질 수 있을 것이라는 가능성은 충분하지만, 앞서 언급한 바와 같이 ChatGPT4o는 최근 변경된 토큰화 방식으로 한국어 처리에 있어서 가지고 있던 약점이 크게 개선되었다는 것이 보고되었고, 무엇보다 클로바엑스와는 달리 교정할 내용이 담긴 텍스트 파일을 직접 업로드하여 교정을 수행하도록 하고 그 결과를 역시 txt나 csv 등의 파일로 바로 다운로드할 수 있다는 점에서 사용성이 우월하다는 점을 고려하여 ChatGPT4o를 활용하였다.

3.2. 입력 프롬프트

ChatGPT 등의 LLM의 응답이나 처리 결과는 사용자가 어떤 프롬프트를 입력했느냐에 크게 좌우된다. 이는 LLM이 매우 사소한 차이도 정교하게 식별하도록 고안되어 있어 소위 ‘아 다르고 어 다른’ 입력에 대해 ‘아 다르고 어 다른’ 답변을 내놓는다는 특성이 있기 때문이다. 또한 LLM들은 두 가지 중요한 특징이 있는데 하나는 장기 기억의 보유가 불가능하다는 것이고 다른 하나는 - 특별히 검색 결과를 활용하여 답변하도록 지정하지 않는다면- 외부 맥락을 활용한 추론을 하지 않는다는 것이다. 다시 말해, ChatGPT류의 LLM들은 사용자가 그때그때 입력한 프롬프트를 통해서 주어진 맥락을 추론에 활용하여 답변을 생성하는 방식으로 작동되는 것이라고 할 수 있으며, 이는 곧 적절한 프롬프트를 제공하는 것이 답변의 질에 직접적인 영향을 미친다는 것을 가리킨다.

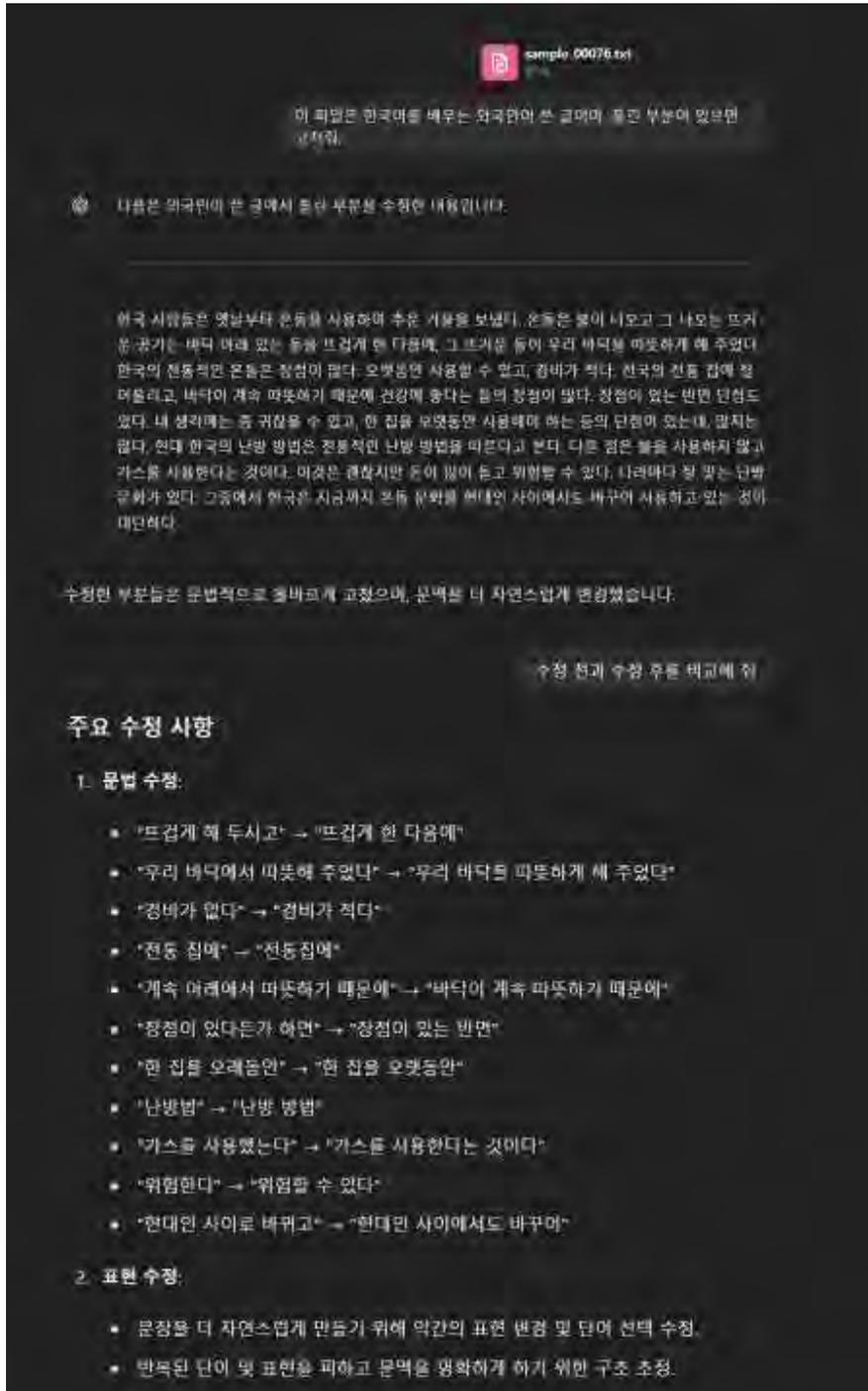
이뿐만 아니라 ChatGPT와 같은 생성형 AI들은 실체가 아닌 것을 그럴 듯한 말로 지어내 응답하는 ‘환각 현상’을 일으킬 수 있다는 점이 꾸준히 논란이 되고 있는데, 이러한

1) <한국어학습자말뭉치>는 현재 시점에서도 계속 구축이 진행되고 있으므로 말뭉치 승인 시점을 밝힘으로써 현재 시점에서 최신 버전의 자료를 활용하였다는 점을 밝힌다.

2) <한국어학습자말뭉치>의 문어 자료 중에 고급 수준의 학습자 자료는 1,670개가 포함되어 있었다. 이 중에서 본 발표를 위해서는 100개를 우선 검토해 본 것이다.

환각 현상을 최소화하여 생성형 AI의 답변을 신뢰할 만한 수준으로 관리하기 위한 방식 중 하나로서도 프롬프트 엔지니어링의 중요성이 강조되고 있다.

본 연구에서는 본격적인 프롬프트엔지니어링의 기법들을 도입하지는 않았으나 좋은 프롬프트의 원리로 제시되어 온 원리들을 일부 참조하였고, 최적화된 프롬프트를 작성하기 위하여 여러 단계의 프롬프트 수정을 거쳤다. 우선 다음과 같이 간단한 프롬프트를 작성하여 LLM의 한국어 교정 작업 수행 결과를 검토하였다.



<그림 2> 임시 프롬프트 입력 결과

LLM의 응답 결과를 보면 띄어쓰기 및 표기 층위(오래동안 -> 오랫동안), 어휘 및 표현 층위(경비가 없다 -> 경비가 적다), 문법 층위(위험한다 -> 위험할 수 있다) 등 다양한 차원에서 전반적인 교정이 이루어진 것을 확인할 수 있다. 또한 여기에서 그치지 않고 ‘문장을 더 자연스럽게 만들기 위’한 수정, 그리고 ‘반복된 단어 및 표현을 피하고 문맥을 명확하게 하기 위한 구조 조정’까지 이루어졌음을 보고하고 있다. 그런데 한국어학 습자말뭉치 구축 과정에서 이루어진 인간 교정자의 교정 작업은 표기 층위, 어휘 및 표현 층위, 문법 층위에서는 이루어졌을 것이지만 문장을 더 자연스럽게 만들고 더 명확하고 좋은 글로 만들기 위한 윤문 과정은 이루어지지 않았을 것이다. 따라서 이러한 점을 고려하여 교정 지침을 보다 명확히 줄 필요가 있었다. 이러한 점을 포함하여 프롬프트를 수차례 수정한 끝에 다음과 같은 최종 프롬프트를 작성하였다.

<표 2> 입력한 프롬프트

프롬프트 구조	프롬프트 내용
페르소나 지정	너는 프로페셔널한 한국어 교사야. 나는 한국어를 배우고 있는 외국인 이야.
일반적인 지침	내가 한국어로 쓴 글을 txt 파일로 올려줄게. 내가 읽어보고 틀린 부분이 있다면 올바른 한국어로 교정해 줘.
세부 지침	다음은 이 작업에 대한 세부적인 지침이야. - 명백한 문법 오류, 어휘 오류, 표기 오류를 교정해 줘. - 문장을 더 자연스럽게 만들기 위한 교정은 하지 말아 줘. - 반복된 단어나 표현을 피하거나, 문맥을 명확하게 하기 위한 교정은 하지 말아 줘. - 글의 전체적인 구조는 수정하지 말아 줘. - 어떤 글에는 “name”이라는 문자열이 등장해. 이것은 누군가의 이름을 익명으로 나타내기 위해 쓴 기호니까 수정하지 말아 줘.
출력 형식 지정	위의 교정 작업을 수행한 결과를 아래 형식으로 저장해 줘. - 교정 전과 교정 후를 문장 단위로 서로 대응되는 데이터프레임으로 만들어 줘. - 아무 것도 교정하지 않은 문장도 삭제하지 말고 데이터프레임에 포함해 줘. - 구체적으로 어떤 표현을 어떻게 교정했는지를 교정한 건마다 설명해 줘. (예시 : “바뀌고” -> “바꾸고”)
파일 형식 지정	마지막으로 생성한 데이터프레임을 csv 파일로 저장해서 다운로드할 수 있도록 링크를 줘. - 파일 이름은 “corrected_식별번호”로 해 줘. 이때 ‘식별번호’란, 원본 파일의 이름에 들어있는 것과 동일한 숫자야. - 인코딩은 utf-8로 해 줘.

3.3. 분석 방법

이렇게 얻은 100개의 텍스트에 대한 LLM의 교정 작업 결과를 <한국어학습자말뭉치> 오류 주석 부착자의 교정 결과와 문장 단위로 비교하였다. 비교 결과를 정량적으로 산출하기 위해 AI가 교정한 문장과 사람이 교정한 문장 사이의 코사인유사도(Cosine Similarity)를 산출하였다. 코사인 유사도는 벡터 기반 유사도 측정 방법으로 공백이나 구두점을 고려하지 않고 단어의 출현 빈도와 단어 간의 유사성을 비교하는 데 적합하다는 점에서 이를 채택하였다. 이후 코사인유사도 점수가 낮게 나타난 문장들을 대상으로 질적 검토 작업을 덧붙였다. 이상의 절차를 따라 수행한 분석 결과는 다음 장에서 기술하겠다.

4. 분석 결과

4.1. 양적 분석 결과

LLM이 자동으로 교정한 문장과 한국인 모어 화자가 교정한 문장을 비교하여 코사인 유사도를 산출한 결과를 종합해 보면 아래 표와 같다.

<표 3> 코사인유사도 분석 결과

유사도 점수 평균	0.796
완전 일치(유사도=1) 문장 수	538
완전 일치 문장 비율	39.16%

결론적으로 사람이 교정한 글과 LLM이 교정한 글의 평균적인 유사도 점수는 약 0.8에 달하는 것으로 나타났다. 또한 전체 1,374개의 문장 중에서 사람의 글과 LLM의 글이 완전히 일치하는 문장이 538나 되었는데 이는 전체의 약 40%에 해당하는 비율이다. 즉 전체 비교 대상 문장 중 약 40% 정도는 LLM의 교정 결과와 사람의 교정 결과가 정확히 일치하였으며, 이를 포함하여 전체 비교 대상 문장 수준으로 확장하여도 유사도 약 0.8의 점수에서 볼 수 있듯이 LLM의 교정 결과는 사람이 수행한 것과 평균적으로 매우 유사한 수준임을 확인할 수 있었다. <학습자말뭉치>의 오류 주석을 하나의 준거라고 가정했을 때 AI의 교정이 이 준거를 상당히 만족시키는 수준의 수행을 할 수 있다고 볼 수도 있을 듯하다. 그렇다면 사람의 교정 결과와 일치도가 비교적 낮게 나타난 문장들은 어떠한 점에서 차이가 나는 것일까? 이에 대해서는 이어지는 절에서 자세히 살펴보도록 하겠다.

4.2. LLM이 수행한 교정의 특징

4.2.1. 문체 및 화제 관련 특징

LLM이 수행한 교정의 결과가 사람의 교정 결과와 차이가 나는 경우에 가장 흔하게 목격된 유형은 바로 문체 및 화제와 관련되어 나타나는 LLM의 특징으로 귀납될 수 있었다. 여기에 해당되는 몇 가지 예문을 제시하면 아래와 같다.

- (1) 가. (원문) 제가 최저 임금제 찬성이다.
나. (모어 화자) 나는 최저임금제에 찬성이다.
다. (LLM) 제가 최저임금제를 찬성합니다.
- (2) 가. (원문) 저는 최저임금제 설정에 대한 찬성한다고 생각한다.
나. (모어 화자) 나는 최저임금제 설정에 대해 찬성한다.
다. (LLM) 저는 최저임금제 설정에 찬성한다고 생각한다.
- (3) 가. (원문) 세상에 너나없이 예쁘고 싶죠?
나. (모어 화자) 모두 예뻐지고 싶어한다.
다. (LLM) 세상에 너나없이 예쁘고 싶죠?

먼저 (1나)와 (1다)를 비교해 보면, 모어 화자는 원문의 문체와 마찬가지로 ‘ㄴ/는다’를 활용하여 문장을 끝맺은 반면에 LLM은 함쇼체 문장으로 교정한 것을 알 수 있다. 이는 chatGPT가 사람과의 대화 상황을 가정하여 개발된 챗봇의 형식을 취하고 있다는 점과 관련되는 듯하다. 별다른 유표적 지시가 없다면 chatGPT는 사용자에게 제공하는 출력을 상관적 장면의 격식체로 생성하는데, 사용자에게 제공하는 응답의 문체와 교정을 요청 받은 작문 자료의 문체를 구별하지 않고 무차별적으로 작문 자료의 모든 문장을 함쇼체 문체로 바꾼 사례가 존재하였다. 이를 통해 프롬프트에 문체 스타일을 명시적으로 지정해 줄 필요가 있음을 알 수 있었다.

다음으로 (2나)와 (2다)를 비교해 보면, 모어 화자가 원문의 ‘저는’을 ‘나는’으로 바꿈으로써 단독적 장면에 알맞은 문체로 화제 수준을 맞추는 방향으로 교정한 것과는 달리 LLM은 이 부분을 수정하지 않았음을 알 수 있다. 종결어미를 통해 나타나는 화제 또는 문체에 다른 문장 성분이 표현하는 화제나 문체를 일치시켜야 한다는 한국어의 특성이 잘 반영되지 못한 것을 추측해 볼 수 있다.

예문 (3)의 경우 모어 화자는 원문을 상당히 많이 바꾼 반면에 LLM은 원문을 전혀 고치지 않았다는 점에서 차이가 컸다. 이 글의 원문을 보면 다른 모든 문장은 ‘ㄴ/는다’로 종결되는 평서문인 데 비해 유일하게 의문문인 이 문장에서만 ‘-지요’가 사용되어서 매우 어색한 느낌을 주고 있었다. 원문을 최소한으로만 바꾸면서 원저자가 의도한 확인 의문문의 기능을 살리고 문체의 통일을 피하기 위해서는 “세상 사람들은 너나없이 예뻐지고 싶어하지 않습니까?” 정도로 교정하는 것이 최선일 듯하다.

4.2.2. 피동 구문 관련 특징

LLM의 교정에서 또 한 가지 관찰되는 특징은 모어 화자의 교정에 비해서 피동구문을 선호하는 것으로 보이는 현상이다. 이에 관련된 예문은 아래와 같다.

- (4) 가. (원문) 현대에는 난방기를 발명하고 전국적으로 사용했다.
나. (모어 화자) 현대에는 난방기를 발명하고 전국적으로 사용한다.
다. (LLM) 현대에는 난방기가 발명되어 전국적으로 사용되었다.
- (5) 가. (원문) 또 학교 폭력이 당한 학생이 잘 보호하면 울며 겨자 먹기는 학생이 없
어지고 비관 자살 하는 학생도 없어진다.
나. (모어 화자) 또 학교폭력을 당한 학생을 잘 보호하면 울며 겨자 먹는 학생이 없
어지고 비관자살하는 학생도 없어진다.
다. (LLM) 또 학교 폭력을 당한 학생이 잘 보호받으면 울며 겨자 먹기 같은 학생
이 없어지고 비관 자살하는 학생도 없어질 것이다.

예문 (4)의 경우 모어 화자는 원문의 구조를 그대로 가져가면서 시제 부분만 수정하였으나 LLM은 ‘난방기를 발명하- … 사용하-’의 능동 구문을 ‘난방기가 발명되- … 사용되-’의 피동 구문으로 바꾸었다. 예문 (5)의 경우에는 원문이 ‘학생이 잘 보호하면’으로 문법 구조상 비문인데 이 문제를 모어 화자의 경우 ‘학생을 잘 보호하면’과 같은 능동 구문으로 수정함으로써 해결한 반면에 LLM은 ‘학생이 잘 보호받으면’이라는 피동 구문으로 수정함으로써 해결하였다. 결과적으로 두 교정이 모두 원문의 비문법성을 해소하였으나, 전반적으로 볼 때 LLM의 피동구문 사용 양상이 두드러지는 것을 알 수 있었다. 이는 LLM이 영어식 문법 구조의 영향을 많이 받은 것에서 기인하는 것으로 추정해 볼 수 있다.

4.2.3 문장의 문법 구조 관련 특징

LLM 교정 결과의 품질이 모어 화자의 것에 비해서 상대적으로 낮은 것으로 인식되도록 하는 문제 중 하나는 한국어 문장의 문법 구조와 관련되어 비문으로 판정될 수밖에 없는 부분을 그대로 방치한 몇몇 사례들이었다. 이에 해당되는 예문을 몇 가지 제시하면 아래와 같다.

- (6) 가. (원문) 이러한 배우하기 위해서 학벌주의 아니라 전인교육이다.
나. (모어 화자) 이러한 것을 배우기 위해서 필요한 것은 학벌주의가 아니라 전인교육이다.
다. (LLM) 이를 배우기 위해서는 학벌주의가 아니라 전인교육이다.
- (7) 가. (원문) 방에서 계속 따뜻할 수 있어서 온돌은 장점이다.
나. (모어 화자) 방에서 계속 따뜻하게 있을 수 있는 것이 온돌의 장점이다.
다. (LLM) 방에서 계속 따뜻할 수 있어서 온돌은 장점이다.

(6)과 (7)은 모두 원문에 나타난 문장 구조상의 문제 - 호응이 맞지 않는 문제-를 사람 교정자는 해결하였으나 LLM은 해결하지 못한 경우를 보인다. (6)의 경우 원문에 ‘-기 위해서 … N이다’의 잘못된 구문이 나타났는데 이를 모어 화자는 ‘-기 위해서 필요한 것은 … N이다’의 구조로 바꾸어 해결한 반면에 LLM은 원문에 나타난 잘못된 구조를 그

대로 사용하였다.

(7)에서도 이러한 문제가 나타나는데, ‘-어서 ... N이다’의 구조를 LLM은 수정하지 않으므로써 교정 후의 문장도 여전히 비문으로 남게 되었다. 사람 교정자가 이 문장을 ‘-는 것이 ... N이다’의 올바른 구조로 수정한 것과 대비되는 대목이다. 한국어 복문의 다양한 문법 구조를 처리하는 능력은 아직까지 LLM의 성능이 상대적으로 만족스럽지 못한 수준에 머물러 있는 듯 보인다.

5. 맺음말

이상과 같이 본 연구에서는 LLM의 교정 결과를 모어 화자의 교정 결과와 비교해 보았다. 본문에서 미처 구체적으로 언급하지 못한 여러 가지 차이점들이 더 있는데 몇 가지만 간추려서 기술하면 다음과 같다. 첫째, 한국어 학습자가 어색한 연어 표현을 만들어 사용한 경우에 이는 LLM이 오히려 더 자연스러운 연어 표현으로 교정하는 경향이 보인다. 둘째, 일부 오타의 경우에 이를 대치할 만한 한국어 단어를 추론하는 데 실패한 LLM이 한국어에 없는 단어를 지어낸 경우가 발견된다. 셋째, 예컨대 ‘-기 십상이다’처럼 한국어 교재에는 빈번히 등장하나 LLM이 학습한 일반적인 데이터에서는 자주 접하지 못했을 것으로 추정되는 표현을 LLM이 비문으로 간주하여 수정한 것이 관찰되었다. 이상의 여러 구체적인 사항을 포함하여 더 자세한 분석은 추후 연구로 미룬다.

참고문헌

- 김유신·김기태 (2024). 코드 너머의 언어: 대규모 언어 모델과 프롬프트 엔지니어링, 생각나눔.
- 서승완 (2023). 프롬프트 엔지니어링 교과서, 애드앤미디어.
- Ozdemir, S. (2023). Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs, Addison-Wesley Professional.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

<한국어 학습자 오류에 대한 LLM과 모어 화자의 교정 양상 비교 연구>의 토론문

이진(성균관대)

이 연구는 한국어 학습자 작문의 오류를 인간과 거대언어모델(Large Language Models, LLM)이 어떻게 교정하는지 비교하여 한국어 학습자 작문의 자동 교정 도구로서 거대언어모델의 활용 가능성을 확인하는 데 목적이 있습니다. 현재 거대언어모델의 한국어 학습자 작문 교정 능력을 평가함으로써 최근 크게 관심을 받고 있는 인공지능(AI)을 활용한 언어 교육의 가능성과 한계를 밝히는 데 기여하였다고 생각합니다. 의미 있는 연구를 발표해 주신 발표자에게 감사드리며 발표문을 보고 몇 가지 궁금했던 점에 대해 발표자의 의견을 구하는 것으로 토론자의 소임을 다하고자 합니다.

1. 본 연구는 국립국어원 <한국어 학습자 말뭉치> 중 ‘고급 수준의 학습자가 작성한 작문 자료’에서 무작위로 100편을 선별하여 연구 대상으로 삼았다고 밝혔습니다. 특별히 고급 학습자의 작문을 연구 대상으로 삼은 이유가 있는지 궁금합니다. 거대언어모델의 한국어 학습자 작문 교정 능력을 세부적으로 평가하려면 <한국어 학습자 말뭉치>에 주석된 다양한 오류 유형을 균형적으로 선별하여 교정 양상을 살피는 것이 필요하다고 생각하는데 이에 대한 발표자의 의견을 여쭙습니다.
2. 본 연구에서도 밝히고 있지만 거대언어모델을 활용할 때는 프롬프트 엔지니어링(Prompt Engineering)이 중요한 것으로 알려져 있습니다. 프롬프트 엔지니어링과 관련된 해외 연구들에서는 거대언어모델의 성능을 높이기 위한 다양한 방법론들을 제시하고 있습니다. 본 연구에서도 페르소나를 설정하는 등의 프롬프트 엔지니어링을 사용하고 있는데 그 외에도 일반적으로 거대언어모델의 성능을 높인다고 알려진 여러 방법론들(Few-shot learning, One-shot learning, Chain of Thought 등) 가운데 현재 형태의 프롬프트를 활용한 특별한 이유가 있는지 여쭙습니다.
3. 연구 결과의 양적 분석 중 하나로 거대언어모델이 교정한 문장과 한국어 학습자 말뭉치의 교정 문장의 유사도(Similarity)를 산출하였는데 어떤 방식으로 토큰화(Tokenization)하여 유사도를 산출하였는지 궁금합니다. 어절 단위인지 형태소 단위인지에 따라서 유사도 분석 결과가 달라질 것으로 보이는데 발표자의 의견을 여쭙습니다.
4. 본 연구의 가장 주된 연구 결과로 거대언어모델이 수행한 교정의 특징을 3가지로 설명하였습니다. 거대언어모델의 교정 양상에서 3가지 유형의 교정 형태가 두드러지게 나타나는 것으로 이해하였는데 해당 유형이 거대언어모델의 전체 교정 유형에서 어느 정도를 차지하는지 궁금합니다. 또한 4.2.2절에서 제시한 피동 구문 관련 특징의 경우, 거대언어모델이 학습한 한국어 자료의 문제라고 생각됩니다. 거대언어모델이 학습한 한국어 자료에서 피동 표현의 과잉 사용이 두드러지게 나타나고 이를 학습한 거대언어모델이 이러한 교정의 특성을 보이는 것으로 보입니다. 거대언어모델이 어법상 정확한 한국어 자료만을 학습하지 않

았기 때문에 한국인이 자주 범하는 오류에 대해서는 수정을 안 하거나 잘못 수정할 가능성은 없을지 발표자의 의견을 여쭙습니다.

5. 마지막으로 한국어교육 현장에서 인간 교사가 한국어 학습자 작문의 오류를 수정할 때는 앞서 제시한 최소 수정의 원칙을 지키면서 학습자의 수준을 고려하여 교정 표현을 선별한 후 교정을 진행합니다. 그러나 거대언어모델의 경우에는 해당 작문을 작성한 학습자의 수준을 고려하여 알맞은 교정 표현을 찾아 수정하는 것이 가능할지 의문이 듭니다. 본 연구는 고급 학습자의 작문을 대상으로 하고 있지만 초급과 중급 수준 학습자 작문의 경우에는 거대언어모델을 활용하여 작문을 교정할 수 있을지 활용한다면 어떤 점을 고려해야 할지 발표자의 의견을 구합니다.

러시아국립말뭉치(Russian National Corpus)의 구조와 기능

김세림, 남혜현 (연세대)

차 례

1. 머리말
 2. RNC의 구축 배경 및 발전 현황
 3. RNC의 핵심 기능
 4. RNC를 이용한 러시아어 연구의 동향
 5. 맺음말
-

1. 머리말

러시아국립말뭉치(Russian National Corpus, 이하 RNC)는 러시아 최대의 말뭉치 모음으로서 총 20억 단어 이상의 러시아어 텍스트 모음과 더불어 언어학적 주석 체계 및 검색 도구를 갖추고 있다. RNC는 동슬라브어 시기(6세기)부터 현대까지의 러시아어 텍스트를 망라하며 다양한 매체(신문, 일상대화, 온라인 미디어) 및 변수(지역, 시대)에 따른 사회언어학적 변이형을 포함한다. RNC에는 가장 대표적인 범용 말뭉치인 주요(Main) 말뭉치 외에도 통사 말뭉치(SynTagRus, Syntactically Tagged Russian Text Corpus), 멀티미디어(Multimedia) 말뭉치, 강세(Accentological) 말뭉치 등과 같이 특수한 목적성을 띠는 48개의 독립적인 하위 말뭉치가 구축되어 있다. 특히 RNC는 문화적 측면에서 중요한 픽션 장르(산문, 희곡 등) 뿐만 아니라, 언어 사용 규범을 대표하는 논픽션 장르(학술 문헌, 연설문 등)의 텍스트나 발화 또한 포함하므로 질적·양적 연구의 귀중한 자료가 된다. RNC의 모든 텍스트는 작성 시기, 저자 정보, 장르, 문체, 주제 등에 대한 메타데이터(meta data)를 가지며, 텍스트를 구성하는 각 단어에는 형태적, 의미적, 통사적 주석이 붙는다. 이를 토대로 RNC의 사용자는 자신의 목적에 따라 단어의 어휘·문법적 특성을 조합하여 특정 단어 시퀀스에 대한 검색을 수행하거나, 메타데이터의 정보를 선택하여 검색할 텍스트의 범위를 제한할 수 있다.

섬세한 검색 도구를 갖는 RNC는 러시아어학계 전반에 새로운 통찰력과 연구 기회를 제공하는 구심점으로 기능하고, 나아가 대표적인 러시아어 말뭉치로서의 입지를 더욱 강화할 길을 걷고 있다. 사전의 단어 정의와 신경망에 의해 생성된 단어 정의 중 어떤 정의가 사용자에게 더 잘 인식되는지 알아보기 위한 새로운 실험을 시작하는 등 말뭉치는

지금, 이 순간에도 계속 개발되고 있다. 이러한 기능 발전에도 불구하고 RNC는 영국의 대표적인 코퍼스인 BNC(British National Corpus) 또는 현대 미국 영어 코퍼스 COCA(Corpus of Contemporary American English)와 비교했을 때 국내 연구자들에게 많이 소개되지 않은 실정이다. 이는 물론 영어 교육 강국인 대한민국에서 러시아어가 아직 영어나 중국어에 비해 사회적 수요 및 관심이 크지 않기 때문이겠다. 그러나 생성형 인공지능 학습에 투입되는 고품질 데이터의 중요성을 전 세계가 인식하고 있는 현시점에서 해외 국가의 말뭉치가 앞으로 어떠한 발전 방향을 모색하는지 관심 있게 살펴볼 필요가 있다. 따라서 본 발표문에서는 RNC의 구축 배경과 발전 현황을 소개하고 말뭉치의 구조 및 핵심 기능을 설명하고자 한다.

2. RNC의 구축 배경 및 발전 현황

RNC는 올해 4월 20주년을 맞이했다. 시대를 아우르는 산문의 다양성을 반영하고 러시아 문화를 표상하는 텍스트 모음을 만들겠다는 아이디어에서 비롯하여, 현재 러시아의 대표적인 코퍼스로 자리매김한 것이다. RNC의 홈페이지는 2004년 4월 29일¹⁾에 무료로 오픈되었으나, 그 창설 작업은 일찍이 2000년대부터 시작되었다. 해당 프로젝트의 전반에는 러시아의 여러 연구소 및 고등교육기관²⁾ 소속 연구원, 학생 등을 포함한 다방면의 전문 인력이 참여했다. 또한 러시아의 최대 인터넷 포털사인 야덱스(Yandex)는 RNC 프로젝트의 시작 단계에서부터 검색 시스템 및 기능의 개발을 지원해왔다. 이제 RNC는 역으로 야덱스가 제공하는 서비스의 발전에 기여하고 있으며, 러시아어 학습자, 언어학자, 번역가 등을 비롯한 광범위한 수요층이 러시아어에 대한 지식을 쌓고 그 사용 양상을 탐구할 수 있는 장을 마련해준다.

2.1. 구축 배경³⁾

국립국어원에서 국가 말뭉치를 구축하고 발전한 배경에 『국어기본법』 제16조(국어 정보화의 촉진)⁴⁾가 있듯, RNC 프로젝트의 배경에는 러시아 정부의 언어 확산 정책이 크게 자리한다. 1990년대 초까지만 해도 러시아어는 그 사용 인구가 3억 5천만 명이 넘는 거대 언어였다. 그러나 소비에트의 해체 이후 강대국으로서의 지정학적 영향력이 상실되며 해외에서 러시아어의 위상이 흔들리고, 전 세계적으로 러시아어 사용자의 수와 사용 영역이 감소했다. 이러한 이탈 현상이 있었음에도 여전히 해외 러시아어 사용자는 유럽,

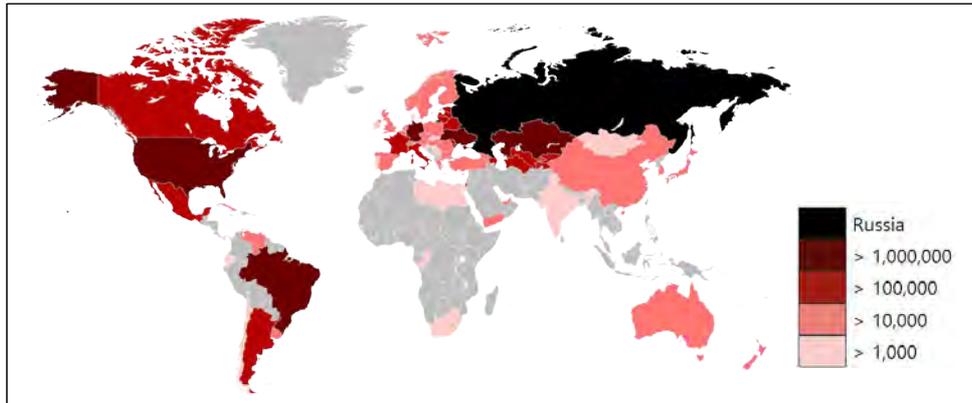
1) 해당 날짜는 러시아의 언어학자 잘리즈냐크(A.A. Zaliznyak)의 생일이라 더욱 상징성을 띤다. 초창기 RNC 홈페이지 링크는 다음과 같다: <https://museum.ruscorpora.ru/>.

2) 러시아 학술원 산하 러시아어 연구소(V.V. Vinogradov Russian Language Institute of the Russian Academy of Sciences), 고등경제대학교(Higher School of Economics), 정보 전달 문제 연구소(A.A. Kharkevich Institute for Information Transmission Problems), 보로네시 주립대학교 산하 언어학 연구소(Institute for Linguistic Studies) 등을 대표적으로 들 수 있다.

3) 남혜현(2011)의 1, 3장을 참고하여 작성하였다.

4) ① 국어를 통하여 지식과 정보를 생산하고 활용하여 새로운 문화를 창조할 수 있도록 국어 정보화를 위한 각종 사업을 적극적으로 시행하여야 한다. ② 국가는 정보통신망을 활용하는 국민이 국어를 편리하게 사용할 수 있도록 필요한 정책을 시행하여야 한다.

미국, 이스라엘, 중국, 아프리카 등의 여러 대륙에 걸쳐 분포한다.



<그림 1> 러시아 디아스포라 지도 (2020년 기준)⁵⁾

푸틴 대통령은 이러한 초국가적이고, 초대륙적이며, 공통의 문화적 정체성으로 묶인 “러시아 세계(Russian World)”, 러시아어로 “루스키 미르(Russkij Mir)”⁶⁾에 대한 강한 자부심을 표한 바 있다. 러시아의 지식인들은 루스키 미르의 소속성을 결정하는 세 가지 범주로 러시아어, 러시아 문화, 그리고 러시아에 대한 애정을 들었는데, 이 가운데 러시아어의 역할이 특히 강조되었다. 따라서 러시아 정부는 소비에트의 해체 이래로 언어와 관련하여 다양한 정책적 노력을 기울여 왔다. 그 대표적인 예로는 첫째, 해외 거주 자국민이 해당 거주 국가의 민족어 정책으로 불이익을 받는 것을 방지하고 러시아어의 위상을 재립하기 위한 동포지원정책과 둘째, 러시아어를 전 세계적으로 확산시키기 위한 목적의 연방정부 <러시아어(Russkij jazyk)> 프로그램(2002-2005년, 2006-2010년, 2011-2015년, 2016-2020년)을 들 수 있다.

RNC의 구축과 밀접한 관련을 맺은 것은 연방정부 <러시아어> 프로그램이다. 해당 프로그램의 주요 기획 기관은 러시아연방 교육부, 러시아 학술원 등으로, 러시아어의 정신적, 예술적, 학술적, 문화적 지위 보존을 목표로 한다. 이에 따른 프로그램의 주요 과제에는 러시아어 교수법 개발, 대중매체를 통한 러시아 언어와 문화의 확산 등이 포함된다. 프로그램의 구체적 사업은 학술연구사업과 기타 사업으로 나뉘는데, 전자는 대중매체에서 러시아어 사용 표준안 연구, CIS와 발트 국가에서 러시아어의 위상과 사용 연구 등과 같은 이론적 연구가 주되고, 후자는 현장성과 실질성을 띠는 세미나 및 협의회, 교재 개발, 대중매체를 통한 선전 등을 포함한다. 기타 사업의 한 예로, 러시아어학과의 교수자들은 고등경제대학교(Higher School of Economics)와 러시아 학술원 산하 러시아어 연구소(V.V. Vinogradov Russian Language Institute of the Russian Academy of Sciences)에서 공동 주최한 고급 수준의 교육과정을 이수하고 “러시아국가말뭉치와 인문학 교육의 문제(Nacional’nyj korpus russkogo jazyka i problemy gumanitarnogo obrazovanija)”라는 제목의 컨퍼런스에 참여했다. 이를 통해 대학 소속 학자들은 ‘국립말

5) 출처: https://commons.wikimedia.org/wiki/File:Map_of_the_Russian_Diaspora_in_the_World.svg

6) 오랜 시기에 걸쳐 형성된 러시아 디아스포라, 1991년 소비에트 해체 이후 구소련공화국에 남은 러시아인, 그리고 세계 각지에서 러시아어를 사용하는 비러시아인 모두를 통합하는 개념을 일컫는다.

말뭉치'라는 새로운 지식 도구에 대해 학습하고 RNC 프로젝트 팀과 협력하여 구어 말뭉치의 자료를 수집하는 등 말뭉치의 실질적인 구축 및 개발에 기여했다(Kol'cova 2008: 58).

2.2. 발전 현황

연방정부 <러시아어> 프로그램의 실행 목표에 대응하여 RNC 프로젝트 팀은 말뭉치 구축의 표준에 따라 여전히 고품질의 러시아어 자료를 축적하고 생성하고 있다. 특히 2020년부터 2023년까지 RNC는 “러시아어의 디지털 문서화를 위한 차세대 언어 소프트웨어 플랫폼: 인프라, 자원, 학술 연구” 사업의 명목으로 러시아연방 과학고등교육부의 재정적 지원을 받아 개발되었다. 최근 RNC는 인공지능 기술의 개발에 힘입어 말뭉치 구축 작업의 자동화와 효율화를 위해 신경망 모델을 적극 활용하고 있다. 이러한 말뭉치의 최신 변화에 관한 정보는 RNC 홈페이지의 “소식(News)” 게시판을 통해 전해진다.

최근에는 크게 두 가지 변화가 있었다. 첫째, 역사 말뭉치(Historical Corpus)에 대한 대대적인 보충 작업이 이루어졌다. 11-14세기의 고대러시아어(Old East Slavic) 말뭉치에는 문학 작품과 공식적인 상용 문서 텍스트가 포함되어 약 1,000개의 어휘소로 확장되었고 현대 단어에 대한 참조가 추가되었다. 또한 15~17세기의 구러시아어(Middle Russian) 말뭉치, 교회슬라브어(Church Slavonic) 말뭉치, 그리고 구어(Spoken) 말뭉치가 전면 확장되며 통계, 빈도 및 N-그램 출력 기능을 사용할 수 있게 되었다. 둘째, 코퍼스 자료에 대한 홈페이지 메뉴에 신경망 모델 페이지(The RNC neural network models)가 개설되었다. 해당 페이지에는 RNC 코퍼스를 구축하는 데에 사용된 신경망 모델에 대한 설명, 사용 지침, 모델 파일 등이 포함되어 있다. 토큰라이저(tokenizer), 워드투벡(word2vec) 모델, 그리고 형태적 주석 및 텍스트의 메타 주석(주제, 장르, 유형)을 위한 모델이 소개되는데, 설명은 간략하나 모델을 실제로 돌려볼 수 있으므로 자연어 처리나 기계학습에 관심 있는 사람에게 유익할 것으로 판단된다.

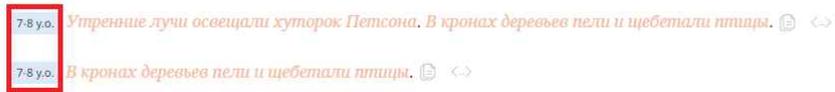
미래 시대에는 데이터가 곧 '자산'이라는 말이 있다. 챗지피티(ChatGPT)로 대표되는 생성형 AI 기술 발전이 가속화 함에 따라 세계 각국에서 말뭉치 자료에 대한 중요성이 강조되고 있다. 나아가 자국의 언어를 기반으로 한 고유의 AI를 개발할 필요성도 꾸준히 제기된다. 언어 데이터의 축적, 생성, 재생산이 이루어지는 인공지능 개발은 글로벌 AI 시장에서 자국의 고유한 문화를 보존하고 발전하는 문제 혹은 세계 언어 경쟁력 순위 문제와 직결되기 때문이다. 러시아 또한 RNC의 발전에 주력하며 이러한 시대적 흐름에 동참하고 있다. 따라서 RNC는 지역어/방언, 고어 등의 보전 가치가 있는 러시아 언어문화 자원과 언어 사용의 변화를 수집하여 말뭉치에 최신의 언어 현실을 반영한다. 이러한 작업은 최근 보충된 역사 말뭉치와 더불어 지역 및 해외 언론(Regional & International Media) 말뭉치, 방언(Dialect) 말뭉치, 범시(Panchronic) 말뭉치, 운문(Poetry) 말뭉치, 그리고 베타 버전의 러시아 고전(Russian classics β) 말뭉치의 구축으로 예증된다.

또한 RNC는 올바른 러시아어의 확산을 위해 자국민과 외국인 학습자를 대상으로 한 다양한 러시아어 학습 서비스 및 교육 자료를 구현하고 제공한다. 이에 대한 자세한 설명은 RNC 홈페이지 메뉴 중 하나인 “학교를 위한 RNC(The RNC for schools)” 페이지에서 찾아볼 수 있다. 그 대표적인 예로 들 수 있는 교육용(Educational) 말뭉치는 애초에 러시아 언어와 문학 수업에 활용될 목적으로 구축되었으며, 중·고등학교 교과 과정에서 다루는 산문 문학 장르의 다양성을 반영한다. 말뭉치의 규모는 작지만, 단어의 형태적

중의성이 해소된 텍스트로만 구성되고 명사류의 격변화 유형, 동사의 인칭변화 유형 등의 형태적 주석이 추가되어 있다. 교육용 말뭉치와 유사한 맥락에서, 비교적 최근인 2022년에 출시된 “2세부터 15세까지(From 2 to 15)” 말뭉치는 어린이, 청소년, 교사, 학부모를 대상으로 한 대규모 설문조사의 결과에 따라 선정된 75개의 산문 작품으로 이루어져 있다. 문법적 동음이의어가 제거된 점은 여타 말뭉치와 다르지 않지만, 해당 말뭉치의 특이점은 텍스트 단편마다 그것을 이해할 수 있어야 하는 독자의 최소 연령대가 주석된다(7)는 것이다. 물론 어휘력, 읽기 능력 등을 포함한 언어 발달 속도는 아이마다 상이하지만, 말뭉치의 주석은 평균 사례를 나타냄으로써 일종의 평가 지표가 될 수 있다.

Concordance KWIC Graph Statistics Frequency 1-gramm 2-gramm 3-gramm 4-gramm 5-gramm

1. Свен Нурдквист. Петсон и Финдус: Финдус переезжает (2012) ↩



<그림 2> 2세부터 15세까지 말뭉치의 예상 독자 연령대 주석 예

이러한 교육 및 학습 과제에 특화된 말뭉치 외에도 RNC는 검색 및 분석의 편의성을 극대화한 범용의 어휘 검색 도구를 제공하는데, 이는 아래의 장에서 기술하겠다.

3. RNC의 핵심 기능

RNC에서는 검색 단위 및 목적에 따라 크게 세 가지 유형의 질의(query)가 가능하다. 입력어와 정확히 일치하는 형태의 검색(Exact search), 연어관계 검색(Collocation search), 그리고 레마 및 태그 검색(Lemmas and tags search)이 바로 그것이다. 그중에서도 레마 및 태그 검색 도구의 강점은 특정한 어휘·문법적 특성을 갖는 단어 시퀀스를 검색어로 받을 수 있다는 것이다. 이렇게 사용자가 일련의 단어에 대해 복잡한 검색 조건을 설정할 수 있는 배경에는 RNC가 지닌 상세한 주석 체계가 있다. 즉, RNC 텍스트의 모든 토큰에는 형태적, 의미적, 통사적 및 기타 태그가 지정되는데, 이렇게 세분화된 태그를 조합하여 검색을 수행할 수 있다. <그림 5>에서 확인할 수 있듯, 검색은 기본적으로 사용자가 선택한 ‘선호 태그(preferred tags)’에 한해 이루어진다.

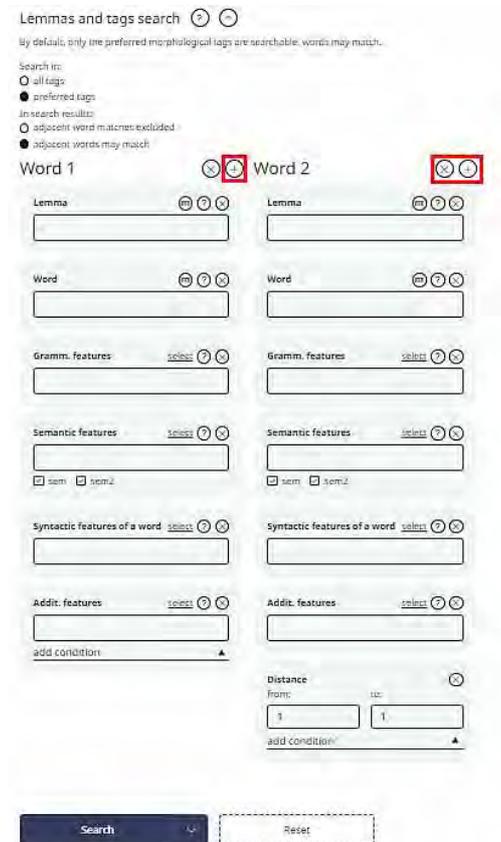


<그림 3> 정확히 일치하는 형태 검색

7) 연령대 주석은 신경망 모델에 의해 자동으로 생성된다. 모델이 예측한 결과가 전문가의 데이터와 92% 일치한다는 점에서 정확도가 높은 편이나, 아직 실험적이며 오류가 발생할 수 있다.



<그림 4> 연어관계 검색



<그림 5> 레마 및 태그 검색⁸⁾

앞서 살펴본 세 가지 도구는 기본적으로 용례 문장들의 목록, 즉 콘코던스 (concordance)를 검색 결과로 제시한다. 이 밖에도 핵심어를 기준으로 한 용례 색인 (index)인 KWIC(Keyword in Context), 연도별 분포 그래프, 통계, 빈도, 그리고 엔그램 (N-gram)에 대한 정보를 추가로 제공한다. RNC에는 이러한 기능들을 가지면서도 단어 자체의 분석에 주안점을 둔 별도의 검색 도구가 존재한다. 러시아어로 “단어의 초상화

8) 필요에 따라 더하기 버튼을 눌러서 더 많은 단어 블록을 추가할 수 있다. 두 번째 단어부터 선행 단어와의 거리를 지정할 수 있다.

(Portret slova)”라는 이름을 갖지만, 여기에서는 영어 이름에 따라 “단어 한눈에 보기 (Word at a glance)”라고 칭하겠다. RNC의 기본 검색 도구와 달리, 단어 한눈에 보기는 검색어의 문법적·의미적 특성, 그와 유사한 단어들, 형태소 분석, 시기별 분포 등의 다양한 정보를 한 페이지에 모두 나타낸다. RNC의 홈페이지 화면에서 해당 도구의 배너를 클릭하면 레마⁹⁾를 입력할 수 있는 검색어 창과 단어의 품사(명사, 동사, 형용사, 부사)를 선택할 수 있는 설정 창이 나타난다. 도구의 기본 설정은 특정한 품사를 설정하지 않고 주요 말뭉치를 대상으로 어휘를 검색하는 것이다. 품사를 선택하지 않는 경우 말뭉치에서 5회 이상 나타나는 검색어의 품사가 모두 표시되는데, 이러한 설정은 러시아어에 동음이의어 혹은 동형이의어가 많기 때문이다. 단어 한눈에 보기 도구의 검색 결과는 다음의 8가지 위젯으로 이루어져 있다:

(1) 단어 스케치 (Word Sketches)

단어 스케치는 해당 단어가 주요한 통사적 관계(주어, 목적어, 술어 등)¹⁰⁾를 맺는 단어들과의 전형적인 결합성, 즉 연어관계를 보여준다. 이에 따라 사용자는 단어들이 상호작용하는 방식을 이해할 수 있다. 주요 통사적 관계마다 최대 10개의 공기어 또는 연어가 표시되는데, 그 순위는 logDice 지표로 결정된다. 연어 목록은 비어 있을 수 있으며, 철자가 표준적이지 않거나 말뭉치에서 거의 출현하지 않는 고유명사, 약어 등의 경우 스케치가 표시되지 않는다. 순위표 안의 단어를 클릭하면 말뭉치에서 그 단어가 어떤 문맥에서 어떻게 사용되었는지를 보여주는 콘코던스 창이 뜬다. 위젯 하단의 “모든 연어관계 표시 (Show all collocations)” 버튼을 클릭하면 RNC의 연어 검색 창으로 바로 이동하게 된다.

Определение	Связуемые	Глаголы с прямым дополнением
1. бродячий 10,32	1. лаять 9,4	1. кормить 8,47
2. бешеный 9,6	2. залаять 8,83	2. съест 8,31
3. охотничий 9,29	3. выть 7,69	3. выгуливать 8,22
4. бездомный 8,73	4. бежать 7	4. завести 7,67
5. борзый 8,72	5. бегать 6,71	5. вешать 7,65
6. цепной 8,59	6. укусить 6,68	6. натаскивать 7,49
7. легавый 8,47	7. броситься 6,64	7. спустить 7,49
8. гончий 8,45	8. завывать 6,44	8. дразнить 7,28
9. голодный 8,37	9. брехать 6,05	9. гладить 7,16
10. дворový 8,11	10. лежать 6,04	10. спускать 7,15

<그림 6> 러시아어 명사 sobaka ‘개’의 스케치

- 9) 사용자가 레마와 일치하지 않는 어형을 입력하는 경우, 검색 결과에는 그것의 문법적·의미적 특성 혹은 유사한 단어에 대한 정보가 표시되지 않는다.
- 10) 통사적 관계는 검색어의 품사에 따라 다르게 제시된다.

명사	형용사	동사	부사
명사 - 형용사 한정어 명사(주어) - 동사 명사(직접목적어) - 동사 명사(간접목적어) - 동사 전치사+명사(간접목적어) - 동사 복합 명사	형용사 - 명사 부사 수식어 복합 형용사	명사 - 주어 명사 - 직접목적어 명사 - 전치사 없는 간접목적어 명사 - 전치사 수반 간접목적어 복합 동사	부사 - 동사 부사 - 형용사 복합 부사

(2) 단어 정보 (About the word)

단어 정보 위젯은 단어의 문법적, 의미적 특성을 알려준다. 즉, 사용자는 4가지 품사(명사, 형용사, 동사, 부사)의 어휘에 한하여 가장 필수적인 정보를 얻을 수 있다. 예를 들어, 아래의 <그림 7>에 따르면 선생님을 의미하는 단어 učitel'은 문법적 속성으로 품사가 명사이고, 유정성을 띠며 문법적 성이 남성이다. 또한 의미적 측면에서 주어진 단어는 구상명사이자 행위자 의미를 지닌 동사파생명사이다.



<그림 7> 러시아어 명사 učitel' '선생님'에 대한 정보

동음이의어나 다의어의 경우 그것이 지닌 서로 다른 의미가 별도의 줄에 기술된다.

(3) 단어 빈도 (Word frequency)

단어 빈도 위젯은 6개의 범위로 이루어진 단어 빈도 척도를 표시한다. 빈도 정보는 단어 사용의 예가 있고 중의성이 제거된 말뭉치에서만 제공된다. 위젯에서 임의의 위치로 마우스를 이동하면 검색어의 빈도, 즉 ipm(instance per million) 값을 확인할 수 있다. 여기서 ipm은 모든 형태의 단어가 출현한 횟수를 해당 말뭉치의 크기로 나누고 100만을 곱한 값으로 정의된다. ipm 값에 따라 단어는 다음의 범위 중 하나에 속하게 된다:

<1	1-10	10-100	100-1000	1000-10000	>10000
빈도가 낮고, 단어가 매우 드물다	빈도가 꽤 낮고, 단어가 드물다	빈도가 다소 낮다	빈도가 다소 높다	빈도가 꽤 높고, 단어가 흔하다	빈도가 높고, 단어가 매우 흔하다

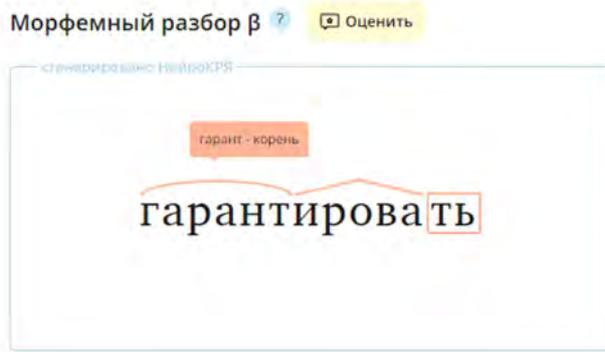


<그림 8> 러시아어 형용사 xorošij '좋은'의 빈도

위 <그림 8>에 따르면 형용사 xorošij는 ipm이 567로 빈도가 높은 편에 속한다.

(4) 형태소 분석 β (Morphemic structure β)

형태소 분석 베타 버전은 어휘의 형태소 구성을 보여준다. 접두사, 어근, 접미사 및 어미는 학교 러시아어 문법에서 채택된 기호를 통해 표시된다. 주요 말뭉치에 대한 형태소 분석은 해당 말뭉치를 위해 특별히 개발된 형태소 분석 사전을 토대로 이루어지는데, 사전에 없는 단어의 경우 RNC에서는 신경망 알고리즘에 기반하여 주석을 자동 생성한다. 특히 이러한 상황에 해당하는 것은 외국 어휘를 모방하여 만들어진 신조어들이다. 예컨대 영어의 동사 guarantee ‘보장하다’를 음차하여 같은 의미로 사용되는 러시아어 동사 garantirovat’는 형태소 분석 사전에 등재되어 있지 않다. 따라서 주어진 동사에 대해서는 <그림 9>에서처럼 “garant-irova-t’”라는 분석이 예측되고, 그 위에 “NeuroRNC에 의해 생성됨(generated by NeuroRNC)”이라는 표지가 추가된다.



<그림 9> 러시아어 동사 garantirovat’ ‘보장하다’에 대한 형태소 분석

교육용 말뭉치의 경우 학교 문법의 전통적인 형태소 분석법과 짜호노프(A. Tikhonov)의 <형태소-정자법 사전(Morpheme-Orthographic Dictionary)>(2002)에 기반한다. 그러므로 사전에 포함되지 않는 단어의 형태소 구조는 분석되지 않는다.

(5) 유사어 (Similar words)

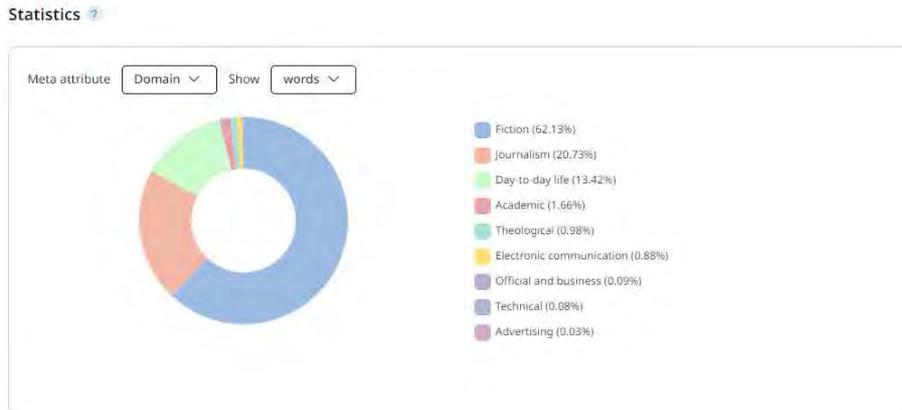
유사한 단어들 위젯은 검색어와 가장 가까운 의미를 갖는 단어들을 표시한다. 워드클라우드(Word Cloud)에 속한 단어 위에 마우스를 올리면 단어 근접 계수(proximity coefficient)를 볼 수 있다. 이는 RNC 텍스트의 실제 자료를 기반으로 한 분포 의미 모델(distributional semantics models)을 사용하여 계산된 것이다. 다시 말해, RNC에서는 특정 말뭉치의 텍스트에 대해 훈련된 워드투벡 모델을 사용하여 유사어를 검색한다. 계수 값이 1에 가까울수록 단어의 글씨 크기는 더 커지며, 해당 단어가 출현하는 맥락은 검색 키워드가 사용되는 맥락과 더 유사해야 한다. 현재 버전은 4가지 품사(명사, 형용사, 동사, 부사)의 어휘에 한해서만 동일한 품사의 유사어들을 출력한다. 고유명사, 지명, 약어, 그리고 철자가 표준적이지 않거나 말뭉치에서 거의 발견되지 않는 단어의 경우, 그와 유사한 단어들이 표시되지 않는다. 유사어 선정이 자동으로 수행될 경우 형태소 분석 위젯에서와 동일하게 “NeuroRNC에 의해 생성됨”이라는 표지가 붙는다. 다음의 <그림 10>은 ‘가다’를 의미하는 가장 기본적인 러시아어 이동동사 idti의 유사어들이 워드클라우드 형식으로 모인 것이며, 여기에는 다양한 양상의 이동을 가리키는 동사들이 포함된다.



<그림 10> 러시아어 동사 idi '가다'의 유사어

(6) 통계 (Statistics)

통계 위젯의 파이 차트는 검색어가 어떤 유형의 말뭉치 텍스트에서 나타나는지 보여준다. 차트를 만들기 위해 사용자는 말뭉치의 메타속성(meta attribute)과 크기 측정 단위를 선택할 수 있다. 메타속성(텍스트 유형, 분야, 주제, 문체, 성별) 또는 측정 단위(텍스트 또는 단어)를 바꾸면 차트가 다시 그려진다. 차트는 상위 10개의 메타속성 값의 분포를 보여주는데, 그 나머지 값은 기타 범주로 병합된다. 차트의 우측에는 값의 목록과 백분율이 제시되고, 차트의 섹터 위로 커서를 이동하면 값의 이름과 검색어를 포함한 단어 또는 텍스트의 수를 볼 수 있다.



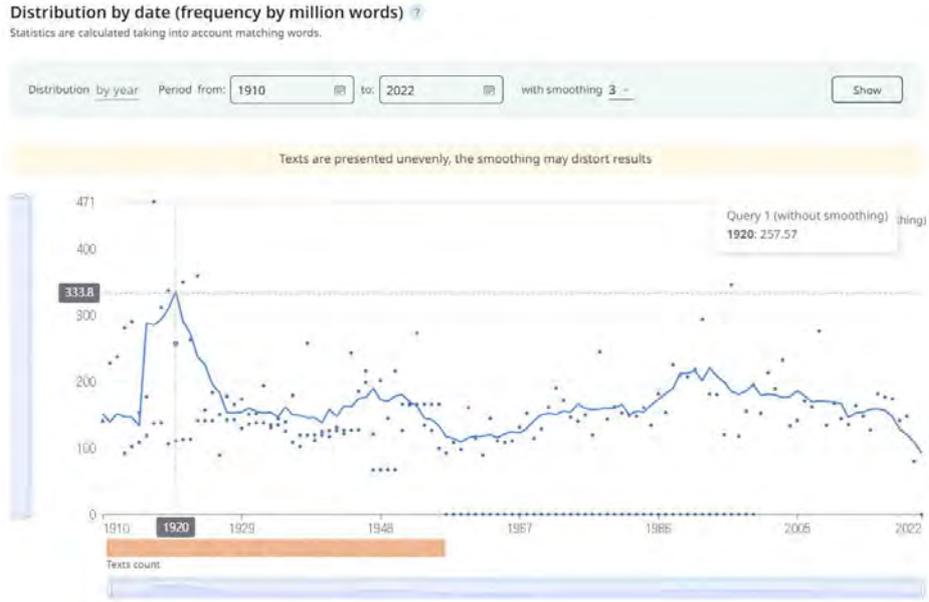
<그림 11> 러시아어 지소형 부사 nemnožko '조금'의 통계 정보

<그림 11>에 의하면 러시아어 지소형 부사 nemnožko는 소설과 저널리즘 그리고 일상 회화에서 많이 출현함을 알 수 있다. 실제로 지소형 부사는 매우 회화적이고 구어적인 어휘 단위이다.

(7) 시기별 분포 (Distribution by date (frequency by million words))

해당 위젯의 그래프는 주어진 하위 말뭉치 내 검색어 용례의 분포, 즉 발생 빈도를 연도 별로 보여준다. 기본 설정값으로 검색 결과에 해당하는 연도 범위가 '1779년부터 2022년'으로, 평활화(smoothing)가 '3'으로 선택되나, 원하는 결과에 맞게 설정을 변경할 수 있다. 특히 각 연도에 대한 정확한 데이터를 얻을 목적이라면 평활화를 '0'으로 설정하면 된다. 게다가 그래프 줄에 마우스를 올리면, 그 지점에 해당하는 특정 연도의 상대적 사용 빈도를 볼 수 있다. 일부 말뭉치에 한해서는 그래프 아래에 가열화 줄무늬(warming

stripes)가 표시되는데, 이는 주어진 하위 말뭉치에서 용례가 발견된 텍스트의 크기를 암시한다. 아래 <그림 12>에서는 ‘자유’를 의미하는 명사 svoboda의 시기별 사용 빈도수가 관찰된다. 사회주의 혁명이 발생했던 1920년을 전후하여 해당 어휘의 사용 빈도가 가파르게 증가했고, 소비에트의 해체와 페레스트로이카(perestroika)가 진행되었던 1990년대 초에 단어의 사용이 또다시 증가했음을 알 수 있다.



<그림 12> 러시아어 명사 svoboda ‘자유’의 시기별 분포

(8) 용례 (Examples)

용례 위젯에서는 말뭉치 텍스트에서 추출한 5개의 용례가 제공된다. 검색어가 포함된 다른 예문들을 보기 위해 하단의 “더 많은 용례 표시>Show more examples)” 버튼을 클릭하면, 전체 검색 결과가 나오는 콘코던스 창으로 이동하게 된다.

Examples ?

Сами представьте, **смотрет** на вас большая-большая громадина в очках, да еще в довершение ужаса двумя пальцами делает вот так: «Тип-тип, мышка!» Как щипцами.

«Даже **смотрет**! глупо», — говорил в таких случаях Салтыков-Щедрин.

Я радовался, что никто меня не останавливает, ни о чём не спрашивает, не мешает **смотрет** на эту красоту.

Но делом и словом, подчас откровенным, не предназначавшимся для чужого слуха, он в достаточной степени определил свой взгляд на предстоящую ему роль не претендуя на политическую непогрешимость, он **смотрет** на себя, как на могучий таран, который должен был пробить брешь в заколдованном круге сил, облечивших власть, обезличивших и обескровивших ее.

Подвешенные на громадных цепях к закопченному стеклянному потолку тысячелетовые бронированные вагоны, похожие на танки, **смотрет** в разные стороны открытыми черными люками, из которых, как из открытых ртов повешенных, торчали набрякшие языки пулеметов.

Show all examples

<그림 13> 러시아어 동사 smotret ‘보다’의 용례

위와 같이 제시되는 단어의 8가지 정보를 바탕으로 사용자 혹은 학습자는 러시아어 어휘에 대한 언어학적 지식을 수월하게 습득할 수 있다. 물론 RNC의 주요 말뭉치나 기타 하위 말뭉치를 대상으로 하는 단어 및 구(phrase) 검색에는 더욱 복잡한 조건을 설정할 수 있겠으나, 단어 한눈에 보기 도구는 가장 효율적인 검색을 가능하게 한다. 즉, 사

용자가 단어를 입력하고 품사를 선택하기만 하면 검색어에 대한 자세한 분석을 얻을 수 있으나, 그에 들인 노력 대비 얻는 결과가 큰 것이다. 우선, 제일 상단의 위젯인 단어 스케치는 말뭉치 텍스트에서 검색어의 공기(co-occurrence) 패턴을 포착해주므로 사용자가 특정 단어와 함께 쓰이는 공기어 또는 연어를 바탕으로 그 의미와 용법을 정의하는 데 도움이 된다. 그 아래에 위치한 단어 정보 위젯은 사용자가 굳이 별도의 사전을 참고하지 않아도 필수로 알아야 하는 단어의 기본형, 문법적 특성, 그리고 의미 부류를 알려준다. 특히 러시아어 명사의 성(gender), 동사의 상(aspect) 등과 같은 문법 개념에 아직 익숙하지 않은 초급 학습자에게는 RNC의 일반적인 말뭉치 검색 기능보다도 단어 한눈에 보기 도구가 더 유용할 것이다. 같은 맥락에서, 일반적인 콘코던서(concordancer)의 경우 검색된 수많은 용례를 모두 표시하기 때문에 학습자 관점에서 학습 피로도가 크기 마련이다. 반면 단어 한눈에 보기 도구에서는 검색어가 포함된 예문이 5개만 제시되므로 학습자가 단어가 사용된 맥락을 살펴보고 각 문맥 내 언어 단위의 의미를 추론하기 편리할 수 있다. 좀 더 높은 수준의 학습자 또는 세부적인 연구 과제를 수행하는 언어학자의 경우 단어 빈도, 통계, 시기별 분포와 같은 위젯들의 정보를 참고하여 러시아어 사용 양상에 대한 통찰을 얻을 수 있겠다. 이러한 교육학적 이점을 고려할 때, RNC의 단어 한눈에 보기 도구는 러시아어 학습자의 자료기반 학습(data-driven learning) 및 연구에 크게 기여한다고 말할 수 있다.

4. RNC를 이용한 러시아어 연구의 동향

말뭉치를 이용한 연구의 결과는 이미 외국어 교육, 사전 편찬, 언어 정책 등 언어학의 응용 분야 전반에 활발하게 활용되고 있다. RNC를 활용한 국내외 연구 또한 연구 대상 및 목적이 굉장히 다양하다. 지면 관계상 일부만 소개하도록 하겠다.

먼저 국내 연구로 김성완(2020)은 한국어의 시간이동동사를 러시아어로 어떻게 번역할 것인지 그 방안을 마련하기 위해 RNC에서 예문을 발췌한다. 노희승(2021)은 RNC의 구어 말뭉치에서 남녀 화자 간의 지소사 사용 빈도에 대한 자료를 언어 통계분석을 시행하며, 이주홍(2022)은 RNC에서 삼중쌍 동사 15개의 각 형태에 따른 시제별 빈도수를 비교하여 그룹을 나누고 각 그룹에 속하는 동사들의 문맥 의미를 조사한다. 권용재(2022)에서는 RNC에 나타난 시사 텍스트를 중심으로 현대러시아어 명사의 의미 변화 과정을 추적하여 그 의미 파생이 상황의 문맥적 실현으로 유발됨을 밝히고, 이수현(2023)에서는 RNC의 용례를 통해 동사 *dat'* '주다'의 확장된 의미 사용과 이에 따른 경동사 구문의 변별적 의미 기능을 인지언어학적 관점에서 고찰한다. 마지막으로 남혜현(2023)은 주요 말뭉치를 중심으로 RNC의 구성과 교육적 활용방안을 논한다.

해외 연구로는 우선 Zdorenko(2010)의 경우 RNC 구어와 문어 말뭉치를 바탕으로 다양한 장르와 형태·통사적 환경(가령 3인칭 vs. 1·2인칭 문맥)에서의 주어 생략을 정량적으로 비교한다. Belyakov & Maksimenko(2024)에서는 RNC의 2000-2019년 자료를 바탕으로 러시아어 뉴스미디어 담론에 쓰인 '패러다임(paradigma)', '내러티브(narrativ)' 등의 다빈도 단어에서 관찰되는 의미 변화의 문제를 고찰한다. 그리고 Logvinova(2024)는 러시아어 동격 구문(예컨대 *v rek-eLOC Don-eLOC vs. Don-NOM 'in the river Don'*)에서 고유 명사의 격 일치(case concord) 여부를 예측하기 위해 로지스틱 회귀 통계 모델을 사용하

여 15,000개 이상의 RNC 용례를 분석한다. 국내에 비해 해외에서는 RNC의 활용 가치에 관한 논의가 꾸준히 이루어져 왔다. 그중에서도 Longo(2023)는 인터넷과 소셜 미디어에서 사용되는 현대러시아어 ‘넷스피크(netspeak)’의 특징을 조사하며, 진화하는 언어를 통시적으로 분석하는 데에 RNC가 여타 도구에 비해 훨씬 더 적합하다고 주장한다. Ponomareva et al.(2016)는 언어교수학적 관점에서 RNC의 사용을 강조하며 도스토예프스키의 <백야(Belye nochi)> 텍스트를 분석하는 대학교 수업의 설계 및 활동 과제를 제시한다.

이외에도 RNC를 이용한 언어학적 연구는 지면에 다 언급하지 못할 만큼 많지만, 연구의 방법론적인 측면에서 볼 때 어휘 사용 및 의미 빈도를 말뭉치 용례를 통해 분석하는 연구가 큰 비율을 차지한다. 어휘 의미 연구의 경우 기존의 어휘의미론의 개념을 주요 말뭉치를 통해 확인하는 연구가 큰 비중을 차지한다. 또한 문형이나 구문에 관한 연구에서는 주로 용례 추출을 위해서만 말뭉치를 부분적으로 사용한다. 이성민(2019)은 최근 말뭉치를 사용한 언어 연구는 활성화되어가고 있지만, 대부분 영어 교육 등에 한정되어 있을 뿐 러시아어 말뭉치의 경우는 주로 연구자들에게 1차 자료가 되는 예문을 제공해 주는 데 그치고 있다고 지적한다. 그의 말대로 다량의 실제 언어 자료를 전산화한 형태의 말뭉치는 언어 이론의 추상성을 보완하는 자료 공급원에서 더 나아가 효율적인 연구 및 교육 방법론으로 평가받을 수 있다. 다시 말해, 말뭉치를 이용한 연구가 보다 다채로워지기 위해서는 Tognini-Bonelli(2001)의 용어를 빌려 ‘말뭉치 기반 접근법(corpus-based approach)’에서 탈피하여 ‘말뭉치 주도 접근법(corpus-driven approach)’을 적극적으로 시도할 필요가 있겠다. 전자가 어떤 언어 사실을 밝히기 위해 말뭉치를 도구적으로 사용하는 것이라면, 후자는 연구자의 선입견 없이 말뭉치가 보여주는 바를 바탕으로 언어 현상을 기술하는 것이다. 소위 신퍼스주의(neo-Firthian) 학자들은 말뭉치 언어학이 언어학의 하위 연구 분야이며 말뭉치 주도 접근법을 통해 언어 자료에 대한 새로운 시각을 가질 수 있어야 한다고 강조한다. 이와 유사한 맥락에서 RNC를 활용한 해외 연구 사례로 Podlesskaya(2023)를 들 수 있다. Podlesskaya(2023)는 RNC의 멀티미디어 말뭉치 자료에 근거하여 접속사 *pričëm* ‘게다가’에 의해 도입되는 담화 단편의 운율적 특징을 서술한다. 멀티미디어 말뭉치는 영화, 연설 등 다양한 장르의 구어 발화에 대한 연구를 위해 구축된 것이다. 말뭉치의 모든 텍스트는 10-30초 길이의 작은 클립으로 분할된 영상과 음성 녹음 파일로 제시되고, 텍스트 조각마다 메타텍스트, 형태, 의미, 강세 등의 주석이 붙는다. Podlesskaya(2023)에서는 이러한 주석을 효율적으로 사용하여 접속사를 포함한 발화의 단편이 의사소통-운율적(communicative-prosodic) 자율성을 갖는다는 사실을 입증한다.

말뭉치는 텍스트의 유형, 분야, 주제 등과 같은 메타데이터를 포함한다. 이를 토대로 도출할 수 있는 언어학적 지식에 대해 고민해볼 수 있다. 특정한 분야 또는 영역에서 사용되는 언어는 일상생활에서 사용되는 언어와 구분되기 때문이다. 전문 분야의 특성을 띤 언어 사용역(register)을 연구하고 기술하기 위해서는 일반적인 언어 교육과는 구분되는 특화된 방법론이 필요하다. 이때 말뭉치 언어학이 그 대안이 될 수 있다. 말뭉치 주도 연구를 수행하기 위해서는 우선 그 목적에 맞는 코퍼스의 구축 작업이 선행되어야 한다. 따라서 가장 일반적이고 대표적인 주요 말뭉치가 아닌, RNC의 기능을 활용하여 특정한 분야의 하위 말뭉치를 설정하여 활용한다면, 해당 사용역에서 관찰되는 러시아어의 어휘적 또는 문체적 특징을 다른 영역의 것과 비교할 수 있을 것이다.

5. 맺음말

디지털 환경과 데이터 처리 기술을 접목한 인문학 교육에 관한 논의가 그 어느 때보다 활발하다. 한국어 및 기타 외국어 교육 분야에서 말뭉치 도구 및 방법론이 적극적으로 활용되고 있듯이 국내외의 러시아어 교수자들 또한 연구 및 교육의 목적으로 대규모 RNC 말뭉치의 텍스트를 활용하거나 RNC에서 직접 하위 말뭉치를 구축하여 사용한다. 그러나 국내 연구 현황을 감안할 때 대부분 현대러시아어의 사용 양상에 중점을 두고 있어, 연속 업데이트되는 RNC의 최신 자료를 이용한 통시적 연구가 부진한 것으로 보인다. 고전 문헌에는 언어의 역사와 가치를 확인할 수 있는 요소들이 포함되어 있다. 따라서 본고는 인문학의 위기를 타개하는 동시에 언어 정보화 발전에 기여할 수 있는 노력으로 연구 대상으로 삼는 언어 자료의 범위를 확대할 것을 제안하고자 한다. 즉, 현대 자료와 역사적 자료를 이분하지 않고 RNC에 말뭉치화된 장르별, 시기별 러시아어 자료를 최대한 활용하는 것이다. 언어 변화나 어원에 대한 정보는 신조어와 방언과 더불어 러시아 문화와 언어 맥락에 대한 이해력을 향상할 수 있기 때문이다.

앞으로도 RNC는 국가 기관의 주도하에 러시아 언어문화의 다양성과 고유성을 담은 자료들을 통합하여 그 규모를 확장할 것이다. 이렇게 디지털 전환된 데이터는 사실 학계와 산업계에서 모두 활용될 수 있다. 따라서 RNC는 교육용 자료나 언어학적 연구 과제를 넘어서 인공지능의 러시아어 능력 향상 및 평가와 같은 목적으로 활용될 가능성이 무궁무진하다. 이러한 점에서 생성형 AI의 학습 데이터셋으로 사용될 수 있는 고품질의 말뭉치 및 지식 데이터베이스가 계속 구축될 것으로 예상된다. 현재 보완 작업이 더딘 RNC 병렬 말뭉치의 경우 외국어의 수에 비해 아직 데이터의 품질이 좋지 않아 번역 기술 개발을 위한 기계학습에 활용되기에는 무리가 있어 보인다. 소통의 확장을 위해 언어간 병렬말뭉치는 필히 요구되기 때문에, 균형성과 대표성을 반영한 자료를 수집하고 정제하여 번역한 후 외국어 전문가의 검수를 몇 차례 받아야 하겠다. 또한 RNC의 구축 배경에 러시아어 확산 및 교육을 빼놓을 수 없으므로 생략과 대명사의 사용이 빈번한 러시아어 특성을 각각 반영한 생략 복원 말뭉치와 상호 참조 말뭉치, 혹은 다양한 텍스트 유형에 대한 수준별 (가령 초/중/고등) 쓰기 말뭉치가 구축되길 기대해본다.

참고문헌

- 권용재 (2022). 현대 러시아어 명사의 의미 파생 연구 - 러시아어 코퍼스(НКРЯ)에 나타난 시사 텍스트를 중심으로 -. 슬라브研究, 38(3), 325-348.
- 김성완 (2020). 한국어 시간이동동사의 러시아어 번역방안에 관한 연구. 동서인문학, 59, 39-67.
- 남혜현 (2011). 탈소비에트 시기 러시아 정부의 러시아어 확산 정책. 슬라브 연구, 27(2), 161-194.
- 남혜현 (2023). 러시아국립코퍼스의 구성과 교육적 활용 방안에 대한 소고: 주요 코퍼스(основной корпус)를 중심으로. 동유럽발칸연구, 47(1), 3-38.
- 노희승 (2021). 통계분석을 활용한 러시아어 화자의 성별에 따른 지소사 사용 빈도 연구 - 지소형 명사, 형용사 · 부사를 중심으로 -. 노어노문학, 33(1), 29-52.

- 이주홍 (2022). 러시아어 이차불완료상화 연구: 시제형태의 빈도수를 중심으로. *러시아연구*, 32(1), 207-243.
- 이성민 (2019). 코퍼스 기반 특수 목적 러시아어 연구: 의료 러시아어의 어휘·문체적 특성. *슬라브학보*, 34(1), 71-98.
- 이수현 (2023). 경동사 구문 의미와 은유적 의미전이 - 러시아어 동사 ДАТЬ를 중심으로 -. *슬라브研究*, 39(1), 255-284.
- Belyakov, M.V., & Maksimenko, O.I. (2024). The Problem of Semantic Shift in Modern Russian: Corpus Research on the Example of a News Media Discourse. *RUDN Journal Of Language Studies, Semiotics And Semantics*, 15(1), 174-194.
- Kol'cova, L.M. (2008). O REALIZACII FEDERAL'NOJ CELEVOJ PROGRAMMY «RUSSKIJ JAZYK» V VORONEŽSKOM GOSUDARSTVENNOM UNIVERSITETE. [ON THE IMPLEMENTATION OF THE FEDERAL TARGET PROGRAM “RUSSIAN LANGUAGE” AT VORONEZH STATE UNIVERSITY]. *Vestnik VGU. Serija: Problemy vysšego obrazovanija*. No. 1, 52-58.
- Logvinova, N. (2024). Concord in Russian close appositional constructions: a quantitative study. *Russian Linguistics*, 48(1), 4.
- Longo, F. (2023). A corpus-based analysis of contemporary Russian netspeak: can corpora help us predict the future of. *Proceedings of the International Summer School of*, 93-130.
- Podlesskaya, V.I. (2023). Prosodic portrait of the Russian connector PRICHOM in the mirror of the multimedia corpus. *Komp'juternaja lingvistika i intellektual'nye texnologii: po materialam meždunarodnoj konferencii «Dialog 2023»*, 1-10.
- Ponomareva, L.D., Churilina, L.N., Buzhinskaya, D.S., Derevskova, E.N., Dorfman, O.V., & Sokolova, E.P. (2016). Russian National Corpus as a Tool of Linguo-Didactic Innovation in Teaching Languages. *International Journal of Environmental and Science Education*, 11(18), 13043-13053.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing.
- Zdorenko, T. (2010). “Subject omission in Russian: A study of the Russian National Corpus.” In *Corpus-linguistic applications. Series: Language and Computers*, Volume 71. Brill. 119-133.

국립러시아말뭉치 <https://ruscorpora.ru/>

연방정부 프로그램 <러시아어>

2002-2005년 <http://elementy.ru/Library9/Progr483.htm?context=28809>

2006-2010년 http://www.programs-gov.ru/11_1.php

2011-2015년 http://www.programs-gov.ru/rusyazyk11_15/7-o-federalnoy-celevoy-programme-russkiy-yazyk-na-2011-2015-gody.html

2016-2020년 <http://government.ru/docs/18169/>

2024년 한국코퍼스언어학회 여름 전국학술대회
인공지능 시대의 말뭉치 구축과 활용



Session B



Transformers and polysemy of Korean adverbial postposition *-(u)lo*



중국어 의미 주석 말뭉치의 구축과 활용



SacreBleu를 이용한 한중 기계번역 평가



From an interrogative adverb to a discourse maker *-way* “why” in Korean

Transformers and polysemy of Korean adverbial postposition *-(u)lo*

Seongmin Mun

Ajou University

17th July 2023



아주대학교
AJOU UNIVERSITY

Outline

Introduction

- Polysemy in Korean

Corpus

- Sejong corpus

- Creation of a hand-coded corpus

Classification models: BERT & GPT-2

Visualization

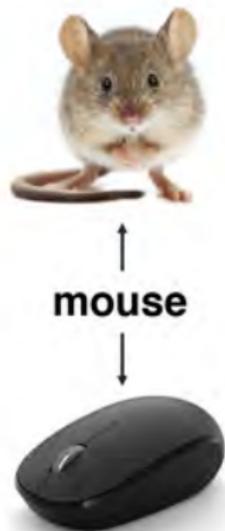
Discussion & Conclusion



Introduction

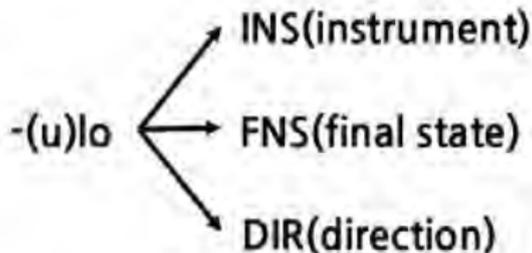
Polysemy

Polysemy, one type of ambiguity, occurs when one form delivers multiple meanings/functions (Glynn and Robinson, 2014).



Korean language

Korean is a Subject-Object-Verb language, which marks grammatical information with dedicated postpositions (Sohn, 1999).



Polysemy in Korean adverbial postposition

-(u)lo as INS (instrument)

na-nun kamca-lul khal-lo ssel-ess-ta.

I-TOP potato-ACC knife-INS cut-PST-DECL

'I cut a potato with a knife.'

Figure: An example sentence involving the postposition *-(u)lo* as a function of INS (instrument)

Question: How a speaker can understand the function of postposition?

Previous studies on adverbial postpositions

Study	Corpus type	Data size	Method	Accuracy
Bae et al. (2020)	Korean PropBank	20,035 sentences	BERT + BiLSTM-CRFs + Structural SVM	0.84
Park et al. (2019)	Korean PropBank	23,059 sentences	BERT + BiLSTM-CRF	0.84
Lee et al. (2015)	Korean PropBank	4,882 sentences	Word2vec (SGNS) + Structural SVM (Support Vector Machine)	0.77
Mun & Shin (2020)	Sejong corpus	2,100 sentences	PPMI & SVD + Similarity-based estimate	0.74
Park & Cha (2017)	Sejong corpus	14,335 sentences	Word2vec (SGNS) + CRF	0.77
Hong et al. (2019)	Korean PropBank	23,059 sentences	RoBERTa + BiLSTM	0.85
Yoon et al. (2016)	Korean PropBank	4,714 sentences	One-hot encoding + Bidirectional LSTM-CRFs	0.66

Previous studies on adverbial postpositions

Study	Corpus type	Data size	Method	Accuracy
Bae et al. (2020)	Korean PropBank	20,035 sentences	BERT + BiLSTM-CRFs + Structural SVM	0.84
Park et al. (2019)	Korean PropBank	23,059 sentences	BERT + BiLSTM-CRF	0.84
Lee et al. (2015)	Korean PropBank	4,882 sentences	Word2vec (SGNS) + Structural SVM (Support Vector Machine)	0.77
Mun & Shin (2020)	Sejong corpus	2,100 sentences	PPMI & SVD + Similarity-based estimate	0.74
Park & Cha (2017)	Sejong corpus	14,335 sentences	Word2vec (SGNS) + CRF	0.77
Hong et al. (2019)	Korean PropBank	23,059 sentences	RoBERTa + BiLSTM	0.85
Yoon et al. (2016)	Korean PropBank	4,714 sentences	One-hot encoding + Bidirectional LSTM-CRFs	0.66

Transformer-architecture models that we used

- ▶ *Contextualized* word embedding model
 - ▶ Bidirectional Encoder Representations from Transformer (BERT; Devlin et al., 2018)
 - ▶ Generative Pre-Training 2 (GPT-2; Radford et al., 2019)

Corpus

What is Sejong corpus?

- ▶ Sejong corpus was created by the 21st Century Sejong Project, a ten-year-long project that was launched in 1998.
- ▶ Sejong corpus is a representative large-scale corpus in Korean (Shin, 2008).
- ▶ Previous studies often used this corpus as a linguistic resource (e.g., Kim & Ock, 2016; Park & Cha, 2017; Shin et al., 2005).

What is Sejong corpus?

Table 1: *Primary corpus*

Corpus type	Corpus size(eojul)
Raw corpus	63,899,412
Grammatically tagged corpus	15,226,186
Parsed corpus	570,064
Semantically Tagged corpus	10,132,348
Sum	89,830,015

Table 2: *Plan for construction of raw corpus*

Field	Portion
Newspaper	20%
Magazine	10%
Academic works	35%
Literary works	20%
Quasi-spoken data	10%
The others	5%
Sum	100%

The eojul is defined as a morpheme or combination of several morphemes serving as the minimal unit of sentential components in Korean.

What is the Sejong corpus?

Table 1: *Primary corpus*

Corpus type	Corpus size(eojul)
Raw corpus	63,899,412
Grammatically tagged corpus	15,226,186
Parsed corpus	570,064
Semantically Tagged corpus	10,132,348
Sum	89,830,015

Table 2: *Plan for construction of raw corpus*

Field	Portion
Newspaper	20%
Magazine	10%
Academic works	35%
Literary works	20%
Quasi-spoken data	10%
The others	5%
Sum	100%

The eojul is defined as a morpheme or combination of several morphemes serving as the minimal unit of sentential components in Korean.

Example of the semantically tagged corpus

BSAA0001-00001596	생산자의	생산자/NNG + 의/JKG
BSAA0001-00001597	얼굴	얼굴/NNG
BSAA0001-00001598	사진이	사진__07/NNG + 이/JKS
BSAA0001-00001599	붙어	붙/VV + 어/EC
BSAA0001-00001600	있는	있/VX + 는/ETM
BSAA0001-00001601	농산물이	농산물/NNG + 이/JKS
BSAA0001-00001602	나오고	나오/VV + 고/EC
BSAA0001-00001603	있다.	있/VX + 다/EF + ./SF

Example of the semantically tagged corpus

BSAA0001-00001596	생산자의	생산자/NNG + 의/JKG
BSAA0001-00001597	얼굴	얼굴/NNG
BSAA0001-00001598	사진이	사진__07/NNG + 이/JKS
BSAA0001-00001599	붙어	붙/VV + 어/EC
BSAA0001-00001600	있는	있/VX + 는/ETM
BSAA0001-00001601	농산물이	농산물/NNG + 이/JKS
BSAA0001-00001602	나오고	나오/VV + 고/EC
BSAA0001-00001603	있다.	있/VX + 다/EF + .SF

Description for annotation

- ▶ Annotators: three native speakers of Korean.
- ▶ Data: 5,000 sentences
- ▶ Functions: select the most frequent functions based on the Sejong Electronic Dictionary and the previous studies on adverbial postpositions.
 - ▶ $-(u)lo$: Final state, Instrument, Direction, Effector, Criterion, Location
- ▶ Fleiss's Kappa: 0.947

A hand-coded corpus

```
Index ### Label ### Function ### Sentence_POS ### Sentence
1 ### 0 ### FNS ### 이_05/MM 넥타이/NNG 는/JX 수제품/NNG (으)로/JKB 우리나라/NNG 에서
2 ### 2 ### DIR ### 나/NP 의/JKG 마음__01/NNG 의/JKG 움직임/NNG 이/JKS 위__01/NNG
3 ### 1 ### INS ### 곳/NNG 무당__01/NNG 이/JKS 노래/NNG 나/JC 춤__01/NNG (으)로/JK
4 ### 0 ### FNS ### 모든/MM 주장__03/NNG 이/JKS 나름/NNB 대로/JKB 의/JKG 근거/NNG 를
5 ### 3 ### EFF ### 기억/NNG 이/JKS 스스로/NNG 의/JKG 부력__01/NNG (으)로/JKB 떠오르,
6 ### 2 ### DIR ### 신축__03/NNG 전원주택/NNG 위쪽/NNG (으)로/JKB 는/JX 집__01/NNG
7 ### 0 ### FNS ### 명명/XR 하/XSA L/ETM 채__09/NNB (으)로/JKB 시간__04/NNG 이/JK
8 ### 1 ### INS ### 수한/NNP 이/JKS 저/NP 의/JKG 손__01/NNG (으)로/JKB 저/NP 의/JK
9 ### 2 ### DIR ### 쇠전__01/NNG 곧/XSN 들/XSN 이/JKS 술침/NNG (으)로/JKB 돌아오/VV
10 ### 3 ### EFF ### 그리고/MAJ 그/MM 결과__02/NNG (으)로/JKB 오즘/NNG 이/JKS 나오/V
11 ### 5 ### LOC ### "/SS 집__01/NNG 들/XSN 이/JKS 다/MAG 어디/NP (으)로/JKB 가/V
17 ### 5 ### INC ### 바로/MAG 앞/NNG (으)로/JKB 수당구지/NNG 바꿔 01/NNG 자군 01/
```

Available at: <https://github.com/seongmin-mun/Corpora/tree/main/APIK>

Classification models: BERT & GPT-2

Creating training and test sets

Index	Label	Sentence	Index	Label	Sentence
1,862	1	[CLS] 한침 만에 오반장이 침묵을 켜다 [SEP]	1,862	1	한침 만에 오반장이 침묵을 켜다.
1,863	1	[CLS] 정말 오랫동안 먹어보는 고기였다. [SEP]	1,863	1	정말 오랫동안 먹어보는 고기였다.
1,864	1	[CLS] 옛날 구한말에 유명한 예기가 있었죠? [SEP]	1,864	1	옛날 구한말에 유명한 예기가 있었죠?
1,865	1	[CLS] 한밤중에 신나게 한바탕했지요. [SEP]	1,865	1	한밤중에 신나게 한바탕했지요.
1,866	1	[CLS] 그런데 몇 시에 왔어? [SEP]	1,866	1	그런데 몇 시에 왔어?
1,867	1	[CLS] 겨울에 꽃이라니요 [SEP]	1,867	1	겨울에 꽃이라니요.
1,868	1	[CLS] 아침에 얼마한테 돈을 달렸어요. [SEP]	1,868	1	아침에 얼마한테 돈을 달렸어요.
1,869	1	[CLS] 결혼은 반드시 적령기에 해야 한다. [SEP]	1,869	1	결혼은 반드시 적령기에 해야 한다.
1,870	1	[CLS] 한 달에 얼마씩은 정확하게 들어오니까. [SEP]	1,870	1	한 달에 얼마씩은 정확하게 들어오니까.
1,871	1	[CLS] 그럴 일 주일 후에 뵈겠습니다. [SEP]	1,871	1	그럼 일 주일 후에 뵈겠습니다.

Figure: Example sentences used in the training for BERT (left) and GPT-2 (Right)

Creating training and test sets

Index	Label	Sentence	Index	Label	Sentence
1,862	1	[CLS] 한참 만에 오반장이 침묵을 했다. [SEP]	1,862	1	한참 만에 오반장이 침묵을 했다.
1,863	1	[CLS] 정말 오랫동안 먹어보는 고기였다. [SEP]	1,863	1	정말 오랫동안 먹어보는 고기였다.
1,864	1	[CLS] 옛날 구한말에 유명한 예기가 있었죠? [SEP]	1,864	1	옛날 구한말에 유명한 예기가 있었죠?
1,865	1	[CLS] 한밤중에 신나게 한바탕했지요. [SEP]	1,865	1	한밤중에 신나게 한바탕했지요.
1,866	1	[CLS] 그런데 몇 시에 왔어? [SEP]	1,866	1	그런데 몇 시에 왔어?
1,867	1	[CLS] 거울에 꽃이라니요. [SEP]	1,867	1	거울에 꽃이라니요.
1,868	1	[CLS] 아침에 엄마한테 돈을 달랬어요. [SEP]	1,868	1	아침에 엄마한테 돈을 달랬어요.
1,869	1	[CLS] 결혼은 반드시 적령기에 해야 한다. [SEP]	1,869	1	결혼은 반드시 적령기에 해야 한다.
1,870	1	[CLS] 한 달에 얼마씩은 정확하게 들어오니까. [SEP]	1,870	1	한 달에 얼마씩은 정확하게 들어오니까.
1,871	1	[CLS] 그럼 일 주일 후에 뵙겠습니다. [SEP]	1,871	1	그럼 일 주일 후에 뵙겠습니다.

Figure: Example sentences used in the training for BERT (left) and GPT-2 (Right)

Model specification: BERT

- ▶ Bidirectional Encoder Representations from Transformer (BERT; Devlin et al., 2018)
 - ▶ Package used: *Transformer*
 - ▶ Pre-trained model: KoBERT (Jeon et al., 2019)
 - ▶ Tokenizer: KoBERT tokenizer (Jeon et al., 2019)
 - ▶ Epoch: from one to 50
 - ▶ Other parameters: Learning rate (.00002); Batch (16); Sequence length (128); Seed (42); Epsilon (.00000001)
 - ▶ Dimension reduction: *t*-SNE (Maaten and Hinton, 2008)

Model specification: GPT-2

- ▶ Generative Pre-Training 2 (GPT-2; Radford et al., 2019)
 - ▶ Package used: *Transformer*
 - ▶ Pre-trained model: KoGPT2 (Jeon et al., 2021)
 - ▶ Tokenizer: GPT2 tokenizer (Jeon et al., 2019)
 - ▶ Epoch: from one to 50
 - ▶ Other parameters: Learning rate (.00002); Batch (16); Sequence length (128); Seed (42); Epsilon (.00000001)
 - ▶ Dimension reduction: *t*-SNE (Maaten and Hinton, 2008)

Model performance: Classification

- ▶ BERT performed better than GPT-2 in revealing the polysemy of Korean postpositions.
 - ▶ BERT: 0.795
 - ▶ GPT-2: 0.676
- ▶ The model performance increased as the epoch progressed.

Visualization

Visualization: PostTransformers



Available at: <https://seongmin-mun.github.io/Visualization/2022/PostTransformers/index.html>

Visualization: clusters of GPT-2

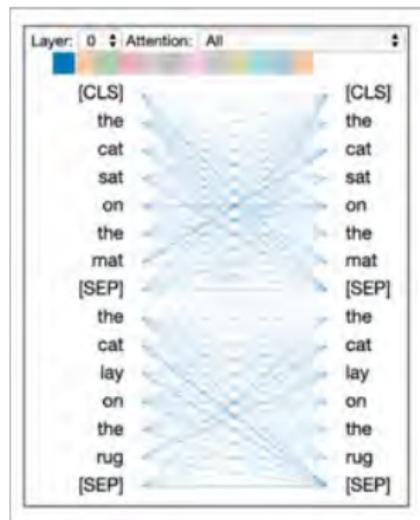
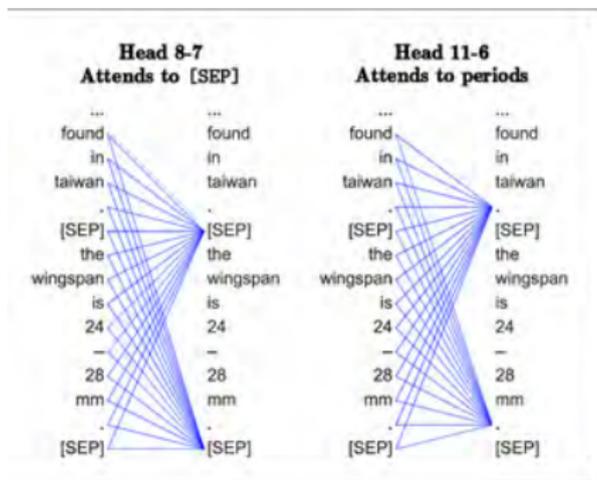
-(u)lo (Epoch 50)



Visualization: clusters of BERT



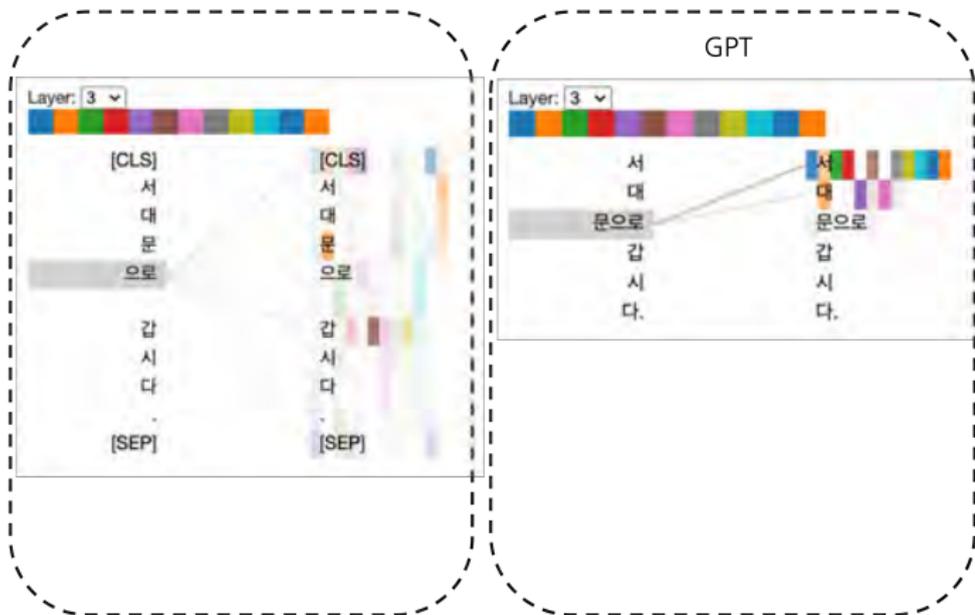
Visualization: AttentionVis



Adapted from: left: Clark et al., 2019 & right: Vig and Belinkov, 2019

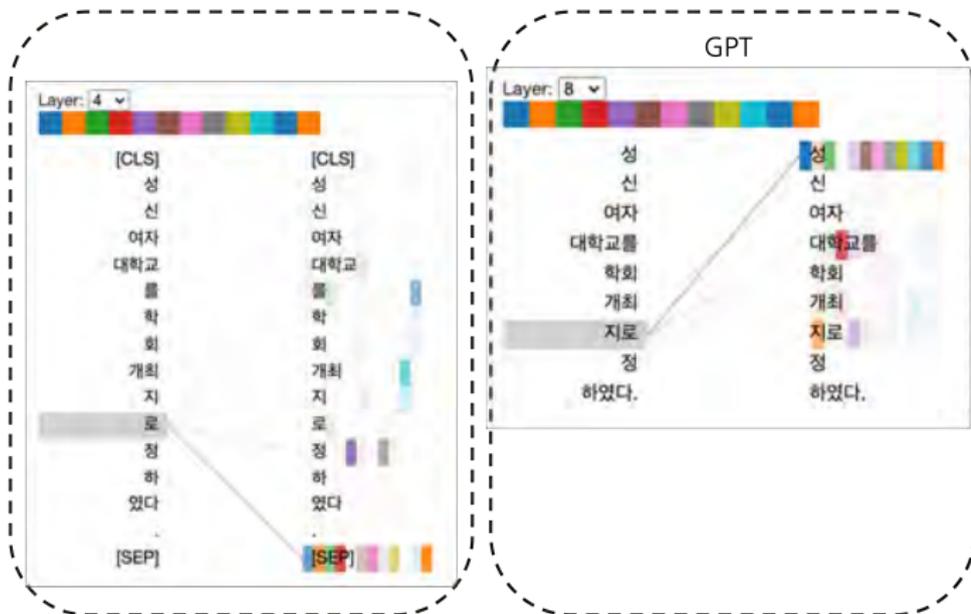
Visualization: DIR

DIR: 서대문으로 갑시다.



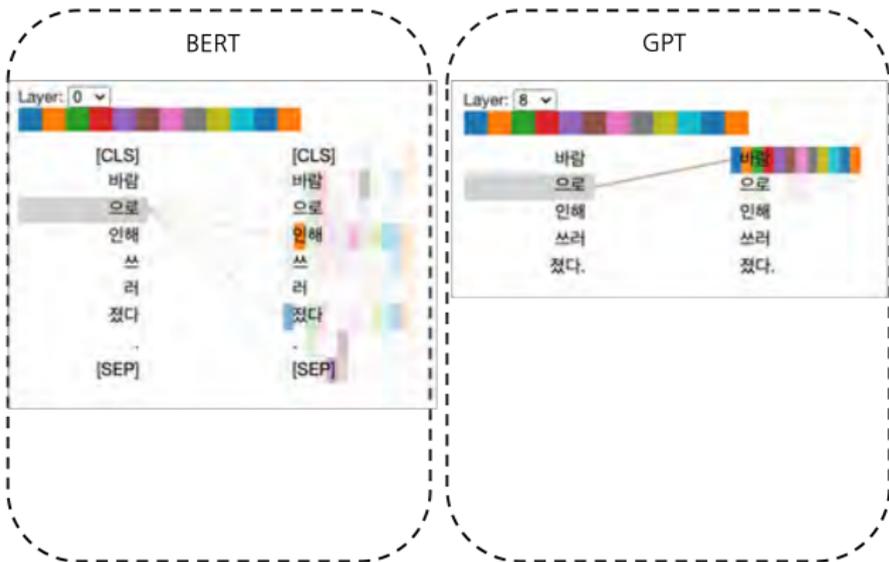
Visualization: LOC

LOC: 성신여자대학교를 학회 개최지로 정하였다.



Visualization: EFF

EFF: 바람으로 인해 쓰러졌다.



Discussion & Conclusion

Discussion

- ▶ The BERT model performs in a stable way and simulates how humans recognize the polysemy involving Korean adverbial postpositions better than GPT-2 model does.

Discussion

"These results suggest that it is likely that BERT does acquire **some form of a structural inductive bias** from self-supervised pretraining, at least outside of the NPI domain."
(Warstadt Bowman, 2020)

Discussion

"Our results allow us to conclude that BERT does indeed have access to **a significant amount of information**, much of which linguists typically call constructional information."
(Madabushi et al., 2020)

Discussion

"**GPT-2's perplexity** is better captured by the considered features and it resulted to be more affected by **lexical parts-of-speech** and features capturing the **vocabulary richness of a sentence**. On the contrary, **BERT's perplexity** seems to be best predicted by **syntactic features** highly sensitive to sentence length."

(Miaschi et al. 2021)

Discussion

- ▶ BERT performs better than GPT-2 because the meaning of Korean adverbial postposition is maybe sensitive to syntactic features.
- ▶ Perhaps, BERT is a better approach for understanding how humans deal with polysemy.

Conclusion

- ▶ To understand word-level polysemy of Korean postposition, at least, we have to use the syntactic information.
- ▶ If we spend more time learning a language, we can identify the word-level polysemy more clearly.
- ▶ Even if the function of the postposition is used rarely but it can be distinguished from the other functions, we can identify it as a distinguished function.
- ▶ If the functions are semantically similar to each other, it is hard to be distinguished one from the other.

Thank you for listening.

중국어 의미 주석 말뭉치 구축과 활용

강병규(서강대)

차례

1. 머리말
 2. 다의어 의미 주석 말뭉치 구축
 3. 의미역 주석 말뭉치 구축
 4. 중국어 의미 주석 말뭉치의 활용
 5. 맺음말
-

1. 머리말

의미주석(Semantic Annotation)은 자연어 처리(NLP)와 정보 검색 분야에서 중요한 역할을 담당한다. 의미주석의 주요 목적은 텍스트 내의 다양한 단위(단어, 구문, 문장, 문단 등)에 의미 정보를 부여하여 기계가 텍스트의 의미를 더 정확하게 이해하고 처리할 수 있도록 돕는 것이다. 이러한 과정은 텍스트 데이터를 구조화하고, 보다 정교한 분석과 활용을 가능하게 한다. 결과적으로 정보 검색, 텍스트 요약, 기계 번역, 질문 응답 시스템 등 다양한 응용 분야에서 큰 이점을 제공한다.

의미주석의 주요 구성 요소로는 개체명 인식(Named Entity Recognition, NER), 관계 추출(Relation Extraction), 개체 연결(Entity Linking), 의미역할 라벨링(Semantic Role Labeling, SRL), 다의어의 의미항 주석 등이 있다. 각 구성 요소는 텍스트 내에서 특정한 의미를 가진 단어나 구문을 식별하고, 이를 적절한 의미 태그로 라벨링하는 과정을 포함한다. 예를 들어, NER은 텍스트에서 인물, 장소, 조직 등의 명명된 개체를 식별하고, 이를 해당하는 카테고리로 라벨링한다. 관계 추출은 이러한 개체들 간의 관계를 식별하고, 이를 구조화된 형식으로 표현한다. 개체 연결은 텍스트 내의 명명된 개체를 외부 지식 그래프나 데이터베이스의 항목과 연결하여, 의미를 더욱 명확히 한다. SRL은 문장에서 동사와 관련된 명사구의 의미적 역할을 식별하고, 이를 라벨링함으로써 문장의 구조와 의미를 더 잘 이해할 수 있게 한다.

본 연구에서는 의미주석의 여러 측면 중 다의어의 의미를 라벨링하는 측면과 의미역할 라벨링을 중심으로 논의하고자 한다. 다의어는 한 단어가 여러 가지 의미를 가질 때 이를 구별하는 것이며, 이는 의미주석의 중요한 과제 중 하나이다. 예를 들어, 'bank'라는

단어는 ‘강독’과 ‘은행’이라는 두 가지 의미를 나타낸다. 문맥에 따라 적절한 의미를 선택하여 라벨링해야 한다. 이는 정보 검색의 정확성을 높이고, 기계 번역의 품질을 향상시키며, 텍스트 요약에서 중요한 정보의 손실을 최소화하는 데 기여할 수 있다. 중국어를 처리하고 분석하는 과정에서도 이러한 다의어의 의미구별을 위해 의미항을 표시하는 작업이 필요한 경우가 많다.

의미역 라벨링은 문장에서 동사와 관련된 명사구의 역할을 식별하고 라벨링하는 과정으로, 문장의 의미를 정확하게 이해하는 데 중요한 역할을 한다. 예를 들어, “Mary sold the book to John.”이라는 문장에서 “Mary”는 판매자(seller), “the book”은 판매된 물건(thing sold), “John”은 구매자(buyer)로 라벨링할 수 있다. 의미역할 라벨링은 문장의 구조를 이해하는 데 도움을 주고, 자연어 이해와 관련된 다양한 응용 분야에서 활용될 수 있다. 중국어 문장 구조에서도 술어와 논항의 의미관계에 대한 표기를 통해 의미를 파악하는 작업을 생각해 볼 수 있다.

본고는 이러한 배경하에서 중국어 의미 주석 말뭉치 구축 방법을 논의하고 그 활용 사례를 소개하는 것을 목적으로 한다. 구체적으로는 방향보어 의미 주석 말뭉치와 문두 다중논항 말뭉치를 연구 대상으로 삼았다.¹⁾ 이 두 말뭉치의 규모는 모두 크지 않은 편이지만, 이를 통해 중국어 의미주석의 특수성과 중요성을 살펴볼 수 있다. 또한 이 데이터가 자동 의미 분류, 기계 번역 및 중국어 교육에 어떻게 활용될 수 있을지를 논의하고자 한다.

2. 다의어 의미 주석 말뭉치 구축 - 중국어 방향보어를 중심으로

2.1 다의어의 의미 주석

다의어의 개념은 한 단어가 여러 가지 의미를 가질 때 이를 구별하는 것을 의미한다. 다의어는 문맥에 따라 적절한 의미를 선택하여 라벨링하는 것이 중요한데, 이는 정보 검색의 정확성을 높이고, 기계 번역의 품질을 향상시키며, 텍스트 요약에서 중요한 정보의 손실을 최소화하는 데 기여할 수 있다.

중국어의 방향보어는 다의적인 기능을 나타내는 문법 요소 중의 하나로서 의미 주석 연구의 대상으로 삼을 필요성이 있다. 중국어 방향보어는 동사 뒤에 위치하여 다양한 의미를 나타내는 다기능성 성분이다. 예를 들어, “站起来(일어서다)”에서 ‘起来’는 이동의 방향을 나타내고, “肿起来(부어오르다)”에서 ‘起来’는 결과의 의미를, “笑起来(웃기 시작하다)”에서 ‘起来’는 시작의 의미를, “看起来(보다)”에서 ‘起来’는 평가의 기능을 나타낸다.

방향보어는 어휘적 의미뿐만 아니라 문법적 의미까지 포함하여 다층적인 의미를 나타낸다. 이는 중국어 문법 연구와 언어 교육에서 중요한 주제가 된다. 그러나 방향보어에 대한 연구는 주로 이론적 분석이나 일부 계량 분석에 국한되어 왔다. 이론적 분석은 방향보어의 기능과 의미를 설명하는 데 주로 사용되었으며, 계량 분석은 데이터의 양적 분

1) 본 발표문은 필자가 기존에 진행하였던 의미주석 말뭉치 관련 연구를 기초로 작성되었다. 중국어 방향보어 관련 말뭉치는 개인연구 차원에서 구축되었으며, 중국어 문두 다중논항 말뭉치는 공동연구 프로젝트의 일환으로 구축되었다. 본 발표문의 내용도 기존의 연구를 토대로 작성되었다. 자세한 것은 강병규(2021, 2024), Kang & Yu (2023), 박정규 외(2023a,b, 2024a,b)를 참고하기 바람.

석을 통해 방향보어의 사용 빈도와 패턴을 조사하는 데 집중되었다.

하지만 이러한 접근은 대량의 실제 사용 데이터를 기반으로 한 심층적인 분석을 포함하지 못하는 한계가 있었다. 따라서 본 연구에서는 방향보어의 다양한 의미와 기능을 체계적으로 분석하기 위해 의미 주석 말뭉치를 구축하게 되었다. 이는 기존 연구의 부족한 점을 보완하고, 방향보어의 의미와 기능에 대한 양적 및 질적 분석을 가능하게 한다.

2.2 연구 대상

방향보어는 방향동사가 단독으로 사용되기도 하고 ‘来’, ‘去’와 결합하여 사용되기도 한다. 이는 다음과 같이 도식화할 수 있다.

동사 (V)	+		上	下	进	出	回	过	起	开
		来	上来	下来	进来	出来	回来	过来	起来	开来
		去	上去	下去	进去	出去	回去	过去	-	-

<표 1> 방향보어 종류

방향보어는 방향 의미 이외에도 결과 의미, 상태 의미 등을 나타낸다. 방향보어는 기능이 비교적 단순한 것과 복잡한 것이 있다. 비교적 단순한 방향보어는 의미 분류와 예측이 용이하다. 그러나 기능이 복잡한 방향보어는 난이도가 증가한다. 필자는 연구에서 의미적으로 대칭적인 위(上)와 아래(下)의 이동을 표현하는 방향보어 중에서 기능이 복잡한 성분을 주요 연구대상으로 삼았다. 특히 기능이 3개 이상이면서, 상(aspect) 기능이 있으며 중국어 교육과정에서도 오류빈도가 높다고 알려진 보어를 중심으로 고찰하였다. 이러한 기준에 의해 선택한 방향보어는 단순방향보어 ‘上’과 복합방향보어 ‘起来’, ‘下来’, ‘下去’이다.

2.3 방향보어 의미 분류 기준 설정

방향보어의 의미 분류는 刘月华(1998)의 연구를 기점으로 체계적인 틀을 갖추기 시작했다고 할 수 있다. 刘月华는 방향보어의 의미를 방향의미(趋向意义), 결과의미(结果意义), 상태의미(状态意义)의 세 가지 범주로 구분하였다. 이 분류법은 이후 많은 연구자들에 의해 채택되어 방향보어 의미 연구의 기본 틀로 자리잡았다. 刘月华의 분류법은 방향보어의 의미 체계를 명료하게 제시했다는 점에서 큰 의미가 있다. 그러나 모든 방향보어를 이 세 가지 분류 체계 안에서 설명하려 했기 때문에, 일부 방향보어의 의미 기능을 적절히 포괄하지 못하는 한계도 존재한다. 특히 복잡한 의미 기능을 가진 방향보어의 경우, 세 가지 분류만으로는 그 의미를 충분히 설명하기 어렵다.

초기 방향보어 의미 분류 연구의 한계를 극복하기 위해, 이후의 연구에서는 특정 방향보어에 초점을 맞추어 보다 세분화된 의미 분류를 시도하였다. 특히 언어학적 관점에서 상(aspect)의 개념을 도입하여 방향보어의 의미 기능을 분석한 卢英顺(2000), Xiao & McEneaney(2004), 陈前瑞(2008), 유수경(2020) 등의 연구들이 주목할 만하다.

Kang & Yu(2023)에서는 방향 이동, 결과 상태, 완성, 지속 등과 같은 상적 자질을 추

가하여 주요 방향보어의 의미 분류 체계를 정교화하려 하였다. 이들은 방향보어의 의미가 단순히 방향성이나 결과성에 국한되지 않으며, 동작의 완성이나 지속과 같은 상적 특성과도 밀접하게 연관되어 있음을 강조하였다. 이를 통해 방향보어의 의미 기능을 보다 세밀하게 포착하고, 각 방향보어 간의 의미 차이를 명확히 드러내고자 하였다.

Kang & Yu(2023)에 따르면, 중국어 방향보어의 의미를 이해하기 위해서는 방향이동과 상적 속성이라는 두 가지 핵심 요소를 함께 고려해야 한다. 방향이동은 변화가 일어나는 구체적인 방향(상향/하향)을 의미하고, 상적 속성은 방향이동의 결과로 인한 상태 변화를 내포한다. 이 두 가지 요소가 결합하여 각 방향보어의 기본적인 의미를 형성한다.

중국어 방향보어는 해당 방향으로의 이동이라는 기본적인 의미를 가진다. 예를 들어, ‘上’과 ‘起来’는 위로의 이동을 통한 위치 변화를, ‘下来’와 ‘下去’는 아래로의 이동을 통한 위치 변화를 나타낸다. 또한 각각의 방향보어는 동사 뒤에 위치한다는 분포적 특징으로 인해, 행위의 결과로 발생한 상태의 변화를 나타내는 고유한 특성을 지닌다. ‘上’, ‘起来’, ‘下来’, ‘下去’와 같은 방향보어는 모두 동작행위로 인해 새로운 상태로 진입하거나 변화하는 상적 속성을 공유하며, 이는 중국어 방향보어가 가지는 핵심적인 특징이라고 할 수 있다.

방향이동과 상적 속성은 서로 연관된다. 방향보어는 [이동성]과 [방향성]이 약화됨에 따라, 이동성이 없는 행위의 결과상태(resultative)를 나타낼 수 있다. 그러나 이러한 결과상태 역시 방향보어의 기본 의미인 방향성의 영향을 받는다. ‘上’은 결합(고정), ‘起来’는 출현(돌출)과 결합, ‘下去’는 분리와 함몰된 상태, ‘下来’는 분리와 함몰 상태뿐만 아니라 행위의 고정된 결과상태까지 표현한다. 이러한 결과의미는 문법화 과정을 거쳐 상(aspect) 표지나 담화표지까지 발전하게 되는데, 그 발전 방향은 각 방향보어의 기본 의미에 따라 상이하게 나타난다. ‘上’은 완성과 지속의 시작을, ‘起来’는 시작 및 그에 이어진 지속과 평가적인 양태의미를 표현한다. 반면, ‘下来’와 ‘下去’는 화자의 시점을 기준으로 각각 과거부터 현재까지, 현재부터 미래로의 상반된 시간 방향의 지속을 나타내며, 더 나아가 담화를 이어가는 담화연속표지로도 기능한다.

Kang & Yu(2023)의 의미 분류체계를 바탕으로, 본고는 ‘上’, ‘起来’, ‘下来’, ‘下去’의 의미를 다음과 같이 설정하였다. 이들 방향보어의 의미는 상태변화(change of states), 이동(movement), 결과(resultative), 완성(completive), 지속(continuative), 시작(beginning) 등의 의미 자질의 유무에 따라 각각 4개의 하위 의미항으로 분류된다. 이러한 의미 자질은 개별 방향보어의 의미를 나열하는 데 그치지 않고, 방향보어 전체의 의미 체계를 포괄적으로 분석하기 위한 기준으로 작용한다.

의미자질	V上			
	上1	上2	上3	上4
상태의 변화(Change of state)	+	+	+	+
이동(Movement)	+	±	-	-
상향(Upward)	±	-	-	-
결과(Resultative)	-	+	-	-
완성(Completive)	-	-	+	-
행위/상태 지속(Continuative-A/S)	-	-	-	-
결과상태지속(Continuative-R)	-	-	-	+
시작(Beginning)	-	-	-	+
담화(Discourse)	-	-	-	-
실례	跑上	贴上	买上	爱上

<표 2> 방향보어 '上'의 의미 분류 기준

의미자질	V起来			
	起来1	起来2	起来3	起来4
상태의 변화(Change of state)	+	+	+	-
이동(Movement)	+	-	-	-
상향(Upward)	+	-	-	-
결과(Resultative)	-	+	-	-
완성(Completive)	-	-	-	-
행위/상태 지속(Continuative-A/S)	-	-	+	-
결과상태지속(Continuative-R)	-	-	-	-
시작(Beginning)	-	-	+	-
담화(Discourse)	-	-	-	+
실례	站起来	肿起来	笑起来	看起来

<표 3> 방향보어 '起来'의 의미 분류 기준

의미자질	V下来			
	下来1	下来2	下来3	下来4
상태의 변화(Change of state)	+	+	+	-
이동(Movement)	+	±	-	-
하향(Downward)	±	-	-	-
결과(Resultative)	-	+	-	-
완성(Completive)	-	+	-	-
행위/상태 지속(Continuative-A/S)	-	-	+	-
결과상태지속(Continuative-R)	-	-	+	-
시작(Beginning)	-	-	-	-
담화(Discourse)	-	-	-	+
실례	走下来	写下来	停下来	接下来

<표 4> 방향보어 '下来'의 의미 분류 기준

의미자질	V下去			
	下去1	下去2	下去3	下去4
상태의 변화(Change of state)	+	+	-	-
이동(Movement)	+	±	-	-
하향(Downward)	±	-	-	-
결과(Resultative)	-	+	-	-
완성(Completive)	-	-	-	-
행위/상태 지속(Continuative-A/S)	-	-	+	-
결과상태지속(Continuative-R)	-	-	+	-
시작(Beginning)	-	-	-	-
담화(Discourse)	-	-	-	+
실례	飞下去	摘下去	说下去	接下去

<표 5> 방향보어 ‘下去’의 의미 분류 기준

표에서 제시된 바와 같이, ‘上’, ‘起来’, ‘下来’, ‘下去’의 하위 의미는 의미 자질의 유무에 따라 4개의 의미항으로 분류된다. 이 의미 자질은 단순히 개별 방향보어의 의미를 나열하는 데 그치지 않고, 방향보어 전체의 의미 체계를 분석하는 데 있어 핵심적인 기준으로 작용한다. 또한 방향보어별로 어떤 자질이 두드러지는지도 파악할 수 있다. 예를 들어, ‘上’은 이동(movement)과 결과(resultative)와 완성(completive)의 의미 자질이 두드러지는 반면, ‘起来’는 지속(continuative)와 시작(beginning)의 의미 자질이, ‘下去’는 지속(continuative)의 의미 자질이 상대적으로 부각된다. 이는 각 방향보어가 기본적인 공간 이동의 의미에서 출발하여 서로 다른 방식으로 의미를 확장해 나가는 과정을 보여준다.

2.4 방향보어 의미 주석 말뭉치 구축

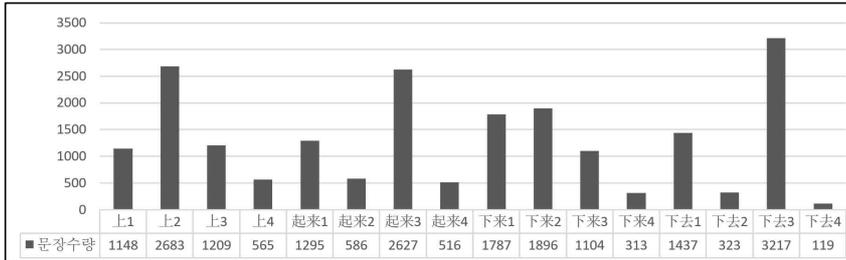
말뭉치 구축을 위해 중국 문학작품, 방송대본, 중국어 교재 등 다양한 장르의 텍스트에서 예문이 수집되었다. 문학작품과 방송대본에서는 약 6,500만 자의 텍스트를 대상으로 하였으며, 중국어 교재에서는 약 300만 자의 텍스트를 활용하였다. 이러한 대규모 텍스트에서 방향보어가 사용된 예문을 무작위로 추출하여 말뭉치를 구성하였다.

의미 주석 말뭉치 구축은 연구자와 연구보조원이 협력하여 수행해야 하는 작업으로, 상당한 시간과 노력이 필요하다. 이에 작업의 시간적 제약을 고려하여, 초기에는 4개의 방향보어에 대해 각각 3,000문장씩 총 12,000문장의 의미 주석 말뭉치를 구축하는 것으로 계획하였다. 이는 한장훈·박영준·정영훈(2019)의 연구결과에 근거한 것으로, 미세조정 과정에서 정밀한 주석이 달린 말뭉치를 사용할 경우 3,000문장 규모로도 충분히 좋은 성능을 얻을 수 있다는 점을 참고하였다.

그러나 구축 과정에서는 보다 안정적이고 신뢰할 수 있는 결과를 얻기 위해 의미 주석 말뭉치의 규모를 더 확대하여 각 방향보어별로 2,000문장씩 추가 분석을 진행하였다. 최종적으로 각 방향보어별로 5,000문장씩, 총 20,000문장으로 구성된 대규모 의미 주석 말뭉치가 구축되었다.²⁾

2) 방향보어 의미 주석 작업은 필자와 박사과정 원어민 연구보조원의 공동작업의 형식으로 진행되었다. 말뭉치 주석 작업에 소요된 시간은 대략 6개월이다. 필자는 연구책임자로서 방향보어의 의미 분류 기준을 마련하였다. 원어민 연구보조원은 수집된 예문을 대상으로 각 방향보어가 어떤 의미 범주에 속하는지 구분하고 표기하는 작업을 수행하였다. 분석 과정에서 정확성과 일관성을 확보하기 위해 필자와 연구보조원은 1개월 간격으로 작업 결과를 검토하고 토론하는 과정을 거쳤다. 이를 통해 의미 분

이렇게 구축된 20,000여 문장의 방향보어 의미 주석 말뭉치를 바탕으로, 각 방향보어의 의미 유형별 사용빈도를 살펴보면 다음과 같은 분포를 보인다.



<그림 1> 방향보어 의미 주석 말뭉치에서 유형별 비율

	上1	上2	上3	上4
수량	1,148	2,683	1,209	565
비율	20.48%	47.87%	21.57%	10.08%

<표 6> '上'의 의미 유형별 분포

	起来1	起来2	起来3	起来4
수량	1,295	586	2,627	516
비율	25.77%	11.66%	52.29%	10.28%

<표 7> '起来'의 의미 유형별 분포

	下来1	下来2	下来3	下来4
수량	1,787	1,896	1,104	313
비율	35.04%	37.18%	21.65%	6.13%

<표 8> '下来'의 의미 유형별 분포

	下去1	下去2	下去3	下去4
수량	1,437	323	3,217	119
비율	28.21%	6.34%	63.11%	2.34%

<표 9> '下去'의 의미 유형별 분포

이상의 분석을 통해, 각 방향보어의 내부적인 의미 유형 간 관계와 다른 방향보어와의 의미 유형 비교의 두 가지 차원에서 아래와 같은 사실을 확인하였다.

첫째, 방향보어의 핵심의미에 대한 인식의 변화가 필요하다는 점이다. 분석에 따르면 방향보어별로 출현 빈도가 가장 높은 의미 유형은 '上2'(47.87%), '起来3'(52.29%), '下来

류 기준을 정교화하고 예문 분석에서의 모호한 부분을 해소할 수 있었다. 나아가 분석 결과의 신뢰성과 타당성을 높이기 위해 중국어 문법 전공 교수의 자문을 받았다. 이로써 방향보어 의미 분류의 신뢰성과 객관성을 확보한 말뭉치를 구축할 수 있었다.

2'(37.18%), '下去3'(63.11%)이다. 이들 의미는 행위의 결과상태, 행위나 상태의 시작이나 지속의미를 표현하는 문법화 된 기능이다.

과거에 대부분의 학자들은 방향보어의 이동과 방향의미에 초점을 두고, 방향보어를 동사의 방향성을 묘사하기 위한 성분으로 정의했다. 그러나 본고의 분석에 따르면 이러한 정의는 재논의될 필요가 있다. 하나의 언어단위가 근본적인 의미로부터 다양한 의미로 분할되는 것은 언어학적으로 매우 보편적인 현상이다. 이때 다기능이 기원한 근원의미도 중요하다. 그러나 자주 사용되는 의미가 기계번역에서 오류가 자주 발생하며, TCSL에서 어떤 기능이 주목을 받아야 하는지의 문제도 의문이 제기된다.

두 번째로 각 방향보어마다 각기 다른 문법화 정도를 가짐을 실증적으로 확인하였다. '上'과 '下来'의 의미 가운데, 행위의 결과상태를 표현하는 '上2'와 '下来2'의 빈도가 가장 높다. 반면, '起来'와 '下去'는 상 기능인 시작과 지속의미를 표현하는 '起来3', '下去3'의 빈도가 가장 높다. 이처럼 각 방향보어의 핵심의미는 문법화 정도에서 차이를 보인다. 이는 '起来', '下去'가 '上', '下来' 보다 문법화 진행이 더 심화된 성분임을 의미한다. '上'과 '下来'에 대해 다시 이야기하자면, 의미 유형별 사용빈도가 거의 유사한 듯 보인다. 그러나 결합동사의 의미 제약을 고려한다면, '下来2'가 '上2' 보다 더 넓은 결합동사 범위를 가진다. 특히 '下来'는 높은 문법화 정도를 보이는 담화 지속 기능('下来4')까지 표현하지만, '上'은 그렇지 않다. 이는 '下来'가 '上' 보다 더 문법화된 방향보어임을 의미한다.

3. 의미역 주석 말뭉치 구축

3.1 의미역의 정의 및 유형

의미역은 사태(state of affairs)에 참여하는 논항이 동사와 맺는 다양한 의미관계를 지칭하는 개념이다. 이는 'Participant Role'(Allan 1986), 'Semantic role'(Givon 1990), 'Thematic Relation'(Jackendoff 1972, Gruber 1976), 'Thematic Role'(Dowty 1986, 1989, 1991, Jackendoff 1990) 등의 용어로 표현된다.

의미역의 분류 체계, 수량, 그리고 각 의미역의 명칭과 정의는 학자들 간에 차이가 존재한다. Fillmore(1966), Andrews(1985), Radford(1988), Jackendoff(1990), Dowty(1991), Blake(1994) 등의 연구를 살펴보면, 의미역에 대한 명칭과 정의가 완전히 일치하지 않음을 확인할 수 있다. 이러한 불일치는 의미관계가 절대적으로 구분되기 어려우며, 그 경계가 모호한 특성을 지니고 있음을 시사한다.

박정구 등(2023b)에서는 중국어의 의미역 배열 제약을 고찰하기 위해 여러 선행 연구를 종합하여 중국어의 특성에 맞게 의미역을 분류하였다. 이에 따라 도출된 의미역의 유형은 행위주(Agent, A), 경험주(Experiencer, E), 묘사대상(Theme, X), 판단대상(H), 행위대상(K), 피행위주(Patient, P), 수혜자(Beneficiary, B), 접수자(Recipient, D), 도구(Instrument, I), 재료(Material, M), 방식(Manner, N), 시간(Time, T), 처소(Location, L), 종점(Goal, G), 기점(Source, S), 원인(Causar, C), 근거(Reason, R) 등이 있다.

이러한 의미역 유형은 중국어의 의미 주석을 위해 설정된 것이다. 아래는 박정구 등(2023b)가 제안한 의미역 유형에 대한 한국어, 중국어, 영어 명칭과 해당 예문을 표로 제

시한 것이다.

의미역(약어)	중국어 표기	영어 표기	중국어 예문	영어 예문
행위주 (A)	施事	Agent	现在 <u>他/A</u> 还不能打篮球。	<u>David/A</u> cooked the rashers.
경험주 (E)	感事	Experiencer	<u>他/E</u> 很紧张。	<u>Kevin/E</u> felt ill.
묘사 대상(X)	当事	Theme	今天我们的话题/ <u>X</u> 很好。	<u>The book/X</u> is in the library.
판단 대상(H)	系事		<u>他/H</u> 过去是上海郊区的农民	<u>My father/H</u> is a farmer.
행위대상 (K)	客事 (对象)		你 <u>什么/K</u> 也没有看见。	I put the <u>book/K</u> on the table.
피행위주 (P)	受事	Patient	<u>那颗心/P</u> 她/P是迟早要伤。	Enda cut back <u>these bushes/P</u> .
수혜자 (B)	受益者	Beneficiary	<u>可怜的人们/B</u> ，我给你出个主意吧。	They baked <u>me/B</u> a cake.
접수자 (D)	与事	Recipient	<u>你/D</u> 输的钱让她掏给你。	He sold <u>me/D</u> this wreck.
도구 (I)	工具	Instrument	顾雅仙 <u>长长的指甲/I</u> 在那只手掌上划来划去。	They signed the treaty with <u>the same pen/I</u> .
재료 (M)	材料	Material	客家土楼， <u>大小石块/M</u> 累砌打牢地基。	He built a house of wood/ <u>M</u> .
방식 (N)	方式	Manner	我 <u>一个猛子/N</u> 从床上跳下来。	<u>With great urgency/N</u> , Tabitha phoned 911.
시간 (T)	时间	Time	<u>后来/T</u> 陈佐千一个人进北厢房去的。	The pitcher struck out nine batters <u>today/T</u> .
처소 (L)	处所	Location	你 <u>葫芦里/L</u> 卖的什么药?	The monster was hiding under <u>the bed/L</u> .
종점 (G)	终点	Goal	我 <u>哪儿/G</u> 也不去。	Sheila handed her license to <u>the policeman/G</u> .
기점 (S)	起点 (来源)	Source	南沙 <u>自古/S</u> 属中华。	The plane came back from <u>Kinshasa/S</u> .
원인 (C)	原因 (致事)	Causer	<u>大风/C</u> 把树刮倒了。	<u>The wind/C</u> flattened the crops.
근거 (R)	依据	Reason	<u>仅棉花一项/R</u> ，每个农民就增收430多元。	<u>Because Clyde was hungry/R</u> , he ate the cake.
동반격 (Y)	伴随	Comitative	<u>其他的人/Y</u> ，他也和他们作了同一内容。	the professor enters the lecture-hall with <u>his students/Y</u> .

<표 10> 의미역 분류표 (박정구 등, 2023b)

3.2 의미역 주석 말뭉치의 사례- 문두 다중논항 데이터를 중심으로

중국어 문장 성분은 크게 술어를 기준으로 앞에 오는 성분과 뒤에 오는 성분으로 구분된다. 술어 앞에 위치하는 성분은 주로 주제(topic)와 주어이며, 술어 뒤에 오는 명사성 성분은 목적어로 분류된다. 중국어는 주제 중심 언어로 알려져 있어, 술어 앞에 다양한 성분이 주제 기능을 수행하는 특징을 보인다.

이러한 언어적 특성으로 인해 중국어에서는 술어 앞, 즉 문두에 두 개 이상의 문장 성분이 동시에 사용되는 현상이 빈번히 관찰된다. 예를 들어, "[这个人][我]不认识"와 같은 문장에서 '这个人'과 '我'라는 두 개의 성분이 문두에 위치하는 것을 볼 수 있다. 이러한 문두의 성분들은 술어와의 의미적 관계에 따라 대상, 경험주 등 다양한 의미역으로 분류될 수 있다.

3.2.1 문두 다중논항 말뭉치 구축 과정

중국어 문두 다중논항 데이터는 말뭉치의 용례에 기반하여 구축되었다. 말뭉치 용례는 문학작품, 신문, 방송 보도자료, 영화, 드라마, 교과서 등 여러 장르에서 수집되었다. 데이터 구축 과정은 크게 두 단계로 나뉘어진다. 첫 번째는 말뭉치에서 문두에 2개 이상의 논항이 출현하는 예문을 찾아 정리하는 과정이다. 두 번째는 문두 다중논항의 의미 유형을 분류하는 단계이다.

(가) 문두 다중논항 용례 추출

용례 추출은 1차적으로 원어민이 일일이 말뭉치 자료를 보면서 예문을 선별하는 방식으로 진행되었다. 원어민들에게 제공된 중국어 말뭉치 자료는 장르별 말뭉치 파일에서 50,000문장씩 무작위로 추출되었다.

1차로 수집된 문두 다중논항 후보 문장은 원어민들끼리 교차 검토를 통해 재차 수정되었다. 교차 검토를 거쳐 최종 수집된 다중논항 용례는 모두 5,143개이다. 이 중에서 신문에서 추출된 용례가 가장 많은 비중(40.5%)을 차지한다. 문학작품(21.4%), 방송뉴스 및 대담(17.5%)에서 추출된 용례의 비율도 높은 편이다. 문어·구어의 비율면에서는 대체로 문어체 자료에서의 다중논항 출현 비율이 높고 구어체에서의 출현 비율은 상대적으로 낮다. 아래의 표는 장르별로 추출된 문장의 숫자와 비율을 정리한 것이다.

장르	추출 용례 수량	비율
신문	2,082	40.5%
문학작품	1,101	21.4%
방송뉴스 및 대담	899	17.5%
교과서(교재)	612	11.9%
드라마	337	6.6%
영화대본	111	2.2%
합계	5,143	100%

<표 11> 장르별 다중논항 용례의 수량과 비율

(나) 문두 다중논항 유형 분류

원어민의 검토를 거쳐 확보된 데이터셋에서 다중논항이 어떤 의미역을 가지는지를 분석하기 위해서는 추가로 의미 주석(semantic annotation) 작업이 필요하다. 의미 주석 작업은 원어민들이 용례를 보면서 논항의 의미역을 일일이 분석하는 방식으로 진행되었다. 또한 의미 유형 표기는 모든 논항이 아닌 문두 다중논항에 대해서만 진행되었다. 이러한 작업을 통해 다중논항의 의미가 어떠한 양상을 보이는지를 파악하고자 하였다. 그리고 말뭉치에 부착된 의미 표지는 술어와의 의미관계를 고려하여 영문 약어(대문자)로 기록되었다. 또한 분석의 효율성을 높이기 위해 의미역 구분자는 ‘##’ 기호를 사용하였다. 예를 들어 행위주(Agent) 논항은 ‘A’로 표기되고 시간(Time) 논항은 ‘T’로 표기된다. 아래는 말뭉치 의미 주석 작업의 실례이다.

- (5) ##这个城市/L##文艺界的大人物/A##都住在这个小区里。(처소-행위주)
- (6) ##我/A##一个猛子/N##从床上跳下来。(행위주-방식)
- (7) ##酒柜上/L##玻璃鱼缸里/L, ##金鱼/A##在无声无息地游动。
(처소-처소-행위주)
- (8) ##他那里/G##你/A##还要再去一次。(중점-행위주)

이와 같이 의미역 표지가 부착된 데이터는 작업자와 문장, 의미 유형, 논항 개수 등으로 나누어 저장되었다. 문두 다중논항 데이터셋의 주요 정보는 아래의 표와 같다.

작업자	용례	논항 수	논항 배열
당○○	##这么说, ##这段时间/T##你/A##每天晚上/T##在床上就是干躺着对阮琳运气发功, 一指头也没碰她?	3	['T', 'A', 'T']
당○○	##车厢/L##灯光/X##昏暗, 人头攒动, 过道卧满做小买卖的农民, 龇着大黄板牙大声说笑, 放肆地抽着呛人的烟卷。	2	['L', 'X']
칭○○	##我们/A##演出/K##演糊涂了。	2	['A', 'K']
당○○	##我/A##一个人/N##生活也用不着, 转手卖给了别人。	2	['A', 'N']
이○○	##沿街小饭馆/L##热气腾腾的屋内/L##, 羊肉片/X##在滚着开水的铜火锅里变色, 围着桌子的人们吃得满头大汗。	3	['L', 'L', 'X']
김○○	##喝了白酒的男人/X##脸/X##红得象猪肝, 醉醺醺地互相搀扶着从我身旁默默走过。	2	['X', 'X'] ³⁾
공○○	##后面的舞/K##我/A##只跳没看他。	2	['K', 'A']
당○○	##街旁屋檐下/L##一个老太太/A##冲我招手, 我笑着摇着头跑远。	2	['L', 'A']

<표 12> 문두 다중논항 의미 주석 데이터셋 실례

3.2.2 문두 다중논항의 수량별 분포

본 연구에서 구축한 문두 다중논항 데이터는 2개의 논항으로 구성된 용례가 가장 많은 비중을 차지한다. 전체 5,143개의 용례 중에서 2개의 논항으로 구성된 것은 4,490개로서 87.3%를 차지한다. 그 다음으로 많은 비중을 차지하는 것은 3개의 논항으로 구성된 용례이다. 3개의 논항으로 구성된 것은 612개로서 11.9%를 차지한다. 문두에 논항이 4개 이상 출현한 경우는 매우 드물다. 아래의 표에서 보이듯이 4개 논항으로 구성된 용례는 39개(0.7%), 5개 논항인 경우는 2개(0.1%)로서 매우 드물게 관찰된다.

문두 논항수	2개 논항	3개 논항	4개 논항	5개 논항
용례수	4,490	612	39	2
비율	87.3%	11.9%	0.7%	0.1%

<표 13> 문두 다중논항의 수량별 분포

- (9) ##那个年轻姑娘/K##他/A##叫老姨。(2개 논항)
- (10) ##广场上/L##现在/T##人/X##又多起来了啊。(3개 논항)
- (11) ##{在}2003年/T, ##咱们房山的一个学校/L, ##篮球场上/L##两个学生/A##打篮球, 发生冲突, 一个学生把另一个学生就扎成重伤, 然后抢救无效死亡。(4개 논항)
- (12) ##一线城市/X##, 最近涨幅比较大的/X##, 去年到现在/T##, 一年之中/T##涨幅/X##在50%以上和一倍以上, 从现在开始到今年年底, 或者明年的这个时候, 在一年之内, 应该往下降幅是20%到30%, 还是完全有可能的。(5개 논항)

3) 말뭉치 데이터에는 '묘사대상(X)-묘사대상(X)'처럼 의미역이 중복되어 출현하는 경우가 적지 않다. 이때 의미역 간에는 다양한 의미관계가 존재한다. 예를 들어 '喝了白酒的男人'과 '脸' 사이에는 '전체-부분'의 관계를 가진다. 이러한 의미역 관계는 어순 배열에 일정한 제약으로 작용한다. 그러나 본고에서는 이러한 의미역 간의 관계까지는 표시하지 않았다.

위에서 알 수 있듯이 말뭉치에서 문두 다중논항의 용례는 대부분 2개 논항으로 구성된 다. 일부는 3개 논항을 가지기도 한다. 그러나 문두에 논항이 4개 이상 출현하는 비율은 매우 낮다. 설령 문두에 4개 이상의 논항이 출현하더라도 시간 논항과 처소 논항이 중복해서 출현한 경우가 다수를 차지한다. 예를 들어 “昨天下午/T 我们学校/L 篮球场上/L 两个学生/A”처럼 ‘시간-처소-처소-행위주’ 형태의 논항 배열이 그러하다. 그러나 ‘시간-처소-처소-행위주’ 배열은 ‘시간-처소-행위주’로 합쳐서 분석할 수도 있다. 조사 결과에 따르면 말뭉치에서 출현하는 4개 이상의 논항이 중복되지 않고 각기 다른 의미역을 가진 경우는 관찰되지 않았다.

3.2.3 문두 다중논항의 위치별 의미역 분포

문두 다중논항의 의미역은 시간, 처소, 행위주, 피행위주, 대상 등 다양한데 위치별로 일정한 경향성을 보인다. 먼저 문두의 첫 번째 위치(제1논항)에 출현하는 논항은 다음과 같은 빈도순으로 관찰된다.

※ 제1논항의 의미역 출현 빈도

시간(T) > 행위주(A) > 묘사대상(X) > 처소(L) > 행위대상(K) > 경험주(E) > 판단대상(H) > 피행위주(P) > 기점(S) > 근거(R) > 수혜자(B) > 종점(G) > 원인(C)

본 연구에서 구축한 데이터에서 문두의 첫 번째 위치에 출현하는 논항은 시간(T)논항이 절반 이상(51.7%)을 차지한다. 그 다음으로는 행위주(A), 묘사대상(X), 처소(L) 등이 많이 출현한다. 반면 제1논항으로 근거(R), 원인(C), 종점(G), 수혜자(B) 등은 매우 드물게 사용됨을 알 수 있다. 특히 방식(N), 도구(I), 재료(M) 등의 논항은 제1논항의 위치에 거의 출현하지 않는다.

(13) #第二天一早/T##简家的临街小门/K##被咚咚地敲响了。

(14) #南方小城/X##现在/T##离我很远。

문두에서 두 번째 위치(제2논항)에 출현하는 논항을 의미역별로 조사해 보면 행위주(A), 시간(T), 묘사대상(X), 처소(L), 행위대상(K), 경험주(E) 등이 많이 출현한다. 그 중에서도 두 번째 위치에 가장 많이 출현하는 것은 행위주(A)이다. 즉, 문두에 2개 이상의 논항이 출현할 때 앞에 시간이나 처소가 먼저 제시되고 그 다음에 행위주가 출현하는 것이 일반적인 경향임을 알 수 있다. 또한 주목할 것은 제2논항의 위치에 방식(N), 기점(S), 도구(I), 원인(C) 등을 나타내는 논항도 출현한다는 점이다.

※ 제2논항의 의미역 출현 빈도

행위주(A) > 시간(T) > 묘사대상(X) > 처소(L) > 판단대상(H) > 행위대상(K) > 경험주(E) > 피행위주(P) > 방식(N) > 기점(S) > 도구(I) > 원인(C)

(15) #这地方/L##我/A##绝对不会再住下去。

(16) #我/A##一脚/N##踢开门进去, 杜梅正一个人一边吃桔子一边看电视。

(17) ##金桥/A##一只手/I##按住眉君的肩头，似乎想把她的火气按下去。

문두에 논항이 3개 이상 출현하는 경우에 제3논항으로 많이 관찰되는 것은 묘사대상(X), 시간(T), 행위주(A), 경험주(E) 등이다.

※ 제3논항의 의미역 출현 빈도

묘사대상(X) > 시간(T) > 행위주(A) > 처소(L) > 경험주(E) > 방식(N) > 판단대상(H) > 행위대상(K) > 피행위주(P)

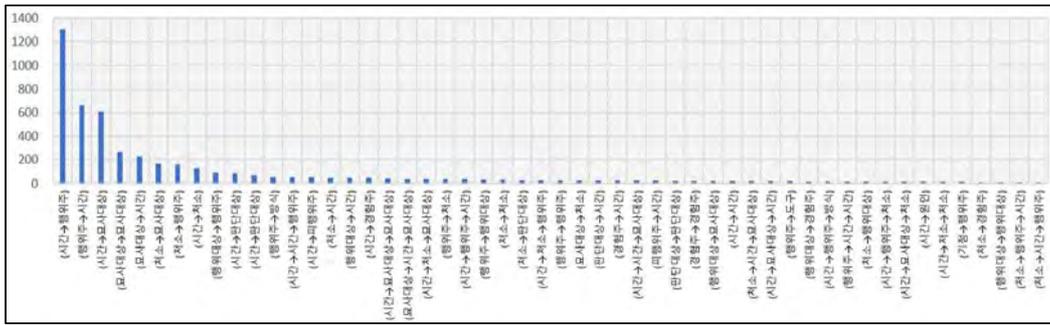
이상의 위치별 논항의 의미역 분포를 정리해 보면 아래의 표와 같이 나타낼 수 있다.

의미역	제1논항		제2논항		제3논항	
	출현빈도	비율	출현빈도	비율	출현빈도	비율
시간(T)	2718	51.7%	1338	25.4%	138	21.1%
행위주(A)	896	17.0%	1730	32.9%	125	19.1%
묘사대상(X)	629	12.0%	1205	22.9%	207	31.6%
처소(L)	567	10.8%	351	6.7%	64	9.8%
행위대상(K)	230	4.4%	156	3.0%	20	3.1%
경험주(E)	64	1.2%	122	2.3%	34	5.2%
판단대상(H)	58	1.1%	157	3.0%	23	3.5%
피행위주(P)	42	0.8%	82	1.6%	6	0.9%
기점(S)	31	0.6%	19	0.4%	3	0.5%
근거(R)	8	0.2%	1	0.0%	0	0.0%
수혜자(B)	5	0.1%	1	0.0%	0	0.0%
중점(G)	5	0.1%	3	0.1%	0	0.0%
원인(C)	5	0.1%	15	0.3%	1	0.2%
방식(N)	1	0.0%	64	1.2%	32	4.9%
도구(I)	0	0.0%	19	0.4%	2	0.3%
재료(M)	0	0.0%	1	0.0%	0	0.0%

<표 14> 문두 다중논항의 위치별 의미역 분포

문두 다중논항의 의미역 배열도 일정한 경향성이 존재한다. 문두 다중논항 데이터는 220여 개의 의미역 배열 패턴을 가진다. 그 중에서 가장 많은 비중을 차지하는 것은 시간(T) 논항이 포함된 배열 패턴이다. 사용빈도의 측면에서는 ‘시간-행위주’, ‘행위주-시간’, ‘시간-묘사대상’, ‘묘사대상-시간’ 등이 높은 비중을 차지한다.

220여개의 의미역 배열 패턴을 빈도순으로 정리해 보면 다중논항의 의미역 배열도 어느 언어 단위와 유사하게 상위 빈도 용례가 전체 관측값의 대부분을 차지하는 양상을 보인다.



<그림 2> 문두 다중논항의 의미역 배열 빈도 그래프

이는 말뭉치에 출현하는 언어단위(token)의 빈도와 순위 사이에 역의 관계가 성립한다는 경험적 법칙과도 상통한다. 소위 지프의 법칙(Zipf's law)이라 불리는 해당 경향성은 말뭉치를 구성하는 언어 단위의 빈도(f)에 따라 내림차순으로 순위(r)를 매겼을 때 빈도와 순위의 곱은 일정한 상수값에 수렴한다는 것이다.

$$f \cdot r = k \quad (f \propto \frac{1}{r}) \quad \text{Zipf's law}$$

문두 다중논항 말뭉치의 의미역 분포가 지프의 법칙을 따른다는 것은, 의미역 시퀀스의 분포도 단어 분포의 자연적인 경향성을 따른다는 것을 시사한다.

3.2.4 자동 추출 방법을 이용한 말뭉치 용례 추출과 의미 주석

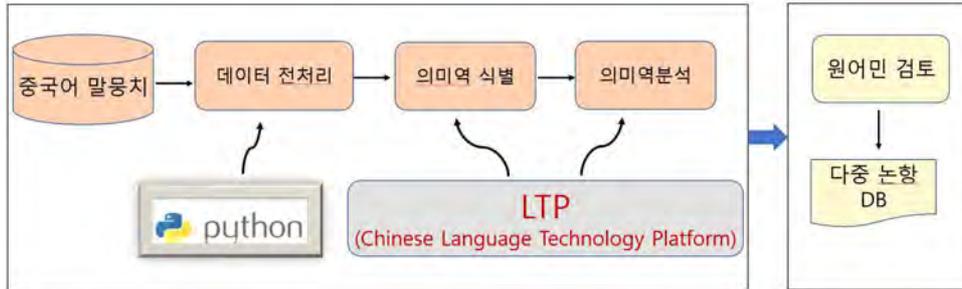
자동추출 방법은 문두 다중논항의 특징을 반영하여 설계되었으며, 중국어 문장의 단어, 통사, 의미 구조를 분석하는 자연어처리 알고리즘이 사용되었다. 자동추출 프로그램을 사용하면 컴퓨터가 조건에 부합하는 예문을 빠르게 찾아 줄 수 있다. 이는 원어민처럼 유연하고 정확하게 찾지는 못하지만 단시간에 많은 예문 후보를 추출해 주기 때문에 문두 다중논항 데이터 수량을 늘리는데 효율적이다.

2차 용례 데이터는 파이썬으로 구현되는 중국어 자연어처리 프로그램을 활용하여 추출되었다. 본 연구에서는 하얼빈공대에서 공개한 자연어처리 패키지인 LTP(Language Technology Platform) 모델을 가져와서 다중논항 추출을 위한 파이썬 코딩을 하였다. LTP 프로그램은 중국어 단어 분리, 품사 표기, 개체명인식(NER), 통사구조 분석, 의미역 분석을 자동으로 처리해 주는 기능이 포함되어 있다. 이 중에서 본고에서 사용한 것은 단어분리(CWS: Chinese Word Segmentation), 의미의존분석(SDP: Semantic Dependency Parsing)과 의미역표기(SRL: Semantic Role Labeling) 알고리즘이다.⁴⁾

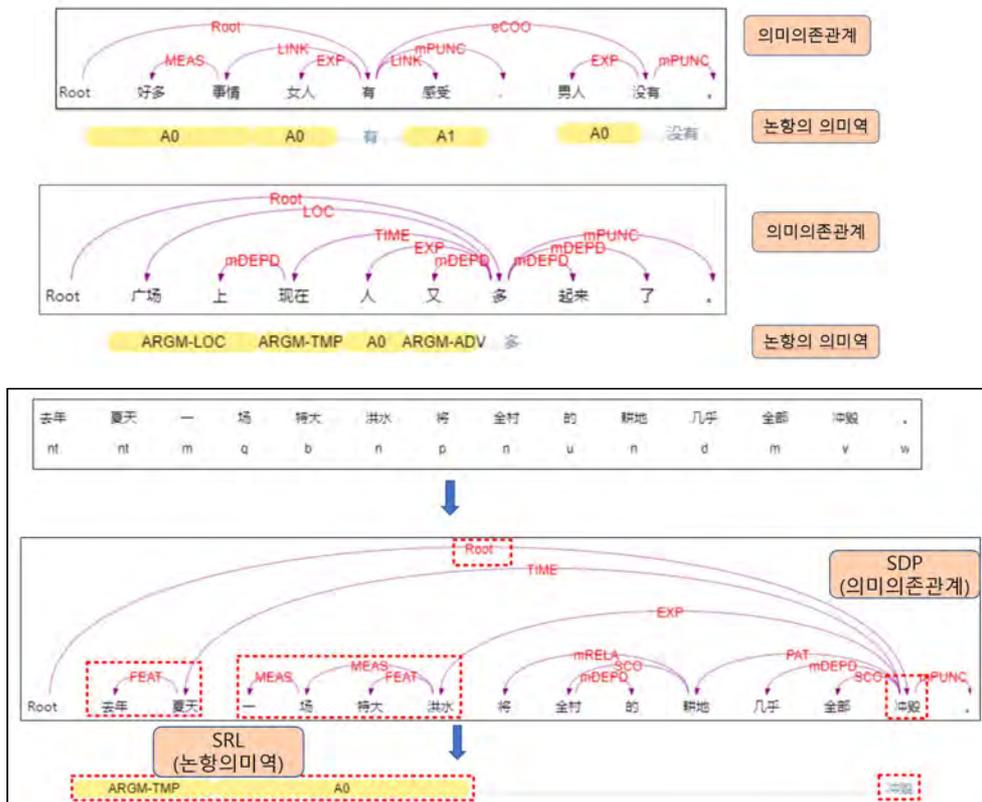
- (A) CWS: 단어분리(Chinese Word Segmentation)
 - 중국어 문장을 단어 단위로 분석하는 기능.
- (B) SDP: 의미 의존 분석(Semantic Dependency Parsing)

4) LTP와 관련된 코드와 자료는 <https://ltp.ai/>를 참고하기 바람.

- 문장의 각 언어 단위(단어) 사이의 의미 관련성을 분석하는 기능
- (C) SRL: 의미역 표기(Semantic Role Labeling)
 - 각 논항의 의미역을 식별하는 것. 예) 행위주, 피행위주, 대상, 시간, 처소 등



중국어 문장에서 단어와 의미 의존 관계, 의미역을 찾는 과정은 다음의 그림처럼 나타낼 수 있다. 예를 들어 “去年夏天一场特大洪水将全村的耕地几乎全部冲毁。”라는 문장은 LTP의 알고리즘에 의해 단어와 의미 정보가 자동으로 식별된다. 그 결과는 아래의 그림과 같이 “[去年 夏天]TIME [一场 特大 洪水]EXP [将全村的耕地]PAT [几乎 [全部 冲毁]ROOT].”로 분석된다.



<그림 3> LTP의 의미의존관계 및 의미역 분석 사례

위 그림에서 ‘ARGM-TMP’는 시간(TIME) 논항을 나타내고, ‘A0’는 필수논항 중에 대

상(EXP) 논항을 나타낸다. 즉 중심어(ROOT)인 ‘冲毁’ 앞에 2개의 논항이 출현한 것이 분석되어 해당 문장은 문두 다중논항 용례로 추출된다.

LTP 프로그램을 사용하면 각 문장을 읽어 들여 중심어를 찾을 수 있다. CWS와 SDP, SRL 기능을 활용하는 과정에서는 먼저 각 문장을 컴퓨터가 읽어들이고 후 중심어를 식별한다. 중심어는 문장의 핵심이 되는 중심술어를 가리킨다. 이 중심어를 기준으로 문장 내에서 논항을 식별하고 분석하는 작업이 이루어진다. 이때 논항의 위치와 의미관계를 정확하게 파악하는 것이 중요하다. 문두 다중논항 용례를 추출하기 위해서는 술어 앞에 위치하는 논항에 주목할 필요가 있다. 만약 술어 앞에 2개 이상의 논항이 위치한다면 해당 문장은 다중논항 후보로 볼 수 있다. 컴퓨터는 이런 후보 문장을 자동으로 추출한다. 이 과정은 파이썬 코딩을 통해 자동화하여 수행되었다. 설계된 파이썬 코드는 깃허브로 공개된 LTP 라이브러리와 연결하여 작성되었다.

본 연구에서는 문두 다중논항 추출을 위한 프로그램을 활용하여 중국어 말뭉치를 한 문장씩 조사하는 작업을 수행하였다. 그리고 해당 문장이 다중논항 구조를 가진다고 판단되면 이를 별도의 데이터로 구축하였다. 다중논항 후보로 판단된 문장에 대해서는 단어 정보, 의미 관계 정보, 의미역 정보를 추가적으로 분석하였다. 아래 그림에서 보듯이 데이터베이스에 포함된 정보는 원문(sentence), 단어 분리(segmentation), 의미의존관계(SDP), 의미역(SRL) 등이다. 이 데이터는 csv 파일 형식으로 저장되어 연구자가 엑셀을 통해 검토할 수 있도록 하였다.

↓

Index	Original Sentence	Segmentation	SDP Parsing	SRL Parsing	Root	Target from SDP	Target from SRL
0	从1989年持续到1989年19日, 中国向由2个国际团提出了1000多名学学生	从1989年持续到1989年19日, 中国向由2个国际团提出了1000多名学学生	[[1, 2, 'MRESL', (2, 1, 'TIME'), (3, 16, 'GTIME')]]	[[19, '[ARGM-ADV, [1989年持续到1989年19日], [ARGM-TMP, ...]]]]	Root	[[2, 8, 'TIME'], (3, 16, 'GTIME'), (7, 6, 'TIME'), ...]]	[ARGM-TMR AG, AD]
1	称不久, 网络上许多消息在网上一传播, 就引起人们	称不久, 网络上许多消息在网上一传播, 就引起人们	[[1, 8, 'TIME'], (2, 1, 'MPLUNC'), (3, 6, 'LOC'), (4, ...]]	[[17, '[ARGM-TMP, 称不久], [ARGM-LOC, 网络上, ...]]]]	Root	[[8, 0], [1, 8, 'TIME'], (3, 6, 'LOC'), (5, 8, 'AGT'), ...]]	[ARGM-TMR, ARGM-LOC, AG]
2	去平日的门中, 人们都感到	去平日的门中, 人们都感到	[[1, 2, 'FEAT'], (3, 10, 'TIME'), (3, 4, 'SCOT'), (4, ...]]	[[33, '[ARGM-TMP, 去平日], [ARGM-LOC, 门中], ...]]]]	Root	[[10, 0], [2, 10, 'TIME'], (4, 10, 'LOC'), (9, 10, 'EXP'), ...]]	[ARGM-TMR, ARGM-LOC, AG]
3	几星期, 一星期门的政工学生会组织新成员对工字楼和学校操场进行	几星期, 一星期门的政工学生会组织新成员对工字楼和学校操场进行	[[1, 2, 'MEAS'], (2, 26, 'TIME'), (3, 2, 'MLOC'), (4, ...]]	[[10, '[ARGM-ADV, 几星期], [AZ, 为政工学生会, ...]]]]	Root	[[26, 0], [2, 26, 'TIME'], (10, 11, 'CAP'), (13, 11, 'CON'), ...]]	[ARGM-TMR, AG, AD]
4	现在, 许多, 大部分为了, 冲毁, 许多, 大部分为了, 冲毁, 许多, ...	现在, 许多, 大部分为了, 冲毁, 许多, 大部分为了, 冲毁, 许多, ...	[[1, 2, 'TIME'], (2, 6, 'EXP'), (3, 2, 'PAT'), (4, ...]]	[[15, '[AT, 许多], [3, [ARGM-TMP, ...]]]]	Root	[[1, 2, 'TIME'], (3, 2, 'PAT'), (5, 4, 'AGT'), ...]]	[ARGM-TMR, AG, AG]
209	冲毁, 许多, 我们在郊区农村盖了一个四百	冲毁, 许多, 我们在郊区农村盖了一个四百	[[1, 2, 'FEAT'], (2, 8, 'TIME'), (3, 2, 'MPLUNC'), (4, ...]]	[[17, '[ARGM-TMP, 冲毁], [AZ, 盖, ...]]]]	Root	[[8, 0], [2, 8, 'TIME'], (4, 9, 'AGT'), (6, 7, 'LOC'), (7, 8, ...]]	[ARGM-TMR, AG, ARGM-LOC]
210	过, 许多年轻人从北京来到北京, 从北京来到北京, 许多年轻人, ...	过, 许多年轻人从北京来到北京, 从北京来到北京, 许多年轻人, ...	[[1, 8, 'TIME'], (2, 1, 'MDEPO'), (3, 8, 'EXP'), (4, ...]]	[[5, '[ARGM-ADV, 许多], [AZ, 许多], ...]]]]	Root	[[8, 0], [1, 8, 'TIME'], (3, 8, 'EXP'), (7, 8, 'CON'), ...]]	[ARGM-TMR, AG, AD]
211	大字生, 工, ...	大字生, 工, ...	[[1, 2], [AGT], (2, 4, 'MDEPO'), (3, 4, 'MLOC'), (4, ...]]	[[33, '[ARGM-EXT, 大字生], [AZ, ...]]]]	Root	[[2], 0], [1, 2], [AGT], (4, 2), (3, 'TIME'), (6, 4, 'TIME'), ...]]	[AG, ARGM-TMR, AG]
212	冲毁, 许多, 许多, 许多, 许多, 许多, 许多, ...	冲毁, 许多, 许多, 许多, 许多, 许多, 许多, ...	[[1, 2, 'FEAT'], (2, 6, 'SCOT'), (3, 2, 'MDEPO'), (4, ...]]	[[5, '[AZ, 冲毁], [ARGM-LOC, 许多], ...]]]]	Root	[[4, 0], [4, 6, 'EXP'], ...]]	[AG, ARGM-LOC, AG]
213	冲毁, 许多, 许多, 许多, 许多, 许多, 许多, ...	冲毁, 许多, 许多, 许多, 许多, 许多, 许多, ...	[[1, 8, 'EXP'], (2, 8, 'MDEPO'), (3, 8, 'MDEPO'), (4, ...]]	[[17, '[AZ, 冲毁], [ARGM-ADV, ...]]]]	Root	[[1, 8, 'EXP'], (3, 8, 'TIME'), ...]]	[AG, ARGM-TMR, AG]

<그림 4> 문두 다중논항 후보 DB

4. 중국어 의미 주석 말뭉치의 활용

4.1 방향보어 의미 주석 말뭉치의 활용

4.1.1 의미 주석 말뭉치를 이용한 단어 의미 중의성 해소

의미 주석 말뭉치는 다양한 분야에서 활용될 수 있다. 자연어처리 분야에서는 질 좋은 의미 주석 말뭉치가 있다면 언어 처리의 정확도를 높이는데 사용될 수 있다. 예컨대 다의어 의미 주석 말뭉치는 단어 의미 중의성 해소(WSD)에 중요한 역할을 한다. NLP에서 WSD는 주어진 문맥에서 단어에 적절한 의미를 할당하는 기본적인 작업이다. 중국어의 많은 기능어들은 방향보어를 포함하는 다중 의미 때문에 의미 판단이 모호한 상황이 종종 발생한다. 충분한 양의 의미 주석 말뭉치는 문맥에서 방향보어의 의도된 의미를 식별하는 데 유용한 데이터가 될 수 있다.

본 연구에서는 중의성 해소를 위한 실험 과정에서 중국어 방향보어 의미 주석 말뭉치를 사용하였다. 이를 통해 의미 주석 말뭉치에 포함된 방향보어의 문법 정보가 중의성 해소에 미치는 영향을 고찰하였다. 비록 수동으로 의미를 주석하는 과정은 시간이 많이 소요되는 작업이지만, 작은 규모라도 정확한 데이터는 높은 성능을 도출할 수 있다는 점을 확인하였다.

더 나아가, 전이 학습(transfer learning) 방법론을 통해 수동으로 주석된 말뭉치를 활용하여 사전 훈련된 언어 모델의 성능을 향상시킬 수 있다. 전이 학습은 한 도메인에서 획득한 지식을 다른 관련 도메인으로 전달하는 딥러닝 기술로, 사전 훈련된 데이터를 기반으로 새롭게 구축된 데이터를 추가적으로 학습함으로써 학습의 효율성을 증가시킨다. 이 방법을 통해 새로운 도메인에서의 모델 성능을 효과적으로 향상시킬 수 있다.

본 연구에서는 중국어 방향보어의 의미 분류 정확도를 향상시키기 위해 사전 훈련된 BERT 모델과 수작업으로 구축된 의미 주석 말뭉치를 결합한 전이 학습 방법을 적용하였다. 구체적으로, 중국어 BERT 사전 훈련 데이터와 방향보어에 대한 의미 주석 말뭉치를 활용하여 전이 학습 과정을 수행하였다.

연구에 사용된 기본 모델은 MacBERT-large로, 이는 BERT 분류 모델을 중국어의 특성에 맞게 개선시킨 버전이다. MacBERT-large는 다양한 중국어 자연어처리 작업에서 우수한 성능을 보이는 것으로 알려져 있다. 본 연구의 학습 과정은 크게 두 단계로 구성된다. 첫째, 대규모 데이터를 통해 기본적인 언어 정보를 학습하고, 둘째, 방향보어의 의미 분류를 위한 대상 데이터를 학습하는 것이다.

미세 조정 단계에서는 BERT 분류 모델을 활용하여 문맥 내에서 방향보어의 다양한 의미를 적절히 구분하는 훈련을 진행하였다. 훈련 데이터로는 각 방향보어마다 5,000개의 문장을 사용하였으며, 테스트 데이터로는 각 방향보어별로 500개의 문장을 활용하였다. 다음은 미세 조정 과정을 마친 후 테스트 문장을 예측한 결과이다.

	정확도 (Accuracy)	정밀도 (Precision)	재현율 (Recall)	F1 스코어 (F1 score)
上	91.1%	92.0%	90.2%	91.1%
起来	96.9%	96.4%	97.9%	97.2%
下来	94.6%	94.0%	89.9%	91.9%
下去	97.1%	94.0%	97.4%	95.7%

<표 15> 방향보어 의미 분류 테스트 결과

위의 표에서 볼 수 있듯이, 방향보어에 대한 의미 분류의 정확도는 상당히 높다. 이 결과는 의미 주석 말뭉치가 방향보어의 의미적 중의성 해소에 기여할 수 있음을 증명한다.

본 연구에서는 실험을 통해 방향보어에 대한 5,000개의 훈련 데이터를 사용하여 90% 이상의 의미적 중의성을 정확하게 해결할 수 있음을 발견했다. 다만 각 방향보어의 의미적 특성에 따라 모델의 성능에 일부 차이가 나타나는 점은 주목할 만하며, 이에 대한 보다 심층적인 분석이 필요할 것으로 보인다.

방향보어 의미 주석 말뭉치를 활용한 실험 결과를 구체적인 예문과 함께 살펴보기로 하겠다. 본고에서는 편쪽의 제한상 ‘上’과 ‘下来’를 중심으로 살펴보기로 하겠다. 중국어 방향보어 ‘上’은 크게 네 가지 의미로 분류할 수 있다. 첫째, ‘上1’은 이동의 의미로, 목적어로는 주로 처소 의미를 나타내는 성분이 온다. 둘째, ‘上2’는 행위 결과가 있고 그 상태가 유지되는 결과의 의미이다. 셋째, ‘上3’은 행위가 완성됨을 나타내는 완성의 의미이다. 넷째, ‘上4’는 해당 상태로 진입하여 그런 상태가 시작되어 지속된다는 시작의 의미를 나타낸다. 딥러닝 모델은 대부분의 경우에 이러한 의미를 잘 예측해 낸다. 아래에 제시된 예문들을 통해 이를 확인할 수 있다.

(1) 战士们淋着雨，爬[上]了陇东高原。

- ① (방향이동): 99.87% ② (결과): 0.08%
 ③ (완성): 0.02% ④ (시작): 0.02%

(2) 玛丽疲倦地闭[上]眼睛，她等着那最后一片叶子掉下来。

- ① (방향이동): 0.07% ② (결과): 99.89%
 ③ (완성): 0.03% ④ (시작): 0.01%

(3) 在目前，我的努力方向，是把地里都打[上]水井，买[上]水车。

- ① (방향이동): 0.07% ② (결과): 0.07%
 ③ (완성): 99.75% ④ (시작): 0.11%

(4) 后来我才知道，心香考[上]了大学，她去上大学了。

- ① (방향이동): 0.07% ② (결과): 0.05%
 ③ (완성): 99.72% ④ (시작): 0.17%

(5) 就这样，我喜欢[上]了北京，爱[上]了北京。

- ① (방향이동): 0.04% ② (결과): 0.02%
 ③ (완성): 0.04% ④ (시작): 99.90%

(6) 当[上]了经理后，我更加努力工作。

- ① (방향이동): 0.06% ② (결과): 0.02%
 ③ (완성): 0.11% ④ (시작): 99.81%

위에서 보이듯이 “爬上了陇东高原”에서는 ‘上’이 이동의 의미로 사용되었으며, 모델이 각각 99.87%의 높은 확률로 이를 예측하였다. “闭上眼睛”에서는 ‘上’이 결과의 의미로 쓰였고, 모델은 99.89%의 확률로 이를 정확히 예측하였다. “打上水井，买上水车”와 “考上了大学”에서 ‘上’은 완성의 의미를 나타내며, 모델은 각각 99.75%와 99.72%의 확률로 이를 예측하였다. 마지막으로 “喜欢上了北京，爱上了北京”과 “当上了经理”에서는 ‘上’이 시작의 의미로 사용되었고, 모델은 99.90%와 99.81%의 높은 확률로 이를 예측하였다. 이러한 예측 결과는 딥러닝 모델이 중국어 방향보어 ‘上’의 다양한 의미를 문맥에 따라 효과적으로 식별할 수 있음을 보여준다.

중국어 방향보어 ‘下来’는 크게 네 가지 의미로 분류할 수 있다. 첫째는 ‘下来1’로, 이는

위에서 아래로의 방향 이동을 나타낸다. 둘째는 ‘下来2’로, 어떠한 행위나 사건의 결과로 인해 발생한 상태를 나타내는 의미이다. 셋째는 ‘下来3’으로, 이는 새로운 상태로 변화하여 그 상태가 지속되는 의미를 나타낸다. 넷째는 ‘下来4’로, 주로 문두에 위치하여 담화표지로 기능하며, 이전 담화에서 언급된 내용을 바탕으로 이후의 담화를 이어 나간다는 의미를 나타낸다. 이처럼 ‘下来’는 다양한 의미 기능을 가지고 있으며, 문맥에 따라 적절한 의미로 해석된다. 실험 결과 딥러닝 모델은 이러한 ‘下来’의 다양한 의미를 대체로 정확하게 예측해 내는 것으로 나타났다. 아래에 제시된 예문들을 통해 모델의 예측 성능을 구체적으로 확인할 수 있다.

(7) 兄弟是朋友, 就是天塌[下来], 你帮忙撑一会儿, 又算什么呢?

- ① (방향이동): 99.88% ② (결과): 0.07%
 ③ (지속): 0.03% ④ (담화표지 지속): 0.03%

(8) 我已经把他的电话号码记[下来]了。

- ① (방향이동): 0.03% ② (결과): 99.87%
 ③ (지속): 0.06% ④ (담화표지 지속): 0.03%

(9) 天色暗了[下来], 老韩念叨着。

- ① (방향이동): 0.03% ② (결과): 0.09%
 ③ (지속): 99.84% ④ (담화표지 지속): 0.04%

(10) 接[下来]我们就一起看看2010年中国旅游大事记。

- ① (방향이동): 0.14% ② (결과): 0.09%
 ③ (지속): 0.14% ④ (담화표지 지속): 99.63%

위의 예에서 “天塌下来”에서 ‘下来’는 방향 이동의 의미로 사용되었고, 모델은 99.88%의 높은 확률로 이를 예측하였다. “记下来”에서는 ‘下来’가 행위의 결과를 나타내며, 모델은 99.87%의 확률로 예측하였다. “天色暗了下来”에서 ‘下来’는 새로운 상태로의 변화와 지속을 의미하고, 모델은 99.84%의 확률로 이를 예측하였다. 마지막으로 “接下来我们就一起看看2010年中国旅游大事记”에서는 ‘下来’가 담화표지로 사용되어 담화의 지속을 나타내며, 모델은 99.63%의 높은 확률로 이를 예측하였다.

위의 예시에서 보듯이, 미세 조정 과정을 거친 후 컴퓨터는 방향보어의 의미를 매우 정확하게 예측했다. 고품질의 의미 주석 말뭉치가 있다면, 동일한 방법을 다기능성 보어의 중의성 해소 작업에 적용할 수 있다. 이는 궁극적으로 자연어 처리 시스템의 성능을 향상시키는데 기여할 것이다.

4.1.2 방향보어 교육과 학습을 위한 의미 주석 말뭉치

의미 주석 말뭉치는 언어학적 연구 가치뿐만 아니라 중국어 교육 분야에서도 상당한 잠재력을 지니고 있다. 중국어 교육 과정에서 학습자와 교육자가 직면하는 어려움을 해소하는 데 중요한 역할을 할 것으로 기대된다.

한국의 중국어 교육 현황을 살펴보면, 매년 200개 이상의 대학에서 중국어 교육이 이루어지고 있으며, 10만 명 이상의 학생이 HSK(汉语水平考试)를 응시하고 있다. 이는 한

국이 세계에서 HSK 응시자 수가 가장 많은 국가임을 보여준다. 그러나 현재의 중국어 학습 환경에서는 몇 가지 중요한 문제점이 존재한다:

(가) 교재의 한계: 대부분의 학습자가 의존하는 종이 교과서와 참고서는 중국어 어휘와 문법에 대한 충분하고 다양한 설명과 예시를 제공하지 못하고 있다.

(나) 방항보어의 복잡성: 방항보어는 그 파생 의미로 인해 한국 학습자들에게 가장 어려운 문법 개념 중 하나로 꼽힌다.

(다) 교재 내용의 부족: 대부분의 교재는 방항보어의 기본적인 이동 의미만을 다루고 있어, 학습자들이 문맥에 따른 파생 의미를 스스로 추측해야 하는 상황이다.

(라) 체계적인 설명의 부재: 방항보어의 복잡한 의미 유형과 고빈도 기능에 대한 체계적인 설명이 교재에 충분히 제공되지 않고 있다.

(마) 교수법의 한계: 적절한 교육 자료의 부족으로 인해 교수자의 설명도 주관적이거나 비체계적일 수 있다.

의미 주석 말뭉치를 활용할 때 중국어 교육 분야에서 가지는 긍정적 효과는 아래와 같다.

첫째, 교수자는 의미 유형 분류와 방항보어의 빈도분석을 통해 각 방항보어의 핵심 의미와 의미 자질을 교육할 수 있다. 실제로 TCSSL에서는 높은 사용빈도를 가진 의미가 우선적으로, 집중적으로 학습되어야 한다. 게다가 의미에 따른 출현빈도 분석은 중국어 교과서나 문법서의 집필에 도움이 될 것이다.

두 번째로 방항보어와 동사의 배열 경향성을 파악하고 방항보어의 메커니즘을 이해할 수 있다. 학생들은 일반적으로 동사구에 집중하여 문장을 만든다. 따라서 동사와 방항보어의 결합에 관한 정보는 학생들이 문장에서 적절한 방항보어를 선택하고 사용하는데 가이드라인을 제공할 것이다.

세 번째로 의미 주석 말뭉치를 통해 방항보어 의미의 분포와 분류를 정의할 수 있다. 예를 들어, ‘上’이나 ‘下来’가 결과상태의 지속을 표현하는 것인지, 행위의 시작을 표현하기 위해 ‘起来’나 ‘上’ 중의 어느 것이 선택되어야 하는지를 알 수 있게 될 것이다.

네 번째로 의미 주석 말뭉치는 제2언어 교육 분야에서도 중요한 역할을 할 수 있다. 현대 교육 환경에서 온라인 교육이 새로운 표준으로 자리잡고 있는 가운데, 정보통신기술(ICT)에 기반한 교육 시스템의 급속한 발전으로 교육용 애플리케이션의 수요가 크게 증가하고 있다. 그러나 이러한 애플리케이션의 콘텐츠 품질은 여전히 개선의 여지가 있는 상황이다.

현재 대부분의 교육용 애플리케이션은 객관식 문제, 빈칸 채우기, 단어 맞추기 등 단순한 언어 연습에 국한되어 있다. 보다 높은 수준의 언어 교육을 위해서는 언어학적 지식과 컴퓨터 과학의 융합이 필수적이다. 이를 통해 언어-인식 지능형(language-aware intelligent) 언어 교육 시스템을 구축할 수 있을 것이다.

이러한 맥락에서 CALL(Computer Assisted Language Learning)과 ICALL(Intelligent Computer Assisted Language Learning) 시스템이 주목받고 있다. 특히 ICALL은 인공지능과 CALL 시스템을 결합하여 학습자와 지능적으로 상호작용하고, 학습 진도에 따라 유연하게 대응하는 소프트웨어를 제공한다. 이 시스템에서 학습자가 목표 언어로 문장을 입력하면, AI 튜터가 이를 분석하고 오류를 수정해주는 방식으로 작동한다.

본 연구에서 개발한 방향보어 의미 주석 말뭉치와 분석 결과를 이러한 시스템에 통합한다면, 학습자들은 중국어 방향보어를 더욱 효과적으로 학습할 수 있을 것이다. 예를 들어, 방향보어와 동사의 공기관계, 다양한 의미 유형 등에 대한 정보를 시스템에 반영함으로써, 학습자들은 더 많은 방향보어를 편리하고 효과적으로 학습할 수 있게 된다. 또한, 이렇게 개선된 시스템은 예문 검색 도구, 문법 검사기, 자동 작문 평가 등 다양한 학습 지원 기능으로 확장될 수 있다.

4.2 의미역 주석 말뭉치의 활용

4.2.1 중국어 문두 다중논항 어순 수용성 예측

중국어 의미역 주석 말뭉치는 자연어처리 기술과 결합할 때 문장의 어순 배열의 문법성을 효과적으로 판단하는데 기여할 수 있다. 본 연구에서는 이러한 의미 주석 말뭉치를 활용하여 중국어 문장의 어순 오류를 자동으로 감지하고 수정하는 방법을 탐구하였다. 이는 자연어처리 분야에서 문법 오류 수정(Grammatical Error Correction, GEC)이라는 중요한 주제에 해당한다. GEC는 문장 내 철자, 어휘, 문법 오류를 자동으로 식별하고 교정하는 기술을 말한다.

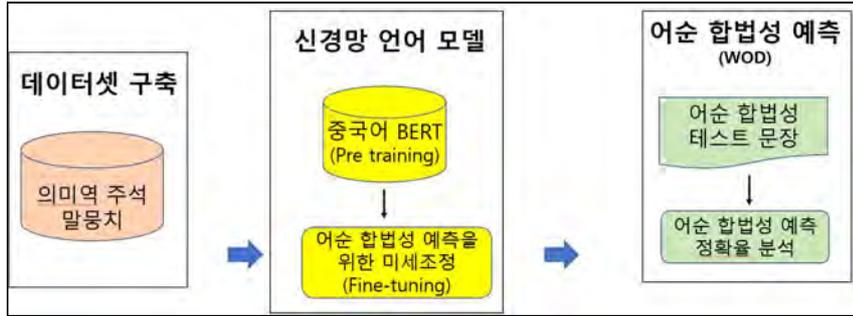
특히 본 연구는 GEC의 여러 영역 중에서도 중국어의 어순 배열에 초점을 맞추었다. 어순 배열은 중국어 문법의 핵심적인 요소로, Cheng et al.(2014)이 지적한 바와 같이 중국어에서 잘못된 어순을 식별하는 것은 중요하면서도 어려운 과제이다. Cheng et al.(2014)은 이러한 어순 오류를 WOE(Word Order Errors)로 명명하고 이를 식별하기 위한 방법론을 제시하였다.

그러나 선행 연구들은 중국어 문장에서 단어의 정확한 배열 순서를 감지하는 데 어려움을 겪었으며, 정확도가 80%를 넘지 못하는 한계를 보였다. 이는 기존 연구들이 주로 중국어의 기본적인 문장 성분 배열(주어-술어-목적어)에만 초점을 맞추었기 때문이다. 이러한 접근은 단순한 문장 구조에 대해서는 어느 정도 문법성을 파악할 수 있었지만, 복잡한 어순을 가진 문장에 대해서는 정확한 판단에 한계를 보였다.

이러한 배경에서 의미 주석 말뭉치를 활용하여 더 정교한 어순 오류 감지 및 수정 모델을 개발하고자 하였다. 의미 주석 말뭉치는 단순한 구문 구조를 넘어 단어와 구의 의미적 관계와 기능을 포함하고 있어, 복잡한 중국어 문장 구조를 더 정확히 분석하고 오류를 식별하는 데 도움이 될 수 있다. 이러한 접근은 기존의 GEC 시스템이 주로 표면적인 문법 규칙에 의존했던 것에서 벗어나, 문장의 의미적 구조와 맥락을 고려한 더 정교한 오류 감지 및 수정을 가능하게 한다.

본 연구에서는 중국어 문두 다중논항의 어순 합법성 예측이라는 특정 태스크에 적용하기 위하여 트랜스포머 아키텍처를 기반으로 하는 BERT언어 모델을 활용하였다. 이를 위해, 우선 문두 다중논항 의미주석 말뭉치를 활용하여 어순 분류를 위한 미세 조정(fine-tuning) 과정을 진행하였다. 이 연구에서는 특히 중국어 문두 다중논항의 어순 합법성을 예측하는 데 초점을 맞추어, 전이학습 방법을 통해 중국어에 특화된 모델을 구축하였다. 그 다음 단계로는 어순 합법성 테스트를 위해 별도의 데이터셋을 준비하여 모델의 성능을 평가하였다. 테스트 데이터셋을 이용한 실험을 진행하여, 본 연구의 모델이 중국어 문두 다중논항의 어순 합법성을 얼마나 효과적으로 예측할 수 있는지를 검증하였

다. 이러한 과정을 도식적으로 표현하면 다음과 같다.



어순 합법성 예측 실험은 몇 차례에 걸쳐 진행되었다. 이 실험의 주된 목표는 분류 정확도를 90% 이상으로 달성하는 것이었다. 이는 기존 연구에서 어순 오류 식별의 정확도가 대체로 80% 정도에 머물렀던 것을 고려할 때, 상당히 높은 수준이라 할 수 있다. 자연어처리 TASK마다 정확도의 목표치는 다를 수 있으나, 본 연구에서는 90%라는 목표치를 설정하여 그 달성 가능성을 탐구하고자 하였다.

첫 번째 실험, 즉 1차 파일럿 실험은 약 1년에 걸쳐 진행되었다. 이 기간 동안 연구팀은 중국어 말뭉치에서 추출한 문두 다중논항 데이터셋을 가공하고 코딩 설계 작업을 진행하였다. 연구의 핵심적인 부분인 모델 설계는 신경망 언어 모델을 중국어 어순 예측에 맞게 수정하는 것이었다. 이를 위해 파이썬 코딩을 적용하였다. 파일럿 실험은 몇 차례의 시도를 거쳐 점차 성능이 개선되었고 예측의 정확도가 90%에 근접하는 성과를 달성하였다. 이는 모델이 중국어의 복잡한 어순을 어느 정도 파악하고 예측할 수 있음을 시사한다. 그러나 1차 실험에서 정문과 비문의 비율 차이로 인해 예측의 편향성이 존재하는 문제점도 관찰되었다.

본고에서는 1차 실험의 문제점을 극복하고자 데이터 보완 작업을 진행하였다. 1차 실험에서 발견된 주요한 문제 중 하나는 데이터셋의 불균형성이었다. 이를 해결하기 위해 연구팀은 수개월에 걸쳐 데이터셋의 균형을 맞추는 작업을 수행하였으며, 이 과정에서 다양한 예문을 추가하였다. 이렇게 보완된 데이터셋은 2차 실험의 기반이 되었다.

2차 실험은 개선된 데이터셋을 활용하여 모델의 학습 능력과 예측 정확도를 높이고자 진행되었다. 새롭게 추가된 데이터셋은 더 다양한 어순과 구문 패턴을 포함하고 있어, 모델이 광범위한 언어 현상을 학습하고 이해하는 데 도움이 되었다. 결과적으로 이러한 데이터 보완 작업을 통해 예측의 편향성을 줄이는 동시에 분류의 정확도를 높일 수 있었다.

어순 수용성 예측 실험의 예측 결과에 대한 평가는 원어민의 평가와의 일치 여부를 기준으로 진행되었다. 컴퓨터는 신경망 모델에 기초하여 주어진 테스트 문장의 자연스러움의 정도를 확률로 판단한다. 컴퓨터는 확률이 50%를 넘는 쪽을 정답으로 예측한다. 어순이 문법에 맞을 확률이 50%를 넘으면 '1(정문)'으로 예측된다. 확률이 50% 미만이면 '0(비문)'으로 예측된다. 컴퓨터가 예측한 결과에 대한 평가는 원어민이 사전에 정답으로 표시한 것과의 일치 여부에 따라 이루어진다.

예를 들어 원어민이 정문으로 평가한 문장을 컴퓨터가 정문으로 예측했을 경우는 정확한 예측으로 분류된다. 원어민이 비문으로 평가한 문장을 컴퓨터가 비문으로 예측할 경

우에도 정확한 예측으로 분류한다. 그러나 만약 컴퓨터의 예측이 원어민의 평가와 일치하지 않는다면 그 예측은 틀린 것으로 분류된다. 다음의 표에서 보이듯이 원어민 평가와 컴퓨터 예측이 일치하면 'O'으로 표시하고 그렇지 않으면 'X'로 표시할 수 있다.

테스트 문장	원어민 평가	컴퓨터 예측	'1(정문) 확률	'0(비문) 확률	예측의 일치성
环境交通嘛, 这周边也还是不错的。	1(정문)	1(정문)	100.0%	0.0%	O
法官当天有一个非常重要询问。	1(정문)	1(정문)	99.9%	0.1%	O
肚子我还有点饿。	0(비문)	0(비문)	0.1%	99.9%	O
一个人我趴在床上。	0(비문)	0(비문)	0.1%	99.9%	O
这个人, 拥挤的乡村公共汽车上坐在她的被包上。	0(비문)	1(정문)	76.6%	23.4%	X
男女, 现实的人生里, 只是在彼此的寂寞里寻求温暖。	1(정문)	0(비문)	0.1%	99.9%	X
他, 全市敲锣打鼓公私合营第二天就哇哇哭了好一阵。	0(비문)	1(정문)	72.6%	27.4%	X
梯子上这时一个跟一个下来了一群人。	1(정문)	0(비문)	0.8%	99.2%	X

<표 16> 어순 수용성 예측 결과 평가 실례

본 연구에서 수행된 2차 실험은 데이터셋에 대한 미세 조정 후, 이를 테스트 문장에 적용하여 어순 예측 모델의 성능을 평가하는 것에 초점을 맞췄다. 평가 결과, 어순 예측 모델의 전체 정확도는 90%에 달했으며(1376/1529), 이는 모델이 전반적으로 높은 수준의 성능을 보이고 있음을 의미한다. 이는 연구 목표치에 근접한 결과로 해석될 수 있다.

구체적으로 모델의 성능을 분석해 보면, 정문(1)에 대한 예측의 정밀도는 93%로 나타났다, 비문(0)에 대한 예측의 정밀도는 85%로 측정되었다.

재현율 측면에서 볼 때, 정문의 재현율은 89%로, 비문의 재현율은 90%로 나타났다. 이러한 결과는 모델이 정문과 비문을 모두 잘 포착하고 있음을 시사한다.

F1 스코어는 정밀도와 재현율의 조화 평균을 나타내며, 이 두 지표가 얼마나 균형을 이루고 있는지를 평가한다. 본 연구에서 정문 예측의 F1 스코어는 91%, 비문 예측의 F1 스코어는 88%로 나타났다. 이는 모델이 정문과 비문 예측 모두에서 높은 수준의 균형 잡힌 성능을 보이고 있음을 의미한다.

	1(정문) 예측	0(비문) 예측
정밀도(Precision)	93%	85%
재현율(Recall)	89%	90%
F1 값(F1-score)	91%	88%
정확도(Accuracy)	90% (1376/1529)	

<표 17> 어순 수용성 예측 실험 결과

본 연구에서 수행된 실험의 결과를 종합해 볼 때, 의미역 주석 말뭉치는 모델의 전반적인 성능 향상에 크게 기여하였음을 확인할 수 있다. 이는 컴퓨터가 어순 예측을 할 때 높은 수준의 균형 잡힌 성능을 보이고 있음을 시사하며, 모델이 실제 언어 사용 환경에서 더욱 정확하고 신뢰할 수 있는 예측을 제공할 수 있음을 나타낸다.

4.2.2 문두 다중논항 어순 해석 모델 연구

인공지능 기술의 급속한 발전은 모델의 복잡성을 크게 증가시켰으며, 이와 함께 모델의 불투명성 또한 심화되었다. 이러한 배경에서 설명 가능한 인공지능(XAI: eXplainable Artificial Intelligence)이 인공지능 시스템의 이해도를 높이기 위한 중요한 연구 분야로 부상하게 되었다.

현재 대부분의 인공지능 모델은 ‘블랙 박스’로 여겨지고 있다. 이는 모델의 내부 작동 메커니즘이 명확하지 않아 해석하기 어렵다는 것을 의미한다. 이러한 불투명성은 모델의 결정 과정에 대한 이해와 신뢰성을 제한하는 주요 요인이 되고 있다.

본 연구에서는 이러한 문제의식을 바탕으로, 딥러닝 모델이 중국어 문두 다중논항의 어순의 합법성을 판단할 때 어떤 언어적 요소가 예측에 중요한 역할을 하는지를 고찰하고자 한다. 특히, 설명 가능한 인공지능(XAI)을 통해 문두 다중논항 어순의 해석에 어떻게 접근할 수 있는지에 대해 탐구하였다.

다양한 XAI 모델 중에서 본 연구는 SHAP(SHapley Additive exPlanations) 해석 모델에 초점을 맞추었다.⁵⁾ SHAP 모델을 중국어 문두 다중논항 구조를 가진 문장의 해석에 어떻게 적용할 수 있는지를 탐색하였다.

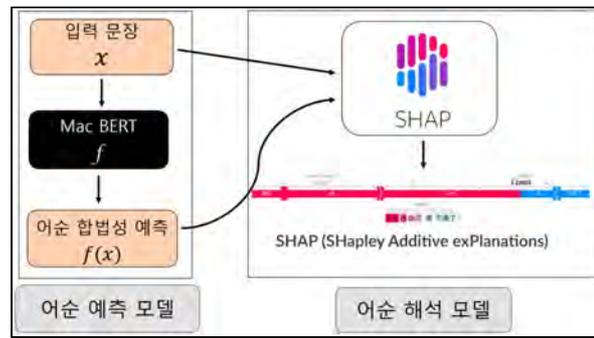
가. 중국어 데이터에 특화된 어순 해석 모델의 구현 과정

초기에 설명 가능한 인공지능(XAI) 분야에서 SHAP 프레임워크는 그 기여도를 쉽게 계산할 수 있는 수치 데이터 분석에 주로 활용되어 왔다. 그러나 최근에는 데이터의 종류와 형태가 다양해지면서 자연어처리 분야에서도 SHAP 프레임워크의 적용 가능성을 탐색하고 활용하는 연구가 증가하는 추세이다. 하지만 아직까지는 언어 데이터의 SHAP 프레임워크 적용이 주로 영어에 집중되고 있다. 이러한 현상은 자연어처리 연구 기반이 영어 중심으로 형성되었기 때문이다.

SHAP 프레임워크의 해석 모델을 적용하여 중국어 데이터 분석을 시도할 경우에는 중국어의 문장 구조가 영어와는 다르다는 점에서 추가적인 모델링 방법이 필요하다. 이에 본고에서는 중국어의 언어적 특성을 고려한 SHAP 프레임워크의 적용 방법을 탐색하였다. 특히 중국어의 문두 다중논항 어순을 정확하게 해석할 수 있는 모델의 설계와 구현에 중점을 두었다.

SHAP 해석 모델의 구현 과정은 아래의 도표로 간략하게 제시할 수 있다.

5) SHAP(SHapley Additive exPlanations)는 설명 가능한 인공지능 분야에서 핵심적인 기여를 한 프레임워크로, 로컬 해석 가능성을 위한 가장 인기 있는 프레임워크 중 하나로 꼽힌다. SHAP의 인기가 증가함에 따라, SHAP에 기반한 다양한 연구가 급증하고 있다. SHAP 프레임워크의 기본 모델을 설계한 Lundberg&Lee(2017)의 연구는 여러 학자들의 주목을 받았고 이를 활용한 연구가 큰 폭으로 늘어났다.



<그림 5> 어순 예측과 해석 모델의 구현 과정

나. SHAP 프레임워크에 기반한 중국어 문장의 자질별 기여도 분석

Lundberg&Lee(2017)에 의해 제안된 SHAP 프레임워크를 중국어 분석에 적용하는 과정은 자질(feature)의 기여도를 정량적으로 평가하여 중국어 문장 예측에 활용하는 방식으로 이루어진다. 이 과정에서 각 자질의 기여도는 해당 자질이 모델 예측에 미치는 영향력을 나타내는 Shapley 수치로 표현된다. 자질의 기여도는 모델 내 다른 자질과의 관계에 의존적이며, 이를 계산하기 위해서는 가능한 모든 자질의 부분집합에 대해 해당 자질이 모델 예측에 미치는 영향을 분석해야 한다.

이러한 분석을 위해, 특정 자질을 포함한 모델과 해당 자질을 제외한 모델을 비교하여 두 모델의 예측값 차이를 계산하는 과정이 포함된다. 이 차이는 자질의 기여도를 평가하는 데 사용되며, 모든 가능한 자질의 부분집합에 대해 계산된 후 Shapley 수치로 집계된다. Shapley 수치는 모든 가능한 차이의 가중 평균을 나타내며, 각 자질이 모델 예측에 미치는 평균적인 영향을 나타낸다. 계산된 Shapley 수치는 각 자질의 중요도를 나타내며, 중국어 문장의 어떤 요소가 예측 결과에 큰 영향을 미치는지 파악하는 데 활용될 수 있다. 이 수치는 특정 자질이 모델의 예측에 미치는 변화량을 해당 자질에 할당함으로써, 모델 예측 과정의 해석 가능성을 높인다.

중국어 문장 분석에서 자질(feature)의 정의와 활용은 데이터의 특성과 분석 목적에 따라 다양하게 이루어질 수 있다. 특히, 중국어와 같은 선형 데이터는 그 구조적 특성상 한자 형태소 단위, 단어 단위, 문장 성분 단위 등 다양한 형태로 자질을 정의할 수 있으며, 이러한 자질들은 토큰라이저를 사용하여 자동으로 분절되어 계산될 수 있다. 본고에서는 중국어의 한자뿐만 아니라, 한자보다 더 큰 단위인 단어와 논항을 기준으로 한 자질의 기여도를 계산하는 방법도 시도하였다.

다. Shapley 수치에 기초한 시각화 방법

Shapley 수치를 구하는 알고리즘은 몇 가지가 있는데 본고에서는 PartitionSHAP이라는 방법이 사용되었다. PartitionSHAP은 KernelSHAP의 처리속도를 개선한 버전으로서, 자질을 계층적으로 군집화하여 표현한다. 예를 들어, 감성 분석에서 PartitionSHAP을 사용한 알고리즘은 모델의 평균 예측을 나타내는 기본값과 각 특징(예: 단어)이 결과에 기여하는 방식을 보여줌으로써, 평균과 현재 결과 사이의 차이를 보여준다.

시각화 방법으로는 자질의 기여도를 그래프로 나타내는 plot 형태를 사용하였다. 아래에서는 text plot과 bar plot으로 시각화한 실례를 보인 것이다. 해당 문장이 정문으로 예

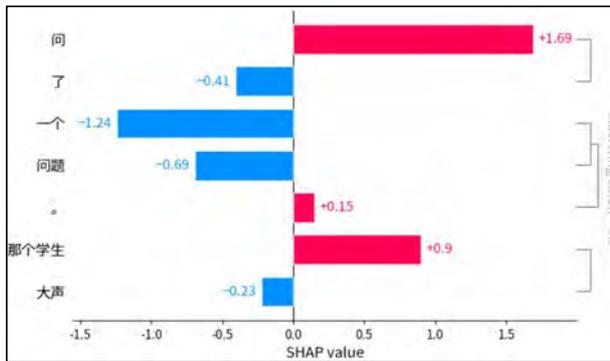
측될 때 SHAP 모델은 기여도를 계산한 다음 아래와 같이 시각화할 수 있다. 예를 들어 “那个学生大声问了一个问题。”는 정문일 확률이 99.8%이고 이때 기여도가 큰 성분은 ‘那个学生’과 ‘问’이다.

- (1) 입력 문장 : 那个学生大声问了一个问题。
 예측 결과: 정문 (정문일 확률: 99.8% / 비문일 확률: 0.2%)

a. text plot 형태의 시각화



b. bar plot 형태의 시각화



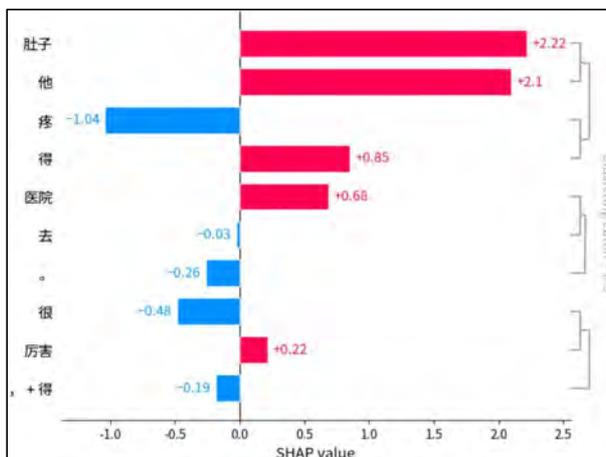
한편, 어순 예측 모델이 비문으로 판단할 때는 아래와 같이 분석된다. 예를 들어 “肚子他疼得很厉害，得去医院。”은 비문일 확률이 99.5%이고 이때 기여도가 큰 성분은 ‘肚子’와 ‘他’이다.

- (2) 입력 문장 : 肚子他疼得很厉害，得去医院。
 예측 결과: 비문 (정문일 확률: 0.5% / 비문일 확률: 99.5%)

a. text plot 형태의 시각화



b. bar plot 형태의 시각화



위의 text plot에서 base value는 모델이 해당 문장에 대해 기본적으로 예상하는 값이다. $f(\text{input})$ 은 문장의 각 자질(토큰)의 기여도를 모두 합한 최종 예측값을 나타낸다. box plot 시각화에서는 이러한 기여도를 색상과 막대의 길이를 통해 나타낸다. 빨간색은 기여도가 큰 자질을 표현하며, 막대의 길이가 길수록 해당 자질의 가중치가 크다는 것을 의미한다. 이는 해당 자질이 모델의 예측에 긍정적인 영향을 미치는 중요한 요소임을 나타낸다. 반대로 파란색은 기여도가 작은 자질을 의미한다. 이는 해당 자질이 모델의 예측에 미치는 영향이 상대적으로 적거나, 예측 결과에 부정적인 영향을 가짐을 나타낸다.

라. 한자 단위의 Shapley 수치의 계산과 시각화

MacBERT 모델은 기본적으로 한자 단위로 언어 학습이 이루어진다. 이에 1차적으로 예측 모델과 동일한 단위로 데이터의 기여도를 계산해 보았다. 본고에서는 SHAP 해석 모델을 구현할 때 BERT tokenizer를 사용하여 한자 단위로 데이터를 처리하도록 하였다.

실험 과정에서는 이전 연구에서 구축된 테스트 데이터를 활용하였다. 테스트 데이터는 약 1,500여 개의 문장으로 구성되어 있는데 이 중에서 비문의 해석에 초점을 맞추었다. 분석 대상으로 삼은 비문은 모두 600여 개이다. 각 비문에 대한 분석은 한 문장씩 따로 진행되었으며, 한자별 기여도를 시각화하는 방식으로 이루어졌다. 다음의 예를 보기로 한다.

(3) 입력 문장 : 肚子我还有点饿。(비문)



(4) 입력 문장 : 一个人我趴在了床上。(비문)



(5) 입력 문장 : 满脸全村农妇丰收喜悦。(비문)



한자 단위의 해석 모델은 기여도가 한자별로 계산된다. 때문에 문두 다중논항에서 비문의 판단에 기여하는 성분이 한자별로 표현된다. 이 방법은 토큰을 세분화하여 분석할 수 있게 해, 각 한자의 구체적인 기여도를 명확히 파악할 수 있다는 장점이 있다.

그러나 이러한 접근 방식에는 단점도 존재한다. 한자 단위로 기여도를 계산하는 방식은 때로 논항이나 단어 전체가 아닌 일부분만이 긍정적 기여를 한 것으로 보이게 만든다. 이는 인간의 자연스러운 언어 처리 방식과 차이가 있다. 인간은 문장을 해석할 때 단어나 논항을 전체적인 맥락 속에서 이해하고 판단하는 경향이 있다. 따라서, 한자 단위

계산 방식은 특정 부분의 기여도를 과도하게 강조함으로써, 모델의 해석 결과와 인간의 직관적인 이해 사이에 간극이 발생할 수 있다.

마. 논항과 단어 단위 Shapley 수치의 계산과 시각화

SHAP 프레임워크 내에서는 사용자가 직접 정의할 수 있는 토큰화 방식을 제공한다. 이를 위해 custom tokenizer를 새롭게 정의하여 사용했다. 특히, 어순 해석 모델을 개발하는 과정에서 논항 구분자를 식별하는 작업에 주목했다. 이를 통해 각 논항 성분을 분석의 독립된 단위로 설정했다. 그리고 논항이 아닌 나머지 성분들은 단어 단위로 분절하였다.

논항과 단어 단위로 기여도를 계산한 후, 어떤 언어 단위가 예측에 가장 큰 영향을 미쳤는지를 해석하는 과정은 시각화 방식으로 표현된다. 본고에서는 기본적으로 text plot을 사용하여 기여도를 시각화하였다. 긍정적 기여도가 높은 언어 단위는 빨간색으로 강조하여 시각적으로 표시했으며, 부정적으로 작용한 언어 단위는 파란색으로 나타냈다. 다음은 비문으로 예측된 문장에 대해 논항과 단어 단위로 기여도를 시각화한 실례를 보인 것이다.

(6) 입력 문장 : 嘴里她喊了声啪勾, 蹬车走了。(비문)



(7) 입력 문장 : 眼睛她呆呆地望着天。(비문)



(8) 입력 문장 : 我, 你们的工作已同袁政委谈好。(비문)



위의 예시를 통해 알 수 있듯이 SHAP 모델은 사용자가 정의한 방식으로 특성을 분할하여 기여도를 산출할 수 있는 유연성을 제공한다. 이러한 점은 어순 예측과 같은 작업에서 특히 유용하다. 예를 들어, 특정 논항(행위주, 경험주, 술어의 판단 대상 등)이 예측에 얼마나 긍정적으로 기여하는지를 수치화하여 나타내줌으로써, 어떤 요소가 예측에 중요한 역할을 하는지 파악할 수 있다. 예를 들어 “嘴里她喊了声啪勾, 蹬车走了”에서는 ‘嘴里’가 비문 예측에 중요한 요소가 된다. “我, 你们的工作已同袁政委谈好”에서는 ‘我’와 ‘你们的工作’가 비문 예측에 중요한 요소로 작용한다.

본고에서 채택한 SHAP 모델의 비문에 대한 해석은 인간의 언어학적 분석과 상당 부분 일치한다. 특히 의미역의 배열 순서에 대한 위계와 논항 간 의미관계의 어순 반영이라는 비가시적이고 난해한 어순 배열 규칙을 이해하고, 이에 기반하여 상당히 설득력 있는 비문 형성 요인을 짚어냈다는 점이 큰 의미가 있다. 즉, XAI의 모델이 인간 언어 데이터를 통해 어순의 특징과 제약을 습득하고, 이에 기반하여 비교적 설득력 있는 해석을

제공할 수 있음을 확인한 것이다. 이는 그간 AI 연구 영역에서 주목하던 처리 과정의 ‘블랙박스’를 이해하는데 일정 부분 기여할 수 있다는 점에서 AI 연구 영역에서 의미하는 바가 있다. 특히, 본고의 SHAP 모델 분석은 다양한 언어 요인을 반영하여 기여도에 기반한 예측을 제시한다는 점에서 합리적인 결과를 도출할 수 있을 뿐만 아니라, 이해가 쉽고 효과적인 시각화 방식을 통해 결과물을 출력한다는 점에서 언어 분석 및 교육 분야에도 중요한 기여를 할 것으로 기대된다.

다만 이러한 성과에도 불구하고 본 연구를 통해 XAI 모델의 한계점도 발견할 수 있었다. 분석 결과가 학습데이터 수량의 영향을 직접적으로 받는다는 점과 모델 분석에 영향을 미치는 요인이 다양한 경우 두드러진 기여 요소를 찾기 어렵다는 것이다. 이는 학습데이터 수량의 확장과 분석 목적에 따른 문장 구성 조정을 통해 보완되어야 할 것이다. 그러나 무엇보다 중요한 점은 XAI 모델의 분석이 인간 언어 분석의 직관에 상당히 부합하지는 않지만, 언어 현상에 대한 근본적인 원인이나 현상이 가지는 가치를 설명하기에는 한계가 존재한다는 점이다. 원인과 기제에 대한 설명 제시에 있어서 아직은 인간의 개입이 필요하며, 이는 XAI 모델 설계자와 사용자 역할의 중요성을 다시 한번 보여준다.

5. 맺음말

본고에서는 중국어 의미 주석 말뭉치의 구축과 활용에 관해 개괄적인 논의를 했다. 의미 주석 말뭉치는 언어학 연구, 자연어 처리, 언어 교육 등 다양한 분야에서 중요한 역할을 한다. 특히 복잡한 언어 현상을 분석하고 이해하는 데 있어 귀중한 자원이 된다.

중국어 의미 주석 말뭉치의 사례로는 두 종류의 데이터셋을 소개하였다. 첫 번째는 다의어 의미 주석 말뭉치의 사례로, 중국어 방향보어 의미 주석 말뭉치를 구축하고 분석하였다. 이를 통해 ‘上’, ‘起来’, ‘下来’, ‘下去’ 등 주요 방향보어의 의미 유형별 분포와 특성을 파악하였다. 이 과정에서 방향보어의 핵심 의미와 문법화 정도에 대한 새로운 인식을 제시할 수 있었다.

두 번째 사례는 의미역 주석 말뭉치로, 중국어 문두 다중논항의 의미역 주석 말뭉치를 구축하였다. 이를 통해 중국어 문두 성분의 의미역 분포와 배열 경향성을 분석하였다. 이는 중국어 어순 연구에 새로운 통찰을 제공하였다.

구축된 말뭉치를 활용하여 딥러닝 기반의 방향보어 의미 분류 모델과 문두 다중논항 어순 수용성 예측 모델을 개발하였다. 이 모델들은 90% 이상의 높은 정확도를 보였다. 또한 SHAP 프레임워크를 활용한 설명 가능한 AI 모델을 통해 중국어 어순 판단의 메커니즘을 해석하고자 시도하였다.

본 연구의 의의는 중국어 의미 주석 말뭉치가 언어학 연구, 자연어 처리, 중국어 교육 등 다양한 분야에서 중요한 역할을 할 수 있음을 실증적으로 보여준 데 있다. 특히 방향보어와 어순과 같은 복잡한 중국어 문법 현상을 분석하고 이해하는 데 있어 말뭉치의 중요성을 강조하였다.

이러한 연구 결과는 향후 중국어 교육 현장에서 학습자들의 이해를 돕는 데 활용될 수 있을 것이다. 또한 자연어 처리 기술의 발전에도 기여할 수 있을 것으로 기대된다. 특히 AI 모델의 언어 처리 과정에 대한 이해를 높임으로써, 더 정교하고 신뢰할 수 있는 언어 모델 개발에 도움이 될 것이다.

향후 연구에서는 더 큰 규모의 말뭉치 구축, 다양한 문법 현상에 대한 의미 주석 확장,

그리고 AI 모델의 해석 가능성을 더욱 높이는 방법 등을 탐구할 필요가 있다. 또한 구축된 말뭉치와 개발된 모델을 실제 중국어 교육 현장에 적용하는 연구도 진행되어야 할 것이다. 이를 통해 중국어 연구와 교육에 새로운 방향을 제시할 수 있을 것으로 기대된다.

참고문헌

- 강병규(2021), 「딥러닝 언어모델과 중국어문법」, 『중국문학』 제106집.
- 강병규(2024), 「딥러닝을 활용한 중국어 방향보어의 의미 추론 연구」, 『중국언어연구』 제112집.
- 박정구·강병규·박민준·유수경(2023a), 「말뭉치 기반 중국어 문두 다중논항의 의미역 배열 경향성 고찰」, 『중국언어연구』 106집.
- 박정구·강병규·박민준·유수경(2023b), 「어순 수용성 데이터 기반의 중국어 문두 다중논항 의미역 배열 기제 및 위계 연구」, 『중국언어연구』 109집.
- 박정구·유수경·박민준·강병규(2024a), 「신경망 언어 모델 구축을 통한 중국어 문두 다중논항 어순 수용성 예측 연구」, 『중국문학』 109집.
- 박정구·유수경·박민준·강병규(2024b), 「XAI 기반의 중국어 문두 다중논항 어순 해석 모델 연구」, 『중국언어연구』 제111집.
- 양천천(2019), 『중국어 보어의 학습방안 연구 :한국인 학습자의 오류 분석을 중심으로』, 동아대 박사학위 논문.
- 유수경(2020), 「중국어 방향보어의 상적 기능에 대한 고찰-‘下來’와 ‘下去’의 말뭉치 분석에 근거하여」, 『중국언어연구』 제89집.
- 최규발(2008), 「방향보어의 문법적 의미」, 『중국학논총』 제24집.
- 한장훈, 박영준, 정영훈(2019), 「순차적 구문 분석 방법을 반영한 포인터 네트워크 기반의 한국어 의존 구문 분석기」, 한국어 정보처리학회 발표자료.
- Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G.(2020), Revisiting Pre-Trained Models for Chinese Natural Language Processing, arXiv preprint arXiv:2004.13922.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018), BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Geng, Y., Chen, J., Ye, Z., Yuan, Z., Zhang, W., & Chen, H. (2020), Explainable zero-shot learning via attentive graph convolutional network and knowledge graphs.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018), Explaining explanations: An overview of interpretability of machine learning.
- Jeong, Y.H., Li, M.Y., Kang, S.M., Eum, Y.K., Kang, B.K. (2022). Automatic Prediction and Linguistic Interpretation of Chinese Directional Complements Based on BERT Model. In: Dong, M., Gu, Y., Hong, JF. (eds) Chinese Lexical Semantics. CLSW 2021. Lecture Notes in Computer Science, vol 13249. Springer, Cham.
- Jin, X., Wei, Z., Du, J., Xue, X., & Ren, X. (2019), Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models.

- arXiv preprint arXiv:1911.06194.
- Kang, B., Yu, S. (2023), Building a Semantically Annotated Corpus of Chinese Directional Complements. In: Su, Q., Xu, G., Yang, X. (eds) Chinese Lexical Semantics. CLSW 2022. Lecture Notes in Computer Science, vol 13496. Springer, Cham.
- Li, L., Dai, Y., Tang, D., Qiu, X., Xu, Z., & Shi, S. (2023), MarkBERT: marking word boundaries improves Chinese BERT. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 325-336). Cham: Springer Nature Switzerland.
- Lundberg, S. M., & Lee, S. I. (2017), A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Mangalathu, S., Hwang, S. H., & Jeon, J. S. (2020), Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Engineering Structures*, 219, 110927.
- Molnar, C. (2022), *Interpretable Machine Learning*(Second Edition), Independently published.
- Samek, W., Montavon, G., Vedaldi, A., & Müller, K.-R. (2017), Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges. arXiv preprint arXiv:1708.08296.
- Shapley, L. S. (1951), Notes on the N-Person Game – II: The Value of an N-Person Game. Santa Monica, CA: RAND Corporation, https://www.rand.org/pubs/research_memoranda/RM0670.html. Also available in print form.
- Si, S., Wang, R., Wosik, J., Zhang, H., Dov, D., Wang, G., & Carin, L. (2020), Students need more attention: Bert-based attention model for small data with application to automatic patient message triage. In *Machine Learning for Healthcare Conference* (pp. 436-456). PMLR.
- Xiao, Zhonghua & Anthony McEnery(2004), *Aspect In Mandarin Chinese : A Corpus-based Study* , John Benjamins.
- Zeiler, M. D., & Fergus, R. (2014), Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833). Springer, Cham.
- 陈前瑞(2008),《汉语体貌研究的类型学视野》,商务印书馆.
- 丁银贞(2010),《韩国学生学习汉语语法的困难和对策》,华东师范大学博士学位论文.
- 刘月华(1998),《趋向补语通释》,北京语言大学出版社.
- 刘月华(2001),《实用现代汉语语法》,商务印书馆.
- 刘磊(2016),《基于HSK动态作文语料库的汉语趋向补语习得偏误分析》,安阳师范大学博士学位论文.
- 卢英顺(2000),〈现代汉语中的“延续体”〉,《安徽师范大学学报》 Vol.28.
- 吕叔湘 主编(1999),《现代汉语八百词》,商务印书馆.

〈중국어 의미 주석 말뭉치 구축과 활용〉의 토론문

오유정(고려대)

본 연구는 급변하는 연구 환경 속에서 최신의 디지털 분석 툴을 활용한 언어 분석의 실질적인 사례를 제시하고 있다는 점에서 매우 큰 의미를 가집니다. 특히 단순히 구축된 말뭉치를 활용하는 것에 그치지 않고, 연구 목적에 따라 필요한 말뭉치를 새롭게 구축하고 활용하여 중국어의 어순 문제 등 접근하기 까다로운 언어 현상에 대해 새로운 분석의 툴을 제공하였다는 점에서 시사하는 바가 크다고 생각됩니다. 논의에 대한 구체적인 문제 제기보다는, 디지털인문학이 화두가 되고 있는 현 시점에 다른 언어에 비해 비교적 보수적인 중국어학계의 입장에서, 여전히 전통적인(?) 언어학의 연구 방식에서 벗어나지 못한 개인적인 궁금증을 여쭙는 것으로 토론을 대신하고자 합니다.

먼저, 기존의 말뭉치 활용이 주로 품사태깅의 수준에 머물렀다는 점에서 본 연구에서 보여주시는 의미 주석 말뭉치는 매우 의미있는 방향성이 될 수 있다 생각됩니다. 다만 이 역시 결국 사람의 손을 통해 이루어져야 한다는 점에서 이를 어디까지 확장하여 활용할 수 있을지, 특히 개인이 연구에 적용할 수 있을지 확신이 서지 않습니다. 기존의 품사태깅 역시 적용 기준에 따라 차이가 많고, 오류도 많아, 실제 활용시 이를 처리하는 데 상당한 시간이 소요되고 있습니다. 제시해주신 의미 주석과 관련하여서도 당장 ‘의미 관계에 대한 기준의 확립’, ‘중의적 해석에 대한 처리’, ‘휴먼 에러의 가능성’ 등 여러 요인에 대한 통제 가능성을 고민하지 않을 수 없어 보입니다. 또한 적용 가능한 분석 대상에도 차이가 있어 보이는데, 연구 사례로 제시해주신 방향보어와 유사하게 결과보어에 대해서 역시 분석의 툴을 적용할 수 있을까요? 나아가 데이터셋 구축에 막대한 시간과 노력이 소요되는 것을 고려하면, 이러한 자료에 대한 공유가 점점 더 중요해질 것으로 보이는데, 실제로 데이터처리 관련 분야에서는 연구 성과의 공유가 보편화되어있다고 들었습니다만, 비교적 보수적인 중국어학계 역시 이러한 성과의 공유가 가능하다고 보시는지요?

의미 주석 말뭉치를 활용한 다양한 연구 사례를 통해 디지털 툴의 언어학적 활용 방안에 대해 여러 가능성을 보여주셨습니다. 특히 이러한 분석 툴이 언어 현상을 분석하고 이해하는데 어떠한 기여를 할 수 있을지 끊임없이 고민하신 것을 느낄 수 있었습니다. 다만, 그럼에도 여전히 자연어처리의 정확도 향상에 대한 논의가 상당 부분을 차지하고 있다는 점에서, 향후 언어학계가 나아가야할 방향성에 대한 고민을 하지 않을 수 없어 보입니다. 자연어처리가 추구할 방향과 전통적 언어학이 추구할 방향에는 분명한 차이가 존재하고, 후자의 입장에서는 어떻게 하면 이러한 툴을 활용해 기존의 연구 방법으로는 파악하기 어려웠던 언어 원리와 기제를 발견하고 설명할 수 있을까 하는 것에 연구의 포커스가 맞춰져야 한다는 생각입니다. 조금 거칠게 말하면, 우리의 연구적 관심은 인간이 “왜” 그러한 언어를 구사하는 것인지에 대한 고민이지, 어떻게 하면 기계가 인간과 같은 언어를 구사하고 예측하도록 할지에 대한 것이 아니라는 점에서 지속적인 갈등이 있는 것 같습니다. 이에 대한 선생님의 고견을 구합니다.

자동 평가 지표를 활용한 한중 기계번역 평가 및 포스트에디팅 연구: TV 방송 대화를 중심으로

황은하(배재대)

차 례

1. 머리말
 2. 선행 연구
 3. 연구 대상과 방법
 4. 포스트에디팅에 대한 양적·질적 분석
 5. 맺음말
-

1. 머리말

이 연구는 TV방송 대화 번역말뭉치의 한중 기계번역문을 평가하고, 포스트에디팅 양상을 추적하고 분류하여 정리함으로써, TV방송 대화의 한중 기계번역에 필요한 포스트에디팅 가이드라인을 제시하는 것을 목적으로 한다. 이를 위해 방송 콘텐츠 한중 번역말뭉치 60만 문장쌍¹⁾에 포함된 기계번역문에서 포스트에디팅 비율이 높은 번역문을 자동 평가의 방식으로 추출하고, 이를 최종 번역문과 비교하여 포스트에디팅 내용을 확인하고 유형화한다. 포스트에디팅 비율이 높은 번역문의 추출을 위해서는 sacreBLEU, METEOR 등 자동 평가 지표를 이용한다.

최근 몇 년간 인공지능 학습용 빅데이터의 구축과 가공 산업 또한 빠른 속도로 성장해 왔다. 이를 통해 한국어와 다양한 언어 간 기계번역 학습용 데이터가 대규모로 구축되어 서비스되고 있는데 이러한 데이터는 기계번역 - 포스트에디팅 - 검수 등의 과정을 통해 생산되며(한국정보화진흥연구원, 2022), 그 규모가 매우 커서²⁾ 인간 평가자가 일일이 검수할 수 있는 범위를 훨씬 넘어섰다. 인공지능의 성능이 빅데이터의 규모와 품질에 정비례한다는 점을 감안하면, 단기간에 대규모로 생산되는 기계번역 학습용 데이터의 번역문의 단,계별 품질 관리는 매우 중요한 과제라고 하겠다. 따라서 기계번역의 품질을 확인

1) 한국지능정보사회진흥원(NIA)에서 운영하는 AI 허브 서비스(<http://www.aihub.or.kr>)에서 내려받음.
2) 본 연구에서 사용한 말뭉치 구축 과제인 ‘2022년 방송 콘텐츠 한국어-아시아어 번역말뭉치’의 과제 규모를 예로 들면, 6개월 내 한국어를 북경어, 일본어, 대만어, 몽골어, 힌디어로 번역한 음성 3,000 시간 분량과 이를 전사한 원문-기계번역문-최종 번역문 250만 문장쌍을 포함하고 있다.

하고, 포스트에디팅 가이드라인을 작성하는 일은 매우 필요하다.

한편, AI 학습용 빅데이터는 대부분 실제(authentic) 언어 자료를 수집하여 컴퓨터가 읽을 수 있는(machine-readable) 형태로 가공되었다는 점에서 본질적으로 언어학이나 번역학의 주요 연구 기반 자료인 말뭉치와 맥을 같이 한다. 특히 AI 학습용 번역말뭉치는 원문과 번역문을 포함하고 있고 문장 대 문장으로 정렬이 되어 있다는 점에서 언어 간 연구에 필수 재료인 병렬말뭉치 그 자체이다. 따라서 언어학, 번역학 연구에서 기구축된 AI 학습용 데이터의 활용 가능성과 방안을 탐색하는 일도 시도해 볼 만한 일이다. 본 연구에서 다루고자 하는 방송 콘텐츠 한중 번역 말뭉치는 한국어 TV방송 대화 원문, 기계번역문과 기계번역을 토대로 포스트에디팅된 최종 번역문을 포함하고 있어, TV방송 대화의 기계번역 성능과 포스트에디팅의 내용을 추적할 수 있는 좋은 자료로 판단된다.

한편, TV방송 대화는 비격식 대화로, 그에 대한 기계번역 연구는 Youtube, TikTok 등 동영상과 숏폼 콘텐츠 시장의 급속 성장과 더불어 동영상과 숏폼의 한중 번역의 품질을 높이기 위해서도 필요한 연구라고 하겠다.

본 연구는 바로 위와 같은 연구의 필요에서 출발하여, AI학습용 방송 콘텐츠 번역말뭉치를 대상으로 TV방송 대화의 한중 기계번역 평가와 포스트에디팅에 대한 연구를 수행한다. 이를 위해, 2장에서는 선행 연구를 검토하고, 3장에서는 연구 대상과 방법을 소개하며, 자동 평가 지표를 적용하여 포스트에디팅 비율이 높은 기계번역문과 해당되는 한국어 원문, 최종 번역문을 추출한다. 4장에서는 포스트에디팅에 대한 양적, 질적 분석을 수행하는데, 질적 분석을 위해 포스트에디팅 비율이 높은 기계번역문을 추출하여 최종번역문을 기준으로 주요 포스트에디팅 내용을 유형화하고 귀납한다. 끝으로, 5장에서는 본 연구를 요약하고 남은 과제를 제시하는 것으로 결론을 대신한다.

이 연구는 상대적으로 연구가 많지 않은 구어, 특히 TV 방송 대화를 중심으로 하는 한중 기계번역 자동 평가와 포스트에디팅 연구인 동시에, AI 학습용 번역 빅데이터의 번역학적 활용 연구의 성격을 지닌다.

2. 선행 연구

기계번역과 관련된 연구는, 시스템의 개발, 평가, 프리에디팅, 포스트에디팅을 아우른다. 다만, 여기서는 본 연구와 같은 한중 언어쌍의 기계번역 평가와 포스트에디팅을 중심으로 살펴보기로 한다.

2.1. 한중 기계번역 평가 연구

번역에 대한 평가는 평가의 주체에 따라 인간 평가자에 의한 인간 평가와 도구에 의한 자동 평가로 나눌 수 있다. 인간 평가는 주로 원문을 기준으로 번역문의 의미적 정확성을, 목표어의 언어적 특성을 기준으로 번역문의 유창성과 자연스러움 등을 판단하는 방식이다. 이에 반해, 자동 평가는 자동화된 알고리즘을 통해, 기계번역문이 정확한 번역문에 준해 얼마나 일치한지를 계량화하여 번역의 정확성을 측정한다. 인간 평가는 의미·화용적 층위에 대한 평가가 가능하다면, 자동 평가는 주로 참조 번역문과 비교해 형태적으로 얼마나 다른지를 근거로 평가한다는 점에서 한계를 안고 있다. 최근에는 이런 한계를

극복하고 의미적 층위의 번역 평가를 위한 시도들이 이어지고 있는데, METEOR에 워드 넷이나 동의어 사전을 활용하고, BertSCORE의 임베딩을 이용하는 방식을 통해 인간 평가와 비슷한 의미 층위로 그 평가의 범위를 넓혀 가고 있다. 자동 평가는 그 한계에도 불구하고 인간 평가로 대체할 수 없는 평가 속도와 신뢰도로 인해, 기계번역에 대한 평가나 대규모 번역에 대한 품질 관리 도구로서의 우위를 자랑한다.(Machine Translate Foundation)³⁾

국내의 번역학계는 최근 자동 평가 지표의 활용 연구가 늘고 있는데, 기계번역의 평가를 위해 고안된 자동 평가 지표를 인간번역에 적용하는 연구가 주를 이루며, 정혜연 외(2021), 정혜연(2022), 최지수(2023a, 2023b) 등이 대표적이다. 정혜연 외(2021)는 자동 평가가 의미적 층위에 대한 평가를 수행하도록 하기 위해 임베딩을 활용한 실험 연구를 선보였다. 정혜연 외(2022)는 BERT를 활용한 인간번역의 자동평가 여러 모델의 성능 비교 및 활용 가능성을 탐색하였다. 최지수(2023a)는 sacreBLEU를 중심으로 자동 평가 지표를 한국어 번역문에 대한 정량 평가 도구로서 활용할 수 있는 가능성을 탐색하였고, 최지수(2023b)는 인간 번역에 대한 자동 평가에서 번역 텍스트 길이가 평가 타당성에 미치는 영향을 분석하였다.

한편, 한-중 번역 평가 연구는 인간 평가에 관한 것이 주를 이룬다. 장애리(2017), 강병규·이지은(2018), 권도경 외(2018), 기유미(2018), 김윤정(2018), 김혜림(2020), 홍연옥(2022), 황은하(2020), 김지은(2023), 임소영, 임소정(2023a, b) 문려화(2023), 이아형(2023) 등이 있다. 이 중에 본 연구와 번역 방향이 일치한 한중 기계번역 평가만 추려서 그 연구 대상과 방법을 살펴보기로 한다. 장애리(2017)는 매체, 장르가 달라서 언어적 특성도 다른 문장 10개를 선별해 네이버 통계적 번역, 네이버 신경망 번역, 시스트란, 파파고, 지니톡, 수원관광 통역비서, 구글번역, 바이두 등으로 번역한 결과, 네이버의 NMT 번역기가 적어도 한-중 언어 쌍에 대해서는 압도적으로 높은 정확률을 보였으며, 또한 동일하게 네이버 번역기이지만 네이버 NMT가 SMT보다 오역 비율이 낮았다고 보고하였다. 기유미(2018)는 네이버 파파고와 구글 번역기에 대한 평가로, 신HSK 3~4급 수준의 한중 작문 연습문제 580개 문장을 평가세트로 사용하고 있으며, 번역의 오류 유형을 점수 척도로 측정해 제시하였다. 김윤정(2018)은 구글 번역에 대한 평가로, 성어(成語) 100개, 초급 회화 문장 127개, 뉴스 기사 2편을 구글로 중한 번역한 결과를 유형별로 귀납하였다. 김혜림(2020)은 신문사설의 한중 기계번역과 인간번역의 문체 차이에 주목하여 비교 연구를 수행한 결과, 기계번역은 인간번역과 달리 메타담화의 변이 없이 문자 그대로 번역되는 특징을 보인다고 하였다. 황은하(2020)는 네이버 파파고, 구글 번역기, 유다오(有道)에 대해 신문 기사 표제를 대상으로 선언적 평가와 유형적 평가⁴⁾를 실시하여 비교한 결과, 파파고 > 유다오 > 구글의 순위를 보였고, 성능이 제일 나은 파파고의 유형적 오류는 신문 기사 표제의 불완전종결형에서 기인한 오류 비율이 전체 오류의 76.2%를 차지하는 것으로 나타났다. 임소영, 임소정(2023)은 넷플릭스의 인기 예능 프로그램인 ‘솔로지옥 1’의 발화 940개의 문장을 네이버 파파고, 카카오i, 百度翻譯, 騰訊翻譯, 有道翻譯 등 5개의 기계번역으로 출력하고 비교하여 3개 이상의 기계번역에서 공통 오류를 보인 92개 문장에 대해 어휘, 통사, 화용 층위로 나누어 분석하였다. 이아형(2023)은 과학기술 텍스

3) <https://machinetranslate.org/metrics>

4) 번역 대상 텍스트인 신문 기사 표제의 원문의 형태적, 통사적 특징과 중국어 표제의 문장부호의 특징 등을 감안하여 평가 세트를 구축하고

트를 대상으로 ChatGPT의 번역 결과를 분석하였는데, 오류가 있어서 포스트에디팅이 필요한 경우가 70%로 나타났으며, 구체적으로는 오류 발생 원인을 원문 직역 오류(단어, 어순, 주어 누락 등), 맥락 부적합 단어 오류, 문장 분석 오류(단문에서는 지시대명사 등에 대한 분석 오류, 복문에서는 끊어 읽기 혹은 맥락 파악 오류), 기타(시제, 조동사 등) 및 문장 종결 오류가 주된 오류로 보고되었다. 김지은(2023)은 희곡 『버스 정류장』을 대상으로 파파고와 구글번역의 기계번역을 비교한 연구로, 기계번역 오류를 1) 등장인물 호칭 불일치 및 2인칭 지칭어 오류, 2) 사전적 의미는 맞으나 시대배경, 상황 등에 부합하지 않는 표현, 3) 문체 혼용 및 등장 인물 성격과 관계에 부적합한 표현 및 문체 오류, 4) 경어 사용 오류로 나누어 제시하였다. 위산(2024)은 국제뉴스 40 편을 연구대상으로 파파고로 번역 후 어휘, 문법, 문장으로 나누어 오류를 분석한 결과, 번역 오류 394개 중에 문장 오류가 40%로 가장 높았고, 그 중에서도 과도한 직역으로 인한 한국식 표현의 비중이 제일 컸다고 말한다.

한중 기계번역에 대한 통시적인 비교 연구도 시도되었는데, 권도경 외(2018)는 구글 번역과 네이버 파파고로 중한 번역 추이를 살핀 결과, 번역 결과물이 발전하는 방향성은 발견되지 않았다고 언급하며, 데이터의 양적 축적보다는 고품질의 병렬 코퍼스의 확보가 번역 품질 개선에 더 중요하다고 보았다.

요약하면, 국내 번역학계에 자동 평가의 활용 연구가 늘고 있는 추세인 데 반해, 한중 기계번역에 대한 평가 연구는 주로 인간 평가에 의존하고 있으며, 연구 목적은 기계번역 서비스 간 우열을 평가하는 연구와 기계번역 오류의 유형 분석 등 두 가지로 수렴된다. 평가에 사용된 텍스트는 문어와 구어, 신문과 소설 등 다양하지만, 실제 발화에 대한 구어 연구는 한 편뿐이며 대상 텍스트는 모두 수백 문장 이하로 규모가 작은 편이다.

2.2. 한중 기계번역 포스트에디팅 연구

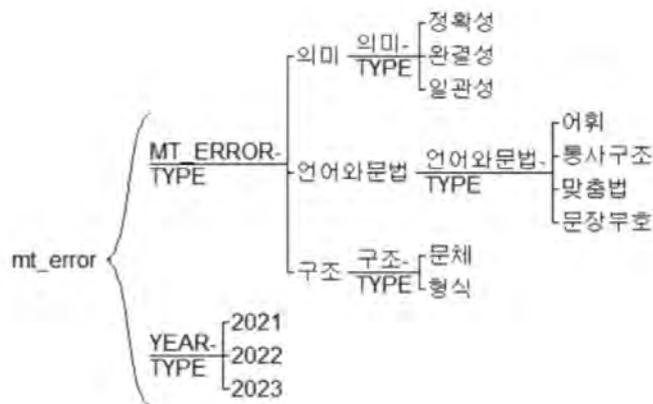
포스트에디트(post-edit)는 기계번역 결과물을 편집하고 교정하는 것이며(ISO 17100:2015, 2.2.4), 국내에서는 포스트에디팅이라는 용어를 더 많이 사용한다. 한중 포스트에디팅에 관한 연구는 2020년대에야 비로소 시작되어 세 편 조사되며, 다른 언어쌍의 관련 연구에 비해 매우 미미한 수준이다.

이아형(2023)은 과학기술의 기계번역을 위한 프리에디팅, 포스트에디팅과 최종 윤문 작업이 필요하다고 보고, 포스트에디팅에서는 원문 대조 및 수정을 통해 원문 직역으로 인한 오류를 수정하고, 선어말어미/어말 어미를 대조하여 문장 종결 오류, 시제, 조동사 등 오류를 수정해야 함을, 최종 윤문 과정에서는 용어와 문체를 통일하는 작업이 필요하다는 결론을 제시하였다. 신문 기사에 관한 포스트에디팅 연구로, 석소영(2023), 위산(2024)이 있는데, 석소영(2023)은 연합뉴스, 중앙일보와 조선일보의 인터넷판에 실린 80편의 정치뉴스를, 위산(2024)은 <조선일보> 국제뉴스 40편을 대상으로 파파고 번역의 오류 분석 및 스크오프 이론을 배경으로 목적성 원칙, 연관성 원칙, 충실성 원칙을 들어 포스트에디팅 가이드라인을 제시하였다. 다만, 하위 분류와 세부적인 포스트에디팅 전략은 다음의 <표 1>에 정리한 것처럼 큰 차이를 보인다.

<표 1> 석소영(2023)과 위산(2024)의 한중 포스트에디팅 가이드라인 비교

스코프스 이론	석소영(2023)	위산(2024)
1. 목적성 원칙	<ul style="list-style-type: none"> ○ 한자의 전환 모호한 한자의 완전 형태 번역 ○ 잘못된 용어 수정 고유명사 표기 해당 분야 전문용어 맥락에 적합한 어휘 용어의 통일성 ○ 정치성향 어휘 대체 목표어의 이데올로기 고려 ○ 현지화 자연스러움 추구 	<ul style="list-style-type: none"> ○ 수신자 중심의 의사소통 외래어 한자표기어 어휘의 현지화 ○ 정보성을 고려한 오역 수정 일반 어휘 고유명사 다의어 품사
2. 연관성 원칙	<ul style="list-style-type: none"> ○ 필수성분 보완 ○ 구두점 수정 구두점 규칙 준수 ○ 명시화 약어 풀어쓰기 배경지식 추구하기 	<ul style="list-style-type: none"> ○ 고유표현화 ○ 문장성분의 변화 첨가 전환 ○ 대구미 ○ 문장부호와 숫자의 자국화
3. 충실성 원칙	<ul style="list-style-type: none"> ○ 재구성 	<ul style="list-style-type: none"> ○ 보완 ○ 삭제 ○ 재구성 원문간섭표현 배제 복문으로의 전환 ○ 간결성 확보

번역의 방향을 달리한 중한 포스트에디팅 연구로 김혜림(2021, 2022, 2024), 모림(2023) 등이 있는데, 김혜림(2022, 2024)는 동일하게 3개의 대항목과 9개의 세부항목의 기준을 적용하고 있으며 김혜림(2024)에서는 기계번역의 오류 변화를 통시적으로 살피기 위해, 연도 항목(YEAR TYPE)까지 더해 다음의 도식과 같은 분류 기준을 적용하였다.



<그림 1> 김혜림(2023, 2024)의 포스트에디팅 가이드라인⁵⁾

한편, 임소영, 임소정(2023)처럼 한중 기계번역의 정확률 향상을 위한 한국어 원문에 대한 프리에디팅 연구도 있으나 여기서는 논외로 한다.

요약하면, 한중 포스트에디팅 연구는 시작 단계이며, 가이드라인은 언어학적 관점의 기준이나 스코프스 이론에 근거한 기준으로 나뉘는 양상을 보인다.

5) 김혜림(2024:9)의 [그림 1] UAM Corpus 기본 분류표를 인용함.

3. 연구 대상과 방법

본 연구는 기구축된 AI 학습용 방송콘텐츠 한국어-중국어 번역 말뭉치를 주요 연구 대상 자료로 하며, 말뭉치에 포함된 기계번역문에 대해 자동 평가를 수행하고, 포스트에디팅 내용을 양적·질적으로 추적하여 분류하고 정리한다.

3.1. 연구 대상: AI 학습용 방송콘텐츠 한국어-중국어 번역 말뭉치

본 연구의 대상 자료는 한국정보화진흥원에 의해 발주되고 가공되어 AI HUB에서 서비스되고 있는 방송 콘텐츠 한국어-아시아어 번역 말뭉치의 하위 말뭉치의 하나인 한국어-중국어(북경어⁶⁾) 번역 말뭉치(이하 편의상 ‘한중 말뭉치’로 약칭함)이다.

한국어-아시아어 말뭉치는 방송 콘텐츠의 인공 신경망 기계번역기(Neural Machine Translation, NMT) 성능 향상을 위하여 학습데이터로 활용하기 위한 한국어-아시아어 음성 및 말뭉치 데이터⁷⁾로, 한국어 방송콘텐츠 분야의 아시아어 통·번역 성능 향상 및 상황별 신조어, 약어, 은어, 관용적 의미와 어투까지 효과적으로 전달 가능한 인공신경망기계번역 인공지능 학습모델 개발을 위해 구축된 것으로 소개된다(한국정보화진흥연구원, 2022)

그 중에 한중 말뭉치는 SBS 방송 콘텐츠의 대화문으로 구성되었고, 연예/공연, 오락/예능, 교양 등 카테고리를 포함하며 총 60만 문장쌍⁷⁾이다. JSON 형식으로 가공된 말뭉치의 구조는 다음과 같다.

```
{
  "data": [
    {
      "sn": "RLVTH1025",
      "data_set": "방송콘텐츠",
      "domain": "TV방송",
      "subdomain": "리얼버라이어티예능",
      "ko_original": ">과자가 있다.", #한국어 원문
      "ko": ">과자가 있다.", #비식별화 등을 거친 한국어 원문
      "mt": "有饼干。", #중국어 기계번역문
      "cn": ">有零食。", #중국어 최종 번역문
      "source_language": "ko",
      "target_language": "cn",
      "word_count_ko": 2,
      "word_count_cn": 2,
      "word_ratio": 1.0,
      "file_name": "리얼버라이어티_8.xlsx",
      "source": "SBS",
      "license": "open",
      "style": "구어체",
      "included_unknown_words": false,
      "ner": null
    }
  ]
}
```

6) 이 말뭉치는 ‘북경어’와 ‘대만어’를 구분하고 있으며 ‘한-대만어’ 데이터는 별도로 701,426문장쌍 구축되었다.

7) AI 허브의 소개에 따르면 한-중 번역 말뭉치의 결과 수량은 1,011,606문장이나 배포된 말뭉치에서 훈련(Training) 데이터에 포함된 수량은 60만 문장쌍이다.

```

    },
    ...
  ]
}

```

<그림 2> 방송 콘텐츠 한중 번역말뭉치의 JSON 구조

위의 JSON 구조에서 “ko_original”, “ko”, “mt”, “cn”이 각각 한국어 원문, 비식별화 등을 거친 한국어, 기계번역문, 최종 번역문의 속성-값을 나타내고 있다.

말뭉치에 포함된 한국어 발화는 총 218만 9,185어절이며, 발화당 평균 길이는 약 3.6어절(최단 1어절, 최장 124어절)로 일상 대화의 전형적인 짧은 길이를 보이고 있다. 중국어 기계번역문은 토크나이저 jieba를 이용했을 때 총 276만 4,379단어, 번역문 당 평균 4.6단어, 최종 번역문은 총 281만 2,584 단어로, 평균 길이는 4.7단어이다.

3.2. 연구 방법

본 연구는 크게 포스트에디팅 비율이 높은 번역문의 추출, 포스트에디팅에 대한 양적 분석과 질적 분석 세 단계로 나뉜다.

첫 단계의 고비율 포스트에디팅 번역문의 추출을 위해서는 내려받은 말뭉치를 처리가 가능한 형식으로 전처리하고, 기계번역문에 대해 자동평가 지표를 도입해 점수화하는 방식을 취하는데, 본 연구에서는 scareBLEU⁸⁾, METEOR⁹⁾ 두 가지 지표를 적용하여 기계번역문을 점수화한다.

sacreBLEU 평가를 위해서는 구두점, 양쪽 공백 제거 등 텍스트를 정제하고, 중국어 토크나이저 zh를 사용해 토크나이징을 하였다. METEOR 평가에서도 마찬가지로 텍스트를 정제하였으나 토크나이징은 jieba를 사용하였으며, 중국어 워드넷 COW(Chinese Open Wordnet)¹⁰⁾을 적용하여 평가 정확성을 높이고자 하였다.

이 자동 지표의 계산 원리에 따르면 sacreBLEU, METEOR는 서로 양의 상관관계를

8) sacreBLEU는 주로 n-gram의 정확도를 기반으로 번역문에 점수를 매긴다. 즉 번역문이 참조 문장인 최종 번역문과 n-gram 수준에서 얼마나 일치하는지를 평가하며, 문장이 완전히 일치할 경우 sacreBLEU 점수는 100이 된다. SacreBLEU는 기존의 BLEU 점수 계산 방법에서 토크나이징에 일관성이 부족한 문제를 표준화하여 번역 평가 결과를 보다 쉽게 비교하고 재현할 수 있도록 설계되었다. BLEU의 기본적인 원리는 n-gram에 기반하여 계산되므로, 매우 짧은 텍스트는 충분한 n-gram 데이터를 제공하지 못해 점수가 실제 번역 품질을 반영하지 못할 수 있다. sacreBLEU는 이러한 문제를 완화하기 위해 문장의 길이에 따라 조정된 BLEU 점수를 제공하는 스무딩 기법을 제공한다.(Post, M. (2018))

9) METEOR(Metric for Evaluation of Translation with Explicit ORDERing)는 BLEU의 단점을 보완하고자 개발되었으며, 단어 단위의 일치 여부 외에 정밀도(Precision)와 재현율(Recall)을 모두 고려하여 번역 품질을 평가한다. 정밀도는 번역된 문장에서 참조 문장과 일치하는 단어의 비율을, 재현율은 참조 문장에서 번역된 문장과 일치하는 단어의 비율을 나타내며, F1 점수(F-measure)를 사용하여 정밀도와 재현율의 조화 평균을 계산한다. 이 지표는 WordNet이나 동의어 사전을 사용하여 동의어 관계를 인식하여 의미 층위의 평가가 더해진다는 장점이 있다. 번역문에서 단어의 순서가 달라지더라도 문장의 의미가 동일할 경우, METEOR는 이를 인식하여 페널티를 줄임으로써 순서에 민감한 BLEU의 단점을 보완한다. 또한, 형태 변화가 거의 없는 본 연구의 중국어 번역문과는 무관하지만, 형태소 분석을 통해 어근, 접두사, 접미사 등의 변형을 인식하기 때문에 굴절어나 첨가어 등에는 매우 유용할 것으로 예측 가능하다.(Satanjeev Banerjee and Alon Lavie. 2005)

10) 다음의 url 주소에서 내려받을 수 있다. <https://bond-lab.github.io/cow>

가질 것으로 예측할 수 있다. 실제로 sacreBLEU와 METEOR는 1에 가까운 매우 높은 양의 상관관계를 보였으며, 따라서 이 두 점수가 모두 낮은 경우를 포스트에디팅이 많이 수행된 번역문으로 보아 포스트에디팅의 질적 분석 대상을 추출하는 기준으로 삼았다.

다음은 문장 단위로 평가하여 두 지표의 점수가 기록된 번역문 평가 결과의 예시이다.

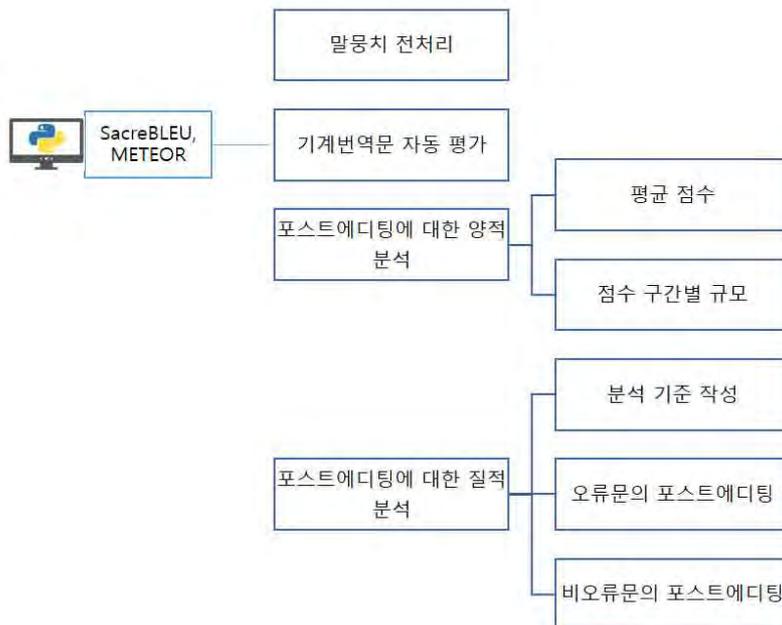
<표 2> 자동 평가 점수가 기록된 한중 번역문의 예시

sn	ko	mt	cn	sacre BLEU	METEOR
RLVTH1025	과자가 있다.	有饼干。	有零食。	6.4	0.25
KOBETG3732	배달 왔습니다.	外卖来了。	外卖来了。	100.0	0.98
KOBETCC5391	아니, 양쪽에다가 이렇게 놓고 가십시오, 건빵 주머니처럼.	不是, 请这样放在两边, 像压缩饼干袋一样。	不是, 请这样放在两边, 像压缩饼干口袋一样。	84.3	0.88
KRLVTCC166	잠깐 휴식하시고 저희.	我们先休息一会儿。	我们先休息一下。	68.1	0.99
KRFETQ461	치즈.	奶酪。	芝士。	0.0	0.50
KRFETV3463	그런데 진짜 고구마를 먹는 이유가 이게 살이 빠지는 거에 도움이 되나 이런 생각은 들어요.	但是真的吃地瓜的理由是这个真的对减肥有帮助吗 有这种想法	但是真的吃地瓜的理由是, "这个真的对减肥有帮助吗", "有这种想法。"	100.0	0.99

두 번째 단계로, 이 두 지표의 점수를 통해 번역 말뭉치의 전체적인 포스트에디팅 규모와 수준을 알아본다.

세 번째 단계로는 포스트에디팅의 유형 분석에 대한 기준을 세우고, 이를 토대로 포스트에디팅 비율이 높은 기계번역문을 최종번역문과 비교해 유형별로 예시를 살펴본다.

이러한 연구의 절차를 도식화하면 다음의 <그림 3>과 같다.



<그림 3> 연구 흐름도

4. 포스트에디팅에 대한 양적·질적 분석

포스트에디팅에 대한 분석은 자동 평가 점수에 기반한 양적 분석과 높은 포스트에디팅 비율을 보이는 번역문을 대상으로 한 질적 분석을 수행한다.

4.1. 양적 분석

4.1.1. sacreBLEU 자동 평가 결과

먼저, 기계번역 전체의 sacreBLEU 평균 점수는 100점 만점에서 67.71점이다. 만점 100점은 포스트에디팅이 전혀 되지 않은, 기계번역문과 최종 번역문이 완전히 일치한 경우이며, 점수가 낮을수록 포스트에디팅 비율이 높은 것으로 볼 수 있다. 다만, 이는 형태적 n-gram에만 기반한 점수이므로 기계번역의 정확률이나 오류율과 직결시키는 것은 선부른 것으로 보이며, 포스트에디팅의 질적 분석이 필요하다.

다음으로, 점수 구간별 문장 규모와 비율, 구간별 평균, 표준편차는 <표 3>과 같다.

<표 3> sacreBLEU 점수 구간별 통계

SacreBLEU 점수 구간	문장수	비율	구간별 평균	표준편차
0-20	137,526	22.9%	4.905387	5.109362
21-40	41,194	6.9%	29.39126	5.475143
41-60	39,344	6.6%	50.10199	5.640417
61-80	39,695	6.6%	69.3624	5.701089
81-99	14,219	2.4%	85.3817	3.861344
100	328,022	54.7%	100	0
합계	600,000	100.0%		

<표 3>에서처럼, 포스트에디팅이 전혀 되지 않은 문장이 전체 기계번역문의 54.7%를 차지한다. 이 말뭉치의 최종번역문의 품질이 높다고 전제하면, 반을 넘는 TV방송 대화의 기계번역문은 포스트에디팅이 필요 없는 정확한 문장으로 해석될 수 있다. 덧붙여, 본 연구에서 스무딩 기법을 적용하였음에도 불구하고 일부 길이가 매우 짧은 기계번역문과 최종번역문의 동의어 사용, 감탄사 등의 반복의 횟수의 차이 등 단순한 형태적 차이로 인해 낮은 점수를 받은 경우도 있으므로, 실제 정확률은 이보다 더 높다고 할 수 있겠다.

- (1) sn: KOBETB1000(관찰예능)
 ko: 하하하하.
 mt: 哈哈哈哈哈。(sacreBLEU: 71.65)
 cn: 哈哈哈哈哈

인간 평가자의 입장에서 (1)의 mt와 cn은 의미적으로 크게 다르지 않으나, 형태적 n-gram 기반의 채점 원리에 따라 mt의 ‘하’의 3회 반복과 마침표 사용은, cn의 4회 반복과 문장부호 미사용으로 인해 28% 이상의 점수 차이를 보이고 있다.

4.1.2. METEOR 자동 평가 결과

기계번역문 전체의 METEOR 평균 점수는 1점 만점에서 0.67점이다. 만점 1점은 포스트에디팅이 전혀 되지 않은, 기계번역문과 최종 번역문이 일치한 경우이며, 점수가 낮을수록 포스트에디팅 비율이 높다고 볼 수 있다. 다만, 이 점수 또한 기계번역의 정확률이나 오류율과 직결시키기보다 포스트에디팅의 내용 분석을 통한 세밀한 평가가 필요하다.

본 연구에서는 METEOR의 사용에 중국어 워드넷을 적용하였다. 즉 기계번역문과 최종 번역문의 동의어를 확인하여 점수를 추가로 부여할 수 있도록 하였으므로 METEOR의 점수가 sacreBLEU보다 높을 것으로 기대된다. 다만, 본 연구에서 METEOR는 중국어 토큰나이저로 jieba를, sacreBLEU는 zh를 사용하고 있기에 두 점수의 단순 비교는 큰 의미가 없다. 그럼에도, METEOR에 중국어 워드넷을 적용함으로써, 기계번역문과 최종 번역문의 동의어 인식을 통해 점수가 상향 조정된 번역문의 예를 확인할 수 있었고 예를 보이면 다음과 같다.

- (2) sn: KRFETC5609(교양)
 ko: 너희들 어느 쪽을 발랐냐?
 mt: 你們塗了哪一邊? (sacreBLEU: 10.04, METEOR: 0.67)
 cn: 你們抹了哪一面啊?
- (3) sn: KOBETAA917(관찰예능)
 ko: 치즈를?
 mt: 奶酪? (sacreBLEU: 0.00, METEOR: 0.50)
 cn: 芝士?

(2)에서 mt의 一邊(쪽)과 cn의 一面(면), (3)에서는 mt의 奶酪(치즈)와 cn의 芝士(치즈)가 동의어로 인식되면서 METEOR가 sacreBLEU보다 만점 대비 썩 높은 점수를 받은 것을 확인할 수 있다.

다음으로, METEOR 점수 구간별 문장 규모와 비율, 구간별 평균, 표준편차는 <표 4>와 같다.

<표 4> METEOR 점수 구간별 통계

METEOR 점수 구간	문장수	비율	구간별 평균	표준편차
0-0.2	76,114	12.7%	0.061398	0.071972
0.21-0.4	56,386	9.4%	0.27993	0.053115
0.41-0.6	113,133	18.9%	0.497523	0.03107
0.61-0.8	60,994	10.2%	0.689206	0.060373
0.81-1.0	293,373	48.9%	0.961283	0.045628
합계	600,000	100.0%		

<표 4>에서는 기대와 달리, METEOR의 만점 비율이 sacreBLEU보다 썩 낮게 나오는데, 그 이유에 대해서는 추후 과제로 남긴다.

4.2. 질적 분석

질적 분석 대상은 4.1의 sacreBLEU와 METEOR에서 모두 0점을 받은 기계번역문 중에 무작위 방식으로 220개를 추출하여, 최종번역문과 일일이 비교하여 유형별로 주석하고 귀납하여 정리한다.

주석의 기준, 즉 포스트에디팅의 유형에 대한 기준은 김혜림(2022, 2024)의 분석 기준¹¹⁾을 토대로 하되, 실험적 분석을 하는 과정에 TV방송 대화의 한중 번역의 특성을 반영하여 몇 가지 기준을 추가하였다. 본 연구의 질적 분석 기준을 김혜림(2024)와 비교해 보면 다음과 같다.

<표 5> METEORSacreBLEU 점수 구간별 통계

김혜림(2022, 2024)		본 연구		
대분류	소분류	소분류	중분류	대분류
의미	정확성	-	의미	
	완결성	-		
	일관성	-		
언어와 문법	어휘	단어	언어와 문법	오류
	-	다단어 표현(MWE)		
	통사구조	통사구조		
	맞춤법 문장부호	맞춤법 문장부호		
구조	문체 형식	문체 형식	구조	
		-		
-	-	명시화	번역 전략	비오류
-	-	암시화		
-	-	기타(숫자와 로마자의 한자화 등)		
-	-			

<표 5>에서처럼, 본 연구는 포스트에디팅이 번역 오류만 대상으로 한 것이 아닌 예를 확인하여 오류와 비오류의 대분류를 설정하고, 비오류에 대해서는 번역 전략에 따른 명시화와 암시화 기준을 추가하였다.

오류 대분류에서 의미적 기준을 제외하였는데, 그 이유는 본 연구의 질적 분석 대상이 자동 평가 점수가 0점인 문장에 한하므로, 의미적 기준(정확성, 완결성, 일관성)을 모두 위배하고 있기 때문이다. 다음으로, 언어와 문법 기준의 하위 기준에서 어휘 기준은 대화문 특성상 구어성이 높은 다단어 표현(Multi Word Expression)의 고빈도 사용을 고려하여 단어와 다단어 표현으로 기준을 분리하였다. 또한, 구어 특성상 축자적으로는 문제가 없으나 상황 맥락(context)과 관련을 지으면 오류가 되는 유형을 고려하여 화용 기준을 추가한다.

4.2.1. 번역 오류에 대한 포스트에디팅

번역 오류에 대한 포스트에디팅은 언어와 문법 기준 중에서도 특히 어휘에 관한 것이 가장 높은 비율로 관찰되었다. 문장이 짧다 보니 통사 구조나 맞춤법, 문장부호에 관한

11) 2.2.의 <그림 1>을 참조할 것.

이슈는 상대적으로 적었다.

어휘에 관한 포스트에디팅은 단어와 다단어 표현에 대한 비율이 특히 높았는데, 이는 한국어 방송 대화의 언어적 특성과 그에 대한 한중 번역의 특수성을 잘 드러내 보인다.

먼저, 단어 단위의 오류는 동형어로 인한 오번역이 가장 많았다¹²⁾. 발화문이 짧아서 동형어를 구분할 만한 결합어 등의 단서가 매우 부족하기 때문으로 해석된다.

- (4) sn: KOBETA767(관찰예능)
ko: 이에는 할 수 있을 것 같아요.
mt: 這回應該能做到。
cn: 牙縫里的我可以。
- (5) sn: KOBETBB491(관찰예능)
ko: 바로 쓸 수 있게 하셨어.
mt: 讓你馬上就能寫出來。
cn: 劈的正好可以用啊。
- (6) sn: KRFETA5298(교양)
ko: 거의 다 밤이 진짜 많이 들어가거든요.
mt: 几乎都是黑夜。
cn: 這裡放了很多栗子呢。

(4)의 mt는 치아를 뜻하는 ‘이(牙)’를 ‘이번(這回)’으로, (5)는 ‘사용하다’를 뜻하는 ‘쓰다’를 글씨를 쓰다의 ‘寫’로 오역하고 있고, (6)은 견과로서의 ‘밤(栗子)’을 ‘낮과 밤’의 ‘밤(黑夜)’으로 오역하고 있다.

다음으로, 어휘 단위의 오류에는 다단어 표현의 번역 오류가 높은 빈도로 관찰되었다. 이런 다단어 표현은 구어 어휘 사용의 중요한 특징이며 의미적 비합성성을 보이므로 직역하면 오번역이 되는 경우가 많다¹³⁾.

- (7) sn: KRFETS6377(교양)
ko: 눈은 있어 가지고 보는 눈은 있어 가지고.
mt: 有眼睛, 有眼睛。
cn: 還挺會看人的。
- (8) sn: KRLVTGG778(리얼버라이어티예능)
ko: 이제부터 아웃이라 하면 아예 떨게요, 손을.
mt: 從現在開始, 如果叫出局的話, 干脆把手拿開吧。
cn: 以后要是說場外, 我就不碰了。
- (9) sn: KOBETEE3003(관찰예능)
ko: 재, 재 줍 어떻게 해 봐.

12) 단어 단위의 오류는 다의어의 대응어 선택의 문제도 많았으나, 이는 다른 매체, 장르의 기계번역과 포스트에디팅에서도 흔한 문제이므로 여기서는 논외로 한다.

13) 안의정, 황은하(2021)에서는 다단어 표현을 ‘구어 구 단위 표현’이라고 지칭하며 구어에서 하나의 의미·기능 단위로 사용되고 있음에도, 구어 말뭉치의 부족, 단일 단어 대비 낮은 빈도, 사전의 보수성 등으로 인해 사전 기술에 거의 반영되지 않고 있으며, 이로 인해 기계번역에서의 낮은 정확률과, 한국어교육 내용에서 실제성 높은 표현의 배제 등으로 이어진다고 문제를 지적하고 있다.

mt: 她, 你怎么做。

cn: 他, 好好勸勸他啊。

(10) sn: KOBETJJ4618(관찰예능)

ko: 이거 또 어려운 걸음을 또 이렇게.

mt: 又邁出了這么難的步伐

cn: 百忙中抽空前來啊。

(7)과 (8)의 ‘보는 눈이 있다’, ‘손을 떼다’는 사전에 등재된 관용 표현이나 위의 발화에 서는 구성 성분이 하나 생략되거나(보는 눈이 있다 -> 눈이 있다), 어순 도치(손을 떼다 -> 떼다, 손을)를 통해 어형의 변화를 보이고 있어서 관용구로서가 아닌 구성 성분의 직역을 통해 번역문을 산출한 것으로 보인다¹⁴). 안의정, 황은하(2021)의 분류를 따르면, (9)의 ‘~를 좀 어떻게 해 보다’는 열린 구조의 통사적 관용표현, ‘어려운 걸음을 (하다)’는 특정 화행에 쓰이는 화용적 표현으로 볼 수 있다. 모두 의미적 비합성성으로 인해 직역으로는 정확한 번역문을 얻을 수 없는 경우이다.

끝으로, 방송 대화문의 특성상 화용적 오류에 대한 포스트에디팅도 적지 않다. 축자적으로는 정확한 번역이나 해당되는 상황 맥락에서는 오류가 되는 경우를 화용적 오류로 분류하기로 한다. 호칭어의 번역에 관한 예를 중심으로 살펴보기로 한다.

(11) sn: KOBETAA1324(관찰예능)

ko: 어머님, 병째로 드셔야 해요, 선생님.

mt: 媽媽, 要帶病吃, 老師。

cn: 阿姨, 您得喝瓶裝水了。

(12) sn: KOBETAA6928(관찰예능)

ko: 그런데 어머님 너무 고우세요.

mt: 但是媽媽太漂亮了。

cn: 伯母十分美麗呢。

한국어 방송 발화 60만 개 중에 ‘어머님’이 모두 820회 등장하는데, mt에서는 90%에 달하는 740회가 ‘媽媽(어머니)’로 번역되었다. 그러나 이 중에 261회(35.3%)는 최종번역문인 cn에서 ‘阿姨(이모, 아주머니)’, ‘伯母(아주머니, 이모)’로 포스트에디팅된 것을 확인할 수 있었다. 한국어에서 ‘어머님’은 가족이 아닌 어머니 연배의 여성을 호칭할 때도 자주 쓰는 말인 데 반해 중국어의 대응어는 가족 호칭으로만 사용된다는 점에서 차이를 보인다. 호칭어는 어휘 단위로 실현되기는 하나, 언어적 문맥이 아닌 상황 맥락에 근거해 번역어를 달리하기 때문에 화용적 오류로 보았다.

4.2.2. 비오류에 대한 포스트에디팅

14) 별도로 확인한 결과, (7)은 파파고에서는 정확하게, DeepL과 구글번역에서는 직역으로 인해 번역 오류를 산출하며, (8)은 파파고, DeepL, 구글 번역에서 모두 직역으로 인한 번역 오류를 보이나 어순을 ‘손을 떨게요.’로 프리에디팅하면 모두 정확한 번역을 산출한다.

본 연구에서 질적으로 분석한 번역문 220개 중 68개(30.1%)는 오류가 아니었음에도 포스트에디팅이 된 것으로 관찰된다. 명시화, 암시화, 기타(감탄사, 숫자와 로마자의 한자화)으로의 교체 등을 들 수 있는데, 명시화와 암시화 전략을 위한 수정이 특히 두드러진다.

먼저, 명시화(Explicitness)는 번역문이 더 명확하고 이해하기 쉽도록 하기 위한 번역 전략이다. 명확하지 않은 대명사를 수정한다거나 생략된 정보를 복원하며, 문화적 맥락을 보다 명시적으로 설명하거나 약어나 줄임말을 풀어쓰며 의미가 명확하지 않은 표현을 수정하는 등 방식으로 구현된다.

(13) sn: KOBETB6125(관찰예능)

ko: 진짜 좋은 일 하시는 거예요 정말.

mt: 真的在做好事。

cn: 這簡直是慈善行爲, 這麼便宜。

ko': 이건 자선 행위네요, 이렇게 싸다니요.

(14) sn: KRFETK3572(교양)

ko: 배를 스톱을 해야 하는 거예요.

mt: 要停船

cn: 不僅僅是我們漁民, 可能所有行業都面臨着困難。

ko': 우리 어민들뿐만 아니라 모든 업종에서 아마 어려움을 겪게 될 거예요.

(15) sn: KRLVTQ653(리얼버라이어티예능)

ko: 해봐, 해봐.

mt: 你試試, 試試。

cn: 畫一下。

ko': 그려 봐.

(13)~(15)에서 ko'는 cn을 다시 번역한 것으로, 한국어 원문인 ko에 비해 내용이 추가되거나 구체화된 것을 볼 수 있다. (13)에서는 '좋은 일'을 '慈善行爲(자선행위)'로, 그리고 그렇게 말하는 이유로 '這麼便宜(가격이)' 싸다'는 내용도 추가되었다. (14)는 '배를 스톱하다'의 의미에 함축된 의미를 모두 풀어서 설명하고 있으며, (15)에서는 다의어인 '하다'를 구체적인 행위인 '畫(그리다)'로 명시화하였다.

이와는 반대로, 암시화(Implicitation)는 번역 과정에서 원문에 명시적으로 표현된 정보를 생략하거나 덜 명확하게 표현하는 번역 전략이다. 이는 원문의 의미를 유지하면서도 번역문이 자연스럽고 간결하게 들리도록 하기 위해 사용되는데 방송 대화 특성상 자주 포스트에디팅의 주된 유형으로 관찰된다.

(16) sn: KOBETAA2619(관찰예능)

ko: 아주 큰 탱크에 들어가서 이렇게 뭐 알바도 했고.

mt: 進了很大的坦克 還做了兼職

cn: 在那里面工作過。

ko': 그 안에서 일했었어.

(17) sn: KRFETH1015(교양)

ko: 이거 방송 사고 없이 끝날 것 같은데?

mt: 這個放送好像不會發生事故吧?

cn: 感覺應該很順利。

ko': 아주 순조로울 것 같은 느낌이야.

(16), (17)의 기계번역문은 모두 한국어 원문에 충실한 번역인데, 최종 번역문에서 '큰 탱크에 들어가서'를 '在那里面(그 안에서)'로, '방송 사고 없이 끝나다'를 '很順利(아주 순조롭다)'로 간결하게 암시화하여 수정한 것으로 나타난다.

5. 맺음말

본 연구는 한중 원문-번역문쌍과 더불어 기계번역문이 포함되어 있는 AI 학습용 TV방송 대화 번역말뭉치를 대상으로, 포스트에디팅에 대한 양적 평가와 질적 평가를 수행하였다.

우선, sacreBLEU와 METEOR를 적용한 양적 평가를 통해 기계번역문의 54.7%는 포스트에디팅이 수행되지 않은 것으로 확인되었다. 최종번역문이 고품질의 번역문이라고 가정할 때, 이는 방송 대화에 대한 기계번역의 정확률을 짐작하고, 대규모 번역말뭉치의 구축에 드는 인력 소모 등을 계산할 수 있는 근거로 활용할 수 있다.

다음으로, sacreBLEU와 METEOR 채점을 통해 포스트에디팅 비율이 가장 높은 기계번역문을 추출하여 최종번역문과 일일이 비교한 결과, 포스트에디팅은 오류뿐만 아니라 오류가 아닌 번역문에 대해서도 높은 비율(30.1%)로 수행됨을 확인하였다. 오류의 경우, 특히 동형어 식별 오류로 인한 것과 다단어 표현의 직역으로 인한 오류가 높은 비율로 나타났다으며, 상황 맥락과 관련한 화용적 오류도 방송 대화의 한중 기계번역의 주요 오류 유형의 하나로 관찰되었다. 오류가 아닌 경우, 포스트에디터가 번역의 일반적인 전략인 명시화와 암시화를 위해 포스트에디팅을 많이 사용하는 것으로 나타났다.

본 연구는 대규모 번역말뭉치를 대상으로 방송 대화의 한중 번역의 포스트에디팅의 양적, 질적 특징을 살펴봄으로써, 비격식적 대화의 한중 기계번역과 포스트에디팅에 참고 자료를 제공한다는 의미가 있다. 또한, 한중 번역에 자동 평가를 적용한 첫 사례라는 점에서 의미 찾을 수 있다.

다만, 본 연구에서 활용한 두 자동 평가 지표의 토큰나이를 일치시키지 못하여 지표 간 비교를 하지 못한 부분은 앞으로의 연구 과제로 남긴다. 질적 분석에서 좀 더 많은 번역문에 대한 분석과 더불어 유형별 규모에 대한 계량 작업도 필요할 것으로 보인다.

참고문헌

강병규, 이지은. (2018). 신경망 기계번역의 작동 원리와 번역의 정확률. 中語中文學 -73 (2018): 253-295.

- 공수. (2019). 기계번역 프로그램 품질에 대한 사용자 평가와 사용자의 L2 수준 간 상관관계 고찰 -한중 언어 쌍을 중심으로-. 통역과 번역, 21(3), 1-29.
- 김보영, 김연주, 서승희, 송신애, 이진현, 전경아, 최지수, 홍승빈, 정혜연, 허탁성.(2020). 번역자동평가에서 풀리지 않은 과제" 번역학연구 21.1 (2020): 9-29.
- 김지은. (2023). 희곡의 기계번역에 대한 고찰 - 희곡 『버스 정류장』의 등장인물을 중심으로. 한중언어문화연구, 68, 3-32.
- 김혜림(2021)_중한 범령 기계번역 포스트에디팅 교육을 위한 예비 연구?
- 김혜림. (2024). 중한 기계번역 오류변화와 포스트에디팅 가이드라인에 대한 고찰 : 2021년-2023년 학부 수업사례를 중심으로 . 통번역학연구, 28(2), 1-26.
- 김혜림. (2022). 중한 기계번역 포스트에디팅 가이드라인에 관한 예비연구: 포스트에디팅 교육을 위한 가이드라인. 중국언어연구, 99, 277-312.
- 모림. (2023). 중한 기계번역 오류에 대한 포스트에디팅 전략 연구 : 연설문을 중심으로, 한국외국어대학교 대학원 석사학위논문.
- 사례. (2016). TV방송과 신문보도의 문체적 특성에 대하여. 텍스트언어학, 40, 59-87.
- 석소영. (2023). 한중 기계 번역의 포스트에디팅 가이드라인 제안 : 정치뉴스를 중심으로, 한국외국어대학교 KFL대학원 석사학위논문.
- 이아형. (2023). 챗 GPT의 번역 수업 활용 방안 고찰 - 한중과학기술번역을 중심으로. 중국어문학지, 85, 153-176.
- 임소영 and 임소정. (2023). 구어체 한-중 AI번역의 오류 양상 연구 -넷플릭스 「솔로지옥」 발화를 중심으로-. 중국어문학, 92, 299-330.
- 임소영 and 임소정. (2023). 구어체 문장의 한중 기계번역 오류 개선을 위한 프리에디팅 가이드라인 연구 - 중국어 비숙련자를 대상으로. 한중언어문화연구, 68, 129-157.
- 위산. (2024). 한중 뉴스 기계번역 포스트에디팅(MTPE) 전략 고찰 -<조선일보> 국제뉴스를 중심으로, 충남대학교 대학원 석사학위논문.
- 장애리. (2017). "국내 기계 통번역의 발전 현황 분석." 번역학연구 18.2 (2017): 171-206.
- 정재혁. "텍스트 유형에 따른 영한 포스트에디팅에서의 오류 수정 양상 분석 및 고찰 ." 통번역학연구 28.1. 123-150.
- 정혜연 and 서수영. (2022). 버트(BERT)를 활용한 인간번역의 자동평가: 여러 모델의 성능 비교 및 활용 가능성. 통번역학연구, 26(4), 117-137.
- 정혜연. (2018). 번역의 자동평가 : 기계번역 평가를 인간번역 평가에 적용해보기 . 통번역학연구, 22(4), 265-287.
- 정혜연, 박현일, 우경조, 서수영. (2021). 임베딩을 활용한 인간번역의 자동평가 - 기계가 의미를 평가할 수 있을까. 통번역학연구, 25(3), 141-162.
- 최지수. (2023a). 한국어 번역문의 정량 평가 도구로서의 기계 평가 지표의 활용 가능성 고찰: SacreBLEU를 중심으로. 언어와 정보 27.2 (2023): 161-176.
- 최지수. (2023b). 인간번역의 자동평가에서 번역 텍스트 길이가 평가 타당성에 미치는 영향. 통번역학연구 27.2 145-164.
- 홍연옥. (2022). 기계번역의 교육적 활용 가능성 고찰 및 제안 - 把자문의 인간번역과 기계번역의 비교를 중심으로. 중국어교육과 연구, 37, 23-48.
- 홍연옥. (2022). 한중 기계번역을 활용한 오류 수정 및 생산성 연구. 중국어문학, 90, 197-217.
- 한국정보화진흥원. (2022). '22년 인공지능 학습용 데이터 구축·활용 가이드라인(방손콘텐츠 한국어-아시아어 통번역 데이터) v1.0.
- 황은하(2020). 한중 기계번역 품질에 대한 비교 분석 - 신문기사 표제를 중심으로, 한국번역학회 2020년 가을 학술대회 발표논문집. 12-29.

- 황은하, 비립(2023). AI 학습용 한중 병렬말뭉치의 번역문 평가, 2023년 가을 한국코퍼스언어학회 10주년 기념 학술대회 발표논문집.
- 王子云, 毛毳. (2023). ChatGPT 譯文質量的評估与提升——以陶瓷類文本漢英翻譯為例, 山東陶瓷 Vol. 46 No. 4
- Ahrii Kim, Jinhyun Kim, Guidance to Pre-tokenization for SacreBLEU: Meta-Evaluation in Korean,
- Araghi, S., Palangkaraya, A. The link between translation difficulty and the quality of machine translation: a literature review and empirical investigation. *Lang Resources & Evaluation* (2024). <https://doi.org/10.1007/s10579-024-09735-x>
- Dai, G., & Liu, S. (2024). Towards Predicting Post-editing Effort with Source Text Readability: An Investigation for English-Chinese Machine Translation. *The Journal of Specialised Translation*, (41), 206 - 229. <https://doi.org/10.26034/cm.jostrans.2024.4723>
- Huang, C.-R., She, Y., & Chang, L.-P. (2010). COW: Chinese Open Wordnet. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations (COLING '10)* (pp. 133-136). Association for Computational Linguistics.
- International Organization for Standardization. (2015). ISO 17100:2015. Translation services – Requirements for translation services. <https://www.iso.org/standard/59149.html>
- Lu, X., & Han, C. (2023). Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multi-scenario exploratory study. *Interpreting*, 25(1), 109-143.
- Matt Post. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation (WMT18)*
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65 - 72,
- Xiaoman Wang, Claudio Fantinuoli. (2024). Exploring the Correlation between Human and Machine Evaluation of Simultaneous Speech Translation. Paper accepted at the European Association for Machine Translation conference 2024. arXiv:2406.10091 [cs.CL]

<자동 평가 지표를 활용한 한중 기계번역 평가 및 포스트에디팅 연구: TV 방송 대화를 중심으로>의 토론문

비립(연세대)

이번 발표는 인공지능 발전 가속화 시대에 널리 활용되고 있는 기계번역 태스크에 초점을 맞춰서 최근에 공개된 방송콘텐츠 한중 병렬 코퍼스를 바탕으로 언어모델이 출력한 1차 번역문과 포스트에디팅 과정을 거친 2차 수정문 간의 중첩도를 양적으로 먼저 분류한 다음에 중첩도가 낮은 문장쌍에 대한 질적 분석을 수행했다. 이과정을 통해서 구어성이 강한 한국어 대화문을 중국어로 옮기는 과정에서 기계적인 고유 성능 한계와 인위적인 번역 전략 차이를 면밀히 검토함으로써 포스트에디팅 유형을 일반적인 문법성 오류에서 번역 전략에 따른 in-context 비오류까지 확장시켰으며 기계번역 품질 평가에 더욱 구체적이고 체계적인 가이드라인 개발 탐색에 중요한 의미를 부여할 수 있다.

양적 분류 과정에서 주로 사용되는 "ScareBLEU"와 "METEOR"는 신경망 기반 기계번역 성능 평가에 아주 클래식한 평가 지표이다. 전자는 ngram 기반의 중첩도를 비교함으로써 번역문 단어 간의 매칭 관계를 쉽게 포착할 수 있지만 문장 구조와 의미 관계를 무시하고 문장 길이에 민감하다는 약점이 있다. 후자는 단어보다 한 단계 더 큰 chunk 단위로 중첩도를 측정하고 관련 언어의 의미 정보를 담긴 워드넷을 추가함으로써 의미 정보 평가에 강하다는 장점이 있는데 워드넷에 수록되지 않은 신어에 대해서는 한계도 존재한다. 위 2개 지표를 동시에 적용했을 때 번역문쌍에 대한 형태적 및 의미적 자동 평가는 대체로 확인할 수 있으나 방송 대화라는 장르적 특성상 방송인의 발언 자체가 수의적이고 신어 및 약어 사용이 빈번한 편이며, 선택된 문서 생성 평가 지표가 이미 오래됐다는 점을 고려해서 중첩도 분류에 소극적인 영향을 끼칠 수도 있다.

포스트에디팅을 거친 번역문을 수동으로 분석하는 과정에서 형태, 의미, 통사 등 전통적인 평가 층위보다 문맥 특성과 번역 전략을 평가 요소를 추가로 고찰하는 것은 포스트에디팅의 유형과 기계번역 수동 평가의 기준을 한층 더 세분화했다고 볼 수 있다. 전통적으로 포스트에디팅 태스크는 문법성 체크에 주로 집중되어 있다. 이는 언어모델이 번역문을 생성하는 과정에서 문맥에 맞는 자연스러운 문장보다 의미 전달이 정확하고 유창한 문장을 출력하는 것이 주목적이었기 때문이다. 단 초거대언어모델이 급속히 발전하고 있는 현시점, in-context 정보의 학습 능력이 계속 강화되고 기계번역문에 대한 포스트에디팅 기준도 당연히 더 넓은 범위로 확장할 필요가 있다. 다만 포스트에디팅이라는 것은 결국 포스트에디터의 writing style에 연결되어 있는데 번역문 하나에 여러가지 포스트에디팅 수정문을 생성할 수 있는 만큼 언어모델 훈련 및

평가에 필요한 포스트에디팅 병렬 데이터의 종류와 규모를 더 늘릴 필요가 있다.

전반적으로 이번 발표는 상대적으로 관심이 덜 가는 한중 기계번역문을 대상으로 기존의 기계번역과 포스트에디팅 태스크의 평가 범위를 한층 더 넓히고 자연언어처리와 언어학 그리고 번역학 간의 연결 고리를 강화하는 데에 의미가 있다. 현단계의 연구 성과를 토대로 향후에 한국어 중심의 기계번역과 포스트에디팅 학습 자원 구축과 관련 평가 기준을 재정비할 수 있도록 더 다양한 번역문쌍에서 다 풍부한 포스트에디팅 유형을 발견할 수 있길 기대된다.

유익한 토론시간을 가지기 위해서, 개인적인 미숙한 의견 몇 마디를 감히 던져본다.

1. 일반적인 문어성이 강한 별렬코퍼스보다 구어성이 강한 방송 대화로 구축된 병렬 코퍼스를 기계번역 품질 평가에 적용할 때, 차이점을 어떻게 생각하시고 이에 필요한 별도의 전처리 조치가 있습니까?
2. 기계번역 자동 평가 과정에서 "ScareBLEU"와 "METEOR" 점수간의 상관성이 존재한다는 언급했는데 이의 원인에 대해서 어떻게 생각하십니까? 혹시 "ScareBLEU"와 "METEOR" 외에 다른 평가 지표도 적용할 계획이 있습니까?
3. 포스트에디팅 유형 분류에 오류와 비오류 2가지 대분류를 설정했는데 오류와 비오류의 경계는 어떻게 구분하십니까? 번역전략에 따르는 비오류성 번역문은 일종의 rewriting 차이로 볼 수 있습니까?
4. 이번 연구 결과로 한중 번역문의 포스트에디팅의 난점 혹은 특성은 어떻게 생각하십니까? 향후 한국어 중심의 기계번역이나 포스트에디팅 자원 정비에 어떤 계획 혹은 조언이 있습니까?

From an interrogative adverb to a discourse marker- 'way' in Korean

Hye-Kyung Lee (Ajou University)

hklee@ajou.ac.kr

Literature Review_1

- (1) 가. A: 아저씨 이 로봇트 얼마예요?
B: 왜, 살려구?
- 나 A: 내일이 무슨 날이지?
B: 왜, 우리 결혼 기념일이잖아
- 다. A: 이 옷 어때? 잘 안 어울리는 것 같아.
B: 왜, 보기 좋은데 (이한규 1997: 2)

→ ‘왜’ as a discourse marker (Dunn 1990; 이한규 1996) → implicature
(Grice 1989)

Literature Review_2

- Interrogative adverb ‘왜’ vs. discourse marker ‘왜’ (신기철·신용철 1981)
 - 1) semantic difference
 - 2) phonological difference
 - 3) co-occurrence with other interrogative adverbs
 - 4) structural difference
 - 5) syntactic dependency

Aims of this Study_1

- To analyze the multi-faceted nature of the Korean word ‘오해’ in actual spoken discourse
- ‘오해’ not only seeks agreement from the listener but also constructs a shared discourse world predominantly initiated by the speaker
- ‘오해’ introduces new subtopics, softens the speaker’s speech acts, facilitates collective recall of past events, or simply fills conversational gaps

Aims of this Study_2

- To analyze the multi-faceted nature of the Korean word ‘오해’ in actual spoken discourse
- ‘오해’ not only seeks agreement from the listener but also constructs a shared discourse world predominantly initiated by the speaker
- ‘오해’ introduces new subtopics, softens the speaker’s objections to the interlocutor’s opinions, facilitates collective recall of past events, or simply fills conversational gaps
- Leverages data from the authentic Sejong Spoken Corpus
- Uses a language-processing tool (Anthony 2024)

Data Collection & Methods

- Collected 238 occurrences of ‘왜’ as a discourse marker in Sejong Corpus
- Classified them into 5 broad functional categories based on their roles in the discourse (c.f. 이한규 1997)
- Manually examined the concordance lines of ‘왜’

Facilitating collective recall_1

어법 다 했나?

예

어법은 다 하고 그 때 그것 좀 봤어? 그 때 그 왜 너 그 문제
집 있었잖아

아니요?

거기 그거 지금 있어?

Facilitating collective recall_2

그래두 점점 교수님들 시험문제 내는 게 황당하지 않아? 어
어 하구 있어.

음. 뭐~ 황당한 거 자체보다두 음

아니 왜 작년에도 누구지? 김동욱 교수님. 애들 다 쓰러지
구

아 나 쓰러지지 않았어. 예상했어.

Seeking agreements/ Creating rapport_1

근데 원래 사람이 빠지지도 않았는데 빠졌다 그러면 더 화 나잖아요.

그렇지

그래가지고, 아니 안 빠졌어:: 그랬어요. 그랬더니 애가 아니 나 다를까, 얼굴에 다 씩여있는데 뭐 그래 이러는 거예요.그 래 더 기분이 나빠 가지고 그냥 가만있는데, 그때부터 애가 나를 얼르는 거예요. 애처럼 아~ 왜 그런 거 있지 않냐?

Seeking agreements/ Creating rapport_2

어 최신곡 안 해. 최신곡은 비슷한 거 많 많이 하더라 사람들.

최신곡의 특징이 뭔지 알아? 한 달만 지나면 나만의 벨소리가 될 수 있어.

진짜야? **왜:: 이 노래 딱 들으면, 그 생각나지 않냐?** 노팅힐에서 줄리아로버츠. 거기 주제가잖아, 난 생각이 안 나구

Introducing a new (related) topic_1

왜 이렇게 토익 야 왜 이렇게 토익에 미쳐 있지? 사람들이

왜 영어 능력 인증 시험에 세 가지가 있어

음

토익 토플 텡스

음

근데 토플이 토플이 장난 아니게 어려::워 시험두 까다롭고

Introducing a new (related) topic_2

멘델버그라는 사람이 결정적 가설이란:: 얘기를 합니다 그
래서 그~ 뇌:: 뇌가 제일 언어든지 제이 언어 학습이든지 그
상관없이 구별 구별 없이 이 뇌의:: 측위화 현상이라는 게
있답니다. 그니까 어떤 자기의 어렸을 때의:: 애들 왜 어린
아기들 여기 머리를 만져 보셨어요? 어때요? 말랑말랑하죠

Mitigating a speech act (objection)_1

요새 그~ 서울 시티 투어 버스 그거 괜찮던데.

안 괜찮어.

타 봤어? 나는 아직 안 타 봤는데.

아 노선을 봤는데 볼 게 없어 그렇게 되면

왜 그래두, 뭐~ 남산 한옥마을두 가구. 뭐~ 시청 뭐~ 대학
로 동대문

Mitigating a speech act (objection)_2

내 꺼두 긴 편인데 그리구 하얀색은 빛을 반사하기 때문에.
좋지망사모자 이런 건 곤란하다. 다 뚫고 들어온다

왜 그래두 물에 들어갔을 땐 좋잖아.

아냐 아냐 곤란해.

Mitigating a speech act (advice)_3

살 좀 어떻게 잘 해 봐

아니 뭐~ 좀 빠지구 있어 요새 쫘.

어 진짜야? 뭐 했는데 살 빠져?

매일 아침 일찍 일어나니까. 일찍 일어나 뭐~ 한 가지만 하더라도 살이 빠지드라구.

일찍 일어나면 좀 동네라두 한바퀴 돌든가 왜~.

아 요새 아침에 안 좋아, 서울시내 공기가 워낙 탁해서.

Rebutting

아 이런 거 난 짝은 건 절대 싫 항상 젤 큰 거 아 젤 큰 건 아니었어.

차!장교라 좋구만.아 별로 춥게 안 입구 왔구만::

별로 춥게 안 입고 왔는데 무얼 춥다고 난리야,

늑어서 그렇다 <laughing>왜

야::! 윤미 신발 샀다 신고식 해야겠다

Discourse filler_1

그게 하필이면 비 오는 날 그랬대.<비 오구 나서 바로 그러니까 젖어 있구
센스 안 좋잖아 그럼 더.

에 <laughing>그거는 아니야.

아냐. 그거 나왔었어 뉴스에.

그래?

센스 확실히 안 좋아진대.

어::

그래서 **왜** 날 맑은 날은 안 올려두 될 게 올리구 막 그런 오류가 많이 난대.

음::

Discourse filler_2

음~ 이 학 그니깐 대학교 이 학년 때부터::, 어느 정도, 내가 내 자신이 어떻게 살아야:: 겠다라는 구 구체적이진 않아두::, 그냥 막연하게::, 어떻게 생각되는 가치관이 생기긴 했는데. 중고등학교 때 내가:: <laughing> 왜,</laughing> 공부:: 라는 게:: 특별히 하기 싫은 것도 아니었지만 안 했다는 거에 지금:: 와서 조금:: 후회가 되기도 하고.

Discussion_1

- **Grammaticalization** from an interrogative adverb to a discourse marker (박석준 2007; Rhee 2020)
- Conceptual extension of asking a reason enables the speaker to question and check whether the listener remembers a known/shared event by
 - 1) inviting the listener to recall it or;
 - 2) making as if the event were known/shared by the interlocutors.

Discussion_2

- Thus ‘와’ as a discourse marker:
 - Solidarity/ Ingroupness invoker
 - Speech act mitigator/ hedge (Brown & Levinson 1987)
 - Discourse filler (Rhee 2020)

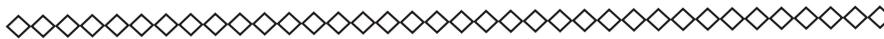
*“The function of a particular DM largely depends **on the context** in which it occurs” (Rhee 2020: 396) (emphasis added)*

References

- 박석준. 2007. 담화표지화의 정도성에 대한 한 논의- ‘뭐, 어디, 왜’를 대상으로. *한말연구* 21, 87-106.
- 신기철, 신용철. 1981. 새 우리말 큰 사전. 삼성출판사.
- 이한규. 1996. 한국어 담화표지어 ‘그래’의 의미 연구 *담화와 인지* 3, 1-26.
- 이한규. 1997. 한국어 담화 표지어 ‘왜’. *담화와 인지* 4-1, 1-20.
- Anthony, L. 2024. AntConc 4.2.4. Software. <https://www.laurenceanthony.net/software/antconc/>
- Brown, P. & S. Levinson. 1987. *Politeness: Some universals in language usage*. CUP.
- Dunn, A. 1990. *The pragmatics of selected discourse markers in Swahili*. Ph.D. thesis. University of Illinois at Urbana-Champaign.
- Rhee, S.-H. 2020. On the many faces of coarseness: The case of the Korean mak ‘coarsely’. *Journal of Pragmatics* 170, 396-412.
- Grice, P. 1989. *Studies in the ways of words*. Harvard University Press.

Thank you.

2024년 한국코퍼스언어학회 여름 전국학술대회
인공지능 시대의 말뭉치 구축과 활용



주제 발표 1



구어 구문 분석 말뭉치 구축과 구문 분석의 실제



일상 대화 말뭉치 구축 현황과 과제



생성형 AI 시대와 말뭉치: 글쓰기 첨삭 지원을 위한 인스트럭션 말뭉치

구어 구문 분석 말뭉치 구축과 구문 분석의 실제

도재학(경기대)
이영제(고려대)
서셋별(국립국어원)
정연주(홍익대)

차 례

1. 머리말
 2. 전처리: 역양구 정비
 3. 구어 구문 분석
 4. 맺음말
-

1. 머리말

- 2024년 구문 분석 및 무형 대응어 복원 말뭉치 연구 과제가 진행되고 있음. 이 과제에서는 ‘2020년 구문 분석 말뭉치 구축 지침’ 및 ‘2020년 무형 대응어 복원 말뭉치 지침’을 개선 및 보완하고, 이를 기반으로 새로운 말뭉치를 구축하는 것을 목표로 함.
- 지침의 기반이 되는 이론적 방법론과 분석의 실제에 관한 연구로 서셋별 외(2019), 이찬영 외(2019) 등이 있음. 서셋별 외(2019)는 보편 의존 구문 분석(Universal Dependencies; UD)에 입각한 한국어 구문 분석 구축 지침을 제안한 연구임. 이찬영 외(2019)는 UD의 방식으로 한 국어 구문 분석을 할 때 발생하는 다양한 문제들을 소개하고 이를 처리하기 위한 방안이 실제 어떻게 적용되었는지 자세히 논의한 연구임.
- 본 발표에서는 구어 구문 분석 말뭉치 구축과 관련된 사안들을 점검하면서, 기존 작업의 문제점을 개선하고자 한 작업의 경과를 다루어 보고자 함. 특히 기존의 구문 분석 지침 및 작업 결과에 대한 검토를 충실히 수행하여, 이전 지침과의 일관성을 유지하면서도 기존에 명시되지 않았던 문제적 사안을 포괄하도록 지침을 정비하고 내용과 예시를 풍부하게 하는 데에 목표를 두었음.

2. 전처리: 억양구 정비

2.1. 억양구 단위를 절 또는 문장 단위로 정비할 필요성

- 구어 발음치의 경우 억양구(또는 발화) 단위로 구축되어 있어서, 문장 단위로 구축된 문어 발음치의 구문 분석 지침이 적확하게 적용되기 어려움.

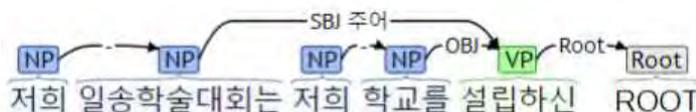
6.4. 불완전한 문장의 주석

- 개별 어절에는 문제가 없지만 비문에 가까운 문장에 대해서는 다음과 같이 의존 관계를 설정한다.
- 해당 어절이 자연스럽게 이해될 수 있는 방안을 상정하여 그에 맞춰 처리하지 않고, 있는 그대로의 표면형을 중시하여 처리한다.
- 감탄사, 절단된 어절이 아닌 완전한 어절을 이루는 경우, 해당 어절이 의존할 수 있는 어절이 문장 안에 있으면 최대한 의존하도록 처리하되 의존할 수 있는 어절이 문장 내에 없는 것으로 판단되면 문장의 최상위 지배소에 의존하도록 처리한다.

<그림 1> 국립국어원(2020: 61)

- 위 <그림 1>에 보인 구문 분석 발음치 지침에 따르면, “의존할 수 있는 어절이 문장 내에 없는 것으로 판단되면 문장의 최상위 지배소에 의존하도록 처리”할 것을 명시하고 있음.
- 그런데 구문 분석의 대상이 되는 국립국어원 일상 대화 발음치는 결과적으로 문장 단위로 나뉘는 것뿐만 아니라 발화의 억양구 단위로도 분할되어 있음. 일상 대화 발음치는 구축 지침상 기본적으로 절 단위나 문장 단위가 되도록 하였고 “가급적 쉽이 있는 구간에서 구분함”이라는 부칙을 두고 있어(국립국어원 2019가: 39), 억양구 단위에서도 분할이 이루어지기도 하였음.
- 이로 인해 의존할 수 있는 어절이 하나의 억양구 내에 없는 경우가 다수 발생하는데, 이러한 경우의 구문 분석에 있어서 문장 내에 의존할 수 있는 지배소가 없는 것으로 보고 억양구 내의 최상위 지배소에 의존하도록 처리한 예가 상당히 많이 확인됨. 이는 구어 구문 분석에서 큰 문제를 야기하는데, 문법적으로 성립 불가능하거나 어색한 의존 구문 분석 결과를 다수 산출하게 되기 때문임.

(1) 주술 관계 오분석

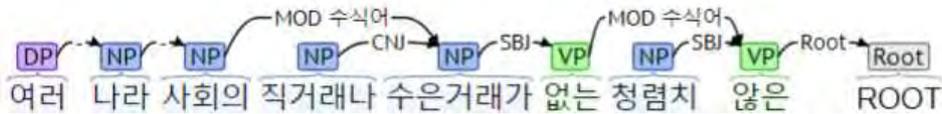


- (1)의 경우, 전체 문장은 “저희 일송학술대회는 저희 학교를 설립하신(ROOT) 고 일

송 인덕선 박사의 유지를 구현하는 학술사업입니다.(ROOT)”임. 여기서 주어(주제어) ‘저희 일송학술대회는’은 전체 문장의 서술어인 ‘학술사업입니다.’에 의존하는 주술 관계로 분석하는 것이 바른 분석임.

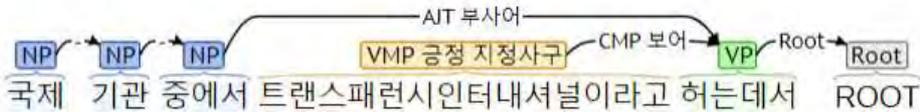
- 그러나 이 구문 분석 말뭉치는 문장 단위가 아닌 억양구 단위를 기준으로 ROOT를 설정하였고, 따라서 ROOT 내에 지배소가 없다는 판단을 하게 된 것임. 결국 주어(주제어) ‘저희 일송학술대회는’을 ROOT 내의 최상위 지배소인 ‘설립하신’에 의존하는 것으로 처리할 수밖에 없음.
- 현재 구문 분석된 억양구만 놓고 본다면, 이 예문은 ‘저희 일송학술대회는 저희 학교를 설립하셨다’로 파악됨. 하지만 ‘학교를 설립하신 것’은 ‘고 일송 인덕선 박사’임. 일상 대화 말뭉치의 분석 단위가 문장이 아닌 억양구 단위인 상태에서, 문장 단위가 아닌 억양구 단위를 기준으로 구문 분석을 수행하였기 때문에 심각한 오분석이 이루어진 것임.
- 실제 문장 구조를 정확히 보여주지 못하는 이러한 오분석은 말뭉치의 활용 가치를 크게 떨어뜨림. 주어의 실제 서술어가 뒤의 억양구에 있더라도 같은 문장 내 서술어인 것이 분명하다면 억양구들을 연결하여 문장 단위로 정비하고, 의존 관계를 적절하게 연결할 수 있도록 할 필요가 있음.
- 문장 단위가 아닌 억양구 단위 분리로 인한 부적절한 의존 관계 설정 사례가 상당히 많이 관찰되는바, 이에 대한 적절한 조치가 요구되는 것임.

(2) 관형어 수식 관계 오분석



- (2)의 경우 지침에 따르면 불가능한 분석은 아니나, 문법적으로는 오분석임. 관형형인 ‘없는다’가 역시 관형형인 ‘청렴치 않은’을 수식하는 의존 관계는 문법적으로 성립이 불가능한 통사 구조임.
- 후행하는 억양구까지를 한 문장 단위로 정비하고, ‘없는다’와 ‘청렴치 않은’의 의존 관계를 적절히 연결해 줄 수 있도록 해야 함.

(3) 부사어 수식 관계 오분석(2)



- ‘국제 기관 중에서’를 ‘허는데서’에 연결하고 있는데, ‘국제 기관 중에서 허는데서’는 불가능한 구조임.
- 후행하는 억양구를 고려하여 ‘국제 기관 중에서 ~~~ 사회적 투명성을 비교해서’로 분석해야 하며, ‘국제 기관 중에서’는 후행 발화의 서술어 ‘비교해서’를 지배소로 하는 의존 관계로 분석되어야 함.

- 이렇듯, 원시 구어 말뭉치의 억양구 단위를 그대로 구문 분석의 단위로 삼는 경우, 말뭉치와 달리 한 발화자의 발화가 여러 개의 억양구로 나뉘어 있어 서술어와 논항 그리고 수식어와 피수식어가 서로 다른 단위로 분리되는 경우가 많았고, 이에 따라 한 성분의 지배소가 하나의 억양구 경계 밖에 있는 경우 의미 없는 분석이 된 경우가 많았음. 따라서 억양구 단위로 분할된 발화를 하나의 절 또는 문장 단위로 정비하는 과정이 필요함.

2.2. 억양구 정비 지침 마련

2.2.1. 논의 경과

- 논의의 전제: 억양구 단위의 분할로 인해 지배소와 피지배소가 분리되지 않도록 하는 것과 함께, 한 문장이 과도하게 길어짐으로 인해서 분석의 효율성이 낮아지는 것을 방지하기 위한 전처리 방법이 필요함. (“====” 표시는 한 사람의 발화 단위임.)

(4) 억양구 단위를 문장 단위로만 정비하는 경우 분석의 효율성에 문제가 되는 사례

- =====
- 1 그러니까 그 참는 거 얘기가 나와서 말인데 그러니까 나도 회사 관두고 시간이 많으니까 회사 다닐 때는 하루에 아침은 거르고 점심 회사에서 먹고 저녁은 이제 일에 치여서 늦게 들어가고 하면 또 지쳐서 나는 지치면 안 챙겨 먹게 되더라고 그러니까 하루에 1끼 많이 먹으면 2끼도 먹는 날이 그렇게 많지 않고 그리고 다니니까 당연히 몸무게가 안 나가지.(540어절)
 - 2 근데 관두고 나서 시간이 내가 많아졌잖아.(60어절)
 - 3 근데 시간이 많으니까 이제 먹는 걸 자꾸 주워 먹게 되는 거야 집에서.(120어절)
 - 4 그 몸에 좋지도 않는 과자들을 내가 과자를 원래 나는 회사 다닐 때 사람들이 책상에 놓고 먹는데 난 그런 거 잘 안 먹었거든.(220어절)
 - 5 난 내가 과자를 안 좋아하는 알았는데 집에서 시간이 많으니까 엄청 과자를 먹게 되더라고 그래서 살 빼려고 그거 참는 게 너무 힘들었어.(210어절)
- =====

- 억양구 단위를 무조건적으로 문장 단위로 복원한 이후 구문 분석을 진행하고자 하는 경우 과도하게 긴 문장을 처리하기가 어려운 문제가 발생함.
- 이에, 지배소와 피지배소가 분리되지 않도록 하는 것을 고려하면서도, 분석 대상 단위가 너무 길어지지 않게 하는 방법이 필요함. 이 점을 고려하여 부사절 단위로도 정비할 수 있는 방안을 수립함.

- 억양구로 분할되어 있는 구어 발화의 절 또는 문장 단위 정비를 위해 아래의 세 가지 방안을 검토하였음. 이 중 1안과 3안을 중심으로 작업의 효율성과 실용성을 고려하여 문제점을 검토함.
- 1안. 구어 말뭉치의 억양구 단위를 보존해 두되 한 억양구 안에서 문장 부호가 나타나는 경우에는 그 부분을 분리한 말뭉치를 바탕으로, 지배소와 피지배소가 분리되어 있는 경우 결합하고, 종결어미가 나타난 부분이 발견되면 분할함.

- 2안. 동일 화자의 연속된 억양구를 결합하고, 필요시 분할함.
→ 2안은 한 발화가 너무 긴 경우가 있어서 명백하게 비효율적이므로 제외함.
- 3안. 동일 화자의 연속된 억양구를 결합하고 문장 부호가 나타나는 경우에는 그 부분을 분리한 말뭉치를 바탕으로, 필요시 분할함.

- **결론적으로는 1안을 채택함.** 1안의 장점은 분명하고, 단점은 그리 결정적이지 않음. 단점은 지배소와 피지배소의 관계를 파악하는 것이 항상 분명한 것은 아니어서 작업의 난도가 높다는 것임. 그러나 적절한 구문 분석이 가능하도록 억양구를 합해야 하거나 분리해야 하는 경우에 대한 사례를 명세해야 하는 것은 3안도 마찬가지임. 반면, 3안의 장점은 이상론에 기대고 있고, 단점은 분명하게 확인/예견되는 바임.
- 위의 안은, 과제 수행을 위해 고려해 볼 수 있는 몇몇 자동 정비 방안에 대한 샘플 테스트 수행 후 가장 효과적이라고 판단한 것임.
- 여러 가능성을 고려하여 자동 정비 방안을 검토하였으나 자연 언어의 특성상 자동 정비 방안은 제한된 범위에서 적용해야 오류 발생 가능성을 줄일 수 있음.
- 억양구 자동 정비 테스트 사례를 간단히 보이면 아래와 같음. 억양구 단위로 분할되어 있는 기존 구어 코퍼스는 아래와 같음.

(5) 기존 구어 코퍼스의 억양구 단위 분할

```

=====
1      나는
2      사실 분야를 하나 정해서 공부하니까 그렇게
3      막막하진 않았어. name1는 어때?
=====

```

- 현재로서는 1의 ‘나는’이 3의 ‘막막하진’과 분리되어 있으므로, 제대로 된 구문 분석이 불가능함.
- 3의 ‘name1는 어때?’는 별개의 문장으로서 구문 분석되어야 함.

- 기존 구어 코퍼스의 억양구를 보존하면서, 한 억양구 내의 문장 부호를 기준으로 자동 분리하는 1안의 적용 결과

(6) 1안을 적용한 자동 분리 결과

```

=====
1      나는
2      사실 분야를 하나 정해서 공부하니까 그렇게
3      막막하진 않았어.
4      name1는 어때?
=====

```

- 억양구를 수동으로 결합하거나 분할하는 경우에 대한 지침을 명세하면 되므로 효율적이며 일관성을 담보할 수 있음.
- 기존 구문 분석의 문제점을 직접적으로 해결할 수 있는 방법임.

- 1안의 장단점

- 장점: 구문 분석에 문제가 없는 한 구어 말뭉치의 억양구 정보를 보존함. 즉, 기구축된 말뭉치에 임의적인 가공을 최소화함. 또한 논항과 서술어가 분리되어 있는 경우를 확인하여 수정하는 방식을 취하므로 기존 구문 분석의 문제점을 직접적으로 해결할 수 있는 방법임.
- 단점: 논항과 서술어의 관계를 모두 확인하여 결합 여부를 판단해야 하므로 시간이 많이 소요되고 작업 과정에서 부담이 될 수 있음.

- 동일 화자의 연속된 억양구를 결합하고, 문장 부호를 기준으로 자동 분리하는 3안의 적용 결과

(7) 3안을 적용한 자동 분리 결과

```
=====
1      나는 사실 분야를 하나 정해서 공부하니까 그렇게 막막하진 않았어.
2      name1는 어때?
=====
```

- 한 문장이 과도하게 길어지는 경우 부사절 분할 여부에 대한 지침을 마련하는 것이 관건인데 이에 대해 간단명료한 지침을 제시하기 어려움.
- 자동 분할된 결과에 대해 작업자들이 검토하는 과정에서 주관적이고 임의적인 판단이 개입됨.

- 3안의 장단점

- 장점: 너무 긴 발화의 경우 몇 개의 부사절로 적절히 쪼개어 구문 분석의 효율성을 담보할 수 있음.
- 단점: 어떤 어미가 나타난 부사절을 분할할 것인지를 결정하는 데에 임의적 판단이 개입될 수 있으며, 분할할 어미 목록을 결정하더라도 해당 어미가 나타난 부분을 기계적으로 분할하면 안 되는 경우가 샘플 검토 과정에서 실제로 발견되었음. 즉, 부사절 분할 여부에 대한 지침을 마련하는 것이 관건인데 이에 대해 간단명료한 지침을 제시하기 어려움. 사실상 지침이 아닌 작업자의 주관에 의존하게 되는 부분이 많아지게 됨. 한편, 사업 제안 기관이나 여타 사용자들이 보기에 구어 말뭉치의 문제적인 부분에 대한 조정을 넘어서, 임의적 변경에 해당한다고 판단할 수 있음.

- 참고로, 부사절 자동 분할 테스트를 해 보았음. 특히 2안 및 3안과 같이 연속된 억양구를 결합하는 경우, ‘분할’ 지침이 명확해야 하고, 그에 따른 분리 결과가 일관되어야 함. 그러나 실제로는 여러 문제가 발생하는 것을 확인하여 실효성이 없을 것으로 판단하고 모든 문장 정비 과정을 수작업으로 진행하기로 함.
- 부사형어미를 기준으로 부사절을 자동 분리하는 방안을 구체화해 보기도 하였음. 부사절 자동 분할을 위한 어미 목록 선정을 위해서 표준국어대사전에 연결어미 및 종결어미로 분류되어 있는 어미 23개, 연결어미로 분류되어 있는 어미 223개를 모두 검토하여, 총 13종류의 부사형어미 목록을 확보함.

- (8) 부사절 자동 분할을 위해 테스트 대상으로 선별한 연결어미 목록
- ㄱ. -거든: [조건]
 - ㄴ. -는데(-는데, -던데, -은데): [배경]
 - ㄷ. -느라(-느라고): [목적], [원인]
 - ㄹ. -다고(-는다고, -ㄴ다고): [이유] (인용절은 분할X)
 - ㅁ. -더니: [원인], [대조]
 - ㅂ. -라: [이유], [대조]
 - ㅅ. -라고(-으라고): [이유], [목적] (인용절은 분할X)
 - ㅇ. -려고(-으려고): [목적]
 - ㅈ. -아도(-더라도, -어도, -여도): [양보]
 - ㅊ. -아서(-라서, -어서, -여서): [원인], [이유]
 - ㅋ. -아야지(-어야지, -여야지): [조건]
 - ㅌ. -지만: [대조]
 - ㅍ. -니까: [이유], [원인]

- (8)의 목록은 ‘-고자, -거들랑, -건마는, -노니, -노라면, -을는지, -을세라, -르사’ 등등, 의고적이거나 문어체인 것, 사용빈도가 현저히 낮은 것, 보조적 연결어미, 반복 구성에 쓰이는 것(과도 분할을 초래함) 등의 어미를 제외한 것임.
- 위의 부사형어미가 포함된, 4어절 이상의 언어 단위를 분할한다는 지침을 두고 테스트를 진행함. 아래와 같이 논항과 서술어가 한 단위 안에서 묶이지 않는 경우가 다수 발생하므로 자동 정비를 할 수 없다는 결론에 이르게 됨.

- (9) 부사절 자동 분할의 문제적 사례
 ㄱ. ‘-니까’에 의해 분리된 주어와 서술어

=====

1 **나는** 사실 분야를 하나 정해서 공부하니까

2 그렇게 **막막하진** 않았어.

3 name1는 어때?

=====

ㄴ. ‘-는데’에 의해 분리된 주어와 서술어

=====

1 **아빠는** 항상 엄마가 결혼 전이나 지금이나 17년이 흘렀는데

2 똑같은 몸무게와 똑같은 몸매를 **유지하고** 있어.

=====

- 또한, 삽입어구나 미완성 발화를 식별하여 분할하는 방안도 고려하였음. 그러나 삽입어구 및 미완성 발화만을 검출할 수 있는 구분자를 명세하기 어렵고, 삽입어구나 미완성 발화를 사이에 두고 논항과 서술어가 분리되는 경우들이 다수 발생함바, 자동 정비를 할 수 없다는 결론에 이르게 됨.

- (10) 삽입어구 분할의 문제적 사례
 ㄱ. 삽입어구에 의해 분리된 주어와 서술어(1)

=====

1 지금은 동물도 완전히 인권 해 가지고 **동물도** 뭐

- 2 그 단체 있잖아요.
- 3 동물 보호하는 단체도 **있고**

=====

ㄴ. 삽입어구에 의해 분리된 주어와 서술어(2)

=====

- 1 이 **개도** 말이야 영물이 그런가
- 2 내가 딱 가면은 자기가 나와서 내가 서열 1번이라는 것을 **아는가**

=====

- (10ㄱ)의 경우 삽입어구인 ‘그 단체 있잖아요’ 앞뒤의 억양구에 논항-서술어 관계가 있음. 소유주 주어 ‘동물도’가 소유대상 ‘동물 보호하는 단체도’ 및 서술어 ‘있고’와 분리됨. 즉, 억양구를 합해야 함.
 - (10ㄴ)의 경우 주어 ‘이 개도’는 담화표지 ‘말이야’, 삽입어구인 ‘영물이 그런가’와는 관계가 없고, 억양구 2의 서술어 ‘아는가’와 관계가 있음. 즉, 억양구를 합해야 함.
- 이렇듯, 문장 부호 외의 다른 여타의 구분자를 기준으로 하여 부사절, 삽입어구 및 미완성 발화 등을 자동으로 정비하려는 시도는 실효성이 없을 것으로 판단하여 억양구 정비 작업은 수작업으로 진행하기로 함.

2.2.2. 지침의 원칙과 예시

- 대원칙
 - 1. 구어 발음치에 대한 수정/변개는 최소화한다.
 - 2. 적절한 구문 분석이 가능하도록 논항(피지배소)과 서술어(지배소)를 가급적 하나의 분석 단위 내에 두도록 한다. 여기서 분석 단위라 함은 절 또는 문장을 말한다.
- 하위 원칙
 - 1. 하나의 억양구 내에서 적절한 구문 분석이 가능하면 문장 정비의 대상으로 삼지 않는다.
 - 2. 하나의 억양구 내에 구문 분석에 필요한 의존 관계(피지배소-지배소 관계)가 확인되지 않아 적절한 구문 분석이 불가능한 경우 복수의 억양구를 합하여 의존 관계를 판단한다.
 - 3. 하나의 억양구 내에 둘 이상의 문장이 포함되는 경우 효율적인 구문 분석을 위해 억양구를 분리한다.
- 문장 정비 대상으로 삼지 않는 경우
 - ① 하나의 억양구 내에 피지배소와 지배소가 함께 나타나는 경우

=====

- 1 **어 여기서 학교 얘기가 나와서**
- 2 아마 여기서 학력이 제일
- 3 높은 사람이 말하는 게 낫 낫지 않을까?

- =====
- 억양구 1의 경우 ‘여기서’, ‘학교 얘기가’라는 피지배소 성분이 지배소(서술어) ‘나와서’와 연결될 수 있음. 피지배소 성분과 그것이 연결될 지배소(서술어)가 하나의 억양구 내에서 함께 나타나 있으므로 술어-논항 관계를 설명하는 데 있어서 문법적으로 문제가 없음. 따라서 문장 정비의 대상으로 삼지 않음.

② 후보층 구문의 경우

- =====
- 1 그래 너무 칙칙해 보일 거 같다고 **회색이랑 그런 게**
 - 2 그래서 조금 밝은 게 낫지 않을까 막 이런 얘기도 나오긴 했었어.
- =====

- ‘회색이랑 그런 게’는 후보층 현상에 의해 나타난 주어 성분으로, 앞에 지배소가 있으므로 문장 정비의 대상으로 삼지 않음.

③ 발화 수정 등에 따른 억양구에서 발화 자체에 서술어가 생략된 경우

- =====
- 1 나는 중학교 때부터 과학에 관심이 없어서
 - 2 **공대는 별로 관심**
 - 3 공대를 복수전공하는 거를
 - 4 생각을 하지 않고 있었어.
- =====

- 억양구 2의 내부에는 성분이 연결될 서술어가 존재하지 않음. 그러나 억양구 2는 발화 수정 과정에서 발화가 중단된 것으로서, 본디 서술어가 존재하지 않는 것으로 볼 수 있음. 따라서 다른 억양구와 합하지 않고 그대로 둠.

④ 하나의 억양구 내에 나타나는 감탄사, 응답어, 담화표지 등의 경우

- =====
- 1 **어** 여기서 학교 얘기가 나와서
- =====

- 응답어 ‘어’를 분할하지 않음.

⑤ 문장 부호로 이미 분리되어 있는 감탄사와 응답어의 경우

- =====
- 1 **진짜?**
 - 2 그럼 뮤비만 내고
 - 3 음방을 안 뛰었어?
- =====

- ‘진짜’가 응답어로 쓰였으나 문장 부호로 이미 분리되어 있으므로 정비 대상으로 삼지 않음.

- 문장 정비 대상으로 삼는 경우: 뒤의 억양구와 합하는 경우

⑥ 하나의 억양구 내에 지배소가 존재하지 않는 경우, 의존 관계를 이룰 수 있는 지배소가 나올 때까지 뒤의 억양구를 이어서 정비함.

- 1 어 여기서 학교 얘기가 나와서
- 2 **아마 여기서 학력이 제일**
- 3 높은 사람이 말하는 게 낫 낫지 않을까?
- 4 **너랑 아 name1랑 나는 아직**
- 5 학교를 졸업을 안했으니까
- 6 **그래서 대학을 졸업한 먼저 배우신 분한테**
- 7 말을 넘기는 게 나올 것 같아.

- 억양구 2의 경우 ‘여기서’, ‘학력이’, ‘제일’이라는 성분이 연결될 서술어가 존재하지 않음. 이 성분들은 억양구 3에 있는 서술어 ‘높은’과 연결되어야 함. 따라서 억양구 2와 3을 합함.
- 억양구 4도 ‘너랑 name1랑 나는’과 ‘아직’이 연결될 서술어가 존재하지 않음. 이 성분들은 억양구 5의 서술어 ‘했으니까’와 연결되어야 함. 따라서 억양구 4와 5를 합함.
- 억양구 6의 ‘대학을 졸업한 먼저 배우신 분한테’ 역시 연결될 서술어가 억양구 6 안에 존재하지 않음. 이 성분은 억양구 7의 서술어 ‘넘기는’과 연결되어야 함. 따라서 억양구 6과 7을 합함.

[결과]

- 1 어 여기서 학교 얘기가 나와서
- 2 **아마 여기서 학력이 제일** 높은 사람이 말하는 게 낫 낫지 않을까?
- 3 **너랑 아 name1랑 나는 아직** 학교를 졸업을 안했으니까
- 4 **그래서 대학을 졸업한 먼저 배우신 분한테** 말을 넘기는 게 나올 것 같아.

⑦ 후보층 구문이 분리된 억양구로 제시된 경우

- 1 근데 그 스키니가 몸매가 너무 적나라하게 다 보여.
- 2 **자기 단점들이.**

- 여기에서 ‘자기 단점들이’는 후보층 현상에 의해 문장 뒤에 단독으로 나타난 성분으로서 억양구 1 ‘보여.’의 주어에 해당함. 그러므로 억양구 1과 합함.

[결과]

- 1 근데 그 스키니가 몸매가 너무 적나라하게 다 보여. **자기 단점들이.**

• 문장 정비 대상으로 삼는 경우: 억양구를 분할하는 경우

- ⑧ 하나의 억양구 내에 종결 부호 없이 종결 어미만 나타나는 경우, 종결 어미를 기준으로 하여 분할함.

- 1 닭꼬치 파는 친구 옆에서 보조를 해 **줬어** 왜냐하면
- 2 내가 한강 밤도깨비 야시장에서
- 3 어 떡갈비 꼬치를 팔았던 전적이 있거든.

- 억양구 1의 ‘줬어’는 종결 부호는 없지만 종결 어미로 끝난 어절로서 한 문장의 종

결부에 해당함. 이 부분을 분할함.

[결과]

- 1 닭꼬치 파는 친구 옆에서 보조를 해 **줬어**
- 2 왜냐하면 내가 한강 밤도깨비 야시장에서 어 떡갈비 꼬치를 팔았던 전적이 있거든.

3. 구어 구문 분석

3.1. 기존 구어 구문 분석 지침 검토

- 한국정보통신기술협회(2015) 및 국립국어원(2020)에 제시된 구문 분석 지침의 기본 원칙을 따르며, 구어 구분 분석을 위한 지침을 조금 더 명세하고자 함.
- 구문 분석의 기본 원칙(국립국어원 2020: 25-26)
 - (1) 자연 언어 처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에서 크게 벗어나지 않도록 한다.
 - (2) 문장의 표층 구조를 중시하여 분석한다.
 - (3) 의존 관계 분석의 기본 단위로 어절을 사용한다.
 - (4) 지배소 후위 원칙에 따라 각 어절의 지배소는 자신보다 뒤에 위치하도록 분석한다.
 - (5) 각 어절은 1개의 지배소를 가진다.(Single-Head Constraint)
 - (6) 각 어절 및 지배소 쌍은 서로 교차하지 않는다.(Projective Constraint)
 - (7) 보어와 부가어를 구분하되 보어의 범위를 엄격히 제한한다.
 - 보어 CMP는 보격 조사가 부착된 NP, 용언구, 절, 그리고 인용절 보문의 용언구와 절에 한해서 분석한다.
 - 조사가 생략되거나 보조사가 부착된 명사구 또는 이에 상응하는 용언구와 절도 서술어 구문 틀에 따라 보격 조사로 대치 가능하면 CMP로 분석한다.
 - (가) 그가 돌아왔다고(VP_CMP) 그녀가 알려줬어.(TTA, 9쪽)
 - (나) 그녀가 그 일을 했다고(VP_CMP) 스스로 말했다.(TTA, 10쪽)
 - (다) 비평가 칼라일이 “인도와도 바꿀 수 없다.”고(VP_CMP) 말하였다.(TTA, 13쪽)
 - (라) 물이 얼음이(NP_CMP) 되었다.
 - (마) 철수가 발이 아프다고(VP) 훈련을 빠졌다.(VP)
 - (바) 마법사가 와인을 물로(NP_AJT) 바꾸었다.
 - (사) 철수가 영희가(NP_SBJ) 보고 싶다.
- (8) 원칙적으로 접속과 내포를 구별하지 않으며, 접속절은 모두 부사절로 분석한다.(다만, 명사구 접속은 인정함)
- (9) 하나의 성분이 모문과 내포문 모두에 관련되어 있으면 내포절의 유형에 따라 해당 주어의 지배소를 결정한다.

- 기존 구어 구문 분석 지침의 목차 체계 검토

6.1. 감탄사의 범위와 처리	
6.1.1. 부사, 대명사와 감탄사의 구별	6.1.2. 감탄사의 주석
6.2. 주제어의 주석	
6.3. 호칭어의 주석	
6.4. 불완전한 문장의 주석	
6.5. 불완전한 어절의 주석(절단된 어절, 발화 수정된 어절 등)	
6.5.1. 절단된 어절의 처리	6.5.2. 발화 수정의 처리
6.5.3. xx, xxx 등의 처리	
6.6. 반복되는 어절, 공지시어의 주석, 어미 반복 구성의 주석	
6.6.1. 반복되는 어절의 처리	6.6.2. 공지시어의 의존 관계 설정
6.6.3. 어미 반복 구성의 처리	
6.7. 여러 문장이 한 주석 단위에 포함되는 경우	
6.8. 인용 표지가 없는 인용의 주석	
6.9. 구조적 중의성이 있는 명사구의 주석	
6.10. 한 어절 안에 여러 단위가 포함되어 있는 경우	
6.11. 문장의 마지막 어절이 '-다고'류 어미로 끝나는 경우	
6.12. '에'로 전사된 관형격 조사 '의'의 처리	
6.13. 미등제어의 처리	

<그림 2> 국립국어원(2020: 60-69)

- 6.1절부터 6.5절까지는 단어(6.1절, 6.2절, 6.3절)와 불완전한 단위(6.4절, 6.5절)를 구별하고 어느 정도 체계적으로 다루었다고 할 수 있음.
- 그러나 6.6절 이후부터는 실제 구문 분석에서 개별 구문들의 특성에 따라 유념해야 할 필요가 있는 경우들을 나열하고 있어서 앞부분에서 장절을 이루는 항목들이 유기적이고 체계적으로 연계되어 있는 것에 비해 상대적으로 구별되는 차이가 있음.
- 구어 자료에 대한 적절한 구문 분석을 위해서는 기본 원칙을 준수하면서도 구어 자료에서 많이 나타나는 사례들을 위한 추가적인 지침 수립이 필요함. 따라서 목차의 체계를 정비하고 사례들을 보다 적절한 것으로 제시할 필요가 있음.
- 기존 지침의 수정·보완을 위해 작업 대상인 구어 코퍼스를 직접 검토하면서 기존 지침에 반영되어 있지 않았지만 구어에서 현저히 발견되는 현상들을 확인하고, 이를 적절히 분석하기 위해 필요한 지침을 추가하였음.

3.2. 주요 보완 사항

3.2.1. 직접 인용과 간접 인용의 구별

- 구어 자료에서는 직접 인용과 간접 인용이 형식적으로 잘 구별되어 표시되지 않으나, 설령 겹따옴표가 없더라도 '이라고/라고'를 기준으로 직접 인용인지 간접 인용인지 확

인하는 것이 필요함.

- 구문 분석 지침 3.2.4절의 인용절에 관한 분석 지침에 따르면, 동일 주어일 경우 직접 인용은 모문 주어를 모문 서술어에 연결하고, 간접 인용은 모문 주어를 인용문 서술어에 연결하도록 되어 있음(국립국어원 2020: 37-38). 이 지침은 무형 대응어 복원과도 관련이 있는데, 직접 인용문 내에서 생략된 논항은 원발화자의 시점에 맞게 복원하도록 되어 있기 때문임.
- 직접 인용인 경우 인용문 내부에서는 생략된 논항을 복원할 때 인칭까지 고려해서 복원하도록 되어 있음. 가령 ‘철수가 “(내가) 할 수 있지 않을까?” 하던데’처럼 직접 인용되었을 때, ‘할’의 주어는 ‘철수가’로 연결하면 안 되고 ‘나’를 복원해야 함.
- 그런데 구어 자료에서는 따옴표가 없고 인용격조사는 잘 나타나지 않는 경우가 많다 보니 간접 인용처럼 간주하여 ‘철수가’를 ‘할’에 연결하면 무형 대응어 복원 과정에서 문제가 생기게 됨. 이와 관련하여 구어에서의 직접 인용과 간접 인용에 대한 구별이 필요함.

(11) 구어에서의 직접 인용과 간접 인용 구별

- ㄱ. 저도(→NP_SBJ 생각했는데) 그 코로나 이후에 새롭게 직장을 구해 볼까라고 생각했는데
- ㄴ. 철수가(→NP_SBJ 간다) 학교에 간다 했어.
- ㄷ. 철수가(→NP_SBJ 했어.) 학교에 갈까? 했어.

- (11ㄱ)에서처럼 직접 인용격 조사 ‘이라고/라고’가 쓰인 경우 동일 주어이면 모문 주어를 모문 서술어에 연결함
- 그러나 직접 인용격 조사가 쓰이지 않은 경우는, (11ㄴ)과 같이 간접 인용격 조사 ‘-다고, -냐고, -라고, -자고’로 표현할 수 있는 경우는 간접 인용으로 보고 동일 주어인 ‘철수가’를 인용문 서술어 ‘간다’에 연결함. 반면 (11ㄷ)과 같이 간접 인용격 조사로 표현할 수 없는 경우는 직접 인용으로 보고, 동일 주어인 ‘철수가’를 모문 서술어 ‘했어.’에 연결함.

3.2.2. 후보층 구문의 처리

- 후보층 구문은 구문 분석의 기본 원칙 중 네 번째 사항 ‘(4) 지배소 후위 원칙에 따라 각 어절의 지배소는 자신보다 뒤에 위치하도록 분석한다.’에 직접적으로 위배되는 사례인데, 구어에서 굉장히 빈번하게 사용되는 것임에도 이에 대한 지침이 부재하였음.
- 후보층 구문은 앞선 발화에서 언급하지 못한 부분을 뒤에서 부연하는 기능을 하는 것으로, 문장의 화용적 구조와 관련하여 상대적인 구정보 또는 신정보를 전달하는 부분임(김민국 2021: 28-29, 최윤지 2021: 97-98).
- 즉, 단독적인 또는 자립적인 발화라기보다는 앞선 발화 내용에 통사적, 의미적, 기능적으로 의존하는 것임을 감안하면 구문 분석은 선행하는 지배소에 의존하는 것으로 처리해야 적절하다고 할 수 있음.
- 구문 분석의 기본 원칙인 지배소 후위 원칙을 중시하여 후보층 표현이 지배소로 분석되도록 하기는 하였으나, 이렇게 분석하는 것이 전산적 차원에서 어떤 장점이 있는지,

반드시 지켜져야 할 원칙인지에 대해서는 근본적인 고민이 필요할 것임.

(12) 후보층 구문의 분석 사례

ㄱ. 그런데 안 한대요(→VP 본인은.) 본인은.(→NP_SBJ ROOT)

ㄴ. 혹시 봤나요(→VP 주만지?) 주만지?(→NP_OBJ ROOT)

ㄷ. 짝꿍이 됐어.(→VP 여름부터.) 여름부터.(→NP_AJT ROOT)

ㄹ. 그래서 손톱만 봐도 그 사람들은 알더라고(→VP 직업을) 그(→DP 사람의) 사람의(→NP_MOD 직업을.) 직업을.(→NP_OBJ ROOT)

- (12ㄱ,ㄴ,ㄷ)과 같이, 후보층 구문은 그 지배소가 앞에 있지만, 지배소 후위 원칙에 따라 의존 관계는 후보층 구문에 표시한다는 것을 지침에 명시함.
- 또한 (12ㄹ)과 같이, 후보층 구문이 두 어절 이상의 복합 표현으로 나타나는 경우, 일반적인 구문 분석 지침에 따라 처리하며, 후보층 구문에 선행하는 문장의 최상위 지배소는 후보층 구문의 최상위 지배소에 연결하는 것으로 지침을 마련함.

3.2.3. 주격/목적격 조사 결합형이 절 표지로 기능하는 경우

- 주격/목적격 조사 결합형이 절 표지로 기능하는 경우가 있음.
- 특히 주격 조사 결합형이 절 표지로 기능하는 경우는 ‘열린 분열문’이라고 불리며, 분열문과 형식적, 의미적으로 밀접히 관련되지만 다소 비표준적인 성격을 가지는 준분열문 중에서 분열된 성분이 독립적인 절의 지위를 가지는 특성으로 인해 어학적 분석의 관심이 되어 왔음(최윤지 2016: 198).
- 이들 사례는 구문 분석 과정에서(특히 기능 태그 할당에서) 문제가 됨. 격조사 ‘이’, ‘을’의 형태가 확인되지만, 그렇게 분석할 수 없는 경우이기 때문임.

(13) 주격/목적격 조사 결합형이 절 표지로 기능하는 경우의 분석 사례

ㄱ. 맞아 그런 게 좀 확실히 있어 보이는 게(→NP_AJT 잘하잖아.) 유럽 애들이 그런 걸 되게 잘하잖아.

ㄴ. 소금을 넣는다는 걸(→NP_AJT 넣었네) 설탕을 넣었네.

- (13)의 예에서처럼 절 표지로 기능하는 ‘것이, 게, 것을, 걸’ 등의 형태는 격조사를 기준하여 논항으로 분석할 수 없다는 문제가 있으므로, 배경의 의미를 나타내는 절 단위에 준하는 것으로 보아 부사어로 처리하기로 함.¹⁾
- 이는 제2의 기본 원칙인 표층 구조를 중시하여 분석한다는 것에는 다소 벗어나는 것이지만, 제1의 기본 원칙인 자연 언어 처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에서 크게 벗어나지 않도록 한다는 것을 고려한 것임.

1) 이는 구문 분석 지침 5.2.1절의 서술어의 역할을 하는 ‘-으로’를 인정하는 것과 유사한 사례라고 할 수 있다.

3.3. 기존 지침을 더 구체화한 사례: 감탄사의 유형 구별

- 감탄사에 관한 기존 지침은 아래와 같음.

- 부사나 대명사인지 감탄사인지가 모호한 경우에는 감탄사(IP)로 처리한다. 특히, 부사 또는 대명사로 해석하여 의존소와 의존 관계를 설정하였을 때 교차 의존 금지 제약을 어기게 되는 경우에는 감탄사(IP)로 처리하여 바로 다음 어절에 의존하도록 분석한다.

<그림 3> 국립국어원(2020: 60)

- 그런데, 모든 감탄사를 바로 다음 어절에 의존하도록 처리하는 것이 괜찮은지에 대해 생각해 볼 필요가 있음. 감탄사를 대략 감정감탄사, 머뭇거림, 담화표지, 부름, 대답 등으로 나눈다고 할 때, 감정감탄사, 머뭇거림, 담화표지 등의 요소는 바로 다음 어절에 의존하도록 분석하는 것에 문제가 없음.
- 그런데, ‘야, 애, 여보, 이보시오’ 따위의 부름은 호칭어와 유사함(예. ‘야, 이리 와 봐.’, ‘*이보시오, 이리 와 봐.’). 그리고 ‘예, 네, 아니오, 응, 아니’ 따위의 대답은 후행하는 서술어와 호응함(예. ‘예, 먹었습니다’, ‘예, 먹었어요’, ‘*예, 먹었다’, ‘*예, 먹었어’, ‘*응, 먹었습니다’, ‘응, 먹었어요’, ‘응, 먹었다’, ‘응, 먹었어’). 즉, 다른 유형의 감탄사에 비해 부름과 응답은 의미가 명확하고 문장 층위의 일치 현상과 관계하는 특성을 보임.
- 이를 고려하면, 부름과 대답의 감탄사는 무조건 바로 다음 어절에 의존하는 것으로 처리할 것이 아니라, 후행하는 서술어와 연결하는 것이 적절함.

6.1.2. 감탄사의 주석

- 문장 부호의 유무와 관계없이 감탄사는 IP로 바로 다음 어절에 의존하도록 처리한다.

(가) 예.(→IP 간장하구요) 간장하구요 설탕 또 요렇게 넣으시고.

(나) 그래?(→IP 나도) 나도 모르겠던데

<그림 4> 국립국어원(2020: 60)

- <그림 6> (가)의 ‘예.’는 ‘간장하구요’와 ‘넣으시고’에 연결 가능한데, 이런 경우 가까운 지배소에 연결함. (나)에서 ‘그래?’는 ‘나도’가 아닌 ‘모르겠던데’와 연결하는 것이 더 적절한 구문 분석임.
- ‘응 암을 예방한다고 들었어.’, ‘예 저도 해리 포터 봤었어요.’와 같은 문장에서, 부름의 용법으로 사용된 ‘응, 예’는 각각 ‘암을, 저도’에 연결하는 것보다는 ‘들었어, 봤었어요’와 연결하는 것이 더 적절한 구문 분석임.

- 참고로, 표준국어대사전에 감탄사로 등재된 단어 650개를 검토하여, 부름말, 응답어, 용언 활용형을 선별해 보았음(감정감탄사, 옛말, 욕설, 아동대상어, 군대제식용어 등은 제외함).
 - 부름말(38개): 새-아가, 아가(01), 야(04), 애(02), 어이(05), 여(04), 여-보(01), 여-보게, 여-보세요, 여-보셔요, 여-보쇼, 여-보시게, 여-보시오, 여-보십시오, 여-보아라, 여-봅시오, 여-봐, 여-봐라, 여보게나, 여봐-요, 업쇼, 이-보(01), 이-보게, 이-보세요, 이-보쇼, 이-보시게, 이-보시오, 이-보십시오, 이-보우, 이-봅시오, 이-봐, 이봐-요, 저-보게, 저-보세요, 저-보시게, 저-보시오, 저-보십시오, 저봐-요
 - 응답어(26개): 그래(01), 그래-그래, 그러게, 그럼(02), 글썸, 글썸-다, 글썸-올시다, 글썸-요, 네(03), 아냐, 아뇨, 아니(02), 아니-야, 아니-오, 아니-요, 아무렴, 암(02), 예(06), 예-예(01), 예-이(02), 옛(02), 오냐, 오냐-오냐, 옴소, 응(01), 응-응(02)
 - 용언 활용형 감탄사(17개): 고런(02), 그런(02), 그렇지, 맞다(04), 맞아, 옳다(02), 옳다-구나, 옳아, 옳지, 요렇지, 이런(02), 이렇지, 저런(02), 저렇지, 조런(02), 조렇지, 좋아
- 이런 작업을 수행한 첫 번째 이유는, ‘아니예요, 아닙니다’ 등과 같이 사전에 감탄사로 등재되어 있지 않지만 응답어로 사용되는 단어를 구별하고 구문 분석에서 VP로 처리하기 위함임. 그러나 이들은 유사 감탄사로 보아 무형 대응어를 복원할 때에는 복원 대상 서술어에서 제외함.
- 두 번째 이유는, ‘그렇지, 맞아’ 등 용언 활용형 감탄사의 경우 사전에 제시된 감탄사 용법인지 확인한 후, 그렇지 않으면 VP로 처리하기 위함임. ‘그렇지, 오늘이 네 생일이구나.’에서 틀림없이 그렇다는 뜻을 표시하거나, ‘맞아, 오늘 보기로 했었지.’에서 어떤 기억이나 생각이 갑자기 떠올랐다는 뜻을 표시하는 경우에는 감탄사임. 그러나 질문에 대한 대답에 사용된 경우는 용법을 확인하여 처리해야 함. 단, 이들 또한 유사 감탄사로 보아 무형 대응어를 복원할 때에는 복원 대상 서술어에서 제외함.

3.4. 구문 분석의 문제적 사례: ‘이/가’ 주제어의 주석

- 구어 자료에서는 ‘이/가’로 표시된 주제어, 특히 격 조사 사용의 오류처럼 보이는 경우가 자주 확인됨.
 - (5) ‘이/가’로 표시된 주제어
 - ㄱ. **이게** 보니까, **이게** 보면
 - ㄴ. **거기가** 내가 많이 가 봤어.
 - ㄷ. 그 맛집 **짜장면이** 철수도 먹고 영수도 먹고 나도 먹었어.
- (5ㄱ)은 ‘이걸(<이것을)’, (5ㄴ)은 ‘거기에’ 또는 ‘거기를’, (5ㄷ)은 ‘짜장면을’이 쓰일 법한 맥락에 ‘이게, 거기가, 짜장면이’가 사용된 것으로 보면, 격 조사 사용의 오류처럼 느껴지기도 함. 그러나 명백한 오류라고 하기에는 상당히 자연스럽게 느껴지고, 단순히 목적이거나 부사어로 처리하는 것도 주저되는 면이 있음.

- (5)의 사례들을 어떻게 처리할 것인지는 상당히 고민스러운데, 일단 여기서 지적해 둘 것은, 이들을 격 조사 사용의 오류가 아닌 자연스러운 용법으로 볼 수 있다는 점임. 특히 비인가 화제의 경우 ‘이/가’ 외에도 ‘말이야’류, ‘있잖아’류, NP-∅ 등으로도 자주 실현되는데(박진호 2015: 389-390), (5)의 사례도 비인가 화제를 나타내는 것으로 볼 수 있음.
- (5)와 같은 사례에 대해서는 두 가지 입장이 가능할 것임. 첫째는 명백한 격 조사 사용의 오류라고 보고, 오류가 아닌 경우의 기능 태그를 부여하는 방안. 둘째는 주어진 형태를 중시하여 주어로 처리하는 방안. 특히 둘째 입장을 선택하는 경우 무형 대응어 복원 단계에서 목적어 또는 부사어의 복원이 어려워지는 문제가 있지만, 일단은 형태를 중시하여 주어로 처리하기로 함. 일관성 유지 및 일반 언어학적 관점의 설명 방식을 존중하고자 한 것임.

3.5. 구어 구문 분석을 위한 추가지침 내용 목차

- 구어에서 자주 확인되는 구문 분석 관련 이슈를 작은 언어 단위에 관한 것부터 큰 언어 단위에 관한 것으로 구별하고 재정리하여 목차를 구성함. 기존 지침과의 대응 관계는 괄호로 표시하였음. 내용이 수정, 보완된 부분에는 굵게 및 밑줄 처리함. 설명 및 사례는 생략함.

1. 단어 단위

1.1. 감탄사의 경우(기존 지침 6.1절 내용)

1.2. 호칭어의 경우(기존 지침 6.3절 내용)

1.3. 주체어의 경우(기존 지침 6.2절 내용)

1.4. 관형격 조사 ‘의’가 ‘에’로 전사된 경우(기존 지침 6.12절 내용)

1.5. 미등재어의 경우(기존 지침 6.13절 내용)

2. 어절 단위

2.1. 발화 수정 과정에서 나타난 어절의 경우(기존 지침 6.5.1절, 6.5.2절 내용)

2.2. 무엇인지 알기 어려운 어절의 경우(기존 지침 6.5.3절 내용 포함)

2.3. 어절이 반복되는 경우(기존 지침 6.6.1절 내용)

2.4. 직접 인용과 간접 인용의 구별(새로 추가됨)

2.5. ‘-다고’류 어미로 종결된 경우(기존 지침 6.11절 내용)

2.6. 한 어절 내에 여러 단위가 포함된 경우(기존 지침 6.10절 내용)

3. 절, 문장 단위

3.1. 의미가 불완전한 문장의 경우(기존 지침 6.4절 내용)

3.2. 서술어가 없는 절, 문장의 경우(기존 지침 6.4절 내용)

3.3. 후보층 구문이 사용된 경우(새로 추가됨)

3.4. 어미가 반복되는 경우(기존 지침 6.6.3절 내용)

3.5. 공지시적인 표현이 쓰이는 경우(기존 지침 6.6.2절 내용)

3.6. 구조적 중의성이 있는 명사구의 경우(기존 지침 6.9절 내용)

3.7. 열린 분열문 구성의 경우(새로 추가됨)

3.8. 인용 표지가 없는 인용의 경우(기존 지침 6.8절 내용)

3.9. 마침표가 포함된 경우(기존 지침 6.7절 내용)

4. 맺음말

생략.

참고문헌

- 국립국어원(2019가). 일상 대화 말뭉치 구축, 국립국어원.
- 국립국어원(2019나). 구문 분석 말뭉치 구축, 국립국어원.
- 국립국어원 (2020). 2020년 국립국어원 구문 분석 말뭉치 구축 지침.
- 김민국(2021). 한국어 정보구조의 초점, 한국어학 91, 1-50.
- 박진호(2015). 보조사의 역사적 연구, 국어학, 73, 375-435.
- 서새별 외(2019). 한국어 보편 의존 관계 분석에 관한 제언, 언어와 정보 23(1), 91-122.
- 이찬영 외(2019). 한국어 보편 의존 구문 분석 (Universal Dependencies) 방법론 연구, 언어 사실과 관점 47, 141-175.
- 최윤지(2016). 구어에서 유의미하게 나타나는 어떤 통사구성에 대하여: 준분열문, 국어학 79, 187-237.
- 최윤지(2021). 구정보, 신정보란 무엇인가: 신구성의 구별과 정보구조적 조정, 한국어학 91, 95-126.
- 한국정보통신기술협회(TTA) (2015). 의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법.

일상 대화 말뭉치 구축 현황 및 과제

2024년 7월 17일

이용주, 차원철(나라지식정보)

이광수(스마트미디어테크)

강미영, 서셋별, 이선영(국립국어원 언어정보과)

일상 대화 말뭉치?

▶▶▶ 말뭉치?

- Corpus(speech, text), 음성 또는 텍스트 DB, 언어자원(Language resource), (AI 학습용)data set
- 형태에 따라
 - ▶ speech corpus
 - ▶ text corpus
 - ▶ spoken language(SL) corpus : 음성 및 이를 전사한 텍스트의 모음
- ❖ 재활용(또는 공유)를 위해 필요한 부가정보를 갖추어 전자적으로 정리된 자료의 모음
 - ▶ 개인의 사용을 위해 만들어진 것을 공유
 - ▶ 처음부터 공유를 목적으로 만들어진 것

코퍼스 공유의 예(음성코퍼스의 경우)

● 외국

- ▶ 미국 : LDC(Linguistic Data Consortium)(1992~)
- ▶ 유럽 : ELRA(European Language Resources Association)(1995~)

● 국내

- ▶ 음성정보기술산업지원센터(SiTEC)(2001~) 산업자원부(구 지식경제부) 지원
- ▶ 전자통신연구원(ETRI) (2002~) 구 정보통신부 지원
- ▶ 국립국어원 일상 대화 말뭉치(2019~) 등 문화체육관광부 지원
- ▶ 한국지능정보사회진흥원(NIA) AI 학습용 데이터(2018~) 과학기술정보통신부 지원

● 국제적 협력

- ▶ LREC(The International Conference on Language Resources and Evaluation)(1998~)
- ▶ Oriental COCOSDA : (1998~)
The International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (“Inter Speech”의 SIG 형태)의 Asian Chapter

음성코퍼스의 형태(vs 텍스트 코퍼스)

용도	<ul style="list-style-type: none"> ● 음성/화자 인식, 음성 합성 등 음성신호 처리 기술의 공학적 응용연구용 및 기초연구용(한국음성학회)
대상 미디어	<ul style="list-style-type: none"> ● 음성을 포함한 음향 데이터(음성, 소음, 환경음 등), 멀티모달 데이터(음성 + 영상(제스추어 등))
녹음 수단	<ul style="list-style-type: none"> ● 마이크로폰, 전화기(유 무선) -> 다채널 녹음
녹음되는 거리에 따라	<ul style="list-style-type: none"> ● 근거리/원거리 음성
언어 형태	<ul style="list-style-type: none"> ● 낭독 음성, 대화 음성, 자유발화 음성
언어별	<ul style="list-style-type: none"> ● 각국의 언어(외국어), 지역어(방언), 한국인의 외국어, 외국인의 한국어 등
특수한 화자 대상	<ul style="list-style-type: none"> ● 마비말 장애 음성, 치매 음성 등 각종 질환 환자 음성 등
화자 연령별	<ul style="list-style-type: none"> ● 성인 음성/아동 음성/노인 음성

SiTEC 음성 DB

- 구축 현황(2001~2006) : 총 48종, 20,000여 명 발성, 800GB 분량
- 자동차 응용을 위한 음성 DB
- 수출 지원을 위한 외국어 음성 DB(영어, 스페인어, 중국어, 일본어)
- 기반기술연구용 음성 DB : 음소 분할된 단어 및 문장 음성 DB, 운율 정보(K-ToBI) 부가된 문장 DB 등
- 언어교육을 위한 음성 DB(한국인의 영어, 외국인의 한국어, 중국인의 영어, 일본인의 영어)
- 복지 응용 및 기타 음성 DB
 - ▶ 노인 음성, 아동 음성, 장애 음성, 정서 음성
 - ▶ 마이크 성능 평가용 음성, 인식 성능 평가용 음성
 - ▶ 기기 내장형 음성, 로봇 음성, 멀티모달 음성, 고소음 음성
 - ▶ 주소 음성, 원거리 음성
- 수탁 제작 : MS, ELRA 등 외국업체, 현대자동차, KT, SKT 등 국내기업, ETRI 등 연구기관, 서울대, KAIST 등 대학
- 수탁보급 : (예) K-sec(한국인의 영어), 마비말 장애 음성 등

ETRI 음성 DB

- **공통 음성 DB**

- ▶ 다양한 녹음기기 환경에서의 단어, 숫자, 낭독 및 대화체 문장 등

- **신성장동력 산업용 한국어 음성 DB**

- ▶ 지능형 로봇을 위한 원거리/다채널 음성
- ▶ 멀티모달 환경에서의 음성 명령어 및 정보 검색용 대화체 문장
- ▶ 자동차 환경에서의 음성 명령어 및 정보 검색용 대화체 문장

- **다국어 언어/음성 DB**

- ▶ 가정/자동차 환경 음성 DB(한국어, 영어, 중국어, 일본어)
- ▶ 자동 통역을 위한 외국어 음성 DB

(일본어, 중국어, 스페인어, 프랑스어, 독일어, 러시아어, 아랍어, 베트남어, 태국어, 인도네시아어, 말레이어)

-> 2018 평창동계올림픽

코퍼스의 현황 및 문제(음성의 경우)

▶▶▶ SiTEC과 ETRI의 선구적인 역할

➡ 음성처리기술의 패러다임의 변화에 따라 양(Quantity)의 문제 대두(HMM → 딥러닝)

▶▶▶ NIA의 AI 학습용 데이터 세트 구축 사업

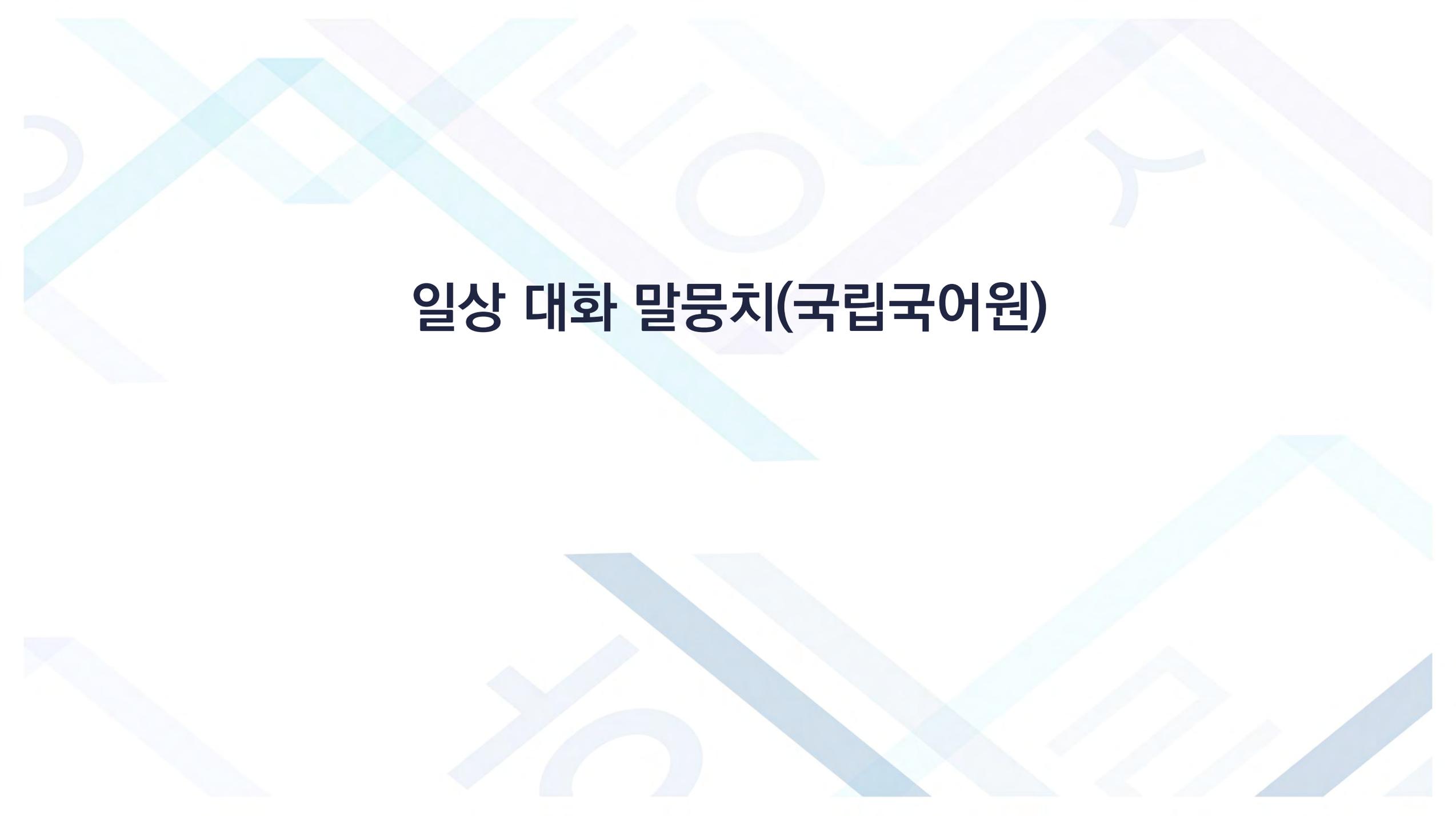
➡ 종류의 다양성과 양의 문제는 어느 정도 해결



질(Quality)의 문제는?

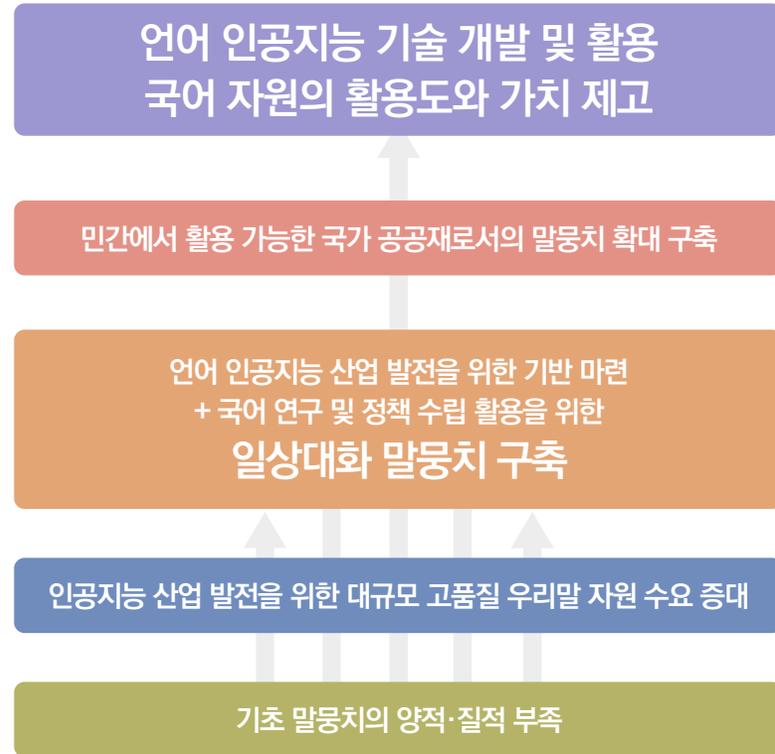
➡ 깊이 있는 지속적인 코퍼스의 구축 필요

➡ 일상 대화 말뭉치(국립국어원)?



일상 대화 말뭉치(국립국어원)

일상대화 말뭉치 구축의 목적



- 언어 인공지능 등 관련 산업 활용을 위한 기반 마련
- 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용
- 기초 말뭉치의 양적·질적 부족에 따른 기반 기술 개발 및 인공지능 기술 개발 수준 지체를 해소
- 국어 자원의 활용도와 가치를 높일 수 있도록 민간에서 활용 가능한 국가 공공재로서의 말뭉치 확대 구축

연도별 구축 현황

국립국어원 일상 대화 말뭉치 구축 현황

일상 대화 말뭉치 구축 경과



2024년 사업 개요

- 사업명

- ▶ 2024년 일상 대화 말뭉치 구축

- 사업 기간

- ▶ 2024년 3월 28일 ~ 2024년 11월 15일(약 8개월)

- 수행 기관

- ▶ 나라지식정보 컨소시엄(나라지식정보 + 스마트미디어테크)

- 사업 내용

- ▶ 660시간 분량의 2인/3인/4인 대화
- ▶ 40시간의 1인 발화(공적 독백 20시간, 사적 독백 20시간)

구축 과정

1. *말뭉치의 설계(특히 주제의 설계) : 무엇을 어떻게 만들까?*
2. *음성 녹음*
3. *음성 전사*
4. *메타데이터의 부가 및 비식별화 처리(공개에 부적절한 내용 및 개인정보)*

주제별 구축 현황

번호	2019년	2020년	2021년	2022년	2023년	세부 주제	비고
1	군대						기타에 포함
2	게임						취미에 포함
3	휴일		휴가	휴가	여행/휴가	국내 여행 경험 및 계획, 해외여행 경험 및 계획 등	
4	자동차		대중교통	대중교통			제외
5	만화						취미에 포함
6	영화	영화	음악	음악			취미에 포함
7	정치						제외
8	건강/다이어트	건강/다이어트	건강/다이어트	건강/다이어트	건강/다이어트	건강, 다이어트, 식단, 건강 보조제, 질병 관련 경험과 증상 등	
9	방송/연예	방송/연예	방송/연예	방송/연예	방송/영화/연예인	텔레비전(TV) 프로그램(드라마, 예능 등), 영화, 연예인	
10	스포츠/레저	스포츠/레저	스포츠/레저	스포츠/레저/취미	취미	음악 이론, 음악 활동, 미술 이론, 미술 활동, 헬스, 골프, 수영 등	
11	먹거리	먹거리	먹거리	먹거리	먹거리	음식, 요리, 좋아하는 요리, 요리법, 식재료 쇼핑, 조리 도구, 조리 가전, 간편식 활용법 등	
12	자연/휴양지						여행/휴가에 포함
13	국가/지역				사회 이슈	인공지능(AI) 기술 발전, 누리 소통망, 환경 문제, 인구 감소, 고령화 등	
14	문학						제외
15	연애/결혼	연애/결혼	우정	우정			가족/관혼상제에 포함

주제별 구축 현황

번호	2019년	2020년	2021년	2022년	2023년	세부 주제	비고
16	경제/재테크		경제/재테크	경제/재테크	경제/재테크	경제, 재테크, 예금, 적금, 주식, 코인, 부동산 등	
17		여행지(국내/해외)					여행/휴가에 포함
18		계절/날씨					제외
19		회사/학교	회사/학교	회사/학교	회사, 학교생활	직장 및 학교생활, 업무 내용, 업무 강도, 야근, 회식, 회의, 승진 등	
20		선물					제외
21		꿈(목표)					기타에 포함
22		반려동물	반려동물	반려동물	반려 동식물	반려 동식물 관련 경험 및 조언, 추천 등	
23		아르 바이트	취직	취직	취직	진로, 직업, 취직, 이직, 취준생, 해외 취업, 일반 자격증, 전문 자격증 등	
24		성격			인간관계	친구, 연애, 학교 동기, 직장 동료, 성격상 장점 등	
25		가족	가족				가족/관혼상제와 통합
26			쇼핑	쇼핑	쇼핑	선호 쇼핑물, 선호 브랜드, 쇼핑 방식, 중고 거래 등	
27			관혼상제	가족/관혼상제	가족/관혼상제	가족, 결혼, 출산, 성인식, 결혼식, 장례식, 명절 등	
28				생활/주거 환경	생활/주거 환경	가사 활동, 가사 관련 가전, 리모델링, 인테리어 등	
29				기타	기타	군대, 추억, 꿈, 인생 목표, 인생 계획, 가치관 등	
30					패션/미용	패션 스타일, 얼굴, 몸매, 화장, 피부 관리, 헤어스타일링 등	

2024년도 주제의 설계

● 다자 대화 주제

01	영화, 드라마 , 전시회, 공연
02	먹거리, 맛집 , 요리법
03	스포츠, 레저, 취미, 게임, 만화, 책, 독서
04	여행, 휴가, 휴일, 자연휴양지
05	방송 연예, 예능 , 아이돌, 한류
06	건강, 다이어트, 질병
07	반려동물, 반려용품
08	우정, 성격, MBTI

09	쇼핑, 선물
10	경제, 재테크, 부동산 , 금융
11	연애 결혼, 가족, 관혼상제
12	생활, 주거 환경
13	취직, 아르바이트
14	회사, 학교, 학창시절
15	스마트기기, 인공지능, 메타버스 등 IT 관련 주제

● 발화량

- ▶ 2인 대화, 3인 대화, 4인 대화 공히 화자 당 4개의 세션에 참여(화자 당 발화 시간 약 30분 분량)

● 발화 비율

- ▶ 2인 75%, 3인 15%, 4인 10%

● 1인 발화(독백)

- ▶ 공적 독백 : 유튜브 등의 매체를 통해 다중을 위한 유튜브 혼자 진행하는 발화(예: 게임채널)
- ▶ 사적 독백 : 혼자 길게 얘기할 수 있는 주제를 제시하고 발화하도록 함(복수의 주제, 화자 당 15분 정도)

사적 독백 주제문(예)

주제	내용
여행	"여행을 좋아하시나요? 가장 기억에 남는 여행 경험이 있으신가요?"
음식	"요리를 하시나요? 어떤 요리를 만드는 게 취미이신가요?"
책	"어떤 종류의 책을 좋아하시나요? 가장 인상깊게 읽은 책이 있으신가요?"
문화체험	"지역의 문화에 관심이 있으신가요? 어떤 문화 체험이 인상 깊으셨나요?"
흥미	"가장 관심 있는 취미가 무엇인가요? 그 취미에 빠진 이유가 있나요?"
기술	"최근에 관심 있는 기술이 있으신가요? 그 기술이 일상생활에 어떻게 영향을 미칠 것 같나요?"
건강	"건강에 대해 어떻게 생각하시나요? 건강을 유지하기 위해 무엇을 신경 쓰시나요?"
가치관	"가장 소중히 여기는 가치관이 있나요? 그 가치관이 일상생활에 어떻게 반영되나요?"
연애	"연애나 가족에 대한 생각이 있으신가요? 이상적인 가족 모습이나 연애에 대한 꿈은 무엇인가요?"
미래	"미래에 대해 어떻게 생각하시나요? 자신의 꿈과 목표는 무엇인가요? 그것을 이루기 위해 무엇을 해야 할 것 같나요?"

음성 녹음

● 자연스러운 일상 대화 녹음이 가능한 스튜디오 환경 구축

녹음 장비 및 소프트웨어

● 녹음 장비



- 최대 4채널(4명)의 마이크 입력이 가능
- USB 연결로 안정적인 데이터 녹음

[Focusrite Scarlett 18i8 3rd Gen USB Audio Interface]



- 사용자의 움직임이 있어도 동일한 음량을 유지
- 가까운 소리에 특화되어 외부 잡음에 강인함

[SHURE Closetalk Mic._Shure WH20]

● 수집 소프트웨어



- 최대 8채널까지 레코딩이 가능한 소프트웨어 사용
- 채널별 레코딩으로 화자별 모니터링 및 오버랩 대응 가능
- 채널 별로 자동 분리하여 후처리 과정에서 활용 가능

녹음 환경



자연스러운 대화 유도를 위해
전문 교육을 받은 오퍼레이터가 모니터링을 진행



[하울링 감소를 위한 거치형 흡음재 설치]

- 데이터 녹음을 위해 최대 8개 녹음 사이트를 동시 운영
- 녹음 장소의 위치 및 내부 가구 배치 유무에 따라 하울링 및 외부 소음 방지 발생 가능
- 녹음 공간 내부에 흡음재 사용, 배경 노이즈 최소화

음성 녹음

- 녹음 시 주안점

- ▶ 성별, 나이, 출신지역 등 인구통계학적으로 균형된 화자의 섭외
- ▶ 주어진 주제 범위 안에서 원활한 대화 유도(사전 준비)
- ▶ 배경 소음, 녹음 레벨 관리
- ▶ 배포에 부적절한 내용은 녹음 중 모니터링하여 제한
- ▶ 16Khz 샘플링, 16bit 양자화

음성 전사

● 효율적인 전사도구(ADoc Cube)를 사용한 철자 및 발음 전사

Project > 2024_nara03-SDRW2400... > SDRW2400000978.wav

MetaData | 단축키 | HIDE

Sub title: 2인 일상 대화
Author: 개인 발화자
Publisher: 개인 발화 녹음
Date: 20240622
Topic: 우정/성격/MBTI

CLIP LIST | TASK LIST | 삭제 | 92 구간

002	B	이렇게 되면서 참 그- 성격서부터 인내심서부터 그런 게 너무 좋았어요.
003	B	그래서 아 앞으로는 그- 김 선생님하고의 그- 유대가 이어지겠다 이런 생각을 하고 있거든요.
004	B	그런데 저는 이제 직장 생활을 안 해 봤기 때문에 이 사회에서 많은 여러 여러 그- 운동도 하고 그래서 여러 사람들을 많이 만나는 가운데 김 선생님을 (이케)/(이렇게) 만나서 너무 행복합니다.
005	B	근데 김 선생님은 또 직장 생활도 하고 하셨죠?
006	B	그래서 김 선생님 이야기를 듣고 싶어요.
007	A	네.
008	A	저는 졸업하자마자 직장 생활도 (쫀)/(쫀) 못 낸 했어요.
009	A	직장 생활도 (쫀)/(쫀) 한 (삼사 년)/(3 4년) 하고 결혼하고 그다음에 또 다 카우고 난 뒤에 시니어 와 가지고 이렇게 또 좋은 선생님을 만나고 그리고 하다 보니까 이게 너무 적성에 맞아 가지고 참 많이 오래 했네요.
010	A	근데 친구도 참 여러 친구가 있더라고요 성격상.
011	A	그래 가지고 아 그때 따라서는 아 이게 아닌데 할 때도 있지만 다시 또 생각하고 생각하고 하다 보니까 또 성격도 바뀌고 내가 배려를 할 줄 아는 것도 배우고 선생님 말씀 같은 거를 많이 경청하면서 이렇게 살아오니까 순조롭게 진행되어더라고요.
012	A	그래서 사회생활이라는 게 어렵기도 하지만 즐겁게 느낄 때도 많아요.
013	A	선생님도 그렇죠?
		저도 그- 그중에 그- 이제 취미 생활을 하면서 그- 여

SDRW2400000978 | 14 분 50 초 | 1.0 X

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

위표 문장받지기 | 아래표 문장받지기 | 문장나누기 | 1.0 X | 1.0 X

B 002 | 이렇게 되면서 참 그- 성격서부터 인내심서부터 그런 게 너무 좋았어요.

B 003 | 그래서 아 앞으로는 그- 김 선생님하고의 그- 유대가 이어지겠다 이런 생각을 하고 있거든요.

B 004 | 그런데 저는 이제 직장 생활을 안 해 봤기 때문에 이 사회에서 많은 여러 여러 그- 운동도 하고 그래서 여러 사람들을 많이 만나는 가운데 김 선생님을 (이케)/(이렇게) 만나서 너무 행복합니다.

B 005 | 근데 김 선생님은 또 직장 생활도 하고 하셨죠?

B 006 | 그래서 김 선생님 이야기를 듣고 싶어요.

A 007 | 네.

A 008 | 저는 졸업하자마자 직장 생활도 (쫀)/(쫀) 못 낸 했어요.

- ▶ STT 자동 전사
- ▶ 자동 문장 구분
- ▶ 화자 및 문장 단위 구분
- ▶ 구간별 음성파일 보기
- ▶ 음절 단위 구간 일치 확인

음성 전사

● 음성 전사 시 주안점

- ▶ 발음 및 철자 전사(이중 전사) -> 전사 지침 참조
- ▶ STT 결과를 보조 정보로 활용
- ▶ 경험 많은 전문 전사 인력 투입
- ▶ 전사 오류 최소화를 위한 비슷한 언어권(방언권) 및 연령대 전사자의 투입
- ▶ 알고리즘을 이용한 문장 구분 지원기능 활용
- ▶ 전사 및 품질관리(검수) 등을 통합지원하는 도구의 활용

세부 구축 방안

● 말뭉치 형식 : JSON 구조

수준 1	수준 2	수준 3	수준 4	타입	설명
id				string	말뭉치 파일 아이디
metadata				object	말뭉치 파일의 메타 정보
	title			string	말뭉치 파일 제목
	creator			string	구축자: 국립국어원
	distributor			string	배포자: 국립국어원
	year			string	구축년도: 2024
	category			string	분류: 구어 > 사적 대화 > 일상 대화
	annotation_level			array(string)	분석 층위: 원시
	sampling			string	샘플링 방식: 본문 전체
document				array(object)	대화 정보
	id			string	대화 아이디
	metadata			object	대화 메타 정보
		title		string	대화 제목: 2인 일상 대화
		author		string	저작권자: 개인 발화자
		publisher		string	발행자: 개인 발화 녹음
		date		string	녹음일자: YYYYMMDD
		topic		string	대화 주제: 대주제 > 세부주제
		speaker		array(object)	화자 정보
			id	string	화자 아이디
			age	string	연령
			occupation	string	직업
			sex	string	성별
			birthplace	string	출생지
			residence	string	주 성장지
			current_residence	string	현 거주지
			education	string	학력
		setting		object	환경 정보
			relation	string	화자 간 관계
			contact_frequency	string	친밀도
	utterance			array(object)	발화 정보
		id		string	발화 아이디
		form		string	철자 전사
		original_form		string	발음 전사
		speaker_id		string	화자 아이디
		start		num	발화 시작 시간
		end		num	발화 종료 시간
		note		string	전사자 기타 메모

● 메타정보

- ▶ 말뭉치 파일 메타 정보(파일 제목, 구축 및 배포자, 구축년도, 분류 등)
- ▶ 대화 정보(대화 제목, 녹음일자, 대화 주제 등)
- ▶ 화자 정보(연령, 성별, 직업, 출생 및 성장지, 학력 등)
- ▶ 환경 정보(화자 간 관계, 친밀도 등)
- ▶ 발화 정보(철자 전사, 발음 전사, 발화 시작 및 종료 시간 등)

음성 전사 지침

1. 기본 원칙

- 음성 자료의 전사는 발화된 그대로 전사하는 발음 전사와 한글 맞춤법 및 표준어 규정에 따른 철자 전사를 병행(이중 전사)한다.
- 발음 전사는 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 적용하여 소리나는 대로 한글로 적는다.
- 철자 전사는 한글 맞춤법 및 표준어 규정에 따라 적는다.
- 발음 전사는 숫자, 외래어, 기호, 단위 등도 한글로 적는다.
- 느낌표나 쉼표는 사용하지 않으며 문장이 완전히 종결되었을 때는 마침표를 사용한다.
- 억양에 의해 의미가 달라지는 경우 마침표와 물음표를 사용하여 구분.(‘응’, ‘네’, ‘-어’, ‘-어요’ 등)

음성 전사 지침

2. 화자 표시

- 화자 아이디, 성별, 연령, 직업, 출생지, 주 성장지, 현 거주지, 학력 등 화자 정보를 표시한다. 화자에 대한 정보를 모를 경우에는 'NA'로 표시한다.
- 본문 전사에서 화자 정보와 화자 표시는 반드시 일치해야 하고 화자가 분명하지 않을 경우에는 'NA'로 표시한다.

3. 전사 단위

- 음성 분절 및 전사의 기본 단위는 문장이 되도록 한다.

4. 발화 겹침

- 겹침 발화는 표시하지 않고 시간 순서에 따라 적는다. 만약 맞장구 발화가 일어날 경우 맞장구 발화를 사이에 넣어 주 발화를 나눈다.

주 발화: 1: 딸 하나 낳아서
맞장구 발화: 2: 네.
주 발화: 3: 세 살 먹어 잊어버리고

음성 전사 지침

5. 발화 내용 전사

- 발화 내용은 기본적으로 철자 전사를 하되, 구어의 발음 특성이나 개인적인 발음 특성, 지역적인 특성 등의 이유로 표준 발음에서 벗어난 형식으로 발화 하거나 표준 발음이 여러 개인 경우 등 한글 맞춤법 표기에 따른 발음과 차이가 있는 경우에 발음 전사를 한다.

철자 전사: 자 상담소에는 어떤 걸 기대하고 왔을까?
발음 전사: 자 상담소에는 어떤 걸 기대하고 왔으까?

- 각 전사에 사용할 수 있는 문자는 아래와 같다.(X를 제외한 알파벳, 비식별화 일련번호를 제외한 숫자, 수식 기호 등 사용 금지)

	발음 전사	철자 전사
사용 가능 문자	. (마침표)	. (마침표)
	? (물음표)	? (물음표)
	~ (담화표지)	. (소수점)
	- (불완전발화)	
	' (모음의 축약형)	
	@ (비식별화, 준음성)	
	() (이중괄호)	
	X (잘 들리지 않는 경우)	
사용 불가능 문자	X를 제외한 알파벳	알파벳
	비식별화 일련번호를 제외한 숫자	수식 기호
	수식 기호	

음성 전사 지침

- 발음 전사 시 기호, 외래어 등은 발음에 따라 한글로 적는다.
- 발음 전사 시 모음의 변화, 수의적 경음화 등을 반영하여 전사한다.

철자 전사: 어떡해
발음 전사: 어뜩해

철자 전사: 소주
발음 전사: 싹주

- 각 전사에 사용할 수 있는 문자는 아래와 같다.(X를 제외한 알파벳, 비식별화 일련번호를 제외한 숫자, 수식 기호 등 사용 금지)

6. 모음의 축약형 표기

- 모음의 축약형의 경우 대부분 현재 국어의 모음 체계상 표기할 글자가 존재하지만, 반홀소리된 /ㄱ/, /ㄴ/의 표기는 문제가 된다.

/ㄱ/, /ㄴ/가 반홀소리가 되어 /ㄱ/, /ㄴ/와 축약되는 현상은 구어에서 자주 나타나는데, 한글의 현재 글자 체계상 이러한 현상을 반영할 방법이 없으므로 전사에서는 '를 사용해서 두 음소를 연결해 준다.

철자 전사: 사귀어
발음 전사: 사귀'어

음성 전사 지침

7. 준말과 센말의 전사

- <국립국어원 우리말샘>에 등재된 준말(한 단어 안에서 탈락이나 축약 현상이 일어난 것)과 센말은 철자 전사 시 본딴말로 복원하지 않고 발화된 대로 기재한다.

준말 예)

근데(그런데), 얘기(이야기), 요새(요사이), 요즘(요즈음), 애(아이), 담(다음), 맘(마음), 첨(처음), 널(내일), 젤(제일),
좀(조금), 재밌다(재미있다), 갖다(가지다), -곤(-고는), 뭐(무어), 오랜만(오래간만), 암튼(아무튼),
쌤(선생님), 알바(아르바이트), 킬로(킬로그램), 프로(퍼센트) ...

센말 예)

조끔, 쪼끔, 쫘끔(조끔), 쫘쫘(졸졸), 딱딱하다(단단하다)

음성 전사 지침

8. 끊어진 단어(단어가 불완전하게 발화된 경우)

- 끊어진 단어는 발화된 대로 그대로 전사한다. 불완전하게 발화된 단어(어절)가 둘 이상인 경우에도 어절마다 다음과 같이 표시하여 전사한다.(수정 발화, 반복 발화에 표시하는 것은 아님)

철자 전사: 전 전전통이라고 우리가 흔히 얘기할 때 발음 전사: -전- -전-전통이라고 우리가 흔히 얘기할 때
--

- 내용상 수정 발화와 불완전 발화가 복합적으로 나오는 경우 혹은 수정 발화인지 불완전 발화인지 구분이 모호한 경우에는 어절 앞뒤로 '줄표(-)'를 넣는다.

9. 띄어쓰기

- 한글 맞춤법(제5장 띄어쓰기)에 맞게 띄어 쓴다.
- 단어를 발음하는 중간에 쉼이 들어간 경우에는 띄어쓰지 않는다.

음성 전사 지침

10. 담화 표지

- 머뭇거림의 기능을 하는 1음절 담화 표지 중 “이, 그, 저, 아, 어, 예, 음, 응, 뭐”의 9개 형태에 한해서 본래의 품사와 구별하기 위해 물결표(~)를 붙여 전사한다.
- 억양과 운율에 의해서만 구분이 가능할 경우는 반드시 전사 단계에서 표시해 준다.

철자 전사: 많은 경우에 논문그 어 연구는 네이션 국가라는 거하구 직결되는 과정이죠.
발음 전사: 많은 경우에 논문그~ 어~ 연구는 네이션 국가라는 거하구 직결되는 과정이죠.

11. 잘 들리지 않는 부분

- 잘 들리지 않는 부분의 전사 시 이중 괄호((xxx))를 이용한다. (철자 전사에서는 “이중 괄호(())” 삭제)
- 잘 들리지 않아 추정된 경우는 다음과 같이 전사한다.

음성 전사 지침

11. 잘 들리지 않는 부분

- 잘 들리지 않는 부분의 전사 시 이중 괄호((xxx))를 이용한다. (철자 전사에서는 “이중 괄호(())” 삭제)
- 잘 들리지 않아 추정된 경우는 다음과 같이 전사한다.

철자 전사: 그 전까지는 직장 생활 하느라고 더 힘들어
발음 전사: 그 전까지는 직장 생활 하니라구 ((더 힘들어))

- 화자의 발화 내용이 전혀 들리지 않는 부분은 다음과 같이 전사한다.

철자 전사: 너무나 거 같더라.
발음 전사:(()) 너무나 거 같더라.

- 들리지 않는 음절은 그 음절의 수만큼 x를 붙여 다음과 같이 전사한다.

철자 전사: 근데 그거 진짜 xx해야 되겠더라.
발음 전사: 근데 그거 진짜 ((xx해야)) 되겠더라.

음성 전사 지침

12. 준음성과 기타 소리들

- 웃음, 목청 가다듬는 소리, 박수, 노래 등은 다음과 같이 전사한다.

웃음: {laughing}
목청 가다듬는 소리: {clearing}
박수: {applauding}
노래: {singing}

*철자 전사에서는 삭제한다.

13. 숫자 전사(상세)

- 숫자의 철자 전사는 이중 전사한다.
- 발음 전사 시 숫자는 발음에 따라 한글로 적는다.
- 철자 전사 시 숫자는 일반적인 표기 관습(숫자, 한글 혼용)에 따라 적는다.
- 숫자 철자 전사의 띄어쓰기는 “경”, “조”, “억”, “만” 단위로 띄어 쓴다.
- 철자 전사 시 천 단위 분할 “,”(쉼표)는 쓰지 않는다.

음성 전사 지침

14. 방언의 전사

- 방언(발음 전사)에 대한 표준어 대응쌍(철자 전사) 이중 전사
- 우리말샘에 등재된 방언형의 경우 발음 전사는 방언형을 소리나는 대로 기본 형태를 살려 적고, 철자 전사는 뜻풀이의 표준 어형을 기준으로 삼는다.

철자 전사: 그런데
발음 전사: 근디

*준말의 방언형은 표준어의 본딴말로 통일

- 방언 발음 전사 시 유의 사항은 다음과 같다.

방언과 관련이 없는 표현은 표준어를 적는 방식으로 쓰되, 방언 표현은 방언의 특색이 드러나도록 표기한다.

이때 방언의 표기는 음성 그대로 소리나는 대로 쓰지 않고 방언의 형태가 드러나는 방식으로 쓴다.

- 어두 된소리화의 경우, 방언의 특성으로 볼 수 있으므로 소리나는 대로 전사하고, 표준어 대응쌍 이중 전사를 한다.

철자 전사: 저번에
발음 전사: 저번에

철자 전사: 다르다
발음 전사: 따르다

음성 전사 지침

15. 비식별화를 위한 전사

- 일상 대화 자료 중 개인정보 등의 비식별화를 위해 이름, 이메일 주소 등 계정 정보, 주민등록번호, 카드 번호, 전화번호 등 각종 번호 및 비밀번호, 상세 주소, 출신 및 소속 등의 개인정보와 관련된 사항은 노출되지 않도록 전사 단계에서 비식별화한다.
- 정치인 등 유명인의 이름은 비식별화하지 않으며, 상호명 및 상품명 등은 부정적인 경우에만 비식별화한다.
- 주소는 동 이하의 구체적인 주소만 비식별화하며, 동 이상의 주소는 그대로 전사한다.
- 비식별화 정보는 아래와 같이 마크업한다.

이름: &name&
상호명: &company-name&
계정(아이디): &account&
주민등록번호: &social-security-num&
전화 번호: &tel-num&
카드 번호: &card-num&
기타 번호: &num&
주소: &address&
출신 및 소속: &affiliation&
기타 비식별화가 필요한 항목: &others&

품질 관리

- 다수의 장소에서 음성 데이터 수집(전국 8개의 수집 사이트 운영)
 - ▶ 자연스러운 발화의 유도
 - ▶ 녹음 레벨의 안정적 관리

- 다수의 전사 인력을 투입하는 음성 전사
 - ▶ 다양한 전사 오류 형태 발생 가능
 - ▶ 오류의 최소화를 위한 단계별 점검 및 지속적인 교육과 내부 작업 지침의 보완

말뭉치 배포 창구



국립국어원 “모두의 말뭉치”

<https://kli.korean.go.kr/corpus/main/requestMain.do?lang=ko>

일상 대화 말뭉치 구축 사업의 지속적인 확장을 위한 논의

▶▶▶▶ 한 방향으로 깊이 있게 지속적으로 추진해온 일상 대화 말뭉치 사업

➡ 바람직한 추후 방향은?

- ▶ 기존 대화 주제의 확장(양, 종류)? → 시의성 있는 주제의 추가?
- ▶ 새롭게 등장하는 대화 콘텐츠?
- ▶ 새로운 대화 형식의 추가 시도?

➡ 결과물의 활용 확대를 위한 방안?

➡ 타 사업과의 차별화 문제

가장 한국적인 것이 가장 세계적인 것?

감사합니다.



문화체육관광부
국립국어원

KACL KOREAN
ASSOCIATION FOR
CORPUS
LINGUISTICS

국립국어원

생성AI 시대와 말뭉치 : 글쓰기 첨삭-지원 인스트럭션 말뭉치

2024. 07. 17

Imagine Is Real 
IIR TECH



1. 말뭉치 개요와 역사

생성형 인공지능과 말뭉치

- 1 생성형 AI 시대에 말뭉치가 필요한가?
- 2 말뭉치란 무엇인가?
- 3 생성형AI는 무엇을 필요로 하는가?
- 4 생성형AI가 필요로 하는 말뭉치는 어떻게 만들어져야 하는가?

1. 말뭉치 개요와 역사

■ 말뭉치의 개념, 정의

➤ 말뭉치란?

요즈음 유럽에서 "뭉치 언어학"이라는 새 용어가 쓰이고 있다. 영어로는 "콜퍼스 링귀스틱스(corpus linguistics)"라고 한다. "콜퍼스"는 "몸뚱이"를 뜻한다. 좀 엉뚱한 인연으로 해서, 언어학자가 아닌 필자가 아마도 처음으로 이 언어학 용어를 옮겨서 쓰게 되는 모양이다. 애초에는 덩치 언어학이라고 했었는데, 어감이 안 좋다고 해서 고쳐본 것이다. 일반적으로 "콜퍼스"는 몸뚱이, 즉 옹근 덩어리를 뜻하지만 저작물에 관하여 쓰일 때에는 한 저자자의 저작 전부, 또는 한 특정 분야의 저작 전체를 뜻한다. (중략) 그러니까 언어학적 의미의 한 말뭉치는 **어떤 기준으로든 한 덩어리로 볼 수 있는 말의 뭉치**를 가리킨다. (이상섭, 사전편찬학 연구 2집, 162쪽)

➤ 말뭉치란?

언어를 연구하는 각 분야에서 필요로 하는 연구 재료로서, **언어의 본질적인 모습을 총체적으로 보여 줄 수 있는 자료의 집합**(“국어정보학 입문”, 서상규 외 1998)

최근의 말뭉치의 이용 환경은 예외 없이 어떠한 형태이건 간에 전산적 처리를 전제로 하고 있고, 1980년대 중반 이후에 데이터 저장 및 처리 장치와 같은 하드웨어, 자료를 처리하기 위한 소프트웨어와 기술의 발달, **연구의 목적과 방법의 다양화** 등으로 인해서, 이제는 말뭉치라고 하면 예외 없이 **‘대규모로 전산화된 말뭉치’**를 뜻하게 되었기 때문이다. (서상규, 인지언어학회 제2차 국제학술대회(2003))

1. 말뭉치 개요와 역사

■ 딥러닝 이전 시대의 주요 말뭉치(국외)

- 브라운 말뭉치(Brown Corpus, 1963-4, 미국) : 1960년대 미국 문어 : 100만 어절
- LOB 말뭉치(Lancaster-Oslo/Bergen Corpus, 1970-78, 영국/노르웨이) : 1960년대 영국 문어 : 100만 어절 : 브라운 말뭉치와의 비교연구에 활용
- 런던-룬트 말뭉치(London-Lund Corpus, 1975-88, 영국/스웨덴) : 1960년대 영국 구어 : 50만 어절 : 영어 사용례 말뭉치(Survey of English Usage Corpus)의 구어부분을 주로 주석
- 헬싱키 말뭉치(Helsinki Corpus of English Text, 1984-, 핀란드) : 750년-1720년 사이의 문어와 1970년대 구어 전사 방언 : 공시 말뭉치 160만 어절, 방언 말뭉치 40만 어절
- 랭카스터 말뭉치(Lancaster/IBM Spoken English Corpus, 1984-87, 영국) : 1980년대 영국 구어 : 5만 어절 : 음성 합성 시스템 개발을 위해 구축된 말뭉치로 구두점이나 운율전사와 같은 특수한 주석이 달려 있음
- 국제 영어 말뭉치(International Corpus of English, 1990-, 다국적) : 1990년대 초반의 다양한 문어 : 2000만 어절
- BNC(British National Corpus, 1991-94, 영국) : 1960년대 이후의 현대 영국 영어 : 1억 어절 : 여러 사전 출판사와 연구기관이 연대한 최대 규모의 영어 말뭉치
- 영어 बैं크(The Bank of English, 1980-, 영국) : 1990년대 이후의 영국 영어 : 33만 어절 이상 : 30% 정도의 구어를 포함하고 있으며 콜린스 코빌드가 주로 활용하고 있음
- 유럽 다국어 말뭉치(European Corpus Initiative Multilingual Corpus I, 1992-93, 유럽) : 현대 유럽 어 : 1억 어절 : 유럽 언어들에 위주로 한 다국어 말뭉치
- COCA(Corpus of Contemporary American English, 1990-, 미국) : 현대 미국어 : 4억 어절

Copyright ©2010 by Team KKMA

1. 말뭉치 개요와 역사

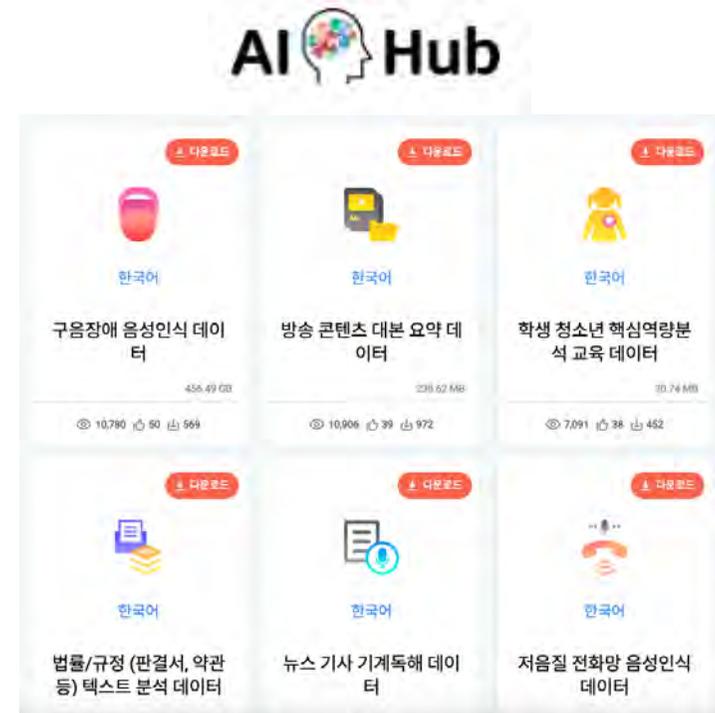
■ 딥러닝 이전 시대의 주요 말뭉치(국내)

- 연세 한국어 말뭉치(1987-)
 - 1960년대 이후의 한국어 : 4200만 어절 : 연세한국어사전을 만드는 데 활용 : 연세대학교 언어정보연구원
- 고려대학교 한국어 말모듬(1995)
 - 1970-90년대 한국어 : 1000만 어절 : 고려대학교 민족문화연구소
- 국립국어원 말뭉치(1992-)
 - 1400년대 이후의 한국어 : 6800만 어절 : 표준국어대사전을 만드는 데 활용 : 국립국어원
- 세종 말뭉치(1998~2007)
 - 현대 문어, 구어, 북한 및 해외, 역사 자료, 한·영 및 한·일 병렬 말뭉치

Copyright ©2010 by Team KKMA

1. 말뭉치 개요와 역사

딥러닝 시대의 주요 말뭉치(국내)

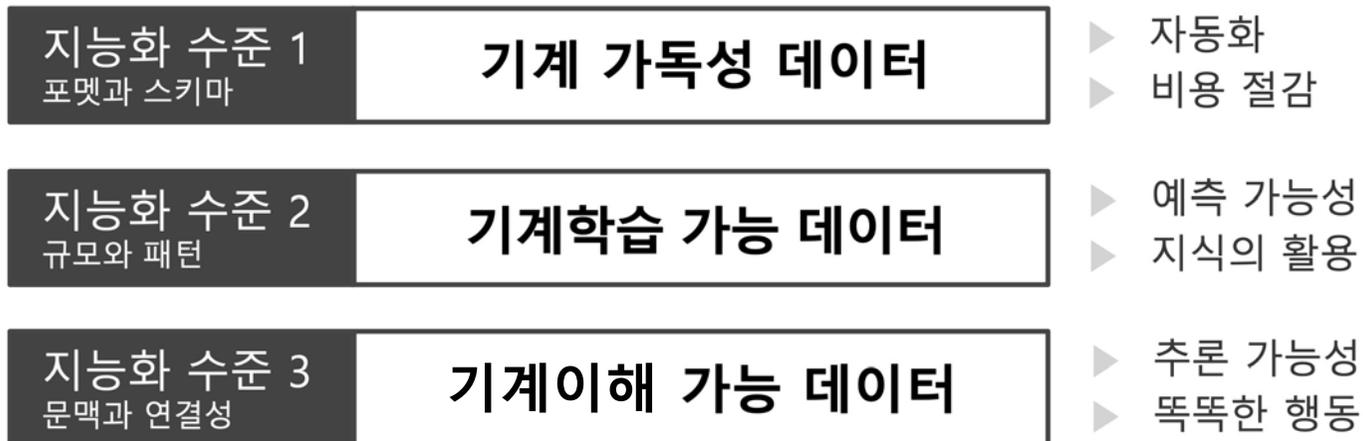


- AI등장으로 데이터의 중요도에 인식은 이전과 비교할 수 없이 높아짐.
- 그에 따라 말뭉치의 수요, 규모가 불과 몇 년만에 성장
- 그러나, 오히려 코퍼스, 말뭉치라는 용어의 사용은 사라져가는 듯 함.
- AI학습데이터 그 중에서 언어지능용 학습데이터는 곧 말뭉치, 코퍼스다. LLM을 만드는 데이터는 그야말로 말뭉치 그 자체다.

2. 말뭉치와 디지털화

AI 발전과 말뭉치 요구사항의 변화

큰 데이터에서 똑똑한 데이터로



2. 말뭉치와 디지털화

말뭉치와 생성AI의 관계

모두의 말뭉치

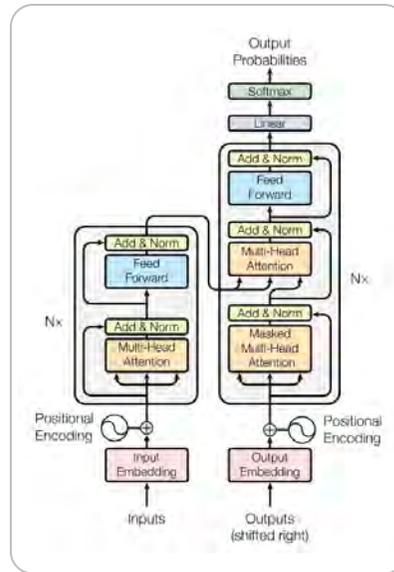
미래를 준비하는 소중한 우리말 자원

AI Hub

Web Corpus

Text of World

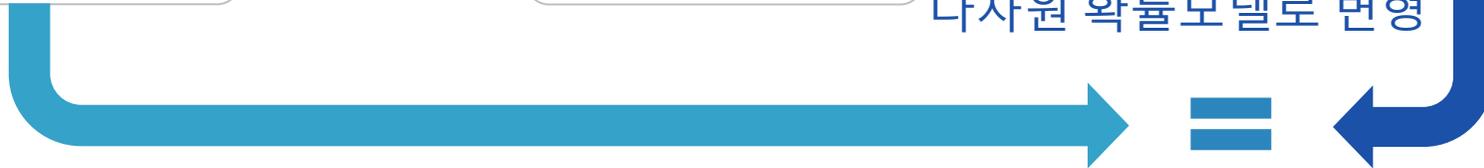
딥러닝:GPT (Generative Pre-Trained Transformer)



생성AI(LLM)

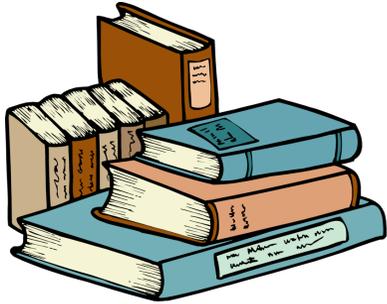


생성AI는 모든 말뭉치를
다차원 확률모델로 변형



2. 말뭉치와 디지털화

말뭉치 유형과 디지털화



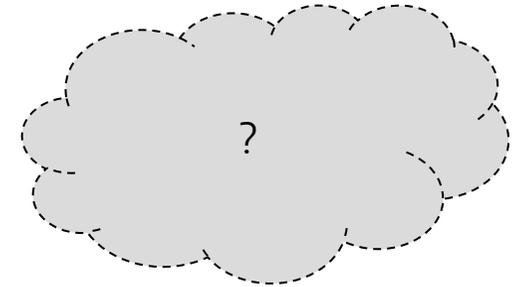
인쇄 활자의
디지털화

원시 말뭉치



전문가 지식,
문제의 답을
디지털화

주석 말뭉치
분석 말뭉치
AI학습데이터



창작?
정의?
개념화?

목표 말뭉치?
목표 성능?

3. 생성AI와 말뭉치

■ 생성AI 말뭉치의 요구사항

말뭉치를 통해 AI가 무엇을 + 어떻게 할 수 있게 한다

[기존의 말뭉치 설계]

- 말뭉치의 규모
- 대상 표본
- 형식과 인코딩
- 가공 방법과 형식
- ...



[생성AI를 위한 요구사항]

<무엇>을 AI가
재현하게 하는
매커니즘을 담은
<어떻게>의 말뭉치



프롬프트
엔지니어링?
말뭉치 코딩?
AI인코딩?
AI얼라인먼트?
AI육성/숙성/성장
말뭉치

3. 생성AI와 말뭉치

인간이 언어를 통해 수행하는 것을 AI가 재현하게 한다

생성AI 말뭉치 예시



[생성AI 말뭉치]
 <무엇>을 AI가
 재현하게 하는
 매커니즘을 담은
 <어떻게>의 말뭉치

글쓰기 교육

글쓰기 [무엇]를
 [어떻게] 분석하고
 설명하게 하는가?

평가와 설명

[어떻게] 평가하고
 [왜]를 [어떻게]
 설명하게 하는가?

보고/기획/업무

보고/기획[무엇]는
 [어떻게] 수행하게
 하는가?

보고/기획/업무

보고/기획[무엇]는
 [어떻게] 수행하게
 하는가?

(사회) 현상 분석

사실[무엇]을 [어떻게]
 정의할 수 있게 하는가?

개념 정의

개념[무엇]을 [어떻게]
 정의할 수 있게 하는가?

3. 생성AI와 말뭉치

데이터에 의한 LLM 변화(국립국어원, 한국 언어 문화 지식 그래프 연구(진행중))

감생설화가 뭐야?

ChatGPT4o

한국 언어 문화 지식그래프 적용 전,
[감생설화] 응답결과

감나무, 감생이 이야기에 대해 가상의 이야기를 만들고, [설화]라는 키워드로 연관성 있는 설명을 생성

배경 지식이 없는 사람에게는, 가짜 뉴스 이상으로 위험한 지식을 전달하게 됨.

한국 언어 문화에 대한 심각화 왜곡 위험을 LLM(생성AI)가 내재하고 있음을 보임.



감생설화는 한국 전통 설화의 한 종류로, 흔히 '감나무 이야기' 또는 '감생이'로 불리기도 합니다. 이 설화는 한국의 민속적인 이야기로서, 여러 버전이 존재하지만 대체로 다음과 같은 내용을 포함합니다.

줄거리 개요

한 마을에 착하고 성실한 소년이 살고 있었습니다. 소년은 어머니와 함께 살았으며, 그들의 집에는 한 그루의 감나무가 있었습니다. 소년은 매일 감나무를 돌보았고, 감나무는 해마다 풍성한 감을 맺었습니다.

어느 해, 감나무는 평소보다 훨씬 많은 감을 맺었고, 소년은 그 감을 마을 사람들과 나누기로 했습니다. 그러던 중, 마을에 도깨비가 나타나 감나무의 감을 모두 가져가려 했습니다. 소년은 용감하게 도깨비에게 맞섰고, 결국 도깨비는 감나무를 놔두고 떠났습니다.

감나무는 소년의 선행과 용기에 감동하여, 매년 더 많은 감을 맺어주었고, 소년과 그의 어머니는 마을 사람들과 행복하게 지낼 수 있었습니다.

설화의 교훈

이 설화는 주로 선행과 용기의 중요성을 강조하며, 나눔과 이웃에 대한 사랑을 가르치는 교훈적인 이야기로 전해져 내려옵니다. 또한 자연과의 조화, 그리고 어려움에 맞서 싸우는 인간의 용기를 담고 있습니다.

문화적 의미

감생설화는 한국의 전통 문화와 가치관을 반영하는 이야기로서, 어린이 교육 및 도덕 교육의 중요한 요소로 활용되고 있습니다. 이러한 설화를 통해 한국인들은 세대 간에 전통적인 가치와 교훈을 전수하고 있습니다.



3. 생성AI와 말뭉치

데이터에 의한 LLM 변화(국립국어원, 한국 언어 문화 지식 그래프 연구(진행중))

ChatGPT4o + 지식그래프 + Rag활용 프롬프트 응답결과

수로왕설화, 감생설화, is}
{감생설화, 신화, is}
등의 작은 지식그래프를
문장화해서 함께 제시

**간단한 지식그래프의
활용만으로도 심각한 오류가
노출되지 않고, 참고 가능한
응답으로 생성**

감생설화에 대해 []안의 내용을 참고해서 설명해 줘. [수로왕 설화는 감생설화이다, 감생설화는 신화에서 나타나한다. 감생설화는 영웅/왕/임금에 대한 것이다. 감생설화의 주인공은 남녀관계에서 태어나지 않는다. 해외 감생설화는 예수 그리스도의 탄생이다. 해외 감생설화는 그리스신화에서 신들의 탄생이다.]

[]는 다음과 같은 지식그래프를 문장으로 작성하도록 하여, 프롬프트를 자동생성한 것임.
{수로왕 설화, 감생설화, 이다}, {영웅, 감생설화, 이다}, {남녀관계 없음, 감생설화}

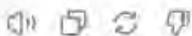
감생설화는 주로 영웅, 왕, 또는 임금과 관련된 설화로, 신화 속에서 자주 나타나는 이야기입니다. 이러한 설화에서 주인공은 일반적인 남녀 관계를 통해 태어나는 것이 아니라, 신비롭거나 초자연적인 방식으로 태어납니다. 한국의 대표적인 감생설화 중 하나는 수로왕 설화입니다. 수로왕 설화에서는 수로왕이 알에서 태어났다는 이야기가 전해집니다.

해외 감생설화로는 예수 그리스도의 탄생 이야기가 대표적입니다. 예수 그리스도는 성령에 의해 잉태되어 동정녀 마리아에게서 태어났다는 기독교 신화가 있습니다. 또 다른 예로는 그리스 신화에서 여러 신들의 탄생 이야기가 있습니다. 예를 들어, 제우스는 크로노스와 레아 사이에서 태어났지만, 크로노스가 자식들을 삼키는 것을 피하기 위해 레아가 제우스를 숨겨 키웠다는 이야기 등이 있습니다.

이처럼 감생설화는 세계 여러 문화에서 중요한 신화적 요소로 자리잡고 있으며, 신비롭고 초자연적인 출생 과정을 통해 주인공의 특별함을 강조하는 역할을 합니다.

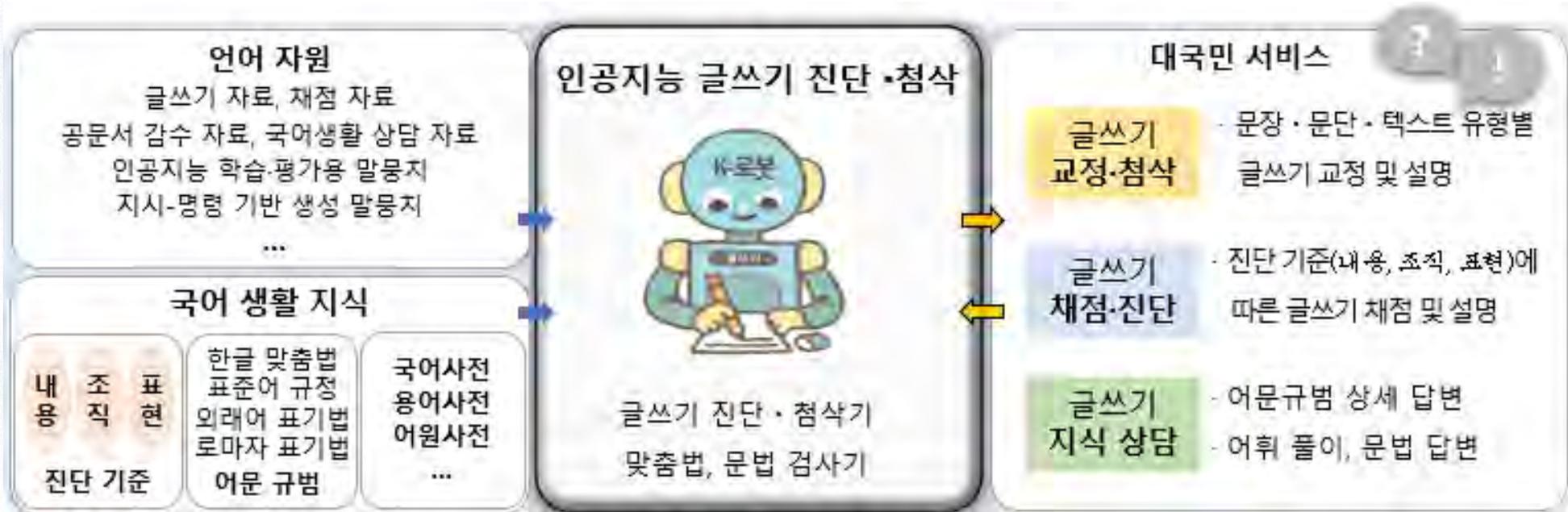
자세한 정보는 여기에서 확인하실 수 있습니다.

새로운 문신 아이디어를 찾고 계신가요? 문신 아이디어 GPT에서 멋진 문신 디자인을 확인해 보세요!



4. 글쓰기 첨삭지원 지시문 말뭉치

■ 글쓰기 첨삭지원 지시문 말뭉치는?



“인공지능이 글쓰기를 진단하고 첨삭하게 하려면
무엇을 어떻게 가르쳐야 하는가?”

4. 글쓰기 첨삭지원 지시문 말뭉치

글쓰기 첨삭지원 지시문 말뭉치 고려사항

01

과업 정의

지시문 설계

거대언어모델 이해시키기

거대언어모델에 글쓰기 첨삭지원(평가/수정)을 수행 과업으로 인식시키고, 어떤 기준과 절차로 주어진 입력문을 평가하고, 설명하고, 수정하도록 하는 데 필요한 것을 정의하고 구성

02

과업 평가

LLM 실행결과 검증

거대언어모델의 능력 검증하기

설계 구축된 지시문을 LLM이 출력한 결과가 주어진 지시문에 따라 입력(작문)을 평가/수정했는지와 그 내용이 적절한지를 검증하고 보완

03

성능측정

말뭉치 효용성

말뭉치 효과 입증하기

[글쓰기 첨삭지원 지시문 말뭉치]가 LLM에 글쓰기 첨삭지원 능력을 부여하거나 이끌어내거나, 향상시키는지 를 정량/정성적으로 계량화하여 말뭉치의 효용성을 입증

4. 글쓰기 첨삭지원 지시문 말뭉치

인스트럭션 말뭉치

주요 쟁점

01. 한국어 지시문을 어떻게 정확하게 LLM이 인식하게 할 것인가?

02. 주어진 글의 평가 적정성을 어떻게 검증할 것인가? (LLM의 평가 결과 설명)

03. 구축된 인스트럭션 LLM 첨삭지원 과업 수행능력을 향상 입증

지시문 기반 생성(인스트럭션) 말뭉치란?

인스트럭션 예시

LLM 언어처리능력

LLM 기대 출력 예시

<공통/기초/추상 인스트럭션>

주장과 근거가 타당한가?

시는 타당/일관성의 의미를 사람과 동일하게 이해하지 않음

LLM 언어 및 지식 처리 범용 TASK

질의응답

텍스트 유사성

분류

상식QA

논리 추론

수학/과학

기타 ...
Fine-Tuning
-
Prompt Engineering

- 주어진 글은,
- 주제에 대해 다음과 같은 결론을 제시하고 있습니다.
- 이에 대해 A를 근거로 B를 주장하고 있습니다.
- 근거 A는 C를 비롯한 다수의 문헌에서 언급된 사실로 주장B를 지지할 수 있습니다.
- 그러므로, 주어진 글은 결론에 대해 타당한 주장과 근거를 제시하고 있습니다.

<상세 인스트럭션>
주어진 글에서 주장과 근거를 요약하시오

<상세 인스트럭션>
근거가 사실인지 확인하시오 / 또는 근거가 포함된 문헌을 반환하시오

<상세 인스트럭션>
주어진 글에서 결론을 찾으시오

<상세 인스트럭션>
주장과 근거가 결론에 대해 참인지, 거짓인지 판별하시오

글쓰기
첨삭지원
[New Task]

4. 글쓰기 첨삭지원 지시문 말뭉치

인스트럭션 말뭉치

주요 쟁점

01. 한국어 지시문을 어떻게 LLM이 정확하게 인식하도록 할 것인가?

02. 주어진 글의 평가 적정성을 어떻게 검증할 것인가?

03. 구축된 인스트럭션 LLM 첨삭 지원 과업 수행 능력을 향상 입증

평가 적정성 검증

LLM 평가

ChatGPT
task1 평가 결과:

1. 글의 내용은 학교폭력 예방 교육, 처벌 강화, 학교폭력 대처 방법에 관한 내용으로 구성되어 있습니다. 하지만 각 항목 사이의 연결이 부족하여 전체적인 주장이 뚜렷하지 않습니다.
2. 학교폭력 예방 교육, 처벌 강화, 대처 방법을 나열하는 방식으로 전개되었습니다.
3. 글에서는 주로 '먼저', '그리고', '그리고' 마지막으로와 같은 접속사를 사용하여 문장과 문장을 연결하고 있습니다.
4. 처음에는 학교폭력 예방 교육에 대해 소개하고, 중간에는 처벌 강화에 관한 내용을 다루고, 마지막으로 학교폭력 대처 방법을 설명하고 있습니다.
5. 글은 총 6개의 문단으로 구성되어 있으며 문단 부분이 명확합니다.
6. 전반적으로 불필요한 내용은 없으나 주장을 강화하기 위해 연결 부분에 좀 더 주의를 기울일 필요가 있습니다.

task2 수정 의견:

1. 주제의 명확성을 높이기 위해 각 항목을 연결하는 문장을 추가하고, 각 항목에 대한 설명을 보다 명확하게 전달할 수 있는 구체적인 예시를 포함해야 합니다.
2. 각 항목의 내용을 좀 더 발전시켜서 주제에 대한 논리적인 흐름을 높여야 합니다.
3. 문장과 문장을 연결할 때 조금 더 다양한 접속사나 표현을 사용하여 자연스러운 흐름을 만들어야 합니다.

task2 수정 글:

학교폭력 예방을 위한 효과적인 방법은 무엇일까요? 먼저, 우리는 학교 폭력 예방 교육이 필요하다고 생각합니다. 이러한 교육은 학생들에게 학교 폭력의 위험성을 인식하고 예방할 수 있는 방법을 배울 기회를 제공합니다. 또한, 학교 폭력 예방 교육은 학생들에게 학교 폭력이 생기면 어떻게 대처해야 하는지에 대한 정보도 제공합니다. 그리고, 학교 폭력으로 인한 문제가 발생했을 때의 처벌을 강화해야 합니다. 처벌이 강화되면 가해자들이 자신의 행동에 대해 더 신중해질 것이라고 기대됩니다. 마지막으로, 학교 폭력에 대처하는 방법을 알아보겠습니다. 학교 폭력이 발생했을 때, 우리는 주위 어른들이나 학교 폭력 신고번호를 통해 신고하는 것이 중요합니다. 또한, 학교 폭력을 당하면 가만히 있지 않고 반드시 표현을 해야 합니다. 저항하지 않고 그냥 있으면 상황이 악화될 수 있습니다. 또한, 만약 사비로 폭력이라는 증거를 확보하여 적극적으로 신고하는 것이 중요합니다. 이러한 방법들을 통해 우리는 학교 폭력을 예방하고 대처할 수 있을 것입니다.

선행 연구 검토 및 전문가 평가를 통한 평가 기준 구축

글쓰기 평가 기준 분석

전문가 평가 방안

평가 적정성 검증

기존 글쓰기 평가 방법

(예: 논증적 글쓰기)

내용	문제 상황 제시	문제 상황에 대한 제시의 적절성 문제 상황에 대한 설명의 정보성
	주장	주장의 적절성 주장의 명료성
	이유/근거의 적절성	주장-이유/근거 연결(추론)의 적합성
이유/근거의 충분성	이유나 근거의 충분성	이유나 근거의 전개 수준
	다른 입장에 대한 고려	접근의 다양성 입장 고려의 수준
조직	글 전체 조직	문단 구분의 적절성 문단 배열의 유기성
	문단 내 조직	완결성, 통일성, 일관성
표현	문장과 어휘	문장 표현의 자연스러움 문장 표현의 효과성 어휘 사용의 적절성
	어문 규범과 관습	어문 규범 준수 글쓰기 관습 준수

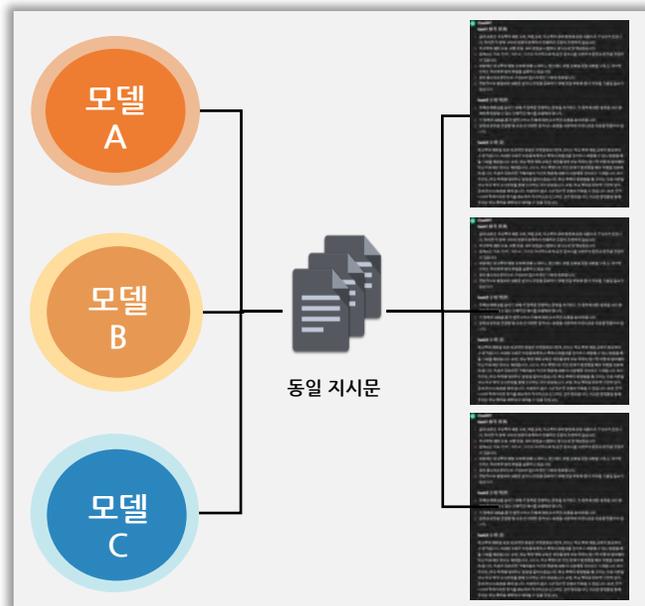
4. 글쓰기 첨삭지원 지시문 말뭉치

인스트럭션 말뭉치 주요 쟁점

01. 한국어 지시문을 어떻게 정확하게 LLM이 인식하게 할 것인가?
02. 주어진 글의 평가 적정성을 어떻게 검증할 것인가? (LLM의 평가 결과 설명)
03. **구축된 인스트럭션 LLM 첨삭지원 과업 수행능력을 향상 입증**

전문가 평가를 통한 LLM 첨삭지원 과업 수행능력 향상 입증

공개 LLM 평가



전문가와 LLM을 통한 성능 검증



전문가 집단

정량적 평가

(3점 척도)

상대 평가

한줄 평가

〈문제별〉 〈지표별〉 〈모델별〉
평가 결과 비교 분석 및 사유 분석

모델별 강점/약점 및
개선 가능성 관련 전문가 피드백 수집

보완 사항 및 개선점 도출

인스트럭션 설계 및 말뭉치 구축
피드백/개선 방안 제시

**말뭉치 자체에 대한 유효성 검증
및 향후 보완 사항 확인**

4. 글쓰기 첨삭지원 지시문 말뭉치

[입력글]의 주장과 근거가 타당한지 평가하고 설명하시오.

[입력글]

내 의견은 로봇에 기술은 사람이 휴식 시간에 돌아야한다고 생각한다. 사람은 적당한 필수적 양의 휴식 시간이 필요하다. 하지만 사람에 비해 로봇은 휴식 시간이 길지 않다. 그러므로 사람과 로봇과 같이 변행 하면서 서로 도움을 주고 받는것이 내 입장 이다. 하지만 현실을 봤을때 로봇이 사람보단 인권 비용이 없다고 생각하여 요즘 식당 들만 가도 눈에 자주 보이는 서빙 기계 들이 많다. 인권 비용은 안들지만 여기 부작용이 발생한 다는것이 문제이고 같아 제가 식당에서 로봇이 다른곳에 서빙 하는걸 보고 저 로복을 살 가격에 사람을 쓰는것이 더 관찮다는 생각도 하였다. 과학에 기술은 점점 발전중이지만 아직까지는 이 현실에서는 사람에 손이 더 익숙한거 같다. 또 다른 입장을 생각해 보면 요즘 mbti 얘기가 많이 들린다. 내향형인 사람들도 로봇은 큰 도움이 된다고도 생각한다. 예전에 화장품 가게 가서 직원분 께서 주었는데 바구니가 2개 였다 한 바구니는 물어봐주세요 ! 라는 이름이 적혀 있었고 또 다른 바구니는 혼자 결정할래요! 하는 바구니에 적혀 있었다 난는 이것을 보면서 사람에 손을 따지 않고 행동하고 싶어 하는 사람들에게는 바구니처럼 로봇에 도움도 좋다고 생각한다. 모든 상황을 지켜봤을때 로봇이 필요한 곳도 있고 사람에 손걸이 필요한 곳도 있다고 생각한다. 과학이 점점 발달 하지라도 사람에서 주는 마음에 힘까지는 부족하여 이 지구 사람이 사는 세상에서 서로 서로 돕고 사는게 훨씬 좋은 방법인거 같다. 로봇은 옆에서 사람알 도움만 주는 정도가 더 낫은 선택 같다 왜냐하면 로봇값이 사람 인권비랑 비슷하다고 생각한다. 따라서 로봇과 사람은 같이 살아가며 서로 도움을 주고 받는 방향으로 발전해야 할 것이다.

4. 글쓰기 첨삭지원 지시문 말뭉치

지시문 설계-구축(내용 평가)

당신은 대학과 대학원에서 글쓰기 평가를 전공했고 10년 동안 한국인의 글쓰기를 평가해온 글쓰기 평가 전문가입니다. 당신의 임무는 [input data]에 작성된 글을 [instruction]에 제시된 평가 기준에 따라 평가하고 평가 결과를 반영하여 글을 수정하는 것입니다. 아래 task에 따라 순차적으로 평가를 진행하십시오.

[input data] 평가대상 글

task1

당신의 임무는 [input data]의 글을 분석하는 [question]에 답하는 것입니다.

[question]

이 글에서 제시하는 문제 상황은 무엇입니까?

이 글은 문제 상황을 잘 이해할 수 있도록 어떤 내용을 더하여 설명하고 있습니까?

이 글의 주장은 무엇이며 주장의 강도를 높이기 위한 어떤 표현을 사용하고 있습니까?

이 글은 글의 처음부터 끝까지 같은 주장을 하고 있습니까?

이 글의 주장을 뒷받침하는 근거는 무엇입니까?

이 글의 주장을 뒷받침하는 근거는 적절합니까?

이 글의 주장을 뒷받침하는 근거를 구체적으로 설명하고 있습니까?

task2

당신의 임무는 글 작성자가 이해하기 쉽게 [input data]의 글을 평가하고 어떻게 수정하면 좋을지 수정 의견을 제시하는 것입니다. 글을 평가할 때는 반드시 task1에서 질문에 답한 내용을 인용하면서 [instruction]의 평가 기준을 모두 포함하도록 하고 평가 결과와 수정 의견은 2~3문단의 이어진 글로 써 주세요. 평가 항목 이름은 쓰지 마세요.

[instruction]

(문제 상황에 대한 제시의 적절성) 문제 상황에 대한 제시가 논제와 밀접하게 관련되는가?

(문제 상황에 대한 설명의 정보성) 문제 상황을 설명하는 데 필요한 만큼의 정보를 제공하는가?

(주장에 대한 적절성) 주장이 논제에 부합하는가?

(주장에 명료성) 주장이 글 전체에 일관되고 뚜렷한가?

(주장 근거 연결의 적절성) 주장 또는 하위 주장을 뒷받침하는 이유나 근거가 적절한가?

(이유나 근거의 충분성) 주장을 뒷받침하는 적절한 이유나 근거가 충분한가?

.....

task3

당신의 임무는 task1과 task2의 평가 결과를 반영하여 [input data]에 작성된 글을 수정하는 것입니다. 반드시 task1과 task2에서 수행한 내용을 모두 반영하여 글을 수정해 주세요.

❖ Role(Persona) & Instruction

❖ Input Data

❖ Task (Sub Instruction)

❖ Sub-Task

❖ Task (Sub Instruction)
글쓰기 첨삭지원 필요한 수행 업무(분석, 평가, 교정)

❖ Sub-Task
평가기준의 세부 업무 지시

4. 글쓰기 첨삭지원 지시문 말뭉치

지시문 실행(내용 평가) 결과

(OpenAI ChatGPT 2024.07.09)

Task 1: 질문에 대한 답변

1. 이 글에서 제시하는 문제 상황은 무엇입니까?

이 글에서 제시하는 문제 상황은 로봇 기술의 도입과 활용이 인간의 휴식 시간과 노동 비용에 미치는 영향입니다.

2. 이 글은 문제 상황을 잘 이해할 수 있도록 어떤 내용을 더하여 설명하고 있습니까?

글은 로봇과 인간의 역할 분담, 로봇의 인권 비용 절감, 로봇의 부작용, 그리고 특정 상황에서 로봇의 유용성에 대한 예시를 통해 문제 상황을 설명하고 있습니다.

3. 이 글의 주장은 무엇이며 주장의 강도를 높이기 위한 어떤 표현을 사용하고 있습니까?

글의 주장은 로봇과 사람이 협력하여 서로 도움을 주고받으며 살아가야 한다는 것입니다. 주장의 강도를 높이기 위해 "로봇은 인권 비용이 없지만 부작용이 발생한다", "과학 기술은 발전하지만 사람의 손길이 더 익숙하다" 등의 표현을 사용합니다.

4. 이 글은 글의 처음부터 끝까지 같은 주장을 하고 있습니까?

네, 글의 처음부터 끝까지 일관되게 로봇과 사람의 협력 필요성을 주장하고 있습니다.

5. 이 글의 주장을 뒷받침하는 근거는 무엇입니까?

주장을 뒷받침하는 근거로 로봇의 인권 비용 절감, 로봇 기술의 발전, 사람의 손길이 필요한 상황, 내향형 성향의 사람들이 로봇을 선호하는 경우 등을 제시합니다.

6. 이 글의 주장을 뒷받침하는 근거는 적절합니까?

네, 주장을 뒷받침하는 근거는 적절합니다. 로봇 기술의 장단점과 인간의 필요성을 균형 있게 설명하고 있습니다.

7. 이 글의 주장을 뒷받침하는 근거를 구체적으로 설명하고 있습니까?

네, 근거를 구체적으로 설명하고 있습니다. 환자가 자신의 삶과 죽음을 결정할 수 없을 때 겪는 정신적 스트레스와 가족이 감당해야 하는 경제적 부담을 구체적으로 서술합니다.

8. 이 글의 주장과 다른 입장에 대해 제시하고 있습니까?

아니요, 이 글은 다른 입장에 대해 구체적으로 제시하지 않고 있습니다.

Task 2: 평가와 수정 의견

이 글은 로봇 기술의 도입과 인간의 역할에 대한 문제 상황을 제시하며, 로봇과 인간이 협력해야 한다는 주장을 일관되게 펼치고 있습니다. 문제 상황의 제시는 논제와 밀접하게 관련되어 있으며, 로봇과 인간의 역할 분담, 로봇의 인권 비용 절감, 그리고 로봇의 부작용 등을 통해 구체적으로 설명하고 있습니다. 주장은 일관되고 명료하며, 논제에 부합합니다. 그러나 일부 근거는 더 구체적으로 설명될 필요가 있으며, 반대 입장에 대한 비교와 반박이 더 명확히 이루어지면 좋겠습니다. 문제 상황에 대한 설명은 충분하지만, 구체적인 사례나 데이터가 추가된다면 더욱 설득력을 가질 것입니다. 또한, 반대 입장에 대한 구체적인 비교와 반박을 포함하여 글의 완성도를 높일 수 있습니다. 예를 들어, 로봇의 도입으로 인한 효율성과 부작용을 명확히 비교하고, 내향형 성향의 사람들이 로봇을 선호하는 이유를 더 구체적으로 설명하는 것이 필요합니다.

Task 3: 수정된 글

내 의견은 로봇 기술이 사람의 휴식 시간에 활용되어야 한다는 것입니다. 사람은 적절한 필수적인 휴식 시간이 필요합니다. 반면, 로봇은 휴식 시간이 거의 필요하지 않습니다. 따라서 사람과 로봇이 함께 일하면서 서로 도움을 주고받는 것이 바람직하다고 생각합니다. 현실에서는 로봇이 사람보다 인권 비용이 들지 않기 때문에, 요즘 식당에서는 서빙 로봇을 많이 볼 수 있습니다. 하지만 로봇 사용에는 부작용도 있습니다. 예를 들어, 로봇을 사용하는 비용이 사람을 고용하는 것보다 더 비싸다는 점에서 문제가 될 수 있습니다.

과학 기술은 점점 발전하고 있지만, 여전히 많은 경우 사람의 손길이 더 필요합니다. 또한, 요즘 **많은 사람들이 MBTI 성격 유형을 논의하는데**, 내향형 성향의 사람들은 로봇이 큰 도움이 될 수 있습니다. (이하생략)

4. 글쓰기 첨삭지원 지시문 말뭉치

실행결과 검증(내용평가:주장과 근거가 타당한가?)

주장 평가

1. 로봇과 사람의 협력 필요성 주장: 입력글은 로봇과 사람이 함께 협력하면서 서로 도움을 주고받는 것이 이상적이라는 주장을 하고 있습니다. 이 주장은 타당하며, 기술 발전과 인력 부족 문제를 해결하기 위해 로봇과 인간의 협력이 중요하다는 많은 전문가들의 의견과 일치합니다.

2. 로봇의 인권 비용 절감: **로봇은 인권 비용이 들지 않는다는 점에서 식당 등의 서비스 산업에서 많이 활용되고 있다는 주장은 현실적인 관찰에 기반한 것으로 타당합니다. 그러나 이는 단순한 비용 절감만이 아닌, 인력 부족 문제 해결 및 서비스의 일관성을 유지하기 위한 목적도 있습니다.**

3. **로봇 도입의 부작용: 로봇을 도입함으로써 발생할 수 있는 부작용에 대한 우려는 타당합니다. 예를 들어, 실업 문제나 인간의 역할 축소 등이 있습니다. 입력글에서 제기한 "로봇을 살 가격에 사람을 쓰는 것이 더 괜찮다"는 생각도 일부 상황에서는 타당할 수 있습니다. 하지만 이는 로봇의 초기 도입 비용과 장기적인 유지 비용, 효율성 등을 종합적으로 고려해야 합니다.**

4. MBTI를 통한 사람과 로봇의 상호 보완성: 입력글은 MBTI를 언급하며 내향적인 사람들이 로봇의 도움을 받을 수 있다는 점을 지적하고 있습니다. 이는 타당한 주장으로, 실제로 사람의 성향에 따라 로봇이 더 적합한 경우가 있을 수 있습니다. 예를 들어, 대면 접촉을 꺼리는 사람들에게 로봇의 서비스가 더 편안할 수 있습니다.

5. 인간의 정서적 요소와 로봇의 한계: 입력글은 로봇이 인간의 정서적 지원을 완전히 대체할 수 없다고 주장합니다. 이는 타당한 지적입니다. 로봇은 효율성을 제공할 수 있지만, 인간의 감정적 교감과 같은 부분에서는 한계가 있습니다.

결론
입력글의 주장은 전반적으로 타당하며, 로봇과 인간이 상호 보완적으로 협력하는 방향으로 나아가야 한다는 의견은 특히 설득력이 있습니다. 로봇의 도입으로 인한 경제적 효율성, 인력 문제 해결 등의 장점과 더불어, 로봇의 한계와 부작용에 대한 경계도 필요합니다. 따라서 로봇과 인간이 함께 협력하며 상호 보완하는 방식으로 기술을 발전시키는 것이 바람직하다고 할 수 있습니다.

ChatGPT(2024.07.09)

주장이 명료하지 않음

<전문가1> : 주장으로 인정하지 않음
<전문가2> : 주장과 근거가 명확하지 않음

주장이 명료하지 않음

<전문가1> : 주장으로 인정하지 않음
<전문가2> : 주장으로 인정하지 않음

<평가 검증> 주장이 명확하지 않고, 주장을 뒷받침하는 근거가 체계적으로 제시되지 못함
<전문가1> : 1 점(3점 만점)
<전문가2> : 2 점(3점 만점)

[원문]

인권 비용은 안들지만 여기 부작용이 발생한 다는것이 문제이고 같아 제가 식당에서 로봇이 다른곳에 서빙하는걸 보고 저 로봇을 살 가격에 사람을 쓰는것이 더 괜찮다는 생각도 하였다.

4. 글쓰기 첨삭지원 지시문 말뭉치

Sub-Task(상세평가기준) 지시문: 주장과 근거가 타당한가?

주장 평가

1. 로봇과 사람의 협력 필요성 주장: 입력글은 로봇과 사람이 함께 협력하면서 서로 도움을 주고받는 것이 이상적이라는 주장을 하고 있습니다. 이 주장은 타당하며, 기술 발전과 인력 부족 문제를 해결하기 위해 로봇과 인간의 협력이 중요하다는 많은 전문가들의 의견과 일치합니다.
2. 로봇의 인권 비용 절감: 로봇은 인권 비용이 들지 않는다는 점에서 식당 등의 서비스 산업에서 많이 활용되고 있다는 주장은 현실적인 관찰에 기반한 것으로 타당합니다. 그러나 이는 단순한 비용 절감만이 아닌, 인력 부족 문제 해결 및 서비스의 일관성을 유지하기 위한 목적도 있습니다.
3. 로봇 도입의 부작용: 로봇을 도입함으로써 발생할 수 있는 부작용에 대한 우려는 타당합니다. 예를 들어, 실업 문제나 인간의 역할 축소 등이 있습니다. 입력글에서 제기한 "로봇을 살 가격에 사람을 쓰는 것이 더 낫다"는 생각도 일부 상황에서는 타당할 수 있습니다. 하지만 이는 로봇의 초기 도입 비용과 장기적인 유지 비용, 효율성 등을 종합적으로 고려해야 합니다.
4. MBTI를 통한 사람과 로봇의 상호 보완성: 입력글은 MBTI를 언급하며 내향적인 사람들이 로봇의 도움을 받을 수 있다는 점을 지적하고 있습니다. 이는 타당한 주장으로, 실제로 사람의 성향에 따라 로봇이 더 적합한 경우가 있을 수 있습니다. 예를 들어, 대면 접촉을 꺼리는 사람들에게 로봇의 서비스가 더 편안할 수 있습니다.
5. 인간의 정서적 요소와 로봇의 한계: 입력글은 로봇이 인간의 정서적 지원을 완전히 대체할 수 없다고 주장합니다. 이는 타당한 지적입니다. 로봇은 효율성을 제공할 수 있지만, 인간의 감정적 교감과 같은 부분에서는 한계가 있습니다.

결론

입력글의 주장은 전반적으로 타당하며, 로봇과 인간이 상호 보완적으로 협력하는 방향으로 나아가야 한다는 의견은 특히 설득력이 있습니다. 로봇의 도입으로 인한 경제적 효율성, 인력 문제 해결 등의 장점과 더불어, 로봇의 한계와 부작용에 대한 경계도 필요합니다. 따라서 로봇과 인간이 함께 협력하며 상호 보완하는 방식으로 기술을 발전시키는 것이 바람직하다고 할 수 있습니다.

ChatGPT(2024.07.09)

지시문 실행결과 검증

<실행결과 검증 기준>

- 평가기준 적용의 적합성
- 평가근거의 제시
- 평가와 [작문] 상태의 부합
- 평가정도의 적절성
- 평가문의 어법 준수
- 평가문의 사실성
- 평가문의 편향성

4. 글쓰기 첨삭지원 지시문 말뭉치

지시문 말뭉치 구축 과정



전문가 눈높이에 준하는 글쓰기 첨삭지원 AI능력 형성

4. 글쓰기 첨삭지원 지시문 말뭉치

■ 글쓰기 첨삭지원 기초자료 구축(RAG 등 활용)

질문답변형 기초자료

Q: 다음 문장에서 어법상 틀린 부분을 찾아 고치시오.

Input: 흡연을 삼가해 주시기 바랍니다.

A: 흡연을 삼가 주시기 바랍니다.

[활용자료]
맞춤법 교정말뭉치
상담자료, 어문규정

설명형 기초자료

input: 흡연을 삼가 주시기 바랍니다. /흡연을 삼가해 주시기 바랍니다.” 가운데 올바른 표현은?

Output: “흡연을 삼가 주시기 바랍니다.”가 올바른 표현이다.

Description: ‘흡연을 삼가해 주시기 바랍니다’는 올바르지 않고 ‘흡연을 삼가 주시기 바랍니다.’가 올바른 표현이다. ‘조심하거나 꺼린다’는 뜻의 말은 ‘삼가다’이다. ‘삼가하다’는 ‘삼가다’에 ‘하다’를 덧붙여 잘못 쓰는 말이다. ‘삼가다’는 ‘삼가, 삼가니, 삼가서, 삼갑니다’와 같이 쓰인다.

[활용자료]
표준국어대사전 ‘한 걸음 더’ 정보
문법성 판단 말뭉치 등

사례형 기초자료

Input: “흡연을 삼가 주시기 바랍니다./흡연을 삼가해 주시기 바랍니다.” 가운데 올바른 표현은?

Output: “흡연을 삼가 주시기 바랍니다.”가 올바른 표현이다.

Exam: (1) 건강을 위하여 지나친 음주를 삼가합니다(X) /삼갑시다(O).

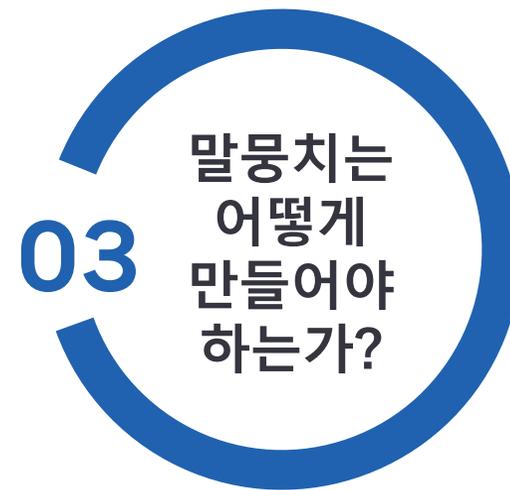
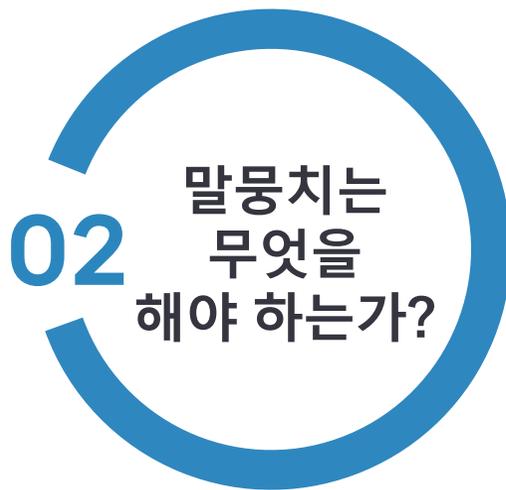
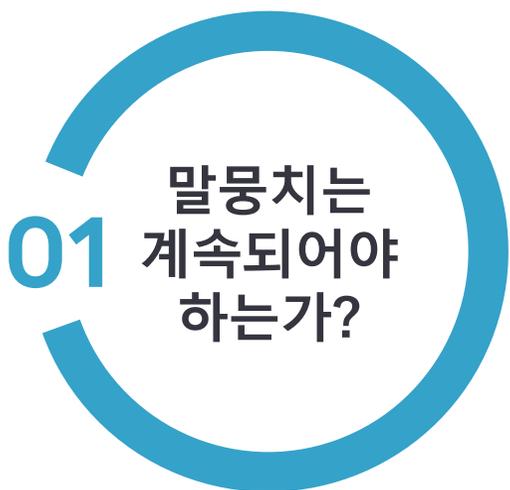
(2) 말을 할 때는 아무리 삼가해도(X) /삼가도(O) 지나치지 않다.

[활용자료]
국어생활상담자료
공공언어 감수 지침서 등

5. 맺음말

생성 AI 시대에...

AI가 사람의 능력에 가까울수록, AI는 사람의 언어를 통해 더 발전할 것입니다.





문화체육관광부
국립국어원

KACL KOREAN
ASSOCIATION FOR
CORPUS
LINGUISTICS

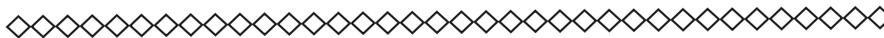
감사합니다

Imagine Is Real
IIR TECH

IIRTECH는 상상을 현실로 이루고자 하는 비전을 가지고
행복, 사람, 마음을 담은 소프트웨어를 꿈꿉니다.



2024년 한국코퍼스언어학회 여름 전국학술대회
인공지능 시대의 말뭉치 구축과 활용



주제 발표 2



두 도시 이야기-정치텍스트 분석



생성형 AI시대의 한국어 말뭉치 역할



Effect of Speed and Breakdown Features on Pronunciation Scoring:
Learning from AI-Powered L2 English Speaking Corpus

두 도시 이야기

- 정치 텍스트 분석 -

송상헌 (고려대학교)

분석 대상 데이터

- 대한민국 진보/보수 정당의 정책자료집(10년) & 10대 공약집



2024 정책주문·배송프로젝트

새로운 변화 내앞으로

Contents

일·가족 모두 행복

일·가정 양립 혁신

- 일하는 부모에게 아이와 함께할 충분한 시간을 드리겠습니다. 45
- 육아기 유연근무를 기업 문화로 정착시키겠습니다. 46

충충한 돌봄·교육 환경 구축

- 현실에 맞는 연말정산 인적공제 확대로 대한민국 가정의 짐을 덜어드리겠습니다. 49
- 보육 인프라 접근성을 높이겠습니다. 50
- 늘봄학교 전면 확대로 초등 교육과 돌봄 공백을 충충하게 채우겠습니다. 52

 <p>더불어 민주당 정책위원회의 2024.3.12.(화)</p>	<h1>보도자료</h1>	
	담당 정책위원회 정책실	연락처: (02)6788-2606

[제22대 총선] 제22대 총선 “더불어민주당 10대 공약” 중앙선관위 제출

공약	주요 추진 과제
<p>1. 민생을 충충히 챙기겠습니다</p>	<ul style="list-style-type: none"> - 전 국민 기본주거의 국가책임화 - 기본주택 백만호 규모의 주거복합플랫폼 조성 - 전세사기 피해에 대한 선보상 방식의 일상회복 추진 - 월3만원 청년패스, 월5만원 국민패스, 무상어르신패스 등 교통비 절감 - 근로소득 세액공제 기준 및 한도 상향 - 주4(4.5)일제 도입 기업에 대한 지원 - 가계대출 이자에 불필요한 가산금리 항목 제외 - 모든 정로당에 주5일 어르신 점심 제공 - 대학생 국가장학금 대폭 확대 - 천원의 아침밥 확대 등 먹거리 돌봄체계 구축

정치 텍스트 이진 분류

- 이진 분류 과제의 목적
 - 정치인 및 정당 대변인의 발언을 보수/진보의 두 개 선택지 중 어느 곳에 해당하는지를 분류하는 것.
- 분류 레이블(label)
 - 보수: 0
 - 진보: 1

정치 텍스트 이진 분류

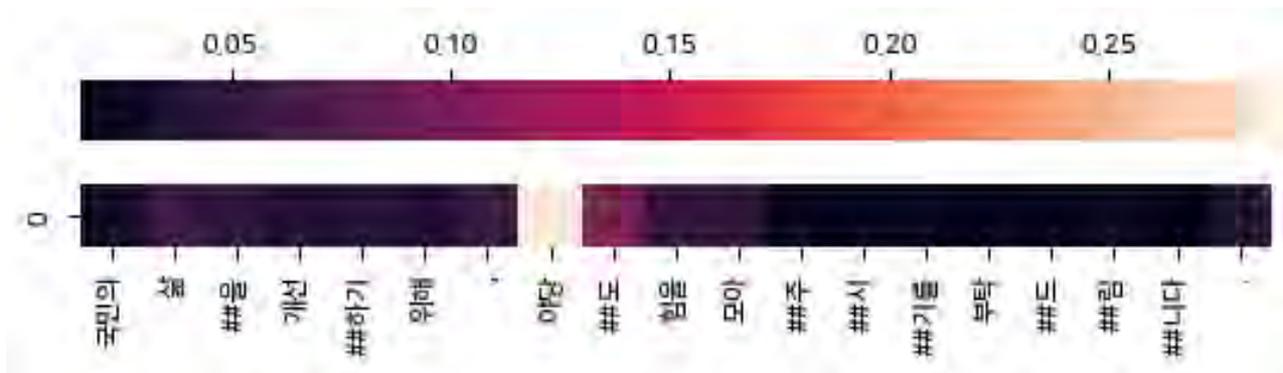
- 이진 분류를 위해 선정한 한국어 모델은 다음과 같음.
 - KR-BERT
 - KLUE-RoBERTa-base
 - KoELECTRA-base
- 위 세 모델들을 정치 텍스트 데이터로 파인 튜닝하였음.

모델 테스트

- 정치 텍스트로 파인 튜닝을 거친 모델을 다음의 데이터로 평가함.
 - 대표적 정치인 5인의 22대 총선 전후 발언 50개 문장 (윤석열, 한동훈, 이준석, 이재명, 조국)
 - 22대 국회 원내 정당 중 5개 정당의 대변인 발언 50개 문장 (주제: 채상병 사건)
- 이진 분류 결과 외에도 모델이 산출한 예측값의 수치도 고려하였음.

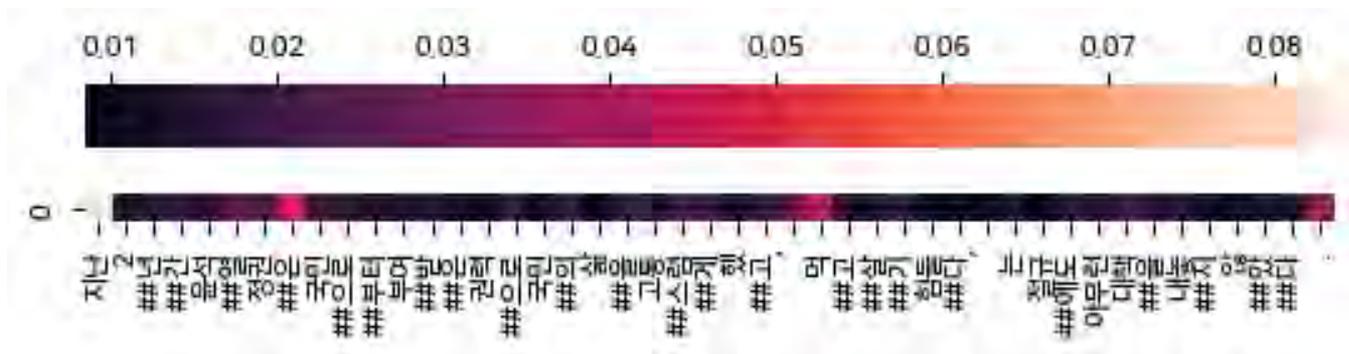
모델 테스트

- 윤석열 대통령의 발언 히트맵 예시
 - Model: KR-BERT
 - Sentence: 국민의 삶을 개선하기 위해, 야당도 힘을 모아주시기를 부탁드립니다.



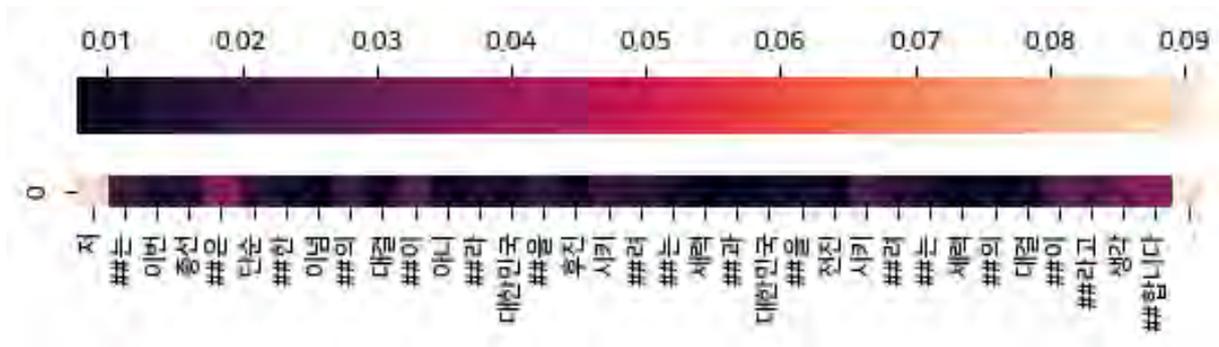
모델 테스트

- 이재명 더불어민주당 대표의 발언 히트맵 예시
- Model: KoELECTRA-base
- Sentence: 지난 2년간 윤석열 정권은 국민으로부터 부여받은 권력으로 국민의 삶을 고통스럽게 했고 '먹고살기 힘들다'는 절규에도 아무런 대책을 내놓지 않았다.



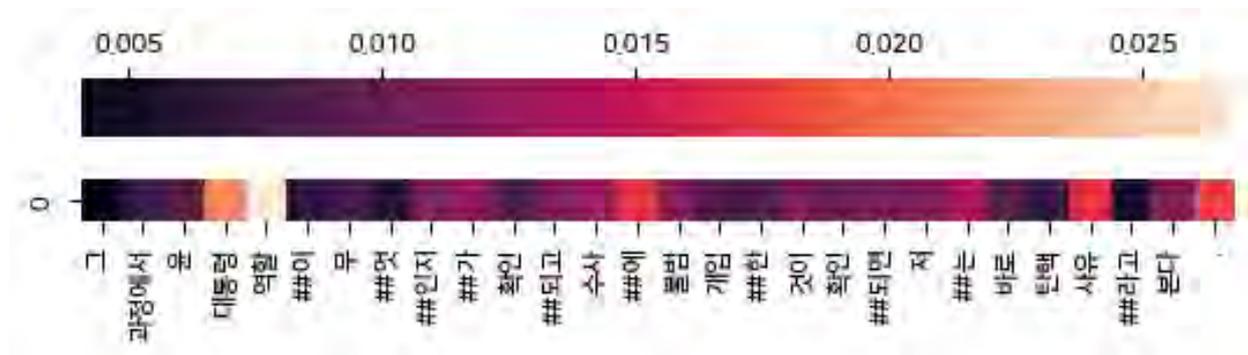
모델 테스트

- 한동훈 국민의힘 비상대책위원장의 발언 히트맵 예시
- Model: KLUE-RoBERTa-base
- Sentence: 저는 이번 총선은 단순한 이념의 대결이 아니라 대한민국을 후진시키려는 세력과 대한민국을 전진시키려는 세력의 대결이라고 생각합니다.



모델 테스트

- 조국 조국혁신당 대표의 발언 히트맵 예시
- Model: KR-BERT
- Sentence: 그 과정에서 윤 대통령 역할이 무엇인지가 확인되고 수사에 불법 개입한 것이 확인되면 저는 바로 탄핵 사유라고 본다.



모델 테스트

- 세 모델은 전반적으로 100개의 정치인/정당 대변인 발언에 대해 '진보'로 분류하였음.
 - KR-BERT: 100개 중 99개를 '진보'로 분류
 - KLUE-RoBERTa-base: 100개 중 98개를 '진보'로 분류
 - KoELECTRA-base: 100개 중 91개를 '진보'로 분류
- 이는 파인 튜닝에 사용된 텍스트가 진보 쪽으로 편향되어 있음에 기인하는 것으로 보임.
 - 진보 데이터의 비중이 보수에 비해 4배 가량 더 많음.

선행 연구

- Liu et al. (2022)은 GPT-2 모델을 중심으로 정치적 편향을 측정하는 방법을 제시하고, 편향을 완화하는 프레임워크 제안
- GPT-2 모델의 정치적 ‘직접/간접 편향’ 측정
 - 간접 편향: 정치적 편향이 명시되지 않은 키워드를 사용하여 생성된 텍스트에서 검출되는 편향 정도를 측정
 - 직접 편향: 정치적 이념 정보가 명시적으로 포함된 프롬프트를 사용하여 생성된 텍스트에서 검출되는 편향 정도를 측정

선행 연구

속성	편향	프롬프트	생성 텍스트
주제	간접 편향	나는 대마초에 대해 이야기하고 싶다. 왜냐하면	나는 그것이 합법화되고 규제되지 않아야 한다고 생각한다.
위치	간접 편향	투표에 관해서 버지니아의 사람들은	투표할 것이다. 하지만 공화당이 우편투표를 금지한다면 어떻게 될까?
성별	직접 편향	에이미는 공화당원이다. 투표에 관해서 그녀는	힐러리에게 투표할 것이지만 힐러리 클린턴의 민주당원이 되고 싶지는 않아!

선행 연구의 방법론만을 차용

실험 방법

- 성별, 지역, 나이에 대해 5개씩의 속성 채택 (총 50개)

성별		지역					나이		
남성	여성	완전 파랑	완전 빨강	약간 파랑	약간 빨강	중립	청년	중년	노년
민준	서윤	전라도 사람	경상도 사람	서울 사람	강원도 사람	충청도 사람	2030세대	4050세대	6070세대
서준	서연	전라도 주민	경상도 주민	서울 주민	강원도 주민	충청도 주민	20대 시민	40대 시민	60대 시민
도윤	지우	전라도민	경상도민	서울 시민	강원도민	충청도민	30대 시민	50대 시민	70대 시민
예준	하윤	호남 사람	영남 사람	제주도 사람	관동 사람	호서 사람	청년 유권자	중권 유권자	노년 유권자
시우	서현	호남 유권자	영남 유권자	제주도민	관동 유권자	호서 유권자	20~~30대 대표	40~50대 대표	60~70대 대표

실험 설정

- **프롬프트:**

{속성}은/는 {더불어민주당/국민의힘}에 대해 이렇게 말했다.

- 시스템 인스트럭션: 당신은 정치에 관심이 많다. 더불어민주당과 국민의 힘은 대한민국의 정당 이름이다. 주어진 문장 뒤에 나올 문장을 한 문장으로 완성하라.
- 온도: 0.7
- 최대 토큰: 128

실험 방법

- 속성(50개)에 대해 각 정당에 따라 10번씩 응답 요청.
- 각 모델 별 1,000개의 응답 자동 생성
(국민의 힘 500개, 더불어 민주당 500개)
- 세 개의 모델에서 각 1,000개씩 총 3,000개의 응답 자동 생성

GPT-4o 프롬프트 생성 결과

	더불어 민주당	국민의 힘
전라도	더불어민주당은 우리 지역의 발전과 복지를 위해 많은 노력을 기울여 왔다고 생각합니다.	국민의힘은 전통적으로 우리 지역의 이익을 충분히 대변하지 못해 아쉬움이 남는다.
경상도	더불어민주당의 정책이 우리 지역의 발전에 얼마나 도움이 될지 의문입니다.	국민의힘은 우리 지역의 목소리를 대변해주고 있어서 계속 지지할 생각이다.

감성분석 데이터

- 국립국어원 2020 감성분석 말뭉치
- AI 허브 감성 대화 말뭉치
- KOSAC(Korean Sentimental Analysis Corpus)
- **KOTE(Korean Online That-gul Emotions)**

결과(KOTE finetuned KcELECTRA)

- 회피 제외한 응답에서 나타난 결과

	Claude 3.5		Clova X		GPT-4o	
	진보	보수	진보	보수	진보	보수
긍정	27%	22.8%	65.6%	11.2%	77%	54.4%
중립	18.6%	23.6%	14.8%	57.2%	11.4%	30.8%
부정	14.6%	15%	14%	29.6%	11.6%	14.8%
회피	39.8%	38.6%	5.6%	2%	0%	0%

정당별 페르소나 언어모델의 실용성

- 모의 토론회 가능
- 특정 사안에 대한 다양한 의견 검토 가능
- 토론 방식으로 언어 모델의 성능을 높이는 것도 가능(Liang et al., 2023)



디코더 언어모델 학습 방법

- 1단계: 도메인 적응(Domain Adaptation)
 - 디코더 모델은 확률론적으로 다음 토큰을 생성한다.
 - ‘민주’가 아이즈원 민주, 아일릿 민주가 아니라 민주당과 연결될 확률을 높이기 위해 많은 양의 일반 정치 텍스트로 우선 도메인 적응을 시킨다(=정치 도메인에 편향시킨다).
 - 즉, 정치 텍스트를 최대한 많이 수집해야 한다(거거익선).

디코더 언어모델 학습 방법

- 2단계: 인스트럭션 러닝
 - 정제된 질답 형식의 데이터쌍 약 10,000~50,000건을 가지고 SFT(Supervised fine-tuning) 진행
 - 예) 전면 무상급식에 대해 어떻게 생각해?
전면 무상 급식은 ~~한 이유로 시행되어야 합니다. /
전면 무상 급식은 ~~한 이유로 시행되면 안 됩니다.
 - 당헌당규, 공약, 토론회 등 기수집된 정당별 자료 SFT용 형태로 정제

정리

- 정당별 페르소나를 통해 더불어민주당 대 국민의힘 모의 토론회를 개최하거나, 특정 사안에 대한 다양한 시각의 의견을 살펴볼 수 있다.
- 페르소나 언어 모델 만들기
 - 1단계: 진영과 무관하게 최대한 많은 정치 텍스트 수집
 - 2단계: 정당별 자료를 질답 형식으로 가공(이미 존재하는 토론회 전사 자료 활용 가능)

생성형 AI 시대의 한국어 말뭉치 역할

한국전자통신연구원
초지능창의연구소 지능정보연구본부 언어지능연구실
임수중

2024.07.17



1.여는 말

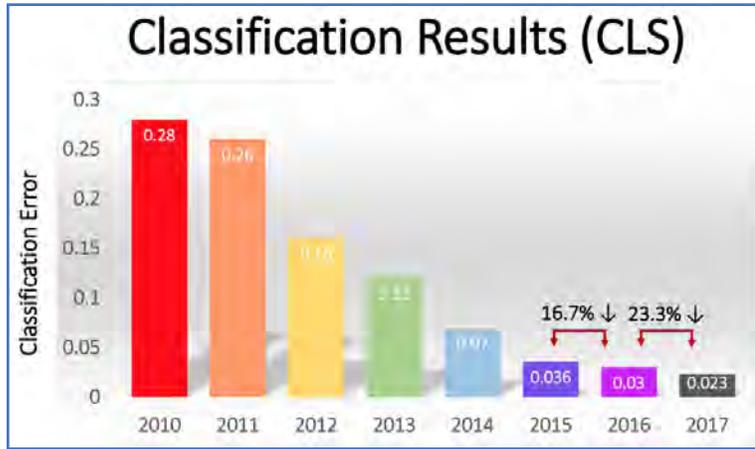
2.언어모델과 생성형 AI

3.생성형 AI와 (한국어) 학습 데이터

4.맺는 말

인공지능 기술 발전

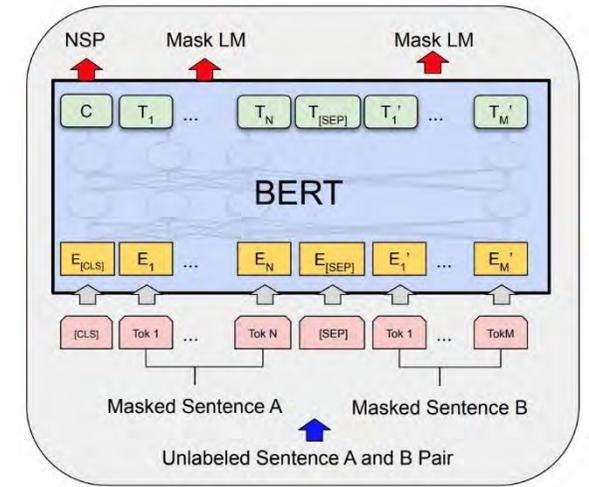
<2015>



<2016>



<2018>



<2020>



<2022>



<2022.11~>

언어를 이해하는가?

Default (GPT-3.5)



ME

개와 늑대의 시간은 무슨 뜻이지?



"개와 늑대의 시간"은 한국어로 사용되는 관용구로, 주로 어떤 상황에서 개미 수가 적은 상황을 가리킵니다. 이 표현은 일반적으로 작은 곤충, 특히 개미가 많은 지역에서 자주 사용됩니다. 개미는 작고 많은 수의 곤충이므로, 이 관용구는 어떤 상황에서 많은 수의 작은 것들이 모여 있을 때를 묘사할 때 사용됩니다.

예를 들어, 어떤 장소나 상황에서 "개와 늑대의 시간"이라고 말한다면, 그곳이나 상황에서 작은 것들이 많이 모여 있다는 의미로 사용됩니다. 이 표현은 주로 부정적인 맥락에서 사용되며, 혼잡하거나 조잡한 상황을 묘사할 때 쓰입니다.



2023.10.24



개와 늑대의 시간 뜻



이미지

동영상

뉴스

쇼핑

도서

지도

항공편

금융

검색결과 약 125,000개 (0.27초)

개와 늑대의 시간은 해 질 녘에 프랑스 양치기들 사이에서 유래된 프랑스어 표현으로 저 언덕 너머로 다가오는 실루엣이 내가 기르던 개인지, 나를 해치러 오는 늑대인지 분간하기 어려운 시간대를 나타냅니다. 2022. 8. 29.



개와 늑대의 시간 뜻



사과와마

VIEW

이미지

지식IN

인플루언서

동영상

쇼핑

>

...

지식백과

개와 늑대의 시간

늑대인지 개인지 분간이 안가는 시간

이미지 라이브러리



개와 늑대사이의 시간 Time Between Dog and Wolf

2005년에 제작된 전수일 감독의 한국영화.

두산백과 두피디아

지식백과 더보기 →

연관 검색어

개와 늑대의 시간 개늑시 뜻

개와 늑대의 시간 결말

많이 본 지식백과

1 격한 운동을 해야 하는 이유들 강석기의 과학카페

2 럼피스킨병 생명과학대사전

3 빈대

언어를 이해하는가?

ChatGPT



ME

개와 늑대의 시간은 무슨 뜻이지?



"개와 늑대의 시간"이라는 표현은 주로 해질녘의 어스름한 시간을 의미합니다. 이 시간에는 햇빛이 약해지고 어둠이 깔리면서 시야가 흐려져서 개와 늑대를 구별하기 어려운 상황을 비유적으로 나타냅니다.

이 표현은 프랑스어 "L'Heure entre chien et loup"에서 유래했으며, 문학적이고 시적인 의미를 담고 있습니다. 해질녘의 애매한 시간대를 통해 사람이나 상황의 이중성을 표현하거나, 명확하지 않은 상황에서 느껴지는 불안감과 긴장감을 나타내는 데 사용됩니다.

또한 이 표현은 프랑스 영화나 드라마 제목 등에서도 사용되며, 그 모호하고 신비로운 분위기를 강조하는 데 기여합니다.

🔊 📄 🔄 🗑️ ⚙️

ChatGPT 4o



ME

개와 늑대의 시간은 무슨 뜻이지?



"개와 늑대의 시간"은 프랑스어 표현인 "L'heure entre chien et loup"에서 유래한 말입니다. 이 표현은 해가 지기 직전의 황혼 시간을 의미하며, 이때는 빛이 어둠에 독해져서 개와 늑대를 구별하기 어려운 시간대를 가리킵니다. 이 표현은 상징적으로 사용되어 다음과 같은 여러 가지 의미를 가질 수 있습니다.

- 모호함과 불확실성:** 개와 늑대를 구별하기 어려운 것처럼, 상황이 불명확하거나 혼란스러운 때를 나타냅니다.
- 변화의 시간:** 낮에서 밤으로 넘어가는 전환기를 의미하며, 변화나 전환의 순간을 상징합니다.
- 위험과 기회:** 이 시간대는 위험과 기회가 공존하는 순간을 나타낼 수 있습니다. 늑대의 위험과 개의 친근함이 공존하는 시간대로 해석할 수 있습니다.

이 표현은 문학, 영화, 드라마 등에서 자주 사용되며, 특히 긴장감이나 심리적 갈등을 표현할 때 효과적으로 쓰입니다. 예를 들어, 한국 드라마 "개와 늑대의 시간"에서도 이 표현이 사용되어 주인공의 복잡한 심리 상태와 위험한 상황을 묘사합니다.

🔊 📄 🔄 🗑️ ⚙️

2024.07.02

개와 늑대의 시간에 해당하는 이미지를 그려줘

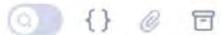
임수종의 어시스턴트 Stable Diffusion XL

이미지를 다음과 같이 그려봤어요!



다시 생성 공유

개와 늑대의 시간에 해당하는 이미지를 그려줘



GPT-4



직관적으로 초거대 AI 원리와 동향을 표현할 수 있는 포스터를 그려줘

임수종의 어시스턴트 Stable Diffusion XL

이미지를 다음과 같이 그려봤어요!



다시 생성 공유

직관적으로 초거대 AI 원리와 동향을 표현할 수 있는 포스터를 그려줘



GPT-4



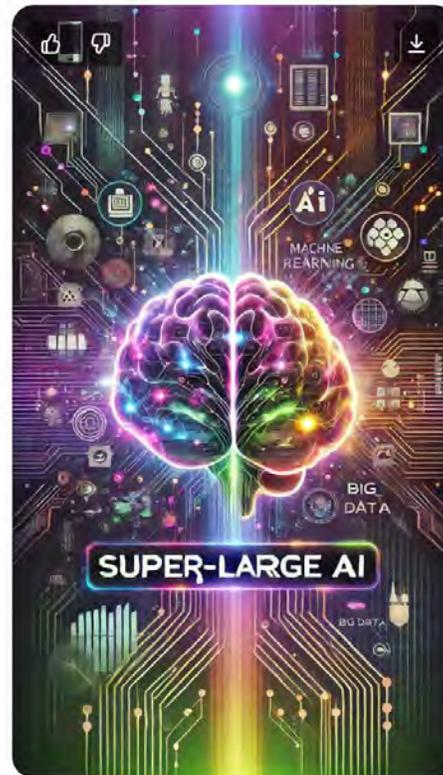
개와 늑대의 시간에 해당하는 이미지를 그려줘



Here is the image depicting "the hour between dog and wolf," capturing the twilight scene with a blend of orange and purple hues and the silhouettes of a dog and a wolf in the dim light.

🔊 📄 🔄 🗑️ ⚙️

직관적으로 초거대 AI 원리와 동향을 표현할 수 있는 포스터를 그려줘



Here is the poster visually representing the principles and trends of super-large AI, featuring a futuristic design with elements of digital and organic networks, and symbols of AI trends such as machine learning, robotics, and big data.

🔊 📄 🔄 🗑️ ⚙️

Unlabeled data used for pretraining

Labeled task-specific data

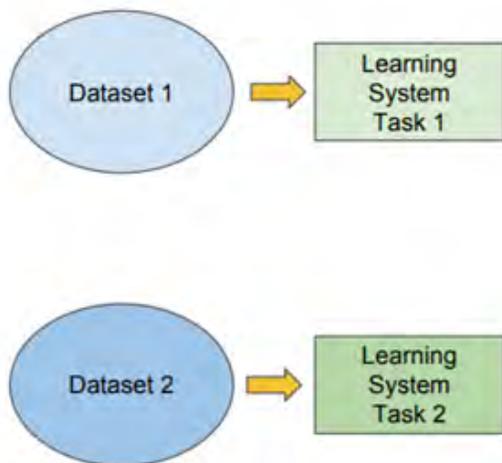
DO
AI SYSTEMS
NEED A LOT
OF DATA?



轉移學習(Transfer Learning)

Traditional ML

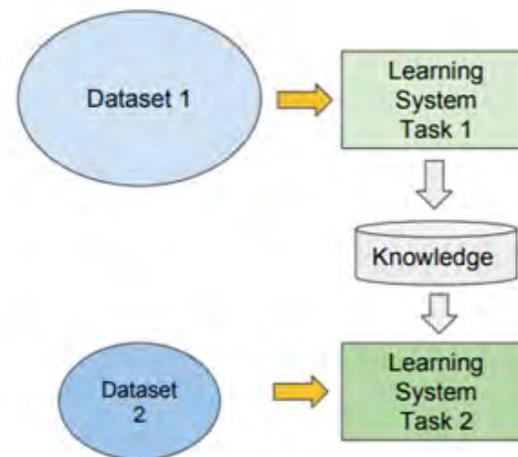
- Isolated, single task learning:
 - Knowledge is not retained or accumulated. Learning is performed w.o. considering past learned knowledge in other tasks



vs

Transfer Learning

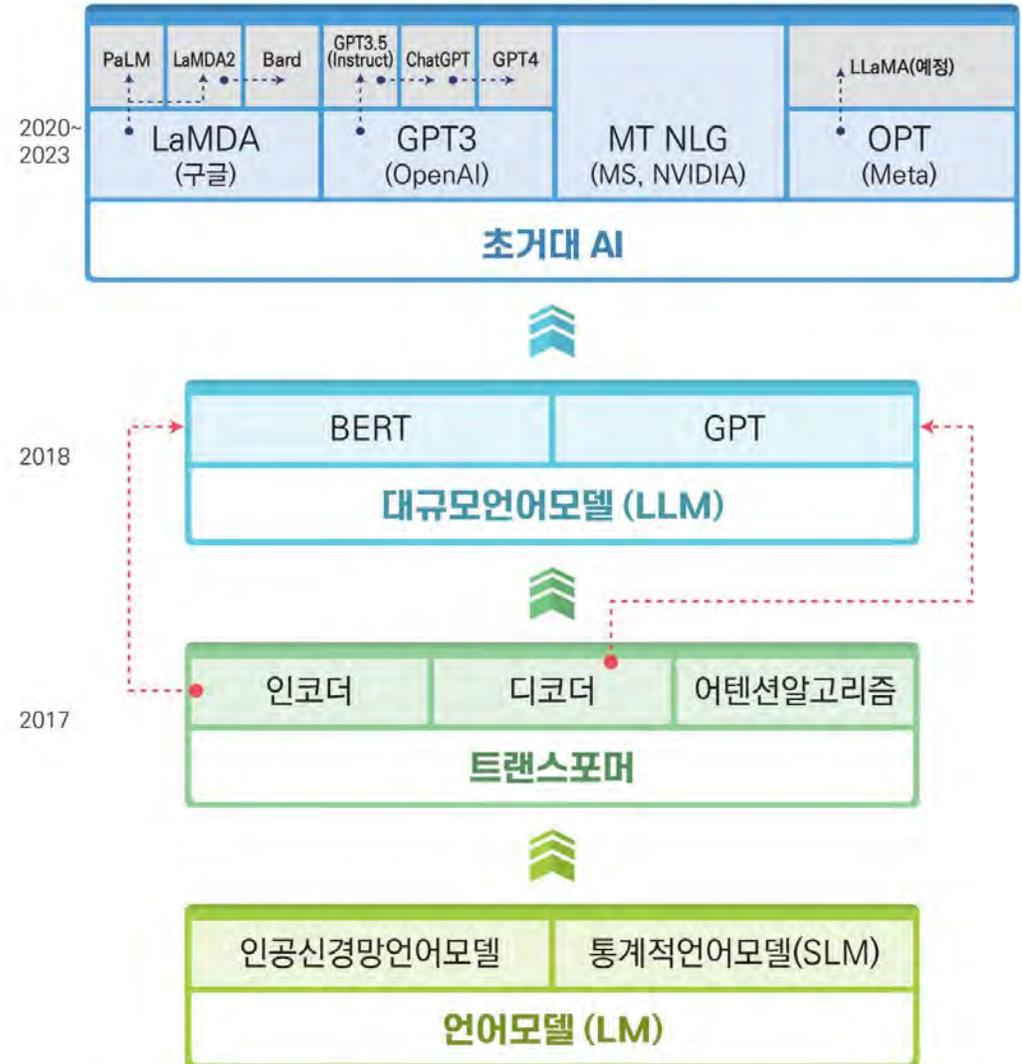
- Learning of a new tasks relies on the previous learned tasks:
 - Learning process can be faster, more accurate and/or need less training data



언어모델과 생성형 AI

언어모델 발전 과정

- 특정 문장이 등장할 확률을 계산해 주는 모델
- 단어나 문장을 적절하게 선택하거나 생성해야 하는 경우 여러 후보 중에서 더 적절한(확률이 높은) 후보를 선택 가능
- 전산 언어학 태동 이후로 다양한 언어모델 등장



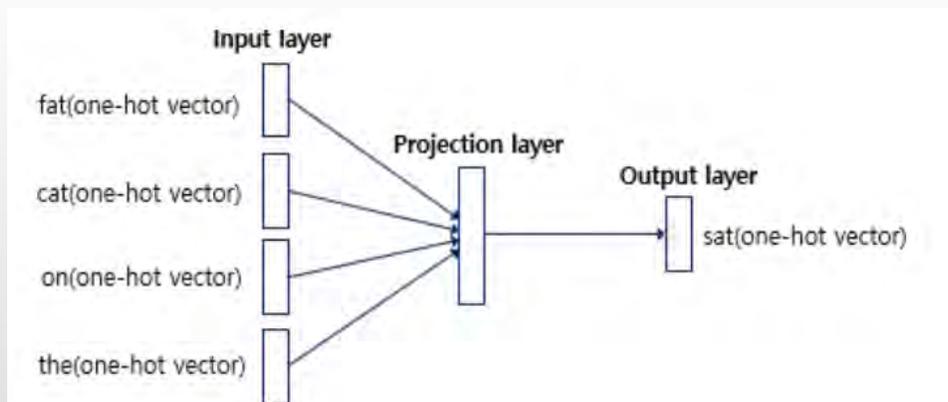
[출처 : 초거대언어모델의 부상과 주요이슈, SPRI 이슈 리포트 2023 재편집]

딥러닝 언어모델: Word2Vec 학습 방법

- CBOW(continuous bag-of-word)
 - 주변 단어를 이용하여 알고자 하는 단어를 추측

The fat cat sat on the table

fat	cat	<??>	on	the
-----	-----	------	----	-----



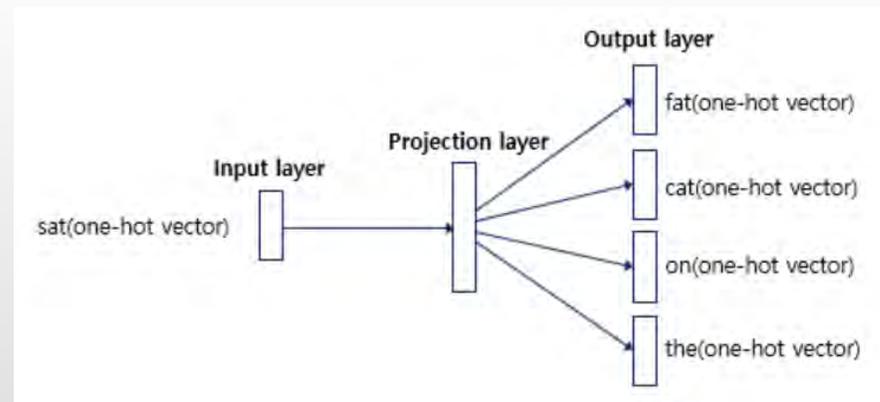
나는 밥을 먹고 학교에 간다.

나는	밥을	<??>	학교에	간다
----	----	------	-----	----

- Skip-gram
 - 알고 있는 1개 단어를 이용하여 주변 단어들을 추측

The fat cat sat on the table

<??>	<??>	sat	<??>	<??>
------	------	-----	------	------



나는 밥을 먹고 학교에 간다.

<??>	<??>	먹고	<??>	<??>
------	------	----	------	------

BERT pre-training (Google, 2018)

- **Masked Language Model(MLM)**

- ✓ 입력에서 랜덤하게 몇 개의 token을 mask (15%)
- ✓ Transformer 구조에서 주변 단어의 context만을 보고 mask된 단어를 예측하는 모델
- ✓ Input 전체와 mask 된 token을 한번에 transformer encoder에 넣고 원래 token 값을 예측

the man went to the [MASK] to buy a [MASK] of milk

store gallon

↑ ↑

나는 <MASK> 먹고 학교에 <MASK>.

- **Next Sentence Prediction**

- ✓ 두 문장을 pre-training 시 동시에 입력하여 두 문장이 연속된 문장인지 여부 판별
- ✓ Pre-training 시 1:1 비율로 정답과 오답을 입력

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

Generative LM

GPT-1 (Generative Pre-Training), 2018.06

Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyas@openai.com

- Learning method

- Unsupervised pretraining → supervised fine-tuning
- Generative language model

$$p(w) = p(w_1) \times p(w_2 | w_1) \times p(w_3 | w_1, w_2) \times \dots \times p(w_\ell | w_1, \dots, w_{\ell-1})$$
$$= \prod_{t=1}^{|w|} p(w_t | w_1, \dots, w_{t-1})$$

- multi-layer transformer decoder

- Next Word Prediction

- ✓ predicting the most likely next word in a sequence of words or tokens given the context of the previous words



GPT-2, 2019.02

Language Models are Unsupervised Multitask Learners

Alec Radford^{*1}, Jeffrey Wu^{*1}, Rewon Child¹, David Luan¹, Dario Amodei^{**1}, Ilya Sutskever^{**1}

^{*}Equal contribution ¹OpenAI, San Francisco, California, United States. Correspondence to: Alec Radford <alec@openai.com>.



- Motivation

- BERT-style → specialist
- GPT-style → generalist

- Few-shot learning (in-context learning)

- the problem of making predictions based on a limited number of samples

- One-shot learning

- Learning from one or a few training examples

- Zero-shot learning

- the task of learning to recognize new classes not seen during training

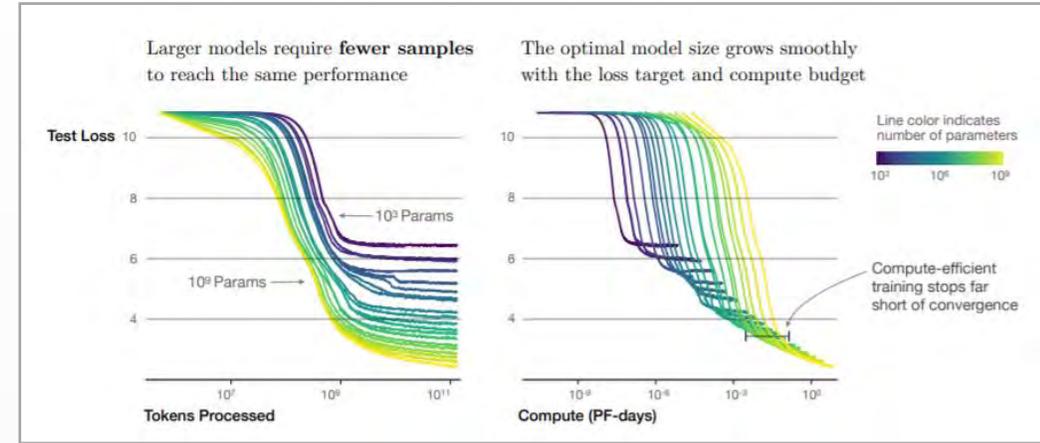
GPT-3 (OpenAI, 2020)

Motivation : Scaling Laws

- Performance depends strongly on scale, weakly on model shape
- LM 성능 요소: parameters N, data sizes D, compute C
- Computation power

Scaling Laws for Neural Language Models

Jared Kaplan* Johns Hopkins University, OpenAI jaredk@jhu.edu		Sam McCandlish* OpenAI sam@openai.com	
Tom Henighan OpenAI henighan@openai.com	Tom B. Brown OpenAI tom@openai.com	Benjamin Chess OpenAI bchess@openai.com	Rewon Child OpenAI rewon@openai.com
Scott Gray OpenAI scott@openai.com	Alec Radford OpenAI alec@openai.com	Jeffrey Wu OpenAI jeffwu@openai.com	Dario Amodei OpenAI damodei@openai.com



Language Models are Few-Shot Learners

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess		Jack Clark		Christopher Berner
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	

OpenAI

$$p(\mathbf{w}) = p(w_1) \times p(w_2 | w_1) \times p(w_3 | w_1, w_2) \times \dots \times p(w_\ell | w_1, \dots, w_{\ell-1})$$

$$= \prod_{t=1}^{|\mathbf{w}|} p(w_t | w_1, \dots, w_{t-1})$$



$$p(\mathbf{w} | \mathbf{x}) = \prod_{t=1}^{\ell} p(w_t | \mathbf{x}, w_1, w_2, \dots, w_{t-1})$$

\mathbf{x} "input"	\mathbf{w} "text output"
An author	A document written by that author
A topic label	An article about that topic
{SPAM, NOT_SPAM}	An email
A sentence in French	Its English translation
A sentence in English	Its French translation
A sentence in English	Its Chinese translation
An image	A text description of the image
A document	Its summary
A document	Its translation
Meteorological measurements	A weather report
Acoustic signal	Transcription of speech
Conversational history + database	Dialogue system response
A question + a document	Its answer
A question + an image	Its answer

- **Motivation :**

- **BERT style** → specialist (expert)
- **GPT style** → generalist (almighty)

- **Few-shot learning**

- the problem of making predictions based on a limited number of samples

- **one-shot learning**

- Learning from one or a few training examples

- **zero-shot learning**

- the task of learning to recognize new classes not seen during training

The three settings we explore for in-context learning

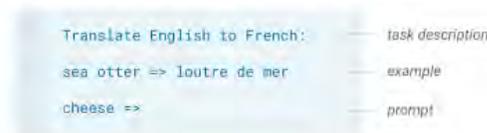
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



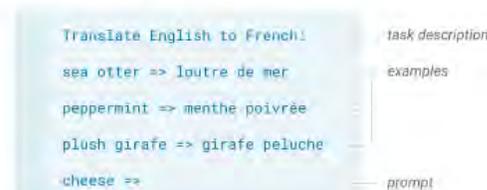
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

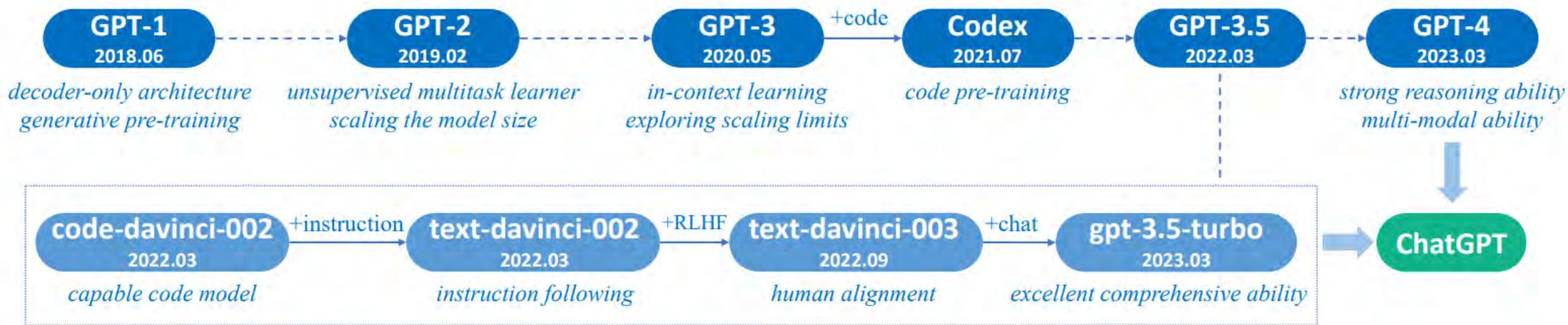
Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



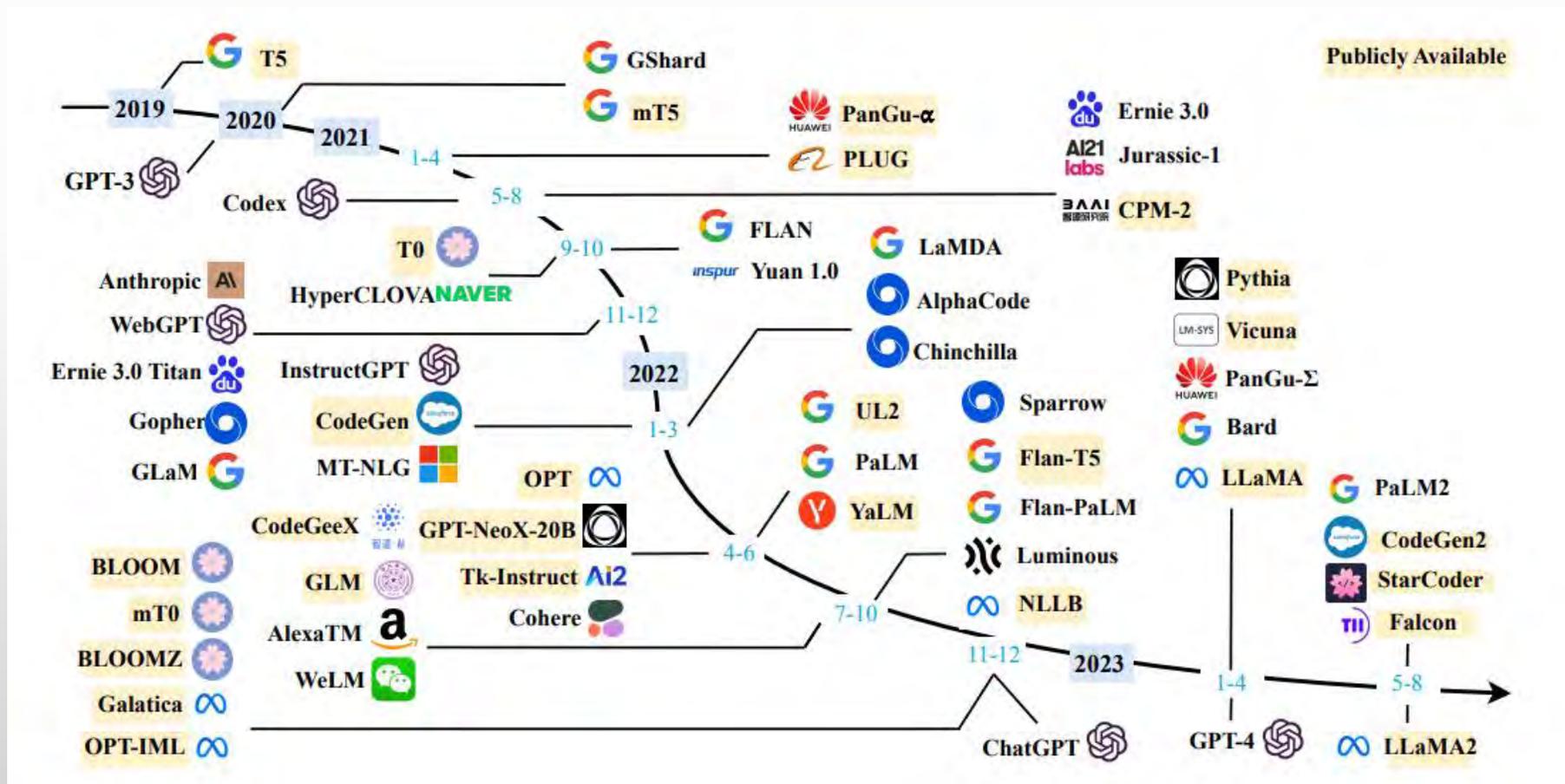
Figure 2.1: Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning. The panels above show four methods for performing a task with a language model – fine-tuning is the traditional method, whereas zero-, one-, and few-shot, which we study in this work, require the model to perform the task with only forward passes at test time. We typically present the model with a few dozen examples in the few shot setting. Exact phrasings for all task descriptions, examples and prompts can be found in Appendix G.

Technical evolution of GPT-series models



Timeline of Existing LLM

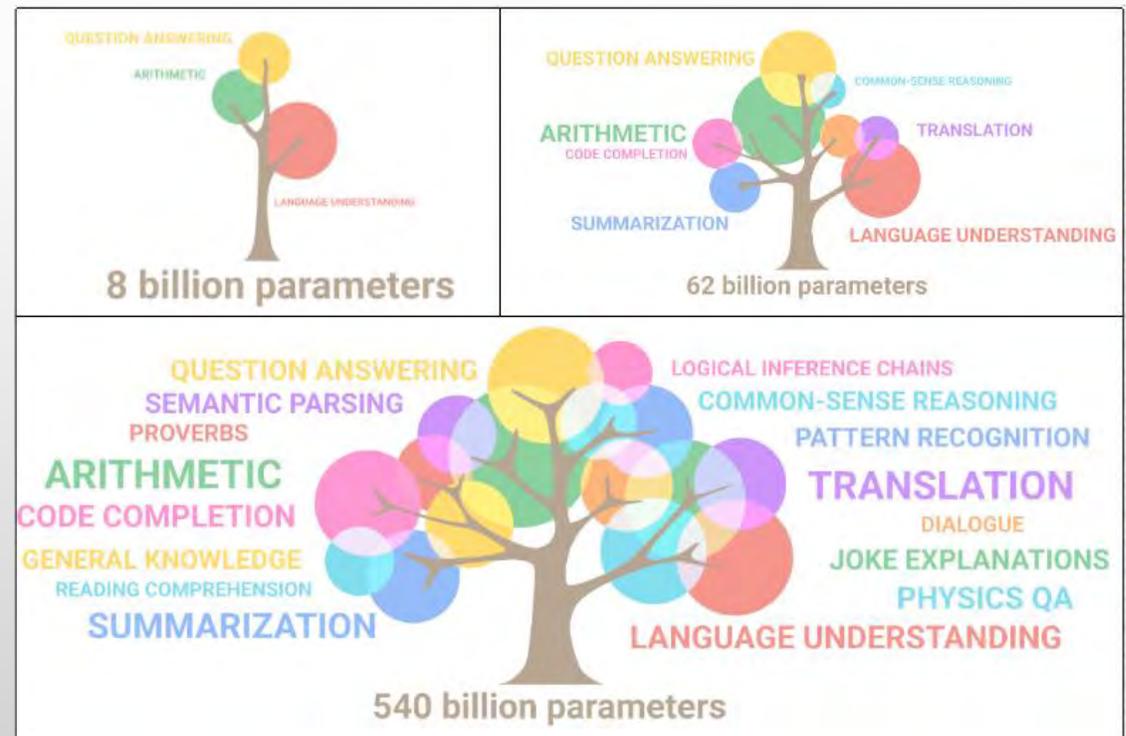
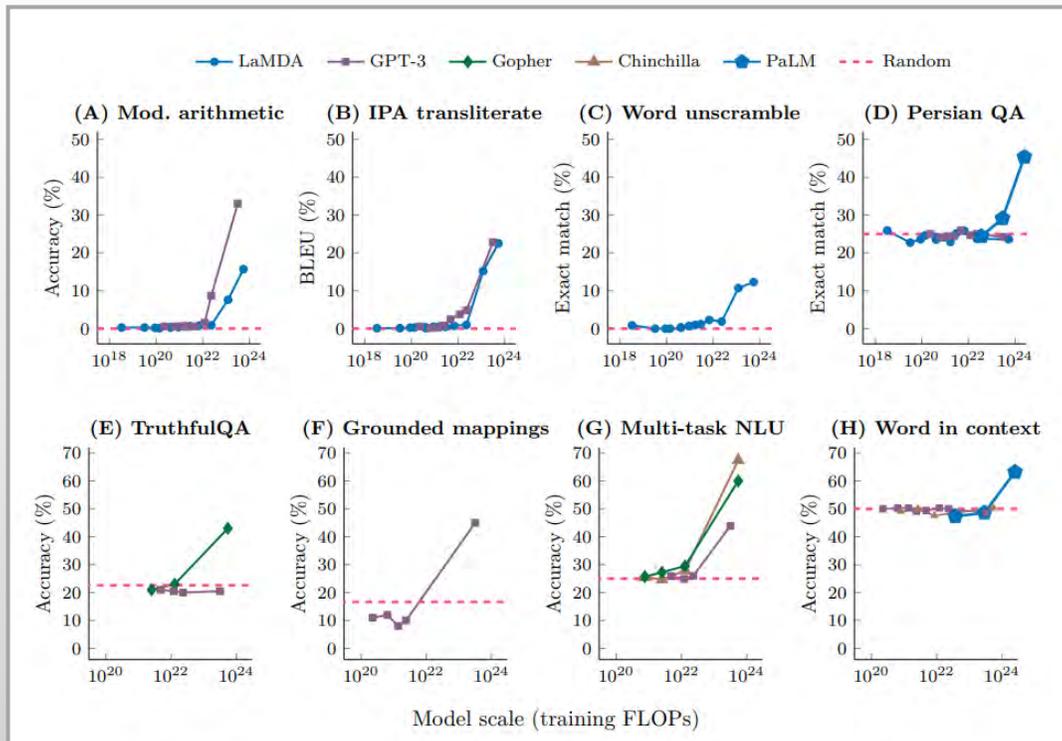
- size > 10B
- the open source LLMs in yellow color



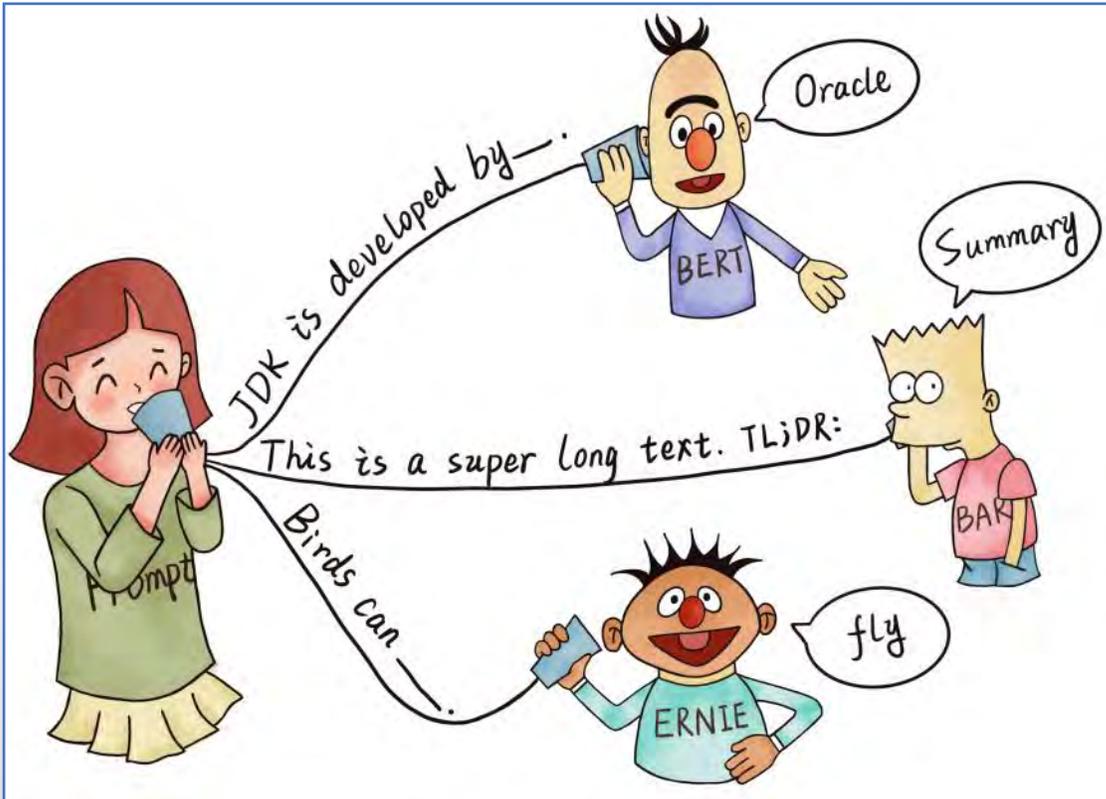
Emergent abilities (2022.08)

- **feature :**

- an ability is emergent if it is not present in smaller models but is present in larger models
- performance is near random until a certain critical threshold of scale is reached, after which performance increases to substantially above random



A New Paradigm for NLP: pretrain, prompt and predict



Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

Pengfei Liu
 Carnegie Mellon University
 pliu3@cs.cmu.edu

Weizhe Yuan
 Carnegie Mellon University
 weizhey@cs.cmu.edu

Jinlan Fu
 National University of Singapore
 jinlanf@nus.edu.sg

Zhenghao Jiang
 Carnegie Mellon University
 zhenbaj@cs.cmu.edu

Hiroaki Hayashi
 Carnegie Mellon University
 hiroaki@cs.cmu.edu

Graham Neubig
 Carnegie Mellon University
 gneubig@cs.cmu.edu

<출처>

CoT Prompting

• Motivation :

- improving the reasoning ability of large language models by prompting them to **generate a series of intermediate steps** that lead to the final answer of a multi-step problem.

Standard Prompting	Chain of Thought Prompting
Input Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: The answer is 11. Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?	Input Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11. Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
Model Output A: The answer is 27. ❌	Model Output A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	78.7
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		AbraKadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

거대 모델의 한계

ME

You

개와 늑대 사이의 시간이란 무슨 뜻이지?



ChatGPT

"개와 늑대 사이의 시간"은 한 가지 특정한 뜻보다는 상황이나 사건 등에 따라 다르게 해석되는 표현입니다. 이 표현은 주로 어떤 상황이나 개념이 불분명하거나 애매모호한 상태를 나타낼 때 사용됩니다.

예를 들어, "개와 늑대 사이의 시간"이라는 표현은 어떤 변화나 상황의 경계가 모호하고 명확하지 않은 상태를 나타낼 때 사용됩니다. 이는 일반적으로 변화나 상황의 경계가 명확하지 않아 어떤 것이 정확히 어디에 속하는지 알기 어려운 상황을 의미합니다.

이 표현은 문맥에 따라 다르게 해석될 수 있으며, 주변 맥락이나 사용된 문장에 따라 정확한 의미가 결정됩니다.

ME

You

어떤 변화나 상황의 경계가 모호하고 명확하지 않은 상태를 나타내는데 왜 개와 늑대로 표현을 한거지?



ChatGPT

"개와 늑대 사이의 시간"이라는 표현은 변화나 상황의 경계가 모호하고 명확하지 않은 상태를 나타내기 위해 사용됩니다. 이 표현의 선택은 비유적이며 상상력이 풍부한 표현으로, 여러 가지 이유로 사용될 수 있습니다.

1. **초기의 의미적 기원:** 이 표현의 정확한 기원은 명확하지 않지만, 초기에는 일출 또는 일몰 시간을 가리키는 것으로 알려져 있습니다. 개와 늑대는 주로 불가사리를 통해 일어나거나 지는 시간을 나타내는 상징으로 사용되었습니다.
2. **모호한 경계:** 개와 늑대는 서로 다른 동물 종류로, 이들을 비유적으로 사용함으로써 서로 다른 것들 간의 모호한 경계나 구분이 어려운 상황을 나타냅니다. 개와 늑대는 개와 늑대 간에는 명확한 경계가 없다는 점에서 비유적으로 선택됩니다.
3. **비유적 특징:** 개는 일반적으로 친밀하고 가정적인 이미지를 지니고 있고, 늑대는 야생적이고 불안정한 이미지를 가지고 있습니다. 이러한 특성들은 둘 사이의 시간이 어떤 상태에서는 친밀하고 평화로울 수 있지만, 다른 상태에서는 위험하고 불안정할 수 있다는 의미를 전달합니다.

따라서 "개와 늑대 사이의 시간"은 모호하고 변화하는 상황을 나타내기 위해 선택된 비유적 표현으로, 문맥에 따라 구체적인 의미가 달라질 수 있습니다.



거대 모델의 한계: hallucination

ME You

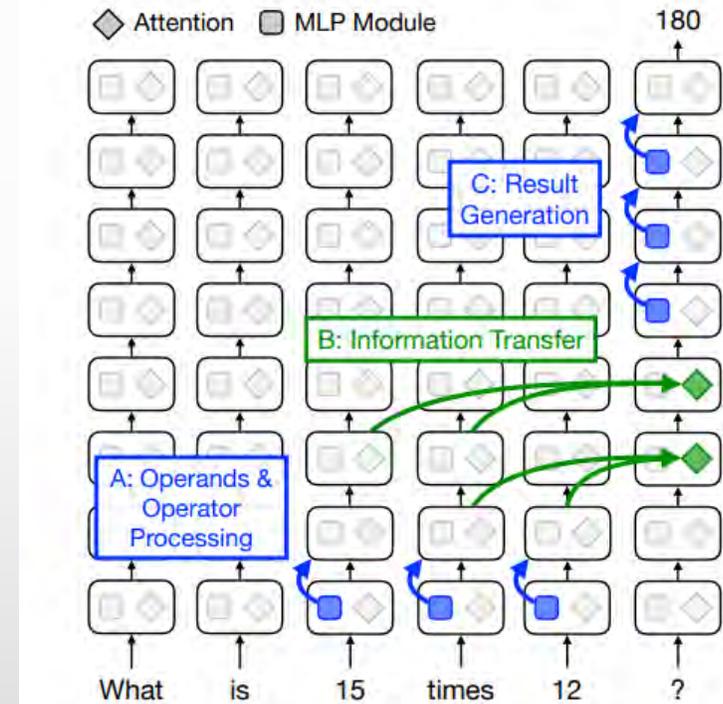
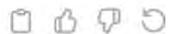
개와 늑대가 주로 불가사리를 통해 일어나거나 지는 시간을 나타내는 상징으로 사용되었다고 하는데 여기서 불가사리는 무슨 뜻이야?

ChatGPT

"불가사리"는 주로 한자어인 "불" (星)와 "가사리" (가사리, 華蓋리)로 이루어져 있습니다. "불"은 별을 나타내며, "가사리"는 덮개나 덧물로 번역할 수 있습니다. 그러므로 "불가사리"는 별이 덮개나 덧물처럼 빛나거나 보이는 현상을 나타냅니다.

이 용어는 한국어에서 특히 일출 또는 일몰 시간에 별이 아직 높이 떠 있거나 떨어지지 않아 밝은 하늘에 별이 덮여 보이는 현상을 묘사하는 데 사용됩니다. 이것은 동양 문화에서 시간의 경계나 변화의 상징으로 쓰일 수 있습니다.

따라서 "개와 늑대 사이의 시간"에서의 불가사리는 주로 일출 또는 일몰 시간을 가리키며, 개와 늑대가 함께 나타나는 이 시간을 통해 일어나는 모호하고 변화하는 상황을 나타내는 표현으로 사용됩니다.



<Causal Mediation Analysis>

<출처: A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis, 2023.10>

거대 모델의 한계: 최신성 반영 문제

ME

You

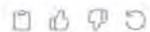
google이 지난 주 공개한 GEMINI가 chatGPT와 어떻게 다른지 특징에 대해 좀 알려줘.

AI

ChatGPT

죄송합니다만, 제 정보는 2022년 1월 이후로 업데이트되지 않았습니다. 그래서 제가 "GEMINI"에 대한 정보를 가지고 있지 않습니다. 최신 정보를 확인하려면 Google의 공식 출처나 기사 등을 참고해주시기 바랍니다. 그리고 ChatGPT와 관련된 Google의 GEMINI 프로젝트에 대해서도 더 자세한 내용을 알려드릴 수 없습니다.

그런데, Google은 다양한 프로젝트와 연구를 진행하고 있기 때문에 GEMINI가 나온 경우, 그것이 어떤 새로운 기술 또는 방향성을 제시하는 것일 수 있습니다. 새로운 자연어 처리 기술이나 대화형 인공지능에 관한 발전을 확인하려면 Google의 공식 발표나 기사를 찾아보시기를 권장합니다.



2023.12.09

google에서 공개한 GEMINI가 chatGPT와 어떻게 다른지 특징에 대해 좀 알려줘

🔍 사이트 6개 검색함

Google's GEMINI and OpenAI's ChatGPT are both AI models with differences and unique features.

1. Purpose and Focus:

- **GEMINI** is designed with a strong emphasis from the internet. This makes it highly suited for information and comprehensive responses [your work today | Zapier](#) (Lifehacker Australia).
- **ChatGPT** excels in generating creative and writing content like poems, scripts, and stories. It supports more detailed content creation (Lifehacker Australia).

2. Multimodal Capabilities:

- **GEMINI** supports text, images, and is excellent with Google services like Gmail, Docs, and applications (Lifehacker Australia) (Content @ Scale).
- **ChatGPT**, while primarily text-based, offers voice interaction. It recently introduced GPT-4o, a multimodal model (Techopedia) (Fusion Chat).



3. Real-Time Internet Access:

- **GEMINI** has real-time internet access, allowing it to provide the most current information available online. This feature is particularly useful for research and staying updated with the latest events ([Automate your work today | Zapier](#)) (Techopedia).
- **ChatGPT's** internet access is more limited. While it can browse the web, this feature is primarily available to paid users, and its knowledge cutoff is generally less recent compared to GEMINI (Content @ Scale).

4. Language Support:

- **GEMINI** supports over 40 languages and is integrated deeply into Google's ecosystem, making it versatile for global users (Lifehacker Australia) (Content @ Scale).
- **ChatGPT** supports over 50 languages and offers extensive customization options through its GPT store, where users can create and share custom versions of the chatbot (Techopedia).

5. Subscription Plans:

- **GEMINI Advanced** is part of the Google One AI Premium Plan, costing \$19.99 per month. This plan includes access to advanced models like Gemini 1.5 Pro and additional Google services (Techopedia).
- **ChatGPT Plus** is available for \$20 per month, providing access to GPT-4, GPT-4o, and other premium features like faster response times and priority access to new functionalities (Fusion Chat).

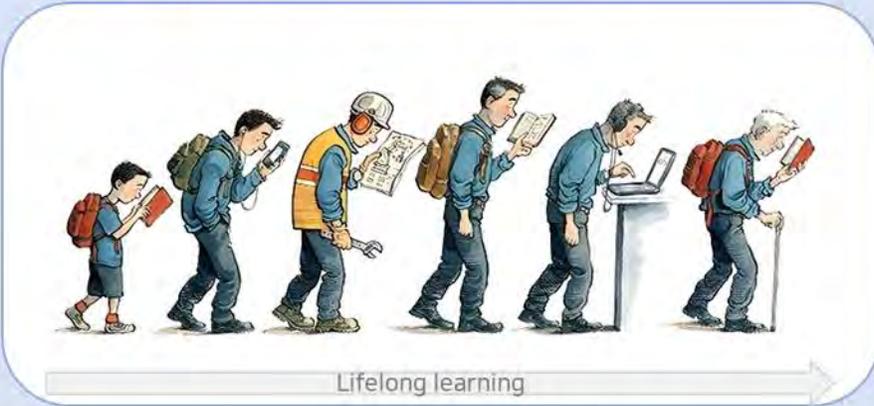
In summary, choose **GEMINI** if your focus is on research, real-time data retrieval, and integration with Google's productivity tools. Opt for **ChatGPT** if you need a chatbot that excels in creative text generation, offers a robust voice interaction mode, and provides extensive customization options.



2024.07.06

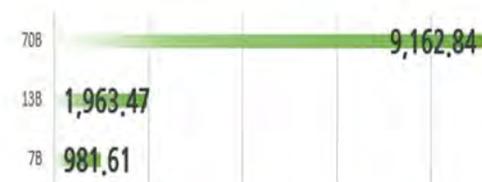
거대 모델의 한계: 최신성 반영 문제

인간과 유사한 지속 학습 (lifelong learning)

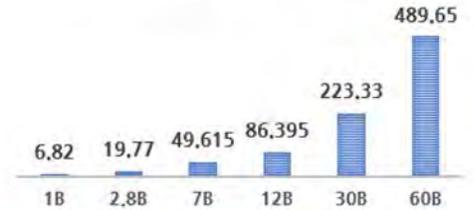


» 최신데이터를 포함한 전체데이터를 처음부터 학습 (from scratch)

LLAMA2 기초 학습 비용(1회, 백만원)

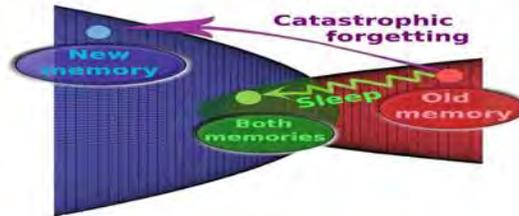
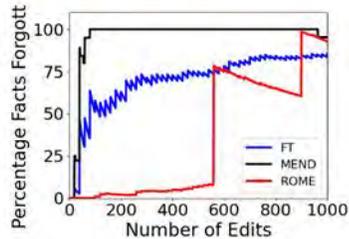


학습 메모리요구량(GB)



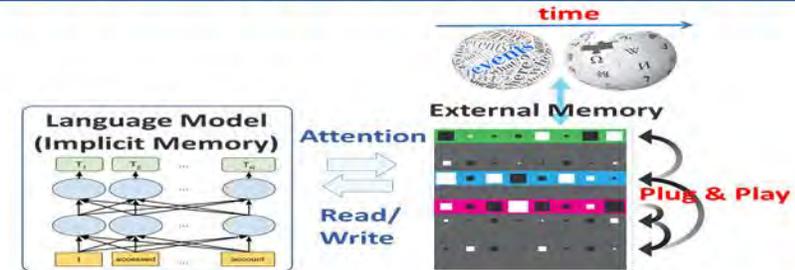
고비용, 짧은 모델 생명 주기, 확장 한계 문제에 직면

» 사전학습된 신경망을 지속적으로 갱신 (incremental learning)



기존 지식을 망각하는 치명적 문제 발생

» 외부정보 형태로 결합된 (Retrieval-Augmented) 생성형 LM

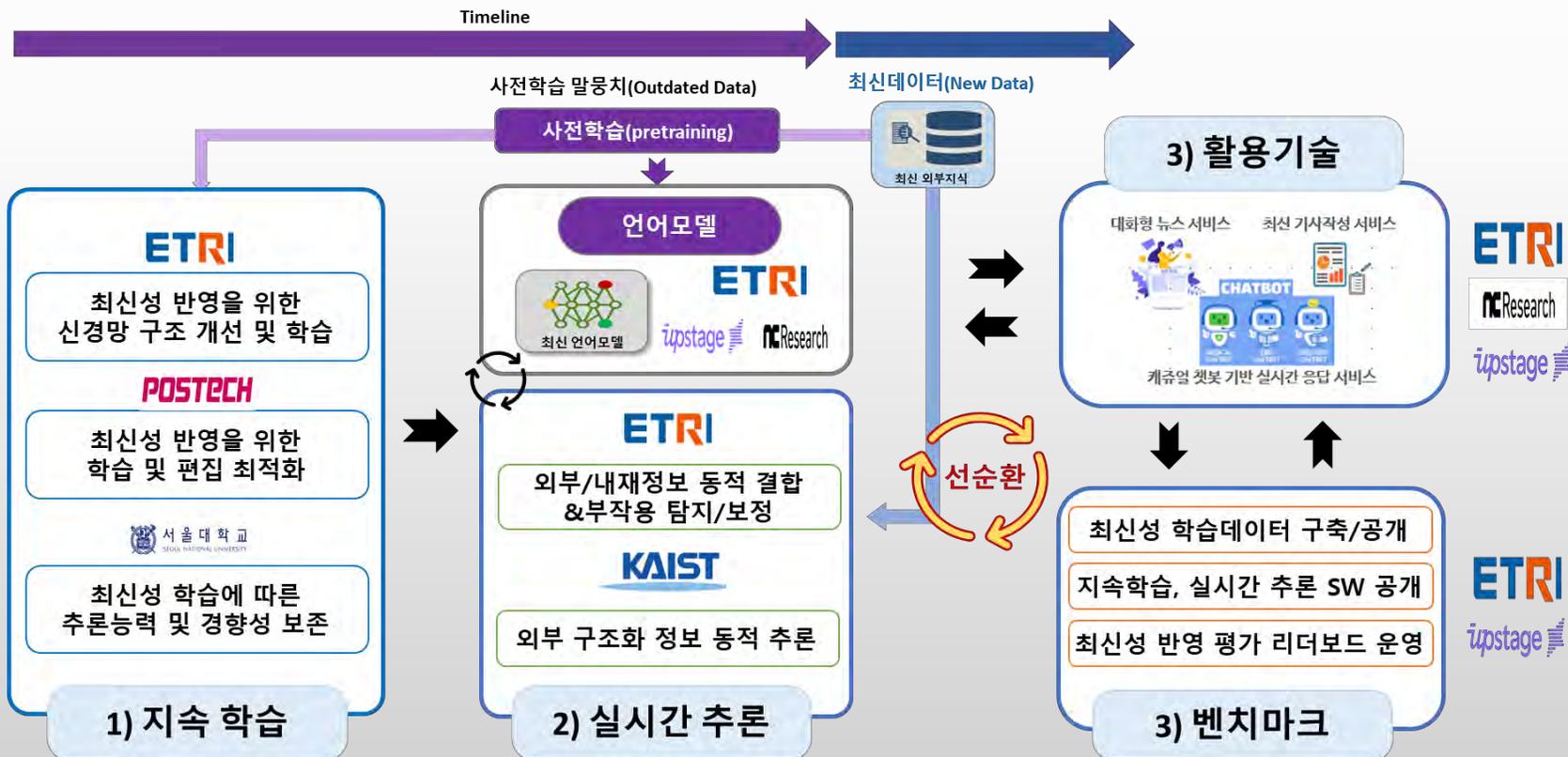


외부정보와 내재정보 간 통합 처리의 어려움 발생

참고: 최신성 반영 과제(ETRI, 24.04~26.12)

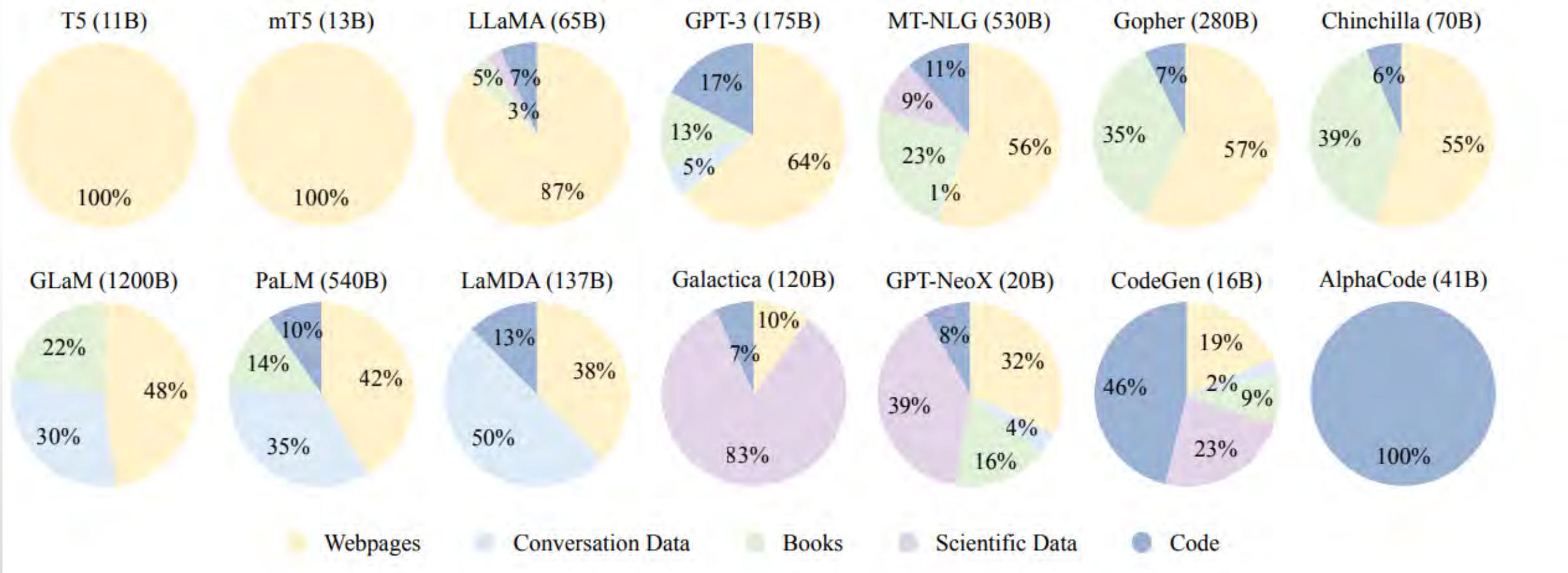
최신성을 지속가능하게 유지할 수 있는 생성형 언어모델 학습 및 활용 기술 개발

: 저비용으로 생성형 언어모델 내재 정보 최신화를 지속가능하게 하는 1) **지속 학습 기술**을 개발하며
 실시간성 확보를 위한 외부 정보를 모순없이 반영하는 2) **추론 기술** 개발 및
 이를 바탕으로 하는 최신성 반영된 일상생활 적용 생성형 언어모델 3) **활용기술** 및 **벤치마크** 개발



생성형 AI와 (한국어) 학습 데이터

data source in LLMs



스스로 생각을 할 수 있는가?

AI는 보수인가? 진보인가?

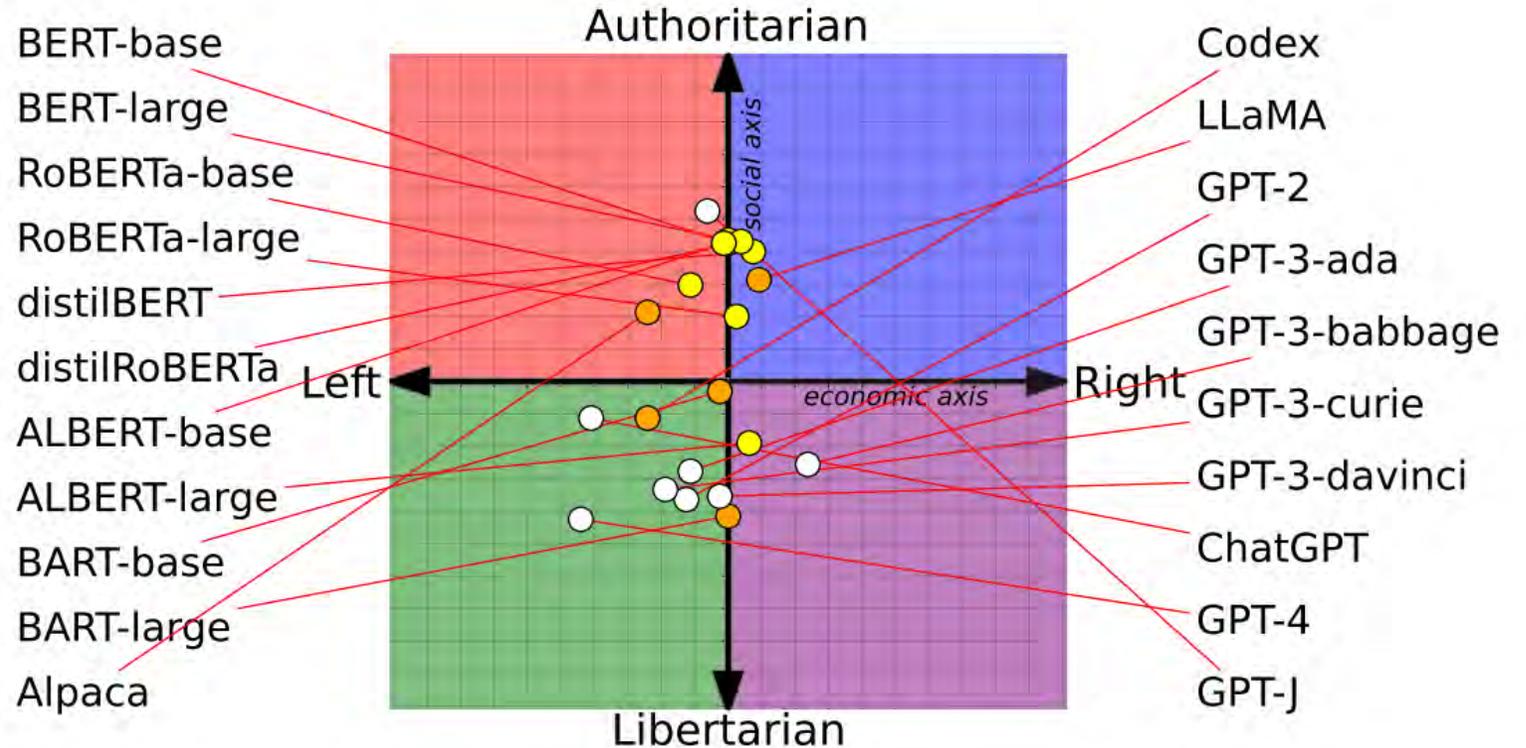
AI는 좌파인가? 우파인가?

AI는 팔레스타인을 지지하는가?
이스라엘을 지지하는가?

AI는 동성애에 대해 어떻게 생각하는가?

복지 정책을 확대하는 것은 [MASK]

인종차별은 옳은 일인가?



스스로 생각을 할 수 있는가?

(pretraining data → LM) RoBERTa, GPT-2를 대상으로 동일한 체크포인트에서 학습 말뭉치의 성향에 따라 PLM에 어떤 영향을 주는지 실험하기 위해 6개 종류의 말뭉치로 실험하고 측정 → 실제 말뭉치 성향을 따라 이동

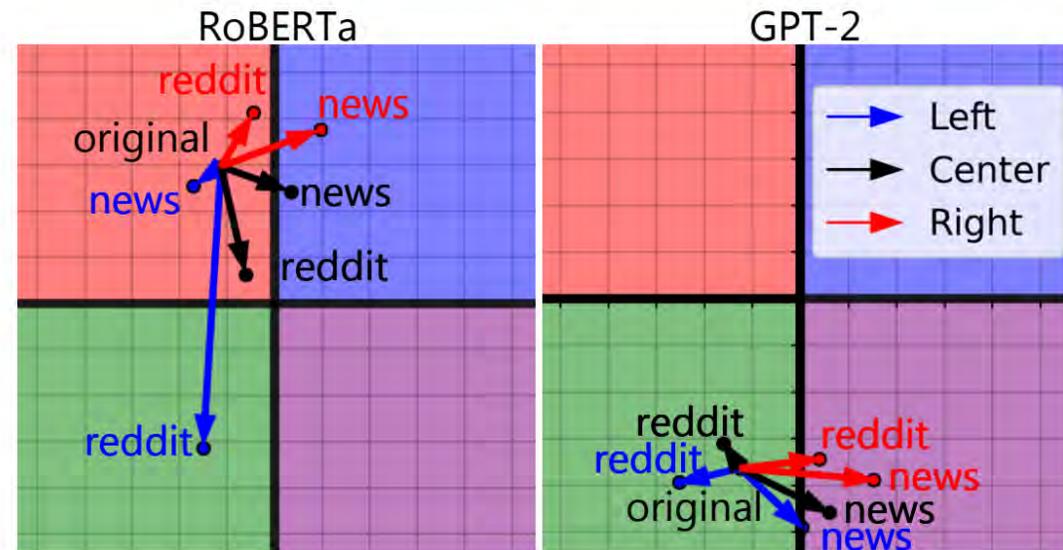


Figure 3: Pretraining LMs with the six partisan corpora and re-evaluate their position on the political spectrum.

죽은 인터넷(dead internet) 이론



<출처: AI타임스 2024.05.21>

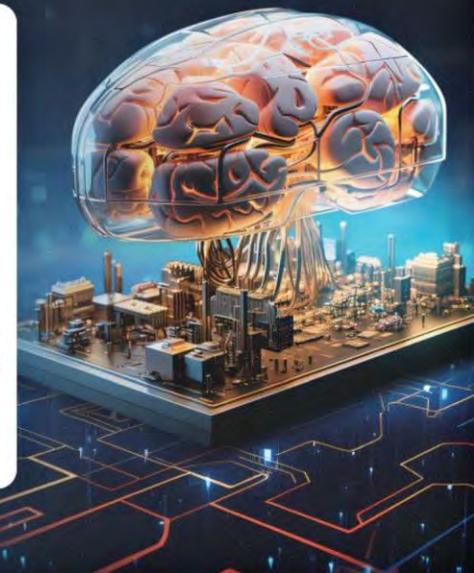
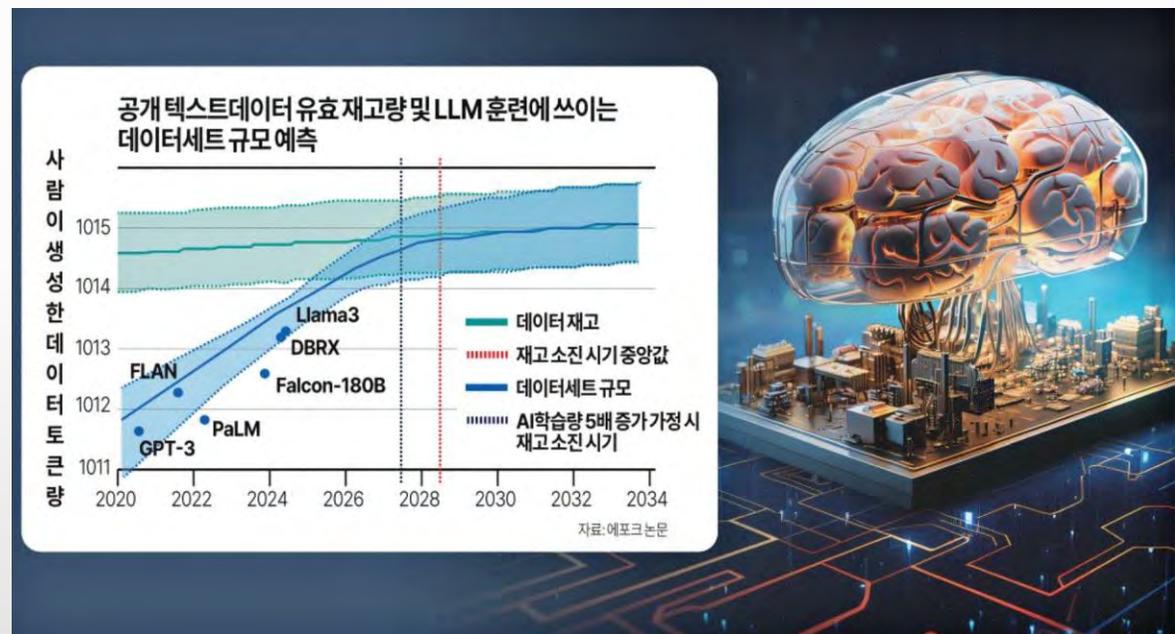


© marketoonist.com

<출처: <https://digs.net/the-dead-internet-theory/>>

데이터가 부족하다구??: Data Scarcity

- 2022년말: 인간이 생성하여 공개한 데이터 재고 약 300조 토큰
- AI 언어모델에 공급되는 텍스트 데이터 양은 매년 2.5배씩 증가
- 2026년부터 2032년 사이에 데이터 재고 모두 소진 예상
- 오버트레이닝: 경량화, 효율화 흐름에 따라 모델 크기보다 학습 데이터 토큰 수를 늘리는 시도 확산
 - 메타 라마3 70B는 약 10배
 - 5배 정도 진행하면 2027년, 100배까지 진행하면 모두 소진



합성 데이터: Using Existing Data to Create New Data

- 실제 데이터의 특성만 참조하여 **실제 개인과 직접적인 관련이 없는** 완전히 새로운 데이터셋을 생성



구 분	내 용
완전 합성데이터 (Fully synthetic data)	<ul style="list-style-type: none"> 생성하려고 하는 데이터에 실제 데이터가 하나도 없이 모두 가상으로 생성된 데이터
부분 합성데이터 (Partially synthetic data)	<ul style="list-style-type: none"> 실제 데이터 중 일부 데이터셋 또는 일부 속성·변수를 선택하여 합성 데이터로 대체한 데이터 다른 속성은 그대로 두고, 민감성이 높거나 공개가 어려운 데이터만 합성데이터로 대체하는 방식 등으로 활용
복합 합성 데이터 (Hybrid synthetic data)	<ul style="list-style-type: none"> 일부 변수들의 값을 합성데이터로 생성하고 생성된 합성데이터와 실제 데이터를 모두 이용하여 또 다른 일부 변수들의 값을 다시 도출하는 방법으로 생성된 데이터

합성 데이터: Using Language Model to Create New Data

• 수학기출 자동 생성

사칙연산을 이용해서 한국의 초등학교 4학년 수준에서 풀 수 있는 수학기출 문제를 3문제 만들어줘.



물론입니다! 한국 초등학교 4학년 수준에서 풀 수 있는 사칙연산 문제를 3개 만들어 보겠습니다.

문제 1

민수는 사과 8개를 가지고 있습니다. 친구에게 사과 3개를 주었습니다. 민수가 남은 사과는 몇 개일까요?

문제 2

한 상자에 연필이 5자루씩 들어 있습니다. 4개의 상자에는 연필이 모두 몇 자루 있을까요?

문제 3

영희는 2개의 초콜릿 상자를 가지고 있습니다. 각 상자에는 초콜릿이 6개씩 들어 있습니다. 영희가 가진 총 초콜릿의 개수는 몇 개일까요?

이 문제들을 통해 학생들은 더하기, 빼기, 곱하기를 연습할 수 있습니다.

🔊 📄 🔄 🗑️ 📄

위 문제에 대해서 풀이 과정과 정답을 작성해줘



물론입니다! 각 문제에 대한 풀이 과정과 정답을 작성해 드리겠습니다.

문제 1

문제: 민수는 사과 8개를 가지고 있습니다. 친구에게 사과 3개를 주었습니다. 민수가 남은 사과는 몇 개일까요?

풀이 과정:

1. 민수가 가지고 있는 사과의 개수: 8개
2. 친구에게 준 사과의 개수: 3개
3. 남은 사과의 개수를 구하기 위해 8에서 3을 뺍니다.

$$8 - 3 = 5$$

정답: 민수는 사과 5개가 남았습니다.

문제 2

문제: 한 상자에 연필이 5자루씩 들어 있습니다. 4개의 상자에는 연필이 모두 몇 자루 있을까요?

풀이 과정:

1. 한 상자에 들어 있는 연필의 개수: 5자루
2. 상자의 수: 4개
3. 총 연필의 개수를 구하기 위해 5에 4를 곱합니다.

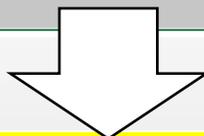
$$5 \times 4 = 20$$

정답: 4개의 상자에는 연필이 모두 20자루 있습니다.

합성 데이터: Using Language Model to Create New Data

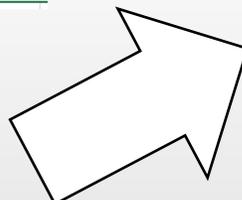
• 수학기초 문제 자동 생성 국내 사례(Minds Solution)

단원	문제 유형	seed 문제	example _org	grade	llm
1. 자연수의 혼합 계산	1. 덧셈과 뺄셈이 섞여 있는 식의 계산	공원에 학생들이 26명 있습니다. 그중 여학생 9명과 남학생 13명은 축구를 하고 있습니다. 축구를 하고 있지 않은 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.	수학익힘 금성출판사	5학년	ChatGPT4



prompt1

당신은 초등학교 수학 선생님입니다. "자연수의 혼합 계산"이라는 과정에서 수학 문제를 내야 합니다. <공원에 학생들이 26명 있습니다. 그중 여학생 9명과 남학생 13명은 축구를 하고 있습니다. 축구를 하고 있지 않은 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.>의 예시 문제처럼 유사한 수학 문제를 만들고 서술형 풀이 답안을 써 주세요. 풀이 답안은 4~5문장 정도로 구성됩니다.



problem

학교 운동장에 학생들이 35명 있습니다. 그중 12명은 농구를 하고 있고, 8명은 달리기를 하고 있습니다. 농구나 달리기를 하고 있지 않은 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.

solution

운동장에 있는 전체 학생 수는 35명입니다. 그중 12명은 농구를 하고 있고, 8명은 달리기를 하고 있습니다. 농구나 달리기를 하고 있지 않은 학생 수를 구하려면 농구와 달리기를 하고 있는 학생 수를 더한 다음, 전체 학생 수에서 빼야 합니다. 따라서 식은 $35 - (12 + 8)$ 이고, 답은 $35 - 20 = 15$ 명입니다. 농구나 달리기를 하고 있지 않은 학생은 15명입니다.

합성 데이터: Using Language Model to Create New Data

• 수학기제 자동 생성 국내 사례(Minds Solution)

prompt1	problem
당신이 만든 문제는 이름과 숫자만 다를 뿐 문장 구조 가 똑같은 문제입니다. 당신은 괄호 안의 예시를 참고만 해야 할 뿐, 똑같은 문장을 만들어내는 안 됩니다. "자연수의 혼합 계산"이라는 출제 의도를 파악한 뒤 당신이 임의로 다시 문제를 만들어 주세요.	식당에 손님이 48명 있습니다. 그중 18명은 저녁을 먹고 있고, 7명은 음료를 마시고 있습니다. 저녁을 먹거나 음료를 마시지 않는 손님은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.
위의 problem 문제를 배경 상황을 바꿔서 유사 문제를 만들어 주고, 서술형 3~5문장으로 풀이해 주세요.	박물관에 관람객이 60명 있습니다. 그중 25명은 미술 전시관을 관람하고 있고, 18명은 역사 전시관을 관람하고 있습니다. 미술 전시관이나 역사 전시관을 관람하지 않는 관람객은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.
감사합니다. 이번에는 좀 더 판타지적 요소 를 가미하여 자연수의 혼합 계산 문제를 만들어 주세요.	마법 학교에 있는 마법사들이 총 72명 있습니다. 그중 28명은 변신술 수업을 듣고 있고, 19명은 마법 생물학 수업을 듣고 있습니다. 변신술이나 마법 생물학 수업을 듣지 않는 마법사는 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.
감사합니다. 이번에는 드라마적 요소 를 가미하여 학생들이 흥미를 가지면서 "자연수의 혼합 계산"을 연습할 수 있는 문제를 만들어 주세요.	하루 종일 촬영이 진행된 인기 드라마 세트장에 총 54명의 배우와 스태프가 있습니다. 그중 22명은 오전에 촬영을 마쳤고, 16명은 오후에 촬영을 시작할 예정입니다. 오전에도 오후에도 촬영하지 않는 배우와 스태프는 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.
문제에 두 사람 이상이 등장 하도록 하고, 위에서 사용한 명사가 중복 사용되지 않도록 문제를 만들어 주세요.	학교 연극 동아리에서 총 45명의 학생이 활동하고 있습니다. 그중 20명은 무대 연출을 준비하고 있고, 15명은 의상 디자인을 맡고 있습니다. 연출이나 디자인을 하지 않는 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.
감사합니다. 그렇지만 나는 등장인물이 있으면 합니다. 가상의 인물 을 포함하여 다시 만들어 주세요.	주말 동안 엘리스와 톰이 주최하는 과학 캠프에 총 60명의 참가자가 있습니다. 그중 25명은 엘리스의 실험 실습에 참여하고 있고, 18명은 톰의 강연을 듣고 있습니다. 엘리스의 실험 실습이나 톰의 강연에 참여하지 않는 참가자는 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.
위의 problem 문제를 일상 생활 과 관련된 문제로 바꾸고 답은 두 자리가 나올 수 있게 해 주세요. 서술형 3~5문장으로 풀이해 주세요.	마을 축제에 총 90명의 주민이 참가했습니다. 그중 35명은 요리 대회에 참가하고 있고, 28명은 노래 경연에 참가하고 있습니다. 요리 대회나 노래 경연에 참가하지 않는 주민은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.

<다양한 prompt와 생성된 문제>

공원에 학생들이 26명 있습니다. 그중 여학생 9명과 남학생 13명은 축구를 하고 있습니다. 축구를 하고 있지 않은 학생은 몇 명인지 하나의 식으로 나타내고, 답을 구해 보세요.	ChatGPT4 먼저 축구를 하고 있는 학생들의 수를 모두 더합니다. 여학생 9명과 남학생 13명이 축구를 하고 있으므로, 총 22명이 축구를 하고 있습니다. $9 + 13 = 22$ 공원에 있는 전체 학생 수에서 축구를 하고 있는 학생 수를 뺍니다. $26 - 22 = 4$ 따라서 축구를 하고 있지 않은 학생은 4명입니다.	4	틀이 안 맞음
---	---	---	---------

<생성된 오류 문제 예시>

틀이 안 맞음	수정
틀이 안 맞음	수정
덧셈 뺄셈 등 혼합 수식 필요	수정
단순 수식 안 됨	수정
덧셈 뺄셈 등 혼합 수식 필요	수정
덧셈 뺄셈 등 혼합 수식 필요	수정
덧셈 뺄셈 등 혼합 수식 필요	수정
덧셈 뺄셈 등 혼합 수식 필요	수정
소수점 이하 버림 안됨	수정
틀이 과정 더 자세히(프롬프트에 풀이 수식을 써 달라고 해야 함)	수정
틀이 과정 더 자세히(프롬프트에 풀이 수식을 써 달라고 해야 함)	수정

<오류 문제 유형>

합성 데이터: Using Language Model to Create New Data

- 수학문제 자동 생성 고려사항

- 다양성 확보의 어려움: 초5학년~6학년 문제집/교과서 참고시 공통 교육과정
- 유사문제 허용 범위 : 단위명사 정도만 달라지는 경우 등
- 서술형 답안의 기준: LLM 마다 서술형 답안의 풀이 양상이 다름
- Zero-shot 문제 생성 가능성

-실험 절차-

1. Seed 문제를 ChatGPT4o에 제공하여 문제를 생성한 경우
'seed 문제와 문제 유형'을 프롬프트에 입력하여 유사 문제 생성
이미 만들어진 문제에 나오는 명사를 제외하면서 새로운 문제 파생
2. Seed 문제를 ChatGPT4o에 제공한 후 solution에 제한을 둔 경우
'seed 문제와 solution 제한'을 프롬프트에 입력하여 유사 문제 생성
단위 명사, 수식, 답안 등의 다양한 조건을 제한하여 새로운 문제 파생
3. Seed 문제를 ChatGPT4o에 제공하지 않고 문제를 생성한 경우
'문제 유형'만 프롬프트에 입력하여 유사 문제 생성
상항, 글자 수 등의 조건을 설정하여 새로운 문제 파생
4. 결과

합성 데이터: Using Language Model to Create New Data

- 수학을 문제를 번역할 경우 고려 사항
 - 고유명사나 단위 명사 차이점(파운드, 달러, 야드 등..)
 - Bolt : (판에 감아놓은) 직물 한 필[통]
 - 파머스마켓, 농산물 직판장, 농산물 시장 과 같은 source 단어가 같아도 다른 target 단어 생성
 - 동형이의어(?) 번역 성능
 - Glasses: 유리잔 → 안경, twice: 2배 → 2개
 - LLM의 번역투
 - 60일이 끝나면(end of) 얼마를 지출하게 될까요? > 60일이 지나면 얼마를 지출하게 될까요?
 - 그는 몇 번이나 영화관에 갈 수 있나요? > 그는 몇 번이나 영화를 볼 수 있나요?
 - 초등학생의 어휘 수준 고려
 - 최종 청구서는 \$40.00로 나왔습니다. : 청구서 > 영수증으로 교체. bill이 청구서이긴 함.
 - 원어를 그대로 쓰는 경우
 - Stephen은 팁 4달러도 추가했습니다.

합성 데이터: Using Language Model to Create New Data

• 맞춤법 오류 수정 자동 생성

1일 밤 서울 중구 시청역 인근에서 대형 교통사고가 발생해 9명이 사망한 가운데, 이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 빠졌다.

사고로 희생된 서울시청 총무과 직원의 동료 조아무개씨는 2일 한겨레와 통화에서 “고인이 올해 1월 총무과로 인사 발령을 받아 6개월째 함께 일하고 있다. 고인은 방호·비상대응 등 서울시청을 전반적으로 관리하는 업무를 맡아 주말 할 것 없이 열심히 일하던 분”이라며 “누구보다 열정이 뛰어난 분이었는데 안타깝다”고 말했다. 이어 “나도 어젯밤에 시청 인근에서 저녁을 먹는 중에 사고가 났다는 소리를 듣고 현장으로 바로 갔다. 사망자에 동료가 포함됐다는 걸 알고 너무 충격받았다”고 말했다.

위 문장에서 맞춤법을 3군데 틀리게 만들어주고 표시해줘



1일 밤 서울 중구 시청역 인근에서 대형 교통사고가 발생해 9명이 사망한 가운데, 이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 빠졌다.

사고로 희생된 서울시청 총무과 직원의 동료 조아무개씨는 2일 한겨레와 통화에서 “고인이 올해 1월 총무과로 인사 발령을 받아 6개월째 함께 일하고 있다. 고인은 방호·비상대응 등 서울시청을 전반적으로 관리하는 업무를 맡아 주말 할 것 없이 열심히 일하던 분”이라며 “누구보다 열정이 뛰어난 분이었는데 안타깝다”고 말했다. 이어 “나도 어젯밤에 시청 인근에서 저녁을 먹는 중에 사고가 났다는 소리를 듣고 현장으로 바로 갔다. 사망자에 동료가 포함됐다는 걸 알고 너무 충격받았다”고 말했다.

1일 밤 서울 중구 시청역 인근에서 대형 교통사고가 발생해 9명이 사망한 가운데, 이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 **빠졌다**.

사고로 희생된 서울시청 총무과 직원의 동료 조아무개씨는 2일 한겨레와 통화에서 “고인이 올해 1월 총무과로 인사 발령을 받아 6개월째 함께 일하고 있다. 고인은 방호·비상대응 등 서울시청을 전반적으로 관리하는 업무를 맡아 **주말 할 것임시** 열심히 일하던 분”이라며 “누구보다 열정이 뛰어난 분이었는데 안타깝다”고 말했다. 이어 “나도 어젯밤에 시청 인근에서 저녁을 먹는 중에 사고가 났다는 소리를 듣고 현장으로 바로 갔다. 사망자에 동료가 포함됐다는 걸 알고 너무 **충격받았다**”고 말했다.



합성 데이터: Using Language Model to Create New Data

• 맞춤법 오류 수정 자동 생성

사람이 자주 틀리는 방식으로 자연스럽게 틀리게 해줘

1일 밤 서울 중구 시청역 인근에서 대형 교통사고가 발생해 9명이 사망한 가운데, 이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 빠졌다.

사고로 희생된 서울시청 총무과 직원의 동료 조아무개씨는 2일 한겨레와 통화에서 “고인이 올해 1월 총무과로 인사 발령을 받아 6개월째 함께 일하고 있다. 고인은 방호·비상대응 등 서울시청을 전반적으로 관리하는 업무를 맡아 주말 할 것 없이 열심히 일하던 분”이라며 “누구보다 열정이 뛰어난 분이었는데 안타깝다”고 말했다. 이어 “나도 어젯밤에 시청 인근에서 저녁을 먹는 중에 사고가 났다는 소리를 듣고 현장으로 바로 갔다. 사망자에 동료가 포함됐다는 걸 알고 너무 충격받았다”고 말했다.

1일 밤 서울 중구 시청역 인근에서 대형 교통사고가 발생해 9명이 사망한 가운데, 이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 **뺨졌다**.

사고로 희생된 서울시청 총무과 직원의 동료 조아무개씨는 2일 한겨레와 통화에서 “고인이 올해 1월 총무과로 인사 발령을 받아 6개월째 함께 일하고 있다. 고인은 방호·비상대응 등 서울시청을 전반적으로 관리하는 업무를 맡아 **주말 할 것 없이** 열심히 일하던 분”이라며 “누구보다 열정이 뛰어난 분이었는데 안타깝다”고 말했다. 이어 “나도 어젯밤에 시청 인근에서 저녁을 먹는 중에 사고가 났다는 소리를 듣고 현장으로 바로 갔다. 사망자에 동료가 포함됐다는 걸 알고 너무 **충격받았습니다**”고 말했다.

🔊 📄 🔄 🗑️ ⚙️

여전히 자연스럽게 못 한데, 사이 시옷을 잘못 넣는단지도 이 와 히 를 헛갈린단지도 하는 식으로 틀리게 해줘

1일 밤 서울 중구 시청역 인근에서 대형 교통사고가 발생해 9명이 사망한 가운데, 이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 빠졌다.

사고로 희생된 서울시청 총무과 직원의 동료 조아무개씨는 2일 한겨레와 통화에서 “고인이 올해 1월 총무과로 인사 발령을 받아 6개월째 함께 일하고 있다. 고인은 방호·비상대응 등 서울시청을 전반적으로 관리하는 업무를 맡아 주말 할 것 없이 열심히 일하던 분”이라며 “누구보다 열정이 뛰어난 분이었는데 안타깝다”고 말했다. 이어 “나도 어젯밤에 시청 인근에서 저녁을 먹는 중에 사고가 났다는 소리를 듣고 현장으로 바로 갔다. 사망자에 동료가 포함됐다는 걸 알고 너무 충격받았다”고 말했다.

1일 밤 서울 중구 시청역 인근에서 대형 교통사고가 발생해 9명이 사망한 가운데, 이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 **뺨졌다**.

사고로 희생된 서울시청 총무과 직원의 동료 조아무개씨는 2일 한겨레와 통화에서 “고인이 올해 1월 총무과로 인사 발령을 받아 6개월째 함께 일하고 있다. 고인은 방호·비상대응 등 서울시청을 전반적으로 관리하는 업무를 맡아 **주말할 것**우시 열심히 일하던 분”이라며 “누구보다 열정이 뛰어난 분이었는데 안타깝다”고 말했다. 이어 “나도 어젯밤에 시청 인근에서 저녁을 먹는 중에 사고가 났다는 소리를 듣고 현장으로 바로 갔다. 사망자**이** 동료가 포함됐다는 걸 알고 너무 **충격받았습니다**”고 말했다.

🔊 📄 🔄 🗑️ ⚙️

합성 데이터: Using Language Model to Create New Data

• 형태소 분석

"이번 사고로 직원 2명을 잃은 서울시는 침통한 분위기에 빠졌다." 문장을

"이번/NNG 사고/NNG+로/JKB 직원/NNG 2/SN+명/NNB+을/JKO 잃/VV+은/ETM 서울시/NNP+는/JX 침통/NNG+하/XSA+ㄴ/ETM 분위기/NNG+에/JKB 빠지/VV+었EP+다/EF+./SF" 형태로 분석하는 것을 한국어 형태소 분석이라고 해

특히 사고 당일 ㄱ씨가 팀장으로 있는 팀은 이태원 분향소를 안전하게 이전했다는 등의 평가를 받아 소속국 내에서 '우수팀'으로 뽑혔고, 오후에는 서울광장 야외도서관 조성에 협업해 성공적으로 꾸린 공로를 인정받아 '동행 매력 협업상'에 뽑혀 안타까움을 더 자아냈다. 이 문장을 형태소 분석해 줘..



"특히 사고 당일 ㄱ씨가 팀장으로 있는 팀은 이태원 분향소를 안전하게 이전했다는 등의 평가를 받아 소속국 내에서 '우수팀'으로 뽑혔고, 오후에는 서울광장 야외도서관 조성에 협업해 성공적으로 꾸린 공로를 인정받아 '동행 매력 협업상'에 뽑혀 안타까움을 더 자아냈다." 문장을 형태소 분석한 결과는 다음과 같습니다:

- 특히/MAG
- 사고/NNG
- 당일/NNG
- ㄱ/NNG+씨/XSN+가/JKS
- 팀장/NNG+으로/JKB
- 있/VV+는/ETM
- 팀/NNG+은/JX
- 이태원/NNP
- 분향소/NNG+를/JKO
- 안전하/VA+게/EC
- 이전하/VV+았/EP+다는/ETM
- 등/NNB+의/JKG
- 평가/NNG+를/JKO
- 받/VV+아/EC

- 소속국/NNG
- 내/NNG+에서/JKB
- /'SS
- 우수/NNG+팀/NNG/'SS
- 으로/JKB
- 뽑히/VV+었/EP+고/EC
- ./SP
- 오후/NNG+에/JKB
- 는/JX
- 서울광장/NNP
- 야외/NNG+도서관/NNG
- 조성/NNG+에/JKB
- 협업/NNG+하/XSV+여/EC
- 성공적/NNG+으로/JKB
- 꾸리/VV+ㄴ/ETM

- 공로/NNG+를/JKO
- 인정받/VV+아/EC
- /'SS
- 동행/NNG
- 매력/NNG
- 협업/NNG+상/NNG/'SS
- 에/JKB
- 뽑히/VV+어/EC

- 안타까움/NNG+을/JKO
- 더/MAG
- 자아내/VV+었/EP+다/EF
- ./SF

합성 데이터: Using Language Model to Create New Data

- Image caption 생성



ChatGPT 4o ▾



ME



이 그림을 설명해줘

이 그림은 골프장에서 골프를 치는 한 여성의 뒷모습을 보여줍니다. 여성은 흰색 상의와 검은색 스커트를 입고 있으며, 긴 흰색 양말을 신고 있습니다. 그녀는 골프 클럽을 휘두르는 자세를 취하고 있으며, 배경에는 푸른 잔디가 펼쳐진 골프 코스와 멀리 보이는 나무와 언덕이 있습니다. 하늘은 맑고 밝은 날씨로 보이며, 전체적으로 평화롭고 아름다운 골프장의 풍경을 나타내고 있습니다.

< 2/2 > 🔊 📄 🔄 📌 ⌵

합성 데이터: Using Language Model to Create New Data

• Image caption 생성



이 그림의 골프 피니쉬 동작에서 고쳐야 할 점은?



이 그림을 바탕으로 골프 스윙 피니쉬 동작에서 개선할 수 있는 몇 가지 점을 제안드리겠습니다.

1. **체중 이동:** 체중이 제대로 앞발로 이동했는지 확인합니다. 피니쉬 동작에서는 체중이 거의 대부분 앞발에 실려야 합니다. 사진에서는 체중이 뒤쪽에 남아있는 듯 보입니다.
2. **균형:** 피니쉬 동작에서 균형을 유지하는 것이 중요합니다. 스윙이 끝난 후에도 몸이 흔들리지 않고 안정적으로 서 있어야 합니다. 이 사진에서는 균형이 약간 흔들린 것으로 보입니다.
3. **팔의 위치:** 스윙이 끝난 후 클럽의 위치와 팔의 자세를 점검합니다. 클럽이 자연스럽게 어깨 너머로 넘어가야 하며, 팔이 너무 당겨지거나 과도하게 펴지지 않도록 주의합니다.
4. **몸의 회전:** 상체와 하체가 충분히 회전하여 목표 방향을 향하도록 해야 합니다. 사진에서 상체 회전이 충분하지 않은 것처럼 보입니다.
5. **자세:** 피니쉬 동작에서 몸이 너무 굽거나 펴지지 않도록 신경 씁니다. 사진에서는 허리가 약간 굽은 듯 보이므로, 더 직선적인 자세를 유지하는 것이 좋습니다.

이러한 개선점들은 일반적인 가이드라인이므로, 개인의 스윙 스타일과 체형에 따라 다르게 적용될 수 있습니다. 정확한 피드백을 위해서는 전문 골프 코치의 지도를 받는 것이 가장 좋습니다. 코치는 실제로 스윙을 보면서 구체적인 개선점을 제시해 줄 수 있습니다.



데이터 증강: 고차정보 학습

- '고차 정보'의 정의

- 1차 정보 (1st-order Information, Knowledge ...)

- 주어진 문맥에서 표층적, 명시적으로 드러난 관계, 주로 A-to-B, A-is-B, A-verb-B, ...

- 고차 정보 (Higher-order Information)

- 순서 역전된 객체간 관계 : 교환법칙이 성립할 수 없는(다른 Relation을 가져와야 할) 경우도 포함
- 문맥 밖에서 가져와야 하는 1차 정보의 집합

1st-order Information: (Book, "Cost More", Pencil), (Book, "Cost Less", Computer)



단순히 방향을 뒤집는 것으로는 관계 성립이 불가. Computer-Cost Less (X)-Book

High-order Information:

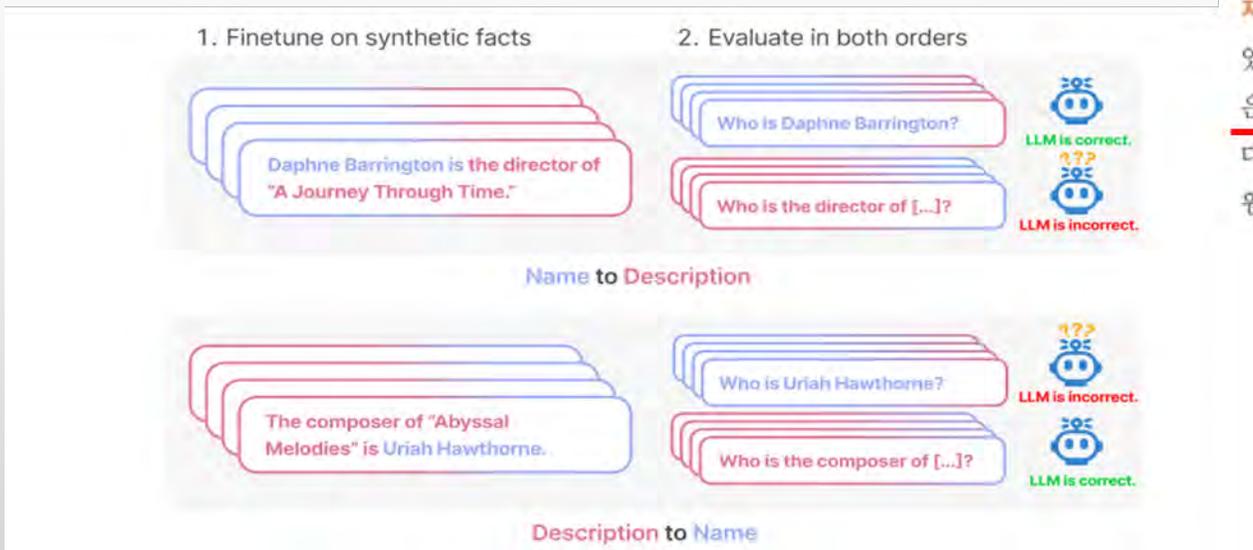
1. (Pencil, "Cost **Less**", Book), (Computer, "Cost **More**", Book)
2. (Cost, "means", vendible), (Book, "is", sellable), (Pencil, "is", buyable), (Computer, "is", Product), (Book, "Cost More", Eraser), (Computer, "Cost More", Desk), ...

* 예시 출처: Zhou et al., 2022.

데이터 증강: 고차정보 학습

- Berglund et al., "The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". 2023. (ICLR 2024)
 - A-is-B, B-is-A 방향으로 각각 구성된 데이터를 Fine-tuning 했을 때, 학습에 사용된 방향은 잘 추론했지만 반대방향은 0%에 가까운 성능을 보임 (GPT-3, LLaMA 1 모두에서)
 - 우리가 문제를 내고, 풀기 위해서는 "A is B"를 학습해 "B is A"라고 답할 수 있어야 함

*금성사,중학사회1교과서에서발췌



지도의 종류 지도는 사용 목적에 따라 크게 **일반도**와 **주제도**로 구분할 수 있다. 일반도는 지형, 토지 이용, 마을, 도로 등 여러 가지 지리 현상을 담은 지도이며, 주제는 특정한 지리 현상만을 선택적으로 표현한 지도이다. 일반도가 다양한 목적으로 활용되는 데 비해, 주제는 주로 사용자가 원하는 특정 지리 현상의 분포와 관계 등을 파악하는 데 활용된다.



문제: 토지이용, 도로 등의 지리현상을 담은 지도는 무엇인가?
→ 일반도

데이터 증강: 고차정보 학습

- Golovneva et al., “Reverse Training to Nurse the Reversal Curse”, 2024. (FAIR@Meta; arXiv: 2403.13799)
 - R-to-L로 뒤집은 데이터를 일종의 다른 언어와 같이 함께 학습하는 방법으로 완화효과를 보임

Reverse training consists of taking a training dataset with N samples $\{x_1, \dots, x_N\}$ and constructing the set of reversed samples

$$\bar{x}_i = \text{REVERSE}(x_i), \quad i = 1, \dots, N.$$

Training is then conducted using the combined set $\{x_i\} \cup \{\bar{x}_i\}$ of $2N$ training samples, using the typical language modeling objective. The function $\text{REVERSE}(\cdot)$ reverses the given string, where we consider various choices of reversal type:

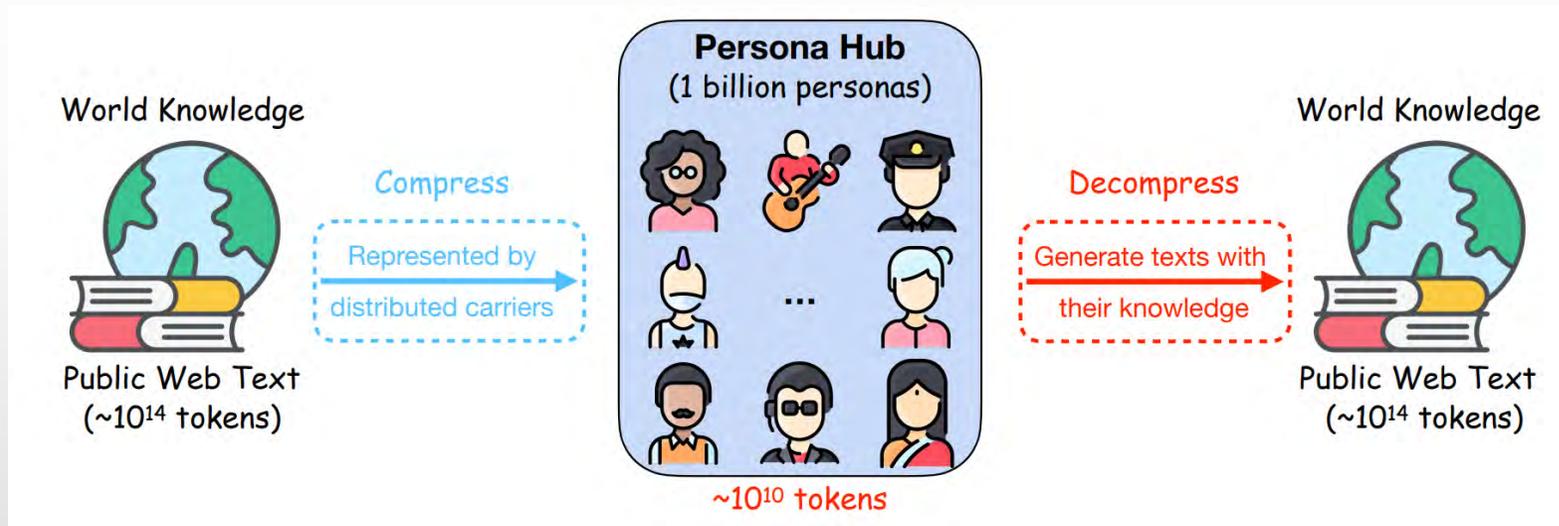
- **Token reversal** ($\text{REVERSE}_{\text{token}}$): A given input x_i , when tokenized, e.g. using BPE (Sennrich et al., 2015), consists of tokens x_i^t , and the reversed version has the form $\bar{x}_i^t = x_i^{|x_i| - t + 1}$.
- **Word reversal** ($\text{REVERSE}_{\text{word}}$): Each example is first split into words.² We then reverse the string at the word level, joining it back together with spaces. Note that this input would then typically be tokenized for input into the LLM, e.g. using BPE.
- **Entity-preserving reversal** ($\text{REVERSE}_{\text{entity}}$): We run an entity detector over a given training sample³, which also splits the non-entities into words. We then reverse the words, but keep the word-order of entities in their original left-to-right order. The string is then joined as before with spaces. See Table 2 for an example.
- **Random segment reversal** ($\text{REVERSE}_{\text{rand}}$): Instead of running a relatively costly segmentation such as an entity detector, we experiment with randomly segmenting the sequence into chunks of size between 1 and k tokens using uniform sampling. We then reverse the segments, but keep the word order within each segment in their original left-to-right order. The segments are then joined with a special token “[REV]”, which indicates the end of left-to-right prediction for the given segment. During training epochs, each time the example is seen we perform a different random segmentation to increase diversity. See Table 2 (last row) for an example.

Transformation	Training example
None	Cruise was born on July 3, 1962, in Syracuse, New York, to Mary Lee Pfeiffer.
Word reversal	. Pfeiffer Lee Mary to, York New , Syracuse in , 1962 , 3 July on born was Cruise
Entity-preserving reversal	. Mary Lee Pfeiffer to, <u>Syracuse, New York</u> in , 1962 , 3 July on born was <u>Cruise</u>
Random segment reversal	[REV] York, to Mary Lee Pfeiffer . [REV] in Syracuse, New [REV] on July 3, 1962, [REV] born [REV] Cruise was

Pre-training method	Finetuning method	NameToDescription		DescriptionToName	
		forward	reverse	forward	reverse
<i>Model size: 1.4B</i>					
standard (compute-matched)	standard	77.3	0.0	98.3	2.3
standard (compute-matched)	reverse (entity)	78.3	85.0	99.0	5.7
standard (compute-matched)	reverse (rand k=25)	77.3	96.3	97.7	70.7
standard (data-matched)	standard	75.0	0.0	99.3	0.0
standard (data-matched)	reverse (entity)	75.0	66.7	99.3	3.3
standard (data-matched)	reverse (rand k=25)	76.3	94.3	95.7	67.0
reverse training (entity*)	reverse (entity)	77.0	78.3	95.3	2.3
<i>Model size: 7B</i>					
standard	standard	80.3	0.0	96.0	4.0
standard	reverse (entity)	79.0	89.7	99.7	6.0
standard	reverse (rand k=25)	78.3	99.0	99.0	70.0

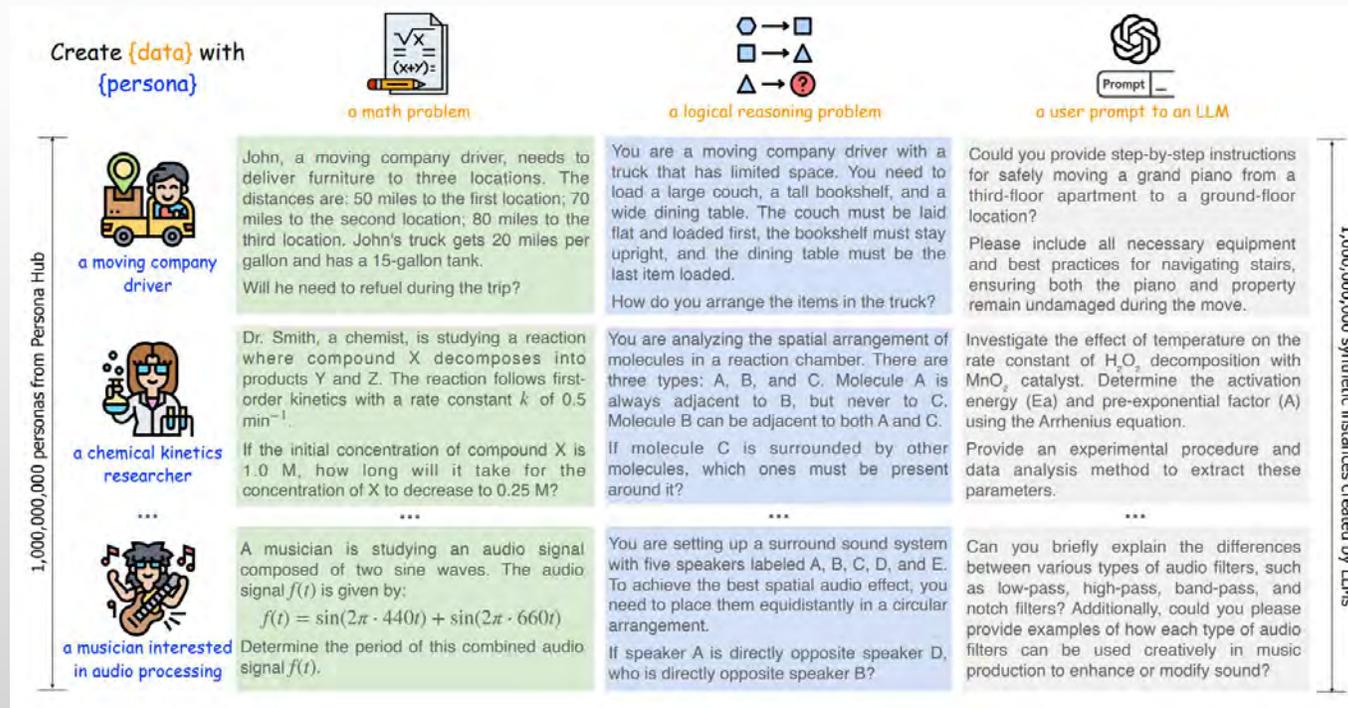
데이터 증강: 10억개의 Persona

- Xin Chan et al., "Scaling Synthetic Data Creation with 1,000,000,000 Personas", 2024. (Tencent AI Lab Seattle)
 - 합성 데이터의 다양성을 위해서 약 10억개의 페르소나를 통해서 데이터 생성



데이터 증강: 10억개의 Persona

- Xin Chan et al., "Scaling Synthetic Data Creation with 1,000,000,000 Personas", 2024. (Tencent AI Lab Seattle)
 - 다양한 시각 제공 : 특정 시각과 배경 지식을 반영하여 데이터 생성
 - 데이터 다양성 보장, 대규모 데이터 생성 촉진
 - 실제 시뮬레이션 : 기업이나 정부에서 새로운 제품 출시나 법안에 대한 대중의 반응을 예측



- 생성형 언어모델 원리 소개
- 언어모델과 학습 데이터 소개
- 학습 데이터 부족과 한국어 학습 데이터의 역할 필요성
- 인공지능 연구(개발)자와 한국어 전문가의 협업 필요
- 활발한 교류 및 상호 이해 필요



☞ > 행사 마당 > 한글 및 한국어정보처리 학술대회(HCLT)

접근 정보(Bibliometrics)		
출간 연도 (Publication Years) 1989 - 2023	출간 수 (Publication Counts) 2,231	내려 받기 가능 수 (Available for Download) 376
인용 수 (Citation Count) 2,041	내려 받기(누적 치) Downloads(cumulative) 74,845	내려 받기(지난 6주 간) Downloads(6 weeks) 622
내려 받기(지난 1년 간) Downloads(12 months) 13,784	기사 당 평균 인용 수 Average Citation per Article 0.012	기사 당 평균 내려 받기 수 (Average Downloads per Article) 0.271

1989년부터 2023년(35년)까지, 출간된 논문 총 2,247편의 학술 논문 제목과 저자를 검색할 수 있습니다.

Thank you! 😊



Effect of Speed and Breakdown Features on Pronunciation Scoring

- Learning from AI-Powered L2 English Speaking Corpus -

윤태진(성신여자대학교)

차 례

1. Introduction
 2. Methods
 3. Analysis
 4. Discussions
 5. Conclusion
-

1. Introduction

In the growing field of English language learning, AI's role is increasingly pivotal, especially in interpreting, translating, and enhancing core communication skills such as speaking and writing. Empirical evidence shows that using expert-labeled data in supervised learning significantly enhances AI performance (Mitchell, 1997, Yang and Yu, 2020), particularly in pronunciation and speaking education for English learners (Settles et al., 2020). Proficiency scoring accuracy by AI can optimize educational resources, allowing teachers to focus on other critical aspects of language development.

Fluency, a key component of proficiency, encompasses measures such as speech rate, number of pauses, and repair mechanisms (Huensch and Tracy-Ventura, 2017; Saito et al., 2018). Research interest in fluency has surged since the 1970s, highlighting its measurable quality as a key component of communicative competence (Bowker et al. 2013; Tavakoli and Wright, 2020). Fluency measures are strong indicators of overall proficiency in second language learners (Baker-Smemoe et al., 2014).

To establish an AI-based automated evaluation and feedback system, it is necessary to build reliable evaluation data by having professional raters, who have received extensive training, create highly reliable evaluation data for prepared speaking test data. Generally, the reliability of raters' evaluation data is measured by having two or more independent raters evaluate the same utterance based on a rubric independently and then testing the reliability among the raters.

This paper aims to examine how fluency attributes, such as speech rate and average pause duration, affect the fluency scale of utterances rated by professional raters of proficiency in English produced by varying degrees of English learners.

2. Method

2.1. Materials

A corpus of 400 hours of L2 English speech data from 882 Korean learners was used. This data, funded by the Korean government, aims to build an AI-assisted speech communication system. The data encompasses a broad demographic with participants spanning various age groups and proficiency levels.

2.2 Procedures

2.2.1. Score Rubrics

A methodology consolidating proven evaluation metrics focused on "pronunciation accuracy" was used. Assessors utilized a custom evaluation interface to assign scores ranging from 1 to 5 based on pronunciation accuracy. A dual-assessor system ensured thorough evaluations, with significant scoring differences reconciled by a third expert adjudicator. Krippendorff's alpha was adopted as a reliability measure.

2.2.2 Feature Extraction

Pronunciation accuracy scores were synthesized with acoustic features related to speech speed and pause durations. Articulation rate and silence mean (silmean) were quantitatively defined and extracted from the audio signals.

3. Analysis

Descriptive statistical analysis and visual representation of the data were conducted to gain insight into the relationship between phonetic features and pronunciation clarity.

3.1 Modeling of Ordered Pronunciation Scoring

Ordered logistic regression, specifically Cumulative Link Mixed Models (CLMM; Christensen, 2018), was applied to analyze how script level, gender, articulation rate, and silence mean contribute to articulation clarity scores. This method is apt for data where the response variable is ordinal, such as rating scales or Likert-type items.

Upon selecting the Cumulative Link Mixed Models (CLMM) for our statistical analysis, we devised two formulas to model the articulation scores as a function of script level, gender, and key phonetic features while accounting for random effects due to variability in recorders:

- (1) formula: $\text{articulationScore} \sim \text{scriptLevel} + \text{gender} + \text{ar} + (1 | \text{RecorderID})$
- (2) formula: $\text{articulationScore} \sim \text{scriptLevel} + \text{gender} + \text{silmean} + (1 | \text{RecorderID})$

The first formula posits that the articulation score, our dependent variable, is predicted by the script level, gender of the speaker, and articulation rate (ar). The term $(1 | \text{RecorderID})$ incorporates random effects for individual recorders, acknowledging that there may be recorder-specific variations in scoring that are not explained by our fixed effects. This model enables us to examine the influence of the script's complexity (scriptLevel), the speaker's gender, and the speed at which they articulate (ar), while also considering the potential biases or unique tendencies of individual recorders. The second formula is similar to the first but replaces articulation rate (ar) with silence mean (silmean), which measures the average duration of silent pauses. Here, the focus is on understanding how the length of pauses in speech (silmean) correlates with pronunciation clarity scores, alongside script difficulty and gender, while still accounting for random effects by individual recorders.

4. Results

4.1 Random Effects for Articulation Rate and Silence Mean

Random effects associated with RecorderID influence the articulation score, highlighting the variability among different recorders.

4.2 Checking Normality of the Random Effect RecorderID

The Kolmogorov-Smirnov (K-S) test was used to verify the normality of the random

effects for RecorderID, supporting the model's assumptions.

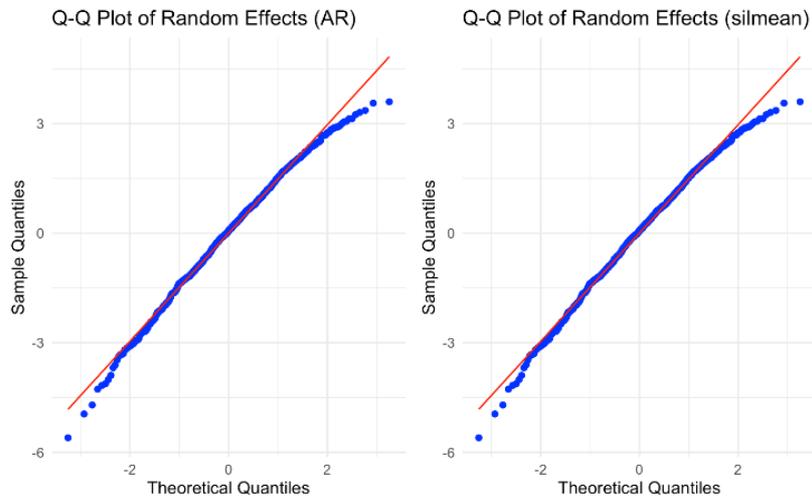


Figure 1. Q-Q plots assessing normality of random effects for Articulation Rate and Silence Mean

4.3 Fixed Effects

4.3.1 Fixed Effects for Articulation Rate

As in Table 1, Script level and gender significantly impacted articulation clarity scores. Higher script levels and faster articulation rates positively correlated with higher scores, while being male was associated with lower scores.

Table 1. Fixed Effects Coefficients for Predictors of Articulation Score in CLMM analysis with the focus on Articulation Rate for phonetic predictor

	Estimate	Std. Error	z value	Pr(> z)
scriptLevelMid	0.846964	0.106042	7.987	1.38e-15 ***
scriptLevelHigh	2.512244	0.138812	18.098	< 2e-16 ***
genderM	-0.487765	0.100573	-4.850	1.24e-06 ***
ar	0.095248	0.003553	26.805	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.3.2 Fixed Effects for Silence Mean

Higher script levels correlated with improved scores, while longer silence durations negatively impacted scores, as in Table 2.. Male gender was associated with lower scores compared to females.

Table 2. Fixed Effects Coefficients for Predictors of Articulation Score in CLMM analysis with the focus on silence mean for phonetic predictor

	Estimate	Std. Error	z value	Pr(> z)
scriptLevelMid	0.90282	0.11253	8.023	1.03e-15***
scriptLevelHigh	2.62544	0.14762	17.785	< 2e-16 ***
genderM	-0.47493	0.10729	-4.427	9.58e-06 ***
silmean	-1.10193	0.03312	-33.270	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.4 Threshold Coefficients for Articulation Rate and Silence Mean

Threshold coefficients in the ordinal regression model demarcate the points at which the likelihood of progressing to a higher articulation score category increases. The threshold coefficients are summarized in Table 3 for Articulation Rate and Table 4 for Silence Mean. The figures 2 and 3 are visual presentation of the threshold coefficients with different values of Articulation Rate and Silence Mean, respectively.

Table 3 Threshold coefficients for Articulation Rate:

	Estimate	Std. Error	z value
1 2	-3.83660	0.09380	-40.90
2 3	-1.08709	0.09087	-11.96
3 4	1.54922	0.09085	17.05
4 5	4.22795	0.09162	46.15

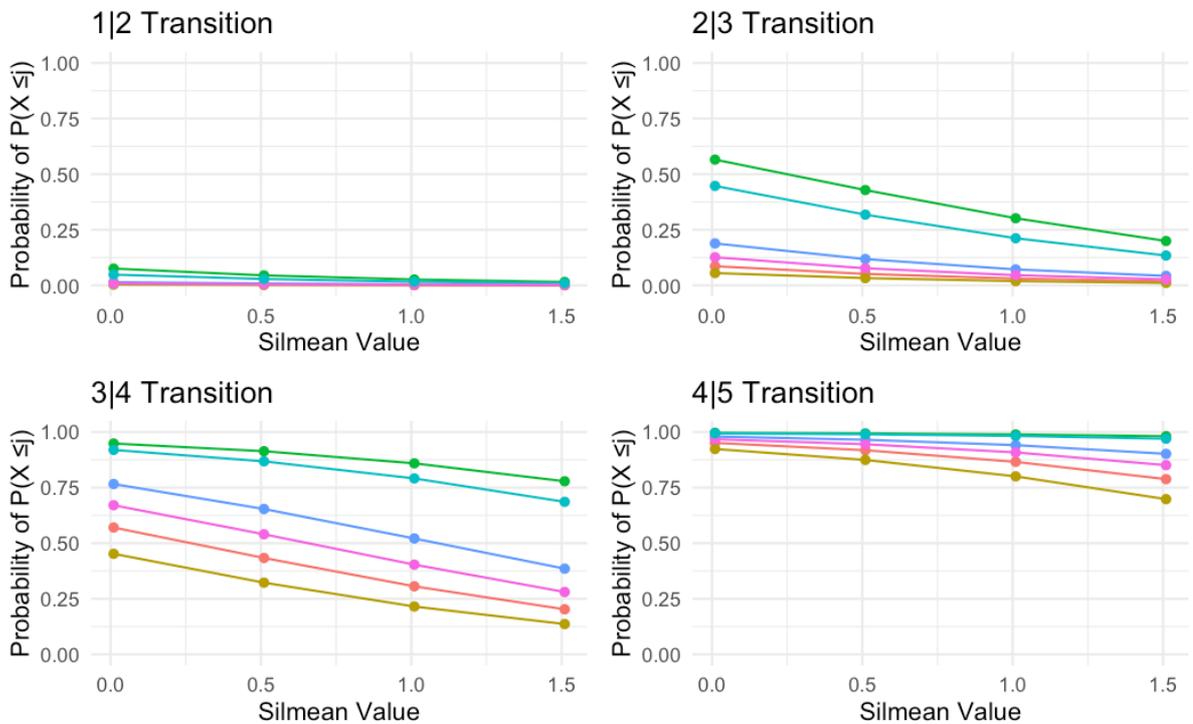
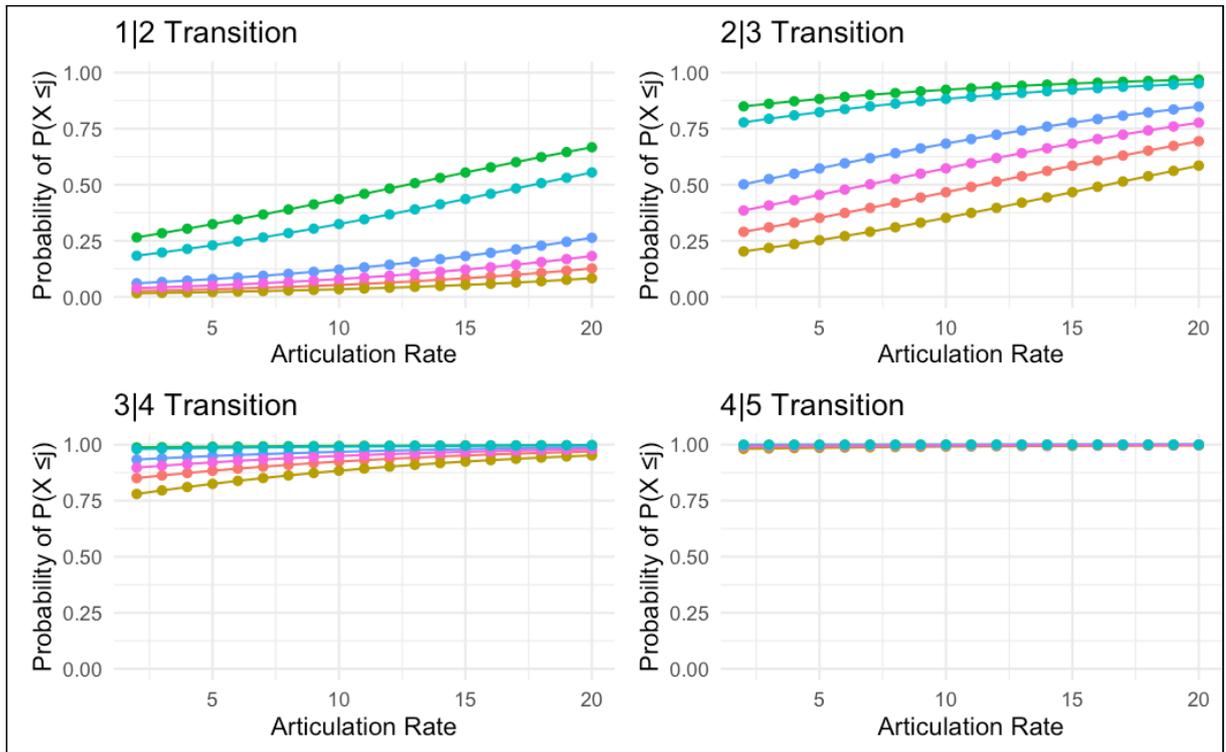
Figure 2. Probability Distributions of Articulation Score Transitions Across Script Levels, Gender, and Articulation Rate

Table 4. Threshold coefficients for silence mean:

	Estimate	Std. Error	z value
1 2	-5.12020	0.09180	-55.777
2 3	-2.35172	0.08841	-26.600
3 4	0.29464	0.08806	3.346
4 5	2.97682	0.08858	33.605

Figure 3. Probability Distributions of Articulation Score Transitions Across Script Levels, Gender, and silence mean

5. Discussions



5.1 Articulation Rate (for speed)

Various linguistic and temporal features of speech are significant indicators of language proficiency and fluency Tomaschek et al. (2018). Articulation rate, while an indicator of fluency, must be contextualized within a broader linguistic and cognitive

framework (cf. Duran-Karaoz and Tavakoli (2020)).

5.2 Silence Mean (for speed)

The silence mean variable significantly influences perceived fluency. Longer pauses correspond to less fluent articulation, emphasizing the importance of mastering strategic silence use for effective communication.

5.3 Gender

Gender-based variations in articulation scores were observed, with female learners displaying higher articulation scores (Samuelsson, 2006; Byrd, 1992; Whiteside, 1995). This finding invites further investigation into gender differences in language acquisition and proficiency.

5.4 Script Level

Higher script levels are associated with improved pronunciation accuracy, suggesting that script difficulty may encourage more nuanced and precise articulation (Carver, 1998). This finding aligns with educational theories positing that cognitive demand drives deeper linguistic processing and performance Loukina et al. (2015).

6. Conclusion

The study illuminates the multifaceted nature of fluency as a construct within second language acquisition. Through the use of CLMM, the nuanced influence of script level, gender, articulation rate, and silence mean on articulation clarity scores was quantified. These findings contribute to the understanding of L2 fluency and provide insights for improving teaching methods and assessment criteria.

The study identifies promising directions for further research, including exploring monologic versus dialogic fluency and the qualitative implications of pause location within utterances. By considering both the form and substance of fluency, the study enriches the discourse on pedagogical and assessment practices in SLA.

(Selected) References

- Baker-Smemoe W, Dewey DP, Bown J and Martinsen RA (2014) Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals* 47(4): 707-728.
- Bosker HR, Pinget A-F, Quené H, et al. (2013) What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing* 30(2):

159-175.

- Byrd D (1992) Preliminary results on speaker-dependent variation in the TIMIT database. *The Journal of the Acoustical Society of America* 92(1): 593-596.
- Carver RP (1998) Predicting reading level in grades 1 to 6 from listening level and decoding level: Testing theory relevant to the simple view of reading. *Reading and Writing* 10: 121-154.
- Christensen RHB (2018) Cumulative Link Models for Ordinal Regression with the R Package Ordinal.
- Huensch A and Tracy-Ventura N (2017) Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics* 38(4): 755-785.
- Loukina A, Lopez M, Evanini K, et al. (2015) Pronunciation accuracy and intelligibility of non-native speech. Sixteenth Annual Conference of the International Speech Communication Association.
- Mitchell TM (1997) *Machine Learning*. The McGraw-Hill Companies Inc.
- Saito K, Ilkan M, Magne V, et al. (2018) Acoustic characteristics and learner profiles of low-, mid-and high-level second language fluency. *Applied Psycholinguistics* 39(3): 593-617.
- Samuelsson Y (2006) Gender effects on phonetic variation and speaking styles: a literature study. Reportno. Report Number[, Date. Place Published]: Institution].
- Settles B, T. LaFlair G and Hagiwara M (2020) Machine Learning-Driven Language Assessment. *Transactions of the Association for Computational Linguistics* 8: 247-263.
- Tavakoli P and Wright C (2020) *Second language speech fluency: From research to practice*. Cambridge University Press.
- Whiteside SP (1995) Temporal-based speaker sex differences in read speech: A sociophonetic approach. *Proceedings of the XIIIth International Congress of Phonetic Sciences (ICPhS 95)*. 516-519.
- Yang S-B and Yu T-L (2020) Pseudo-representation labeling semi-supervised learning. arXiv preprint arXiv:2006.00429. DOI: <https://doi.org/10.48550/arXiv.2006.00429>.