JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 5, pp. 1339-1348, May. 2025

사전 학습 언어모델(BERT)의 자동사에 대한 내적표현 탐색

김 자 경[1] · 김 성 은[1] · 오 동 석[2*]
[1]경북대학교 영어영문학과 박사
[2]경북대학교 영어영문학과 조교수

Probing the Internal Representation of Intransitive Verbs in BERT

Ja-Kyoung Kim[1] · Sung-Eun Kim[1] · Dong-Suk Oh[2*]
[1]Ph.D. Department of English Language and Literature, Kyungpook University, Daegu 41566, Korea
[2]Assistant Professor, Department of English Language and Literature, Kyungpook University, Daegu 41566, Korea

[요 약]

본 연구는 BERT 언어 모델이 자동사의 의미 구조와 주어의 유생성 간의 관계를 어떻게 표현하는지 프로빙 기법을 통해 분석하였다. 이를 위해 자동사를 의미적 핵심성에 따라 Core, Less Core, Periphery로 나누고, 각 자동사 유형이 생물 및 무생물 주어와 결합할 때 나타나는 내부 표현의 차이를 탐색하였다. BERT 모델의 토큰 기반 및 문장 기반 임베딩을 이용해 동사 유형 간 분류 성능을 분석한 결과, 모델의 중간층(layer 6~9)에서 가장 높은 정확도와 명확한 구분 능력이 나타났다. 또한 자동사 유형은 모델 내부에서 비선형적으로 표현되는 것으로 확인되었다. 이러한 결과는 BERT와 같은 사전 학습 언어모델(Pre-trained Language Models, PLMs)이 인간의 언어적 직관과 유사하게 자동사의 의미 구조와 주어의 유생성을 효과적으로 인코딩하고 있음을 보여준다. 본 연구는 향후 인간의 언어 처리와 인지적 타당성을 반영한 언어 모델 설계를 위한 이론적 토대를 제공할 것이다.

[Abstract]

This study investigates how the BERT language model captures the relationship between the semantic structure of intransitive verbs and the animacy of their subjects through probing techniques. To achieve this, intransitive verbs were categorized into Core, Less Core, and Periphery based on their semantic centrality. The internal representations were then analyzed when each verb type was combined with wuther animate or inanimate subjects. By examining classification performance across verb types using both token-based and sentence-based embeddings from BERT, the results showed the highest accuracy and clearest distinctions in the model's intermediate layers (layers 6 to 9). Furthermore, the findings indicate that BERT's internal representations of intransitive verb types exhibit a non-linear structure. These results demonstrate that pre-trained language models (PLMs) like BERT effectively encode the semantic structure of intransitive verbs and subject animacy in a way that aligns with human linguistic intuition. This study provides a theoretical foundation for the future development of language models that incorporate human language processing and cognitive plausibility.

색인어 : 자동사, 주어의 유생성, 프로빙, 사전 학습 언어모델, 내적 표현
Keyword : Intransitive Verbs, Subject Animacy, Probing, BERT, Internal Representation

http://dx.doi.org/10.9728/dcs.2025.26.5.1339

Received 24 March 2025; **Revised** 14 April 2025
Accepted 24 April 2025

***Corresponding Author; Dong-Suk Oh**
Tel: +82-53-950-7155
E-mail: inow3555@knu.ac.kr

Ⅰ. 서 론

영어의 동사는 일반적으로 자동사(intransitive verbs)와 타동사(transitive verbs)로 나누어진다. 논항 구조(argument structure)에 따라 목적어를 요구하는 동사는 타동사로, 목적어 없이 주어만을 필요로 하는 동사는 자동사로 분류된다. 이러한 자동사는 비능격 동사(unergative)와 비대격 동사(unaccusative)로 세분화하는 비대격 가설(Unaccusative Hypothesis)이 제안되었다[1]. 비능격 동사는 주어가 행위자(agent)의 의미역을 가지며 자발적인 행위를 나타내는 반면(walk, run, laugh 등), 비대격 동사는 대상(patient 또는 theme)을 주어로 가지며 주어가 사건이나 상태의 변화 영향을 받는 것을 나타낸다(fall, arrive, break 등). 이러한 의미역의 차이는 주어의 유생성(animacy)과도 관계가 깊다. 예를 들어 비능격 동사는 (1a)처럼 주어가 생물(animate)일 때 자연스러운 해석을 보이며, 비대격 동사는 (1b)처럼 주어가 무생물(inanimate)일 때 문장 처리가 더 용이해지는 경향이 있다[2]. 이는 비대격 동사가 본질적으로 대상의 의미역을 가지는 의미적 특성 때문이며, 생물 주어가 행위자(agent)로 해석되려는 인간의 인지적 경향과 충돌하기 때문이다. 따라서 자동사의 유형별로 주어의 유생성(animacy)에 따른 문장 처리 난이도와 인식이 달라질 수 있다[3].

1. (a) 비능격 동사 : Tom laughed.
   (b) 비대격 동사 : The window broke.

1-1 연구의 필요성

최근 자연어 처리(Natural Language Processing, NLP) 분야에서는 트랜스포머(transformer) 기반 모델이 다양한 언어 과제에서 높은 성능을 보이고 있으며, 그중 BERT와 같은 사전학습 언어 모델(Pre-trained Language Models, PLMs)이 다방면으로 사용되고 있다[4]. 한편, 인간의 언어적 직관과 유사한 능력을 보여주는 대형 언어 모델(Large Language Models, LLMs)도 등장하여 언어 현상의 인지적 타당성을 탐구하는 데 중요한 역할을 하고 있다[5]. 이와 같은 모델은 다양한 언어 과제에서 뛰어난 성능을 보이며 인간의 언어 처리 과정과 유사한 능력을 보여주고 있으나, 모델이 특정 언어적 현상이나 개념을 실제로 얼마나 정확히 이해하고 표현하는지에 대해서는 여전히 명확히 밝혀진 바가 없다. 그 중에서도 특히 주어의 유생성(subject animacy)과 자동사의 의미적 특성(semantic properties of intransitive verbs)을 모델이 얼마나 효과적으로 학습하고 있는지에 관한 체계적인 연구는 아직 부족한 실정이다. 이는 인간이 특정 동사의 논항 구조(argument structure)와 주어의 유생성 사이의 의미적 관계를 직관적으로 인지하고 해석하는 것과 모델의 능력을 직접적으로 비교 평가할 필요가 있음을 시사한다.

최근 NLP 연구에서는 탐색 기법(probing methods)을 통해 언어 모델이 구문적 및 의미적 특성을 어떻게 표현하고 있는지 분석하는 시도가 활발히 이루어지고 있다. 탐색 기법이란 모델의 내부 표현(representations)에 숨어 있는 특정 언어적 정보를 평가하기 위해 고안된 방법론으로, 일반적으로 BERT와 같은 모델의 임베딩(embeddings)을 입력으로 받아 품사(parts of speech), 단어 의미(word meanings) 등과 같은 세부적인 언어적 정보를 예측하는 탐색 모델(probe)을 별도로 훈련하여 분석한다[4],[6]. 이러한 탐색 모델은 다양한 언어 현상과 과제에 대해 높은 정확도를 보이며, 이를 통해 기존의 대형 언어 모델이 구문적 및 의미적 정보를 내부적으로 어떻게 인코딩하고 있는지 구체적으로 평가할 수 있음을 보여준다[5].

한편, 탐색 기법의 또 다른 형태인 검색 분류기(search classifiers)는 모델의 출력 임베딩에 내재된 활성화 값을 학습하여 특정 언어적 범주나 특성(예: 구문적 태그나 의미적 범주)을 보다 정밀하게 식별하고 분석하는 데 사용된다[7],[8]. 최근 들어서는 층별(layer-wise) 분석 기법이 도입되어 transformer 기반 모델의 각 층이 언어적 정보를 어떻게 분산하여 표현하고 있으며, 이러한 정보가 어떤 위계적 구조(hierarchical structure)를 형성하고 있는지에 대한 깊이 있는 통찰을 제공하고 있다[4],[9]-[11]. 이러한 층별 분석 방법론은 모델 내부의 언어적 지식 분포와 학습된 표현의 인지적 타당성(cognitive plausibility)을 평가하는 데 중요한 근거로 작용할 수 있다.

또한, 언어 모델이 인간의 언어적 직관을 얼마나 정확하게 반영하는지 분석하는 것은 실제 디지털 콘텐츠 서비스 측면에서도 매우 중요한 의미를 지닌다. 최근 챗봇(chatbot), 가상 비서(virtual assistant), 대화형 추천 시스템(interactive recommender system) 등 다양한 서비스가 NLP 모델을 활용하여 개인화되고 정교한 사용자 경험을 제공하고 있다. 그러나 만약 이러한 언어 모델이 주어의 유생성이나 자동사의 의미와 같은 미묘한 언어적 특징을 충분히 학습하지 못한다면, 사용자가 인지하는 서비스의 자연스러움과 신뢰성은 크게 저하될 수 있다. 예컨대 챗봇 서비스에서 사용자와 대화 시 문맥에 맞지 않는 표현이나 부자연스러운 문장 구조가 반복될 경우, 사용자 경험과 만족도는 급격히 감소할 수 있다. 따라서 본 연구가 제안하는 언어학적 이해도 분석을 통해 모델의 인지적 타당성을 높이는 것은 궁극적으로 디지털 콘텐츠 서비스의 품질 향상과 사용자 친화적 경험을 실질적으로 구현하는 데 필수적인 과제라 할 수 있다.

따라서 본 연구는 탐색 기법을 활용하여 BERT와 같은 transformer 기반의 사전학습 언어 모델(PLMs)이 주어의 유생성과 자동사의 의미적 특성을 실제 인간과 유사한 수준으로 이해하고 표현하는지 분석하고자 한다. 이를 통해 사전학습 언어 모델이 언어적 직관과 인지적 타당성을 어느 정도 반영하고 있는지 보다 정밀하게 평가하고자 한다.

1-2 연구의 목적

본 연구는 Sorace의 분리 자동사 위계(Split Intransitivity Hierarchy, SIH)[12]를 이론적 기반으로 설정하고, Core, Less Core, Periphery 등 자동사의 유형에 따라 주어의 유생성(subject animacy)이 transformer 기반 언어 모델의 문장 처리에 미치는 영향을 체계적으로 분석하는 데 목적이 있다. 구체적으로, BERT 모델이 각 자동사 유형과 결합된 주어의 유생성을 실제 인간과 유사한 방식으로 처리하고 표현하고 있는지 탐색(probing) 기법을 통해 평가하고자 한다. 이를 위해 본 연구는 모델 내부 층(layer)별로 주어의 유생성과 자동사의 의미적 특성이 어떻게 구체적으로 인코딩되어 있으며, 이러한 정보가 층간 위계적으로 어떻게 분포하고 있는지를 분석하여, 인간의 문장 처리 과정과 비교한 유사점 및 차이점을 규명한다. 최종적으로는 본 연구의 결과를 바탕으로 인간의 언어적 직관을 보다 정확히 반영하는 차세대 언어 모델 개발을 위한 구체적이고 실증적인 이론적 기반을 마련하는 데 연구의 목적이 있다.

이를 구체적으로 실현하기 위하여 다음과 같은 연구 질문을 설정하였다.

RQ1. 인간의 문장 처리 과정에서 주어의 유생성(subject animacy)이 자동사의 유형(Core, Less Core, Periphery)에 따라 어떻게 다르게 영향을 미치는가?

RQ2. BERT 모델의 내부 층(layer)별로 자동사의 유형에 따른 주어의 유생성 정보가 인코딩되는 양상은 어떻게 달라지는가?

이를 통해 본 연구는 자동사의 의미 구조와 주어의 유생성 간의 관계를 명확히 규명하고, 인간과 언어 모델 간의 문장 처리 방식의 유사성과 차이를 체계적으로 분석하여, 향후 언어 모델이 인간의 언어적 직관을 더욱 효과적으로 반영할 수 있는 방법에 대한 논의를 확장하고자 한다.

II. 이론적 배경

2-1 자동사의 분류와 의미적 접근

Sorace는 자동사의 의미적 특성을 더욱 체계적으로 설명하기 위해 분리 자동사 위계(Split Intransitivity Hierarchy, SIH)[12]를 제안하였다. 이 위계는 자동사를 행위성(agentivity)과 목적성(telicity)이라는 두 가지 의미적 기준에 따라 Core, Less Core, Periphery의 세 가지 하위 범주로 나눈다. 행위성은 주어가 얼마나 자발적이고 능동적인 행위자인지를 나타내고, 목적성은 동사가 나타내는 사건이 명확한 종결 상태를 지니는지의 여부를 나타낸다. 예를 들어, Core 동사(arrive, disappear 등)는 비행위적이면서 목적성이 강한 사건을 나타내고, Less Core 동사(sink, grow 등)는 목적성이나 행위성이 상대적으로 불분명한 중간적 특성을 가진다. 반면 Periphery 동사(exist, remain 등)는 상태를 나타내며 목적성과 행위성이 모두 낮다. Sorace의 연구에 따르면, 자동사의 하위 범주는 제2언어 학습자의 언어 습득 과정에서 나타나는 오류 빈도, 특히 과수동화(overpassivization) 오류와도 밀접한 관련이 있다[13],[14].

표 1. 분리 자동사 위계
Table 1. Split intransitivity hierarchy

Change of location	(arrive, leave)	Core	Unaccusative
[Directed motion]	(rise, ascend, descend, advance)		
Change of state	(wilt, become, bloom, decay, rot)	Less Core	
Appearance	(appear, disappear, happen, arise)		
Continuation of preexisting condition	(stay, remain, suffer, continue)	Periphery	
Existence	(exist, suffice, vanish, be, belong, seem, please)		(more variation)
Uncontrolled process			
[Emission]	(rattle)		
[Involuntary reaction]	(tremble)	Periphery	
Controlled motional process	(swim, run)	Less Core	
Controlled nonmotional process	(work, thrive, talk)	Core	Unergative

한편, 무엇이 살아있는지 여부는 인간 인지에서 중심적인 역할을 한다. 이러한 생물성(animacy)의 특별한 지위는 인간 언어에서도 반영된다. 예를 들어, 여러 언어에서 인간이나 동물과 같은 생물적 지시어(animate referents)는 무생물적 대상보다 문장에서 더 앞서 언급되는 경향이 있다[15],[16]. 또한, 생물적 대상은 문장에서 보다 두드러진 문법적 역할을 차지하는데, 특히 문장의 주어 역할을 더 자주 맡는다. 예를 들어, 영어 화자들은 생물적 대상과 무생물적 대상이 포함된 문장을 기억할 때(예: The music soothed the child), 생물적 대상을 주어로 잘못 기억하는 경향이 있다. 이는 수동태 문장을 생성하는 결과로 이어질 수 있다(예: The child was soothed by the music)[17]. 또한, 화자들은 대상(patient)이 생물일 때 무생물일 때보다 더 자주 수동태 구문을 생성하는 경향이 있다[18]. 이러한 연구 결과들은 생물성이 언어 생성(language production)에서 핵심적인 역할을 한다는 점을 시사한다.

자동사의 해석 과정에서도 주어의 유생성은 중요한 역할을 한다. 같은 자동사라도 주어가 생물인지 무생물인지에 따라 문장의 의미적 자연성과 문법적 허용 가능성이 달라진다. 예컨대, "The man arrived."는 자연스러운 문장으로 받아들여지지만, "The rock arrived."는 문법적으로는 가능할지라도 의미적으로는 부자연스럽게 느껴진다. 이처럼 주어의 유생성은 문장의 의미적 자연성과 해석 방향성에 영향을 미치는 주요 요소이다[2]. 이처럼 주어의 유생성은 문장의 의미적 자연성뿐 아니라 문장의 의미적 해석 방향성에도 중요한 영향을 준다.

2-2 언어학 기반 프로빙(Linguistically-Informed Probing)

자연어 처리(NLP) 분야의 모델 중 하나인 BERT(Bidirectional Encoder Representations from Transformers)는 문장을 양방향(bidirectional)으로 읽어 단어와 문맥에 대한 정보를 심층적으로 학습하는 특징을 가진다. 전통적인 언어 모델들이 문장을 한 방향으로만 처리하여 단어의 의미를 고정적이고 제한적으로 이해했던 것과 달리, BERT는 Transformer 아키텍처의 Self-Attention 메커니즘을 통해 문장 내 단어 간의 관계와 맥락 정보를 더욱 정확하게 포착할 수 있다[19]. 이러한 학습 방법을 통해 BERT는 단어의 의미뿐 아니라 주변 맥락을 기반으로 문장 내 미세한 의미 차이를 인지하고, 복잡한 언어 현상을 학습할 수 있게 된다.

Chen과 Gao는 BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, DeBERTa 등 다양한 사전학습 언어 모델을 대상으로 프로빙(probing) 실험을 통해 논리적 추론(logical inference)에 필수적인 언어학적 정보의 인코딩 정도를 분석하였다. 연구 결과, 이들 모델이 어휘 의미 정렬(lexical semantic alignment)이나 모순 신호(contradiction signature) 같은 단순한 의미 정보는 효과적으로 인코딩하고 있지만, 단조성(monotonicity)과 같은 복합적인 언어 현상에 대해서는 상대적으로 낮은 성능을 보였다. 또한 이러한 복합적 정보는 추가적인 파인튜닝(fine-tuning)을 통해 성능 개선이 가능한 것으로 나타났다[20].

Katinskaia와 Yangarber는 BERT 및 RoBERTa와 같은 트랜스포머 기반의 사전 학습 언어 모델이 러시아어 동사의 상(verbal aspect) 범주를 내부적으로 어떻게 인코딩하는지 프로빙 기법을 통해 분석하였다. 연구는 특히 완결성(boundedness)에 따른 의미론적 특성이 동사 상의 선택에 어떻게 영향을 미치는지 조사하였다. 실험 결과, 이들 모델은 주로 최종 층에서 동사 상 정보를 효과적으로 인코딩하는 것으로 나타났으며, 의미적으로 완결된(bounded) 동작과 미완결된(unbounded) 동작 간의 표현이 상호 대비되는 방식으로 표현됨을 확인하였다. 이러한 결과는 언어 모델이 언어학적 이론과 일치하는 방식으로 복잡한 의미적 정보를 내부에 반영하고 있음을 시사한다[21].

Kissane, Schilling, and Krauss의 연구는 대형 언어 모델(LLMs)이 동사-구성 요소(verb-particle combinations)인 다단어 동사(multi-word verbs)를 내부적으로 어떻게 표현하는지를 분석하였다. 연구진은 BERT의 계층별 표현을 분석하여 구동사(phrasal verbs)와 전치사 동사(prepositional verbs) 간의 차이를 탐색하였으며, 이를 위해 프로빙(probing) 기법과 일반화 판별 값(Generalized Discrimination Value, GDV)을 활용하였다. 실험 결과, BERT의 중간 층에서 두 동사 유형의 가장 높은 분류 정확도가 나타났으며, 이는 언어 모델이 특정 계층에서 구문적 차이를 보다 명확하게 학습함을 시사한다. 하지만 GDV 분석에서는 선형적으로 구별되는 패턴이 뚜렷하지 않아, LLM의 내부 표현이 단순한 선형 구조가 아니라 비선형적으로 구성될 가능성이 있음을 보여주었다[22]. 본 연구는 이 선행연구 논문의 실험 방법을 바탕으로 실험을 진행하였다. 기존 연구[22]가 구동사와 전치사 결합형 표현(take off, give up 등)에 초점을 두었다면, 본 연구는 단일 자동사를 중심으로 Sorace(2000)의 의미적 위계(Core-Periphery)와 주어의 유생성이라는 복합 요인을 분석하였다. 또한 다양한 크기의 BERT 모델을 비교 분석함으로써, 모델 규모에 따른 표현 인코딩 특성을 종단적으로 검토하였다는 점에서 큰 차이를 보인다.

Ⅲ. 연구 방법

3-1 실험 데이터

본 연구의 실험에 사용된 동사는 논문[12]가 제안한 자동사의 위계(Core, Less Core, Periphery)에 따라 각 범주별로 대표성이 있는 동사 15개씩을 선정하여 총 45개의 자동사를 실험 대상으로 설정하였다. 각 동사는 논문[1]과 [12]에서 인용하였다. 실험을 위한 총 문장 수는 9000개로 설정하였다.

표 2. 실험에 사용된 단어
Table 2. The words used in the experiment

Core	Less core	Periphery
arrive	wilt	appear
leave	become	disappear
rise	bloom	happen
ascend	decay	arise
descend	rot	stay
advance	melt	remain
come	freeze	suffer
depart	evaporate	continue
enter	vaporize	exist
fall	solidify	suffice
drop	crystallize	vanish
sink	darken	be
float	yellow	belong
slide	germinate	seem
slip	sprout	please
Total 45		

각 동사별로 유생성 조건(animate/inanimate)에 대해 각각 100개의 문장을 생성하여, 동사당 총 200개의 문장을 만들었다. 문장은 ChatGPT-4에게 요청하여 완성하였으며, 생성된 문장에 대해 영어학 박사 2인이 문법성, 의미적 타당성, 자연스러움을 기준으로 수용성을 독립적으로 평가하였다. 평가자 간 일치도는 Cohen의 Kappa 계수(κ = 0.782)로 산출되었으며, 이는 상당한 수준의 일치도에 해당한다. 불일치한 7.6%의 문장에 대해서는 사후 논의 또는 제3의 평가를 통해 최종 결정을 내림으로써 신뢰도를 확보하였다. 그 결과 각 범주(15개 동사)는 3,000개의 문장(15개 동사 × 100개 × 2가지 주어 유형)을 가지게 되었으며, 이를 모두 합산하면 총 9,000개의 실험 문장이 구성되었다. 이는 논문[5]와 [9]와 같은 주요 probing 연구들에서 제시한 조건당 100개 내외의 표본 수 기준을 충족하는 수준으로, 이와 같이 구성된 자료를 바탕으로 안정적인 분류 성능 평가와 층별 표현 분석이 가능하였다. 또한, 전체 문장의 평균 길이는 약 6.68단어였다. 자동사의 유형별로 살펴보면, core 동사는 평균 6.84단어, less core 동사 역시 평균 6.84단어였으며, periphery 동사는 평균 6.37단어로 나타났다. 이처럼 세 범주 간 문장 길이의 차이가 크지 않아, 전체 데이터는 길이 면에서 균등하게 구성되었음을 확인할 수 있었다.

3-2 임베딩 추출

본 연구는 특징 추출 모델(feature extraction model)로 Transformer 기반 모델[23]인 BERT[19] 사용하여 맥락 임베딩(contextual embeddings)을 생성한다. 본 실험에서는 다양한 크기의 BERT 모델들이 사용되었으며, 구글에서 공식적으로 제공하는 'google/bert_uncased' 계열의 모델들을 기반으로 총 다섯 가지 모델이 선정되었다. 실험에 포함된 모델은 크기가 작은 순서부터 BERT tiny, BERT mini, BERT small, BERT medium, 그리고 표준 크기인 BERT-base이다. 이 모델들은 레이어(layer)의 수와 히든 차원(hidden dimension)의 크기에 따라 구체적으로 구분된다.

구체적으로, BERT tiny 모델은 2개의 레이어와 128개의 히든 차원을 가지며, BERT mini는 4개의 레이어와 256개의 히든 차원을 갖는다. 다음으로, BERT small 모델은 4개의 레이어와 512개의 히든 차원을 가지며, BERT medium 모델은 8개의 레이어와 512개의 히든 차원을 가진다. 마지막으로 표준 크기의 BERT-base 모델은 가장 크며, 12개의 레이어와 768개의 히든 차원을 가진다. 본 연구에서는 이와 같이 다양한 크기와 복잡성을 가진 모델들을 사용하여 모델 크기와 성능 간의 상관관계 및 각 모델이 언어적 현상을 표현하는 능력을 체계적으로 분석하고자 한다.

본 연구에서 사용된 BERT 모델들은 층(layer)의 수, 은닉 차원(hidden size), 어텐션 헤드(attention heads) 개수에 따라 성능이 다르게 나타난다. 작은 모델일수록 계산 비용이 낮고 빠르지만 표현 학습 능력이 제한적이며, 큰 모델일수록 더 정교한 맥락 정보를 학습할 수 있다. 특히, BERT Base 모델은 12개 층으로 구성되어 있어 가장 깊은 표현 학습이 가능하다. 각 실험 문장에 대해, 본 연구는 두 가지 수준에서 임베딩을 추출한다. 동사별 분석을 위해 우리는 동사에 해당하는 토큰 임베딩을 추출하며, 문장의 전체 맥락을 포착하기 위해 문장 내 모든 토큰 임베딩의 평균을 계산한다. 이때 일부 동사는 예를 들어 disappear처럼 subtoken 단위로 분리되는 경우가 있으므로, 해당 단어가 둘 이상의 subtoken으로 구성될 경우 각 subtoken의 임베딩을 평균하여 단일 임베딩으로 사용하였다. 이를 통해 토크나이즈 방식의 차이가 분석에 미치는 영향을 최소화하고, 모든 동사에 대해 일관된 방식으로 표현을 구성하였다.

3-3 분류 모델

본 연구에서는 로지스틱 회귀(Logistic Regression, LR)와 서포트 벡터 머신(Support Vector Machines, SVM)을 분류 모델로 사용하였다[24]. 로지스틱 회귀는 입력 변수에 대해 가능한 결과의 확률을 모델링하는 선형 분류 모델이며, 서포트 벡터 머신은 데이터의 클래스를 구분하는 초평면(hyperplane)을 찾는 지도학습 기반 모델이다. 두 모델 모두 실험 설정은 이전 논문[22]의 방법을 그대로 따랐다.

3-4 Generalized Discrimination Value(GDV)

본 연구에서는 데이터의 클래스 간 분리 정도를 평가하는 지표로 일반화 판별 값(Generalized Discrimination Value, GDV)을 사용하였다[25]. GDV는 각 클래스 내부의 데이터들이 얼마나 밀집되어 있고, 클래스 간에는 얼마나 명확히 구별되는지를 정량적으로 측정하는 방법으로, 다음과 같은 과정을 통해 계산된다.

우선 데이터 포인트 $X_{n=1..N} = (x_{n,1}, \cdots, x_{n,D})$를 각 차원별로 표준화(z-scoring)하여 정규화 한다.

$$s_{n,d} = \frac{1}{2} \cdot \frac{x_{n,d} - \mu_d}{\sigma_d} \tag{1}$$

여기에서 $\mu_d = \frac{1}{N}\sum_{n=1}^{N} x_{n,d}$는 평균을 의미한다.

이후, 클래스 내 평균거리와 클래스 간 평균거리를 계산하고, 두 값의 차이를 통해 GDV를 산출한다.

$$GDV = \frac{1}{\sqrt{(D)}}[< \overline{(d)}(C_l) > - < \overline{(d)}(C_l, C_m) >] \tag{2}$$

반면, 계수 $\frac{1}{\sqrt{(D)}}$는 GDV의 차원 불변성을 유지하기 위해 도입되며, 여기서 D는 차원의 수를 의미한다. 여기서 D는

데이터의 차원 수이다. 결과적으로 GDV 값이 작을수록(더 음수일수록) 클래스 간의 분리 성능이 뛰어남을 의미한다. 세부적인 GDV 계산과 실험 설정은 이전 논문[22]의 방식을 그대로 따랐다.

Ⅳ. 실험 결과

4-1 토큰 기반 분류(Token-Based Classification)

토큰 기반 분류에 대한 실험 결과는 다음과 같다.

먼저 Core verbs(핵심 동사)의 경우 모델의 성능은 전반적으로 매우 높은 정확도를 보였다. BERT tiny(2개 층) 모델의 경우 초반 층부터 이미 약 0.97 이상의 높은 정확도를 보였으며, mini 모델(4개 층)에서는 층이 깊어지면서 약간의 성능 향상 후 거의 완벽한 정확도(0.99)에 도달했다. BERT small(4개 층), medium(8개 층), base 모델(12개 층)은 첫 번째 층부터 정확도가 이미 매우 높았으며, 이후 층에서 지속적으로 안정적이고 높은 성능을 유지했다. 특히 표준 크기의 BERT-base 모델은 초기 층부터 거의 완벽한 정확도를 유지하여, Core 동사의 의미 정보가 하위 층에서도 이미 매우 효과적으로 표현되고 있음을 확인할 수 있었다.

Less Core verbs(덜 핵심적인 동사)에서도 초기 층에서 이미 상당히 높은 정확도를 보였으나, tiny 모델은 하위층에서 상대적으로 다소 낮은 정확도를 나타냈다. 그러나 mini 모델 이상의 중간 규모 모델부터는 초반 층에서 이미 매우 높은 정확도를 기록했고, 층이 깊어질수록 더욱 안정화되면서 최종적으로는 약 0.99에 근접한 정확도를 나타냈다. 이는 Less Core 동사 정보가 Core 동사에 비해 인코딩 난이도는 다소 높지만, BERT 모델의 중간 규모 이상에서는 비교적 빠르게 효과적으로 학습됨을 의미한다.

마지막으로 Periphery verbs(주변적 동사)는 세 범주 중 가장 난이도가 높은 유형임에도 불구하고, tiny 모델을 제외한 나머지 모델에서는 초기 층부터 상당히 높은 정확도를 나타냈다. mini 모델 이상에서는 층이 깊어질수록 정확도가 꾸준히 상승했고, 특히 medium과 base 모델에서는 빠르게 높은 정확도에 도달하여 상위층에서 안정적으로 유지되는 양상을 보였다. 이 결과는 더 크고 깊은 모델일수록 Periphery 동사의 추상적이고 복잡한 의미 정보를 더욱 효과적으로 포착할 수 있음을 보여준다.

4-2 문장 기반 분류 (Sentence-Based Classification)

문장 기반 분류에 대한 실험 결과는 다음과 같다.

먼저 Core verbs(핵심 동사)의 경우, BERT tiny 모델(2개 층)은 하위층(0층)부터 비교적 높은 정확도(약 0.90 이상)를 기록하였으며, 상위층(1층)에서는 정확도가 더욱 높아져

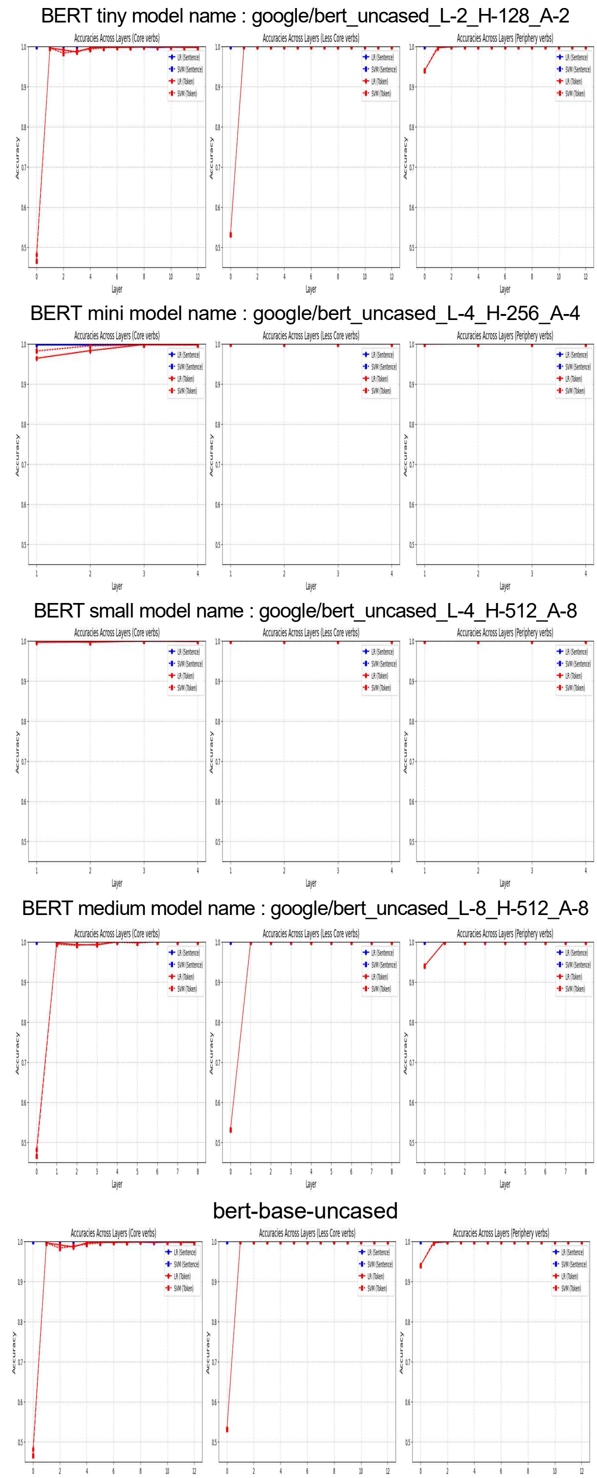

그림 1. 분류 정확도
Fig. 1. Classification accuracies

거의 완벽한 수준(0.99)에 도달하였다. BERT mini 모델(4개 층) 또한 초기 층에서 이미 매우 높은 정확도를 보였고, 2층 이후로는 두 분류기 모두 거의 완벽한 정확도를 유지했다. 특

히 BERT small, medium 및 base 모델에서는 초반부터 매우 높은 성능을 기록했으며, 2층 이상부터는 정확도가 0.99에 근접해 안정적으로 유지되는 모습을 나타냈다. 이러한 결과는 핵심 동사의 의미적 특성이 문장 전체의 맥락과 결합하여 초기 층부터 이미 효과적으로 인코딩되고 있음을 시사한다.

Less Core verbs(덜 핵심적인 동사) 범주에서는 BERT tiny 모델의 초기 층(0층)에서 두 분류기의 정확도가 상대적으로 낮은 수준을 나타냈으나, 1층에서 급격히 정확도가 향상되며 안정화되었다. mini 모델에서도 초반 층(0~1층)의 정확도는 상대적으로 낮았지만, 2층 이상에서 급격한 상승을 보이며 최고 성능에 도달한 후 유지되었다. 특히 small, medium 및 base 모델에서는 초반 층부터 이미 높은 정확도를 기록하였으며, 상위 층에서는 거의 완벽한 정확도를 나타내며 매우 안정된 성능을 유지했다. 이는 덜 핵심적인 동사의 의미 정보가 모델의 규모가 증가할수록 보다 효과적으로 표현됨을 시사한다.

마지막으로 Periphery verbs(주변적 동사) 범주에서는 가장 큰 변화가 관찰되었다. BERT tiny 모델에서는 초기 층에서 상당히 낮은 정확도를 기록했으나, 상위층에서 빠르게 증가하는 양상을 보였다. mini 모델 또한 초반 층에서 다소 낮은 성능을 보였지만, 2층부터 큰 폭으로 향상되어 3~4층에서 거의 완벽한 정확도에 도달했다. 특히 BERT small, medium 및 base 모델에서는 초기 층부터 높은 성능을 기록했고, 상위 층에서 매우 빠르게 최고 성능에 도달하여 안정적인 정확도를 유지하였다. 이러한 결과는 주변적 동사가 다른 범주에 비해 초기 인코딩 난이도가 높지만, 모델의 크기가 커지고 층이 깊어질수록 더욱 효과적으로 문장 맥락 정보와 결합하여 인코딩됨을 의미한다.

4-3 데이터 분리 가능성을 위한 GDV 값

본 연구에서는 BERT 모델의 각 층(layer)에서 자동사 유형(Core, Less Core, Periphery)에 대한 데이터의 분리 가능성을 정량적으로 평가하기 위해 일반화 판별 값(Generalized Discrimination Value, GDV)을 분석하였다. GDV 값이 작을수록(더 음수일수록) 클래스 간의 구분이 명확함을 나타낸다.

먼저 Core verbs(핵심 동사) 범주의 경우, BERT tiny 모델(2개 층)은 전반적으로 클래스 내(intra-class)와 클래스 간(inter-class)의 GDV 값 차이가 크지 않았으며, 두 층 모두 상대적으로 높은 GDV 값(덜 음수)을 기록하여 클래스 간의 분리 가능성이 낮은 것으로 나타났다. 반면 BERT mini 모델(4개 층)부터는 층이 깊어질수록 클래스 내 거리(intra-class)는 안정적으로 낮아지고 클래스 간 거리(inter-class)는 높아지는 경향이 관찰되어 GDV 값이 층에 따라 점차 더 음수로 변화하면서 데이터의 분리 가능성이 향상되는 것을 확인할 수 있었다. 특히 BERT small(4개 층), medium(8개 층), base(12개 층) 모델에서는 층이 깊어질수록 GDV 값이 지속적으로 낮아져(더 음수) 클래스 간의 명확한 구분이 나타났다. 특히 표준 크기의 BERT-base 모델에서는 중간층(약 6~8층)에서 가장 낮은 GDV 값을 기록하여, 이 층에서 데이터의 분리 가능성이 가장 뛰어남을 시사했다.

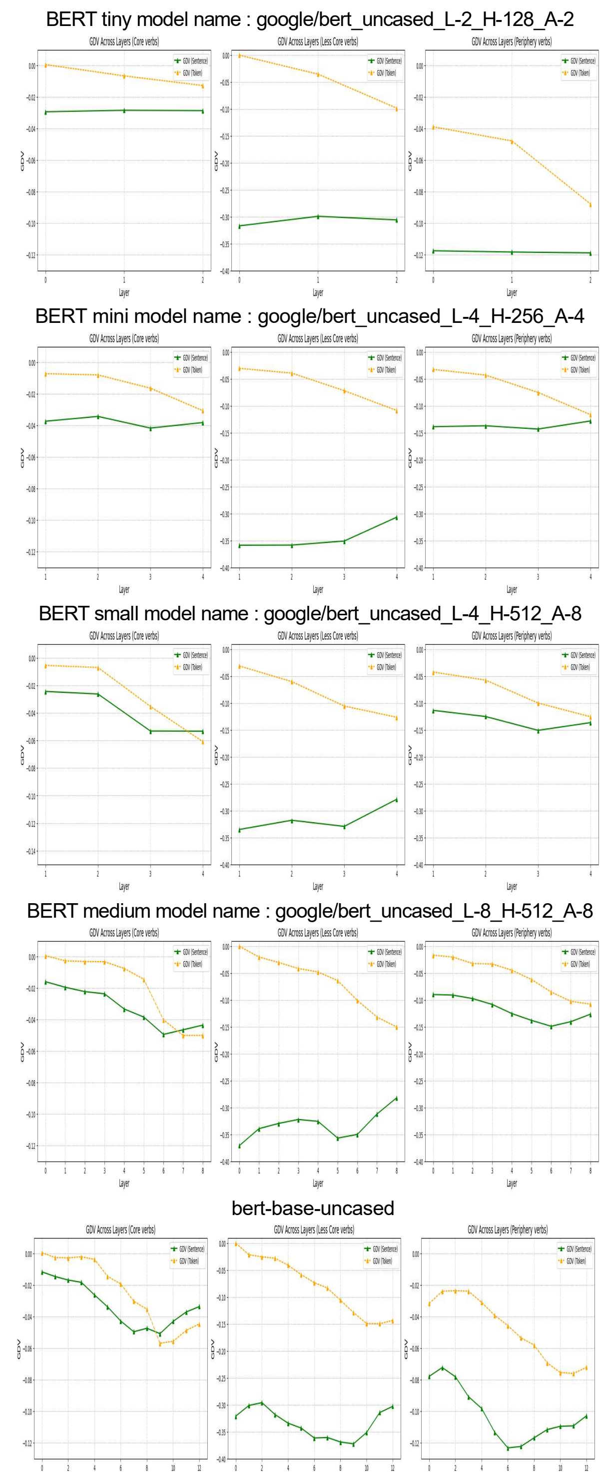

그림 2. 클래스 간 데이터 분리 가능성을 나타내는 GDV 값
Fig. 2. GDV values indicating data separability between classes

Less Core verbs(덜 핵심적인 동사) 범주에서도 유사한 경향이 나타났으나, 전반적으로 Core 범주보다 GDV 값의 절대값이 다소 작아 상대적으로 구분이 덜 명확했다. BERT tiny 및 mini 모델에서는 층이 깊어지면서 GDV 값의 개선이 다소 미미했지만, small 모델부터 GDV 값의 개선이 명확히 관찰되었다. 특히 BERT medium 및 base 모델에서는 층이 깊어질수록 GDV 값이 지속적으로 감소하며(더 음수 방향), 클래스 간 데이터 분리 가능성이 중간층 이후 더욱 명확히 나타났다. BERT-base 모델의 경우 중간층(약 6층 전 후)에서 GDV 값이 가장 낮아지는 것을 확인할 수 있었다. 이는 덜 핵심적인 동사의 의미 정보가 중간 규모 이상의 모델에서 보다 효과적으로 인코딩됨을 나타낸다.

마지막으로, Periphery verbs(주변적 동사) 범주는 GDV 값이 세 범주 중 가장 높게 나타나 상대적으로 데이터 분리 가능성이 낮았다. 특히 tiny와 mini 모델에서는 층이 깊어질수록 클래스 간 분리 가능성이 소폭 증가했으나 GDV 값의 개선이 크지 않았다. 하지만 small 모델부터 GDV 값의 감소가 명확히 나타나기 시작했으며, medium 모델에서는 층이 깊어짐에 따라 클래스 내 거리가 감소하고 클래스 간 거리가 증가하면서 GDV 값이 지속적으로 개선되었다. 특히 base 모델에서는 중간층에서 GDV 값이 가장 낮은 지점(가장 음수)을 기록하여, 주변적 동사의 의미 정보 역시 중간 규모 이상의 BERT 모델 중간층에서 효과적으로 인코딩됨을 시사했다.

종합적으로 GDV 분석 결과, 모든 자동사 범주에서 모델의 규모와 층이 증가할수록 데이터 분리 가능성이 명확히 증가하였다. 특히 BERT의 중간층(대략 6~9층 전후)에서 클래스 내 거리는 작아지고 클래스 간 거리는 커지는 최적의 GDV 값을 보였다.

V. 결 론

본 연구는 트렌스포머 기반 언어 모델인 BERT가 자동사의 의미 유형(Core, Less Core, Periphery)과 주어의 유생성(animate/inanimate)이라는 두 가지 복합적인 언어 정보를 내부 표현에 어떻게 반영하는지를 분석하였다. 이를 위해 토큰 단위와 문장 단위의 임베딩을 추출하고, 프로빙 기법과 Generalized Discrimination Value(GDV)를 활용하여 층별 표현의 구분 가능성과 구조적 특성을 비교하였다.

실험 결과, BERT의 중간층(layer 6~9)에서 가장 높은 분류 정확도와 분리 가능성이 관찰되었으며, 이는 트랜스포머 모델이 단어의 표층적 빈도나 위치 정보를 넘어서, 자동사의 의미적 성격이나 주어와의 의미적 결합성 같은 문장 내 구조 정보를 비교적 체계적으로 인코딩하고 있음을 시사한다. 특히 자동사의 의미적 위계(Sorace, 2000)와 주어의 유생성이라는 의미적 요인이 논항 구조와 함께 중간층 표현에서 분명하게 드러났다는 점은, 언어 모델이 의미 중심의 문장 처리 기제를 일정 부분 반영하고 있음을 보여주는 실험적 근거로 해석될 수 있다.

GDV 분석에서는 전반적으로 선형 분리도가 높지 않았으나, 분류 정확도는 일관되게 높게 나타났으며, 이는 해당 정보가 명시적으로 분리되어 있지는 않더라도 고차원의 비선형적 표현 공간에서 일정한 규칙성에 따라 구조화되어 있음을 시사한다. 이러한 결과는 언어 모델의 내부 표현 구조가 단순한 어휘 연관성이나 문맥 예측을 넘어, 문장 내 논항 구조 및 의미적 계층성과 같은 높은 차원의 언어 정보를 일정 수준 반영한다는 가능성을 열어준다.

본 연구는 다중단어 동사를 중심으로 한 선행 연구[22]와 방법론적으로는 유사하지만, 분석 대상의 깊이와 해석 관점에서 뚜렷한 차별성을 지닌다. 기존 연구가 구동사 및 전치사 결합 표현을 중심으로 한 표면적 구조 분석에 초점을 맞추었다면, 본 연구는 단일 자동사의 의미적 위계와 주어 유생성이라는 보다 추상적인 의미 중심 변수들이 내부 표현에서 어떻게 구분 가능한 방식으로 반영되는지를 계층별로 정밀하게 분석하였다. 이는 단어 수준의 통계적 처리를 넘어, 의미론적 해석과 구문적 결합 양상이 PLMs 내부에서 어떻게 드러나는지를 밝힌 실증적 시도라 할 수 있다. 또한 본 연구는 다양한 크기의 BERT 모델을 대상으로 실험을 반복함으로써, 모델 규모가 표현 계층 구조에 미치는 영향을 함께 검토하였다.

다만, 본 연구는 BERT 계열 모델에 한정된 실험 결과를 바탕으로 하고 있으며, 분류 정확도 및 GDV 중심의 해석이라는 제한점을 지닌다. 이에 따라 본 연구의 결과를 모든 언어 모델에 일반화하는 데에는 한계가 있다. 향후 연구에서는 RoBERTa, DeBERTa, GPT-3와 같은 구조 및 규모가 상이한 언어 모델뿐 아니라, 문장 임베딩에 특화된 SimCSE, LaBSE, E5 등 contrastive 학습 기반 모델들을 포함하여 비교 분석을 수행함으로써, 모델 학습 방식과 문장 의미 표현 간의 관계를 보다 정밀하게 평가할 필요가 있다. 또한 GDV 외에도 층별 probing 정확도 변화, attention 가중치 분포의 시각화, 표현 간 유사도 분석(예: CCA, SVCCA) 등 다양한 보완 지표를 활용하여 모델 내부 표현의 해석 가능성과 신뢰도를 높일 수 있을 것이다. 아울러 실험 문장의 의미 및 통사적 특성의 통제 수준을 더욱 정교하게 설정하여 조건 간 confound 요인을 최소화하고, 문장 구조의 일관성을 확보함으로써 분석 결과의 타당성을 강화하는 것도 중요한 과제로 남아 있다.

본 연구는 자동사의 의미 구조와 주어 유생성이라는 전통적인 언어학적 개념이 신경망 기반 언어 모델의 표현 구조에서 어느 수준까지 인코딩되고 있는지를 계층별로 실증적으로 탐색하였다는 의의를 가진다. 이는 자연어 처리와 언어학, 심리언어학을 연결하는 교차 영역 연구의 하나로, 언어 모델 내부 표현의 해석 가능성과 인간 언어 처리와의 연결 가능성을 탐색할 수 있는 새로운 분석 틀을 제시하였다. 향후에는 인간의 문장 처리 과정(예: 읽기 시간, ERP 반응 등)과의 비교 연구를 통해 언어모델의 인지적 정합성을 보다 체계적으로 평가 하는데 기여할 수 있을 것이다.

참고문헌

[1] D. M. Perlmutter, "Impersonal Passives and the Unaccusative Hypothesis," in *Proceedings of the Annual Meeting of the Berkeley Linguistics Society,* pp. 157-190, September 1978. https://doi.org/10.3765/bls.v4i0.2198

[2] M. Vernice and A. Sorace, "Animacy Effects on the Processing of Intransitive Verbs: An Eye-Tracking Study," *Language, Cognition and Neuroscience*, Vol. 33, No. 7, pp. 850-866, 2018. https://doi.org/10.1080/23273798.2017.1421316

[3] M. García García, B. Primus, and N. P. Himmelmann, "Shifting from Animacy to Agentivity," *Theoretical Linguistics,* Vol. 44, No. 1-2, pp. 25-39, 2018. https://doi.org/10.1515/tl-2018-0002

[4] P. Ramezani, A. Schilling, and P. Krauss, "Analysis of Argument Structure Constructions in the Large Language Model BERT," *Frontiers in Artificial Intelligence*, Vol. 8, 2025. https://doi.org/10.3389/frai.2025.1477246

[5] A. Conneau, G. Kruszewski, G. Lample, L, Barrault, and M. Baroni, "What You can Cram into a Single Vector: Probing Sentence Embeddings for Linguistic Properties," arXiv preprint arXiv:1805.01070, 2018. https://doi.org/10.48550/arXiv.1805.01070

[6] M. Immertreu, A. Schilling, A. Maier, and P. Krauss, "Probing for Consciousness in Machines," arXiv:2411.16262, 2024. https://doi.org/10.48550/arXiv.2411.16262

[7] D. Hupkes, S. Veldhoen, and W. Zuidema, "Visualisation and 'Diagnostic Classifiers' Reveal How Recurrent and Recursive Neural Networks Process Hierarchical Structure," *Journal of Artificial Intelligence Research*, Vol. 61, pp. 907-926, 2018. https://doi.org/10.1613/jair.1.11196

[8] H. Sajjad, N. Durrani, and F. Dalvi, "Neuron-Level Interpretation of Deep NLP Models: A Survey," *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 1285-1303, 2022. https://doi.org/10.1162/tacl_a_00519

[9] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, ... and E. Pavlick, "What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations," arXiv:1905.06316, 2019. https://doi.org/10.48550/arXiv.1905.06316

[10] P. Krauss, J. Hösch, C. Metzner, A, Maier, P. Uhrig, and A. Schilling, "Analyzing Narrative Processing in Large Language Models (LLMs): Using GPT4 to Test Bert," arXiv:2405.02024, 2024. https://doi.org/10.48550/arXiv.2405.02024

[11] A. Banerjee, A, Schilling, and P. Krauss, "Exploring Narrative Clustering in Large Language Models: A Layer-wise Analysis of BERT," arXiv:2501.08053, 2025. https://doi.org/10.48550/arXiv.2501.08053

[12] A. Sorace, "Gradients in Auxiliary Selection with Intransitive Verbs," *Language*, Vol.76, No. 4, pp. 859-890, 2000. https://doi.org/10.2307/417202

[13] A. Sorace, "Unaccusativity and Auxiliary Choice in Non-native Grammars of Italian and French: Asymmetries and Predictable Indeterminacy," *Journal of French Language Studies*, Vol. 3, No. 1, pp. 71-93, March 1993. https://doi.org/10.1017/S0959269500000351

[14] A. Sorace, "Incomplete vs. Divergent Representations of Unaccusativity in Non Native Grammars of Italian," *Second Language Research*, Vol. 9, No. 1, pp. 22-47, 1993.

[15] E. Feleki and H. Branigan, "Conceptual Accessibility and Serial Order in Greek Speech Production," in *Proceeding of the 21st Annual Conference of the Cognitive-Science-Society*, Vancouver: Canada, pp. 96-101, 1999. https://doi.org/10.4324/9781410603494

[16] P. De Swart, M. Lamers, and S. Lestrade, "Animacy, Argument Structure, and Argument Encoding," *Lingua*, Vol. 118, No. 2, pp. 131-140, 2008. https://doi.org/10.1016/j.lingua.2007.02.009

[17] J. L. McDonald, K. Bock, and M. H.Kelly, "Word and World Order: Semantic, Phonological, and Metrical Determinants of Serial Position," *Cognitive Psychology*, Vol. 25, No. 2, pp. 188-230, 1993. https://doi.org/10.1006/cogp.1993.1005

[18] F. Ferreira, "Choice of Passive Voice Is Affected by Verb Type and Animacy," *Journal of Memory and Language*, Vol. 33, No. 6, pp. 715-736, 1994.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (long and short papers) Minneapolis: MN, pp. 4171-4186, June 2019.

[20] Z. Chen and Q. Gao, "Probing Linguistic Information for Logical Inference in Pre-Trained Language Models," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence,* Vol. 36, No. 10, pp. 10509-10517, June 2022. https://doi.org/10.1609/aaai.v36i10.21294

[21] A. Katinskaia and R. Yangarber, "Probing the Category of Verbal Aspect in Transformer Language Models," arXiv:2406.02335, 2024. https://doi.org/10.48550/arXiv.24

06.02335
[22] H. Kissane, A. Schilling, and P. Krauss, "Probing Internal Representations of Multi-Word Verbs in Large Language Models," arXiv:2502.04789, 2025. https://doi.org/10.48550/arXiv.2502.04789
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, "Attention Is All You Need," *31st Conference on Neural Information Processing Systems*, Long Beach: CA, Vol. 30, pp. 5998-6008, 2017. https://doi.org/10.5040/9781350101272.00000005
[24] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory,* Pittsburgh: PA, pp. 144-152, July 1992. https://doi.org/10.1145/130385.130401
[25] A. Schilling, A. Maier, R. Gerum, C. Metzner, and P. Krauss, "Quantifying the Separability of Data Classes in Neural Networks," *Neural Networks*, Vol. 139, pp. 278-293, July 2021. https://doi.org/10.1016/j.neunet.2021.03.035

김자경(Ja-Kyoung Kim)

2022년 : 경북대학교 대학원 영어영문학과 (석사)
2025년 : 경북대학교 대학원 영어영문학과 (박사)

2022년~2025년: 경북대학교 영어영문학과 박사
※관심분야 : 인공지능 활용 교육, 영어교육, 자연어처리, 언어심리학 등

김성은(Sung-Eun Kim)

2021년 : 경북대학교 대학원 영어영문학과 (석사)
2025년 : 경북대학교 대학원 영어영문학과 (박사)

2021년~2025년: 경북대학교 영어영문학과 박사
2025년~현 재: 부산외국어대학교 영어학부 강의초빙교수, 인천대학교 영어영문학과 강사
※관심분야 : 인공지능 활용 교육, 영어교육, 프롬프트 엔지니어링, 자연어처리 등

오동석(Dong-Suk Oh)

2016년 : 서강대학교 대학원 컴퓨터공학과 (석사)
2020년 : 고려대학교 대학원 컴퓨터학과 (박사)

2016년~2018년: 다이퀘스트 자연어처리 연구원
2018년~2019년: NHN Ent. 자연어처리 연구원
2019년~2020년: Human-inspired AI Research 인공지능 연구원
2023년~2023년: KT Large AI Alignment Project 팀 선임연구원
2023년~현 재: 경북대학교 영어영문학과 조교수
※관심분야 : 신뢰가능한 인공지능, 뉴로-심볼릭 인공지능, 자연어처리 등