JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 1, pp. 41-49, January 2026
https://doi.org/10.9708/jksci.2026.31.01.041

Interpretability for Korean Language Models: Evidence from Attention Visualization

Jung-Kyu Shin*, Beak-Cheol Jang**

*Student, Graduate School of Information, Yonsei University, Seoul, Korea
**Professor, Graduate School of Information, Yonsei University, Seoul, Korea

[Abstract]

This study investigates how the agglutinative nature and morphological complexity of Korean are reflected in language model (LM) internal representations by fine-tuning KLUE RoBERTa Base on the NER task and conducting qualitative and quantitative analyses of attention maps. Our methodology includes a stable training design based on subword–label alignment and masking respecting character-level annotations, attention weight extraction, attention strength visualization, and pattern-specific attention distribution quantification. The analysis reveals three patterns: span-internal cohesion, where entity tokens attend to span boundaries; boundary alignment, where post-entity particles tagged O function as boundary cues; and long-distance dependencies, where distal arguments form semantically coherent links. These findings suggest that Korean linguistic characteristics are structurally organized at the attention layer and head level. This work enhances the interpretability of Korean LMs and establishes a foundation for interpretability research applicable to diverse downstream tasks.

▸ **Key words**: Language Model, Transformer, Attention, Visualization, Korean, Interpretability

[요 약]

본 연구는 한국어의 교착어적 특성과 형태론적 복잡성이 언어모델(LM) 내부 표현에 어떻게 반영되는지를 규명하고자 KLUE RoBERTa Base 모델을 NER 태스크에 파인튜닝한 뒤 어텐션 맵을 정성적 및 정량적으로 분석한다. 연구 방법은 문자 단위 라벨을 반영한 서브워드-라벨 정렬 및 마스킹 기반의 안정적 학습 설계, 어텐션 가중치 추출, 토큰 간의 어텐션 강도 시각화, 그리고 패턴별 어텐션 분포 산출로 구성된다. 어텐션 맵을 분석한 결과, 내부 응집(엔티티 내부 토큰이 스팬의 시작/말단에 집중), 경계 정렬(엔티티 직후 조사(O)의 경계 신호 작용), 장거리 의존성(원거리 논항의 의미적 결속) 패턴이 관찰된다. 이는 한국어의 언어적 특성이 어텐션 레이어와 헤드 수준에서 구조적으로 조직됨을 시사한다. 결론적으로 본 연구는 한국어 LM의 해석 가능성을 제고하고 다양한 과업에 적용할 수 있는 해석 가능성 연구의 기반을 마련한다.

▸ **주제어**: 언어 모델, 트랜스포머, 어텐션, 시각화, 한국어, 해석 가능성

• First Author: Jung-Kyu Shin, Corresponding Author: Beak-Cheol Jang
*Jung-Kyu Shin (jkshin0903@yonsei.ac.kr), Graduate School of Information, Yonsei University
**Beak-Cheol Jang (bjang@yonsei.ac.kr), Graduate School of Information, Yonsei University
• Received: 2025. 11. 12, Revised: 2025. 12. 14, Accepted: 2025. 12. 29.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 트랜스포머(Transformer) 기반의 대규모 언어 모델(LLM)은 자연어 처리(NLP) 분야에서 혁신적인 발전을 가져왔으나[1], 이러한 모델이 복잡한 언어적 지식을 어떻게 처리하고 최종 예측을 도출하는지에 대한 내부 작동 메커니즘 해석(Mechanical Interpretibility)은 여전히 난제로 남아있다[1]. 특히, 기존의 LLM 해석 연구 및 벤치마크 개발 노력은 대부분 영어권 데이터와 모델(예: BERT, GLUE)에 집중되어 발전되어 왔으며[3], 이로 인해 다른 언어, 특히 유형론적 특성(Typological Features)이 크게 다른 비영어권 언어에 대한 모델의 이해력은 상대적으로 간과되어 왔다.

한국어는 어간과 접사, 조사, 어미 등의 결합으로 문법적 기능이 실현되는 교착어이며[3][4], 이는 단어 경계가 명확한 영어와는 근본적으로 다른 형태론적 복잡성을 야기한다[3]. 또한, 한국어는 비교적 자유로운 어순을 가지며, 담화 맥락에 대한 의존도가 높은 의미론적 특성을 가지기 때문에[2], 단순히 영어 모델의 해석 프레임워크를 적용할 경우, 한국어 LM이 이러한 고유한 언어적 특징을 어떻게 표상하고 활용하는지 심층적으로 이해하기 어렵다.

이에 본 연구는 한국어 중심의 PLM인 KLUE RoBERTa Base 모델을 활용하여 한국어 NLP 벤치마크인 KLUE[3]에서 제공하는 개체명 인식(NER) 태스크를 중심으로 한국어의 형태론적, 의미론적 특성이 LM의 내부 어텐션 맵(Attention Map)에 어떻게 반영되는지 정성적 및 정량적으로 분석하고자 한다. 단순 어텐션 맵 시각화를 넘어, 어텐션 가중치 추출, 토큰 간의 어텐션 강도 시각화, 그리고 패턴별 어텐션 분포 산출을 통해 모델의 예측에 실질적으로 이바지하는 어텐션 분포를 확인한다. 상호작용적 시각화 툴인 BertViz과 히트맵을 활용하여 모델의 내부 표현에 인코딩된 한국어의 언어학적 지식을 시각적으로 검증한다.

본 연구의 주요 기여점은 다음과 같다.

- KLUE 벤치마크라는 표준화된 환경에서 한국어 LM의 내부 작동 방식을 정량적 및 정성적으로 분석하는 방법론을 제시한다.
- 한국어 특화 모델인 KLUE RoBERTa Base의 어텐션 헤드(Attention Head)들이 내부 응집, 경계 정렬, 장거리 의존성 패턴을 통해 구문론적 지식과 의미론적 지식을 학습하는 방식을 직관적으로 해석한다.

II. Preliminaries

1. Related works

Transformer 기반의 LLM이 NLP 분야에서 뛰어난 성능을 보임에 따라, 모델이 언어적 지식을 내부 은닉 상태에 어떻게 표상하고 활용하는지 이해하려는 해석 가능성(Interpretability) 연구가 활발히 진행되어 왔다. 이러한 연구 동향은 주로 내부 표현 추출(Probing)을 통한 언어학적 지식의 정량적 검증과, 구조적 모델링 및 메커니즘적 해석 가능성 탐구라는 축을 중심으로 발전하였다.

1.1 Probing Internal Linguistic Knowledge

사전 학습된 LLM의 특정 레이어에서 추출된 벡터 표현이 실제 언어학적 지식을 인코딩하고 있는지 확인하기 위해 프로빙 분류기(Probing Classifier)를 활용하는 기법이다.

구문론적 지식의 계층적 발견. 선형 프로빙(Linear Probing)을 포함한 분석을 통해 BERT와 같은 Transformer 모델의 문맥화된 임베딩이 구문 구조 및 의존 구문 관계에 대한 정보를 인코딩하고 있음이 입증되었다[5]. 특히, 이러한 연구는 LLM이 임베딩 레이어에서 형태론적 정보를 포착한 후, 하위 레이어에서는 구문 청크 및 구문 구조와 같은 고전적인 NLP 파이프라인 지식을 재발견하는 계층적 학습 구조를 보인다는 점을 밝혔다[6]. 또한, Probing은 모델이 주어-동사 일치와 같은 특정 통사적 의존 관계를 학습하는 능력을 평가하는 데에도 활용되었다[7].

의미론적 구성 과정 탐사. Probing 기법은 구문론적 능력을 넘어 모델의 합성 의미론적 과정 이해도를 검증하는 데 확장되었다. 예를 들어, 명사-명사 복합어의 의미(예: "KITCHEN CHAIR")와 관련된 주제 관계 지식을 LLM이 얼마나 잘 표상하는지 탐사하는 연구가 수행되었으며, 이를 통해 모델의 토큰 벡터가 동일한 주제 관계를 공유하는 복합어를 구분할 수 있음이 확인되었다[1]. 나아가, LLM이 추상 의미 표현과 같은 전통적인 의미론적 구조를 얼마나 잘 이해하고 추론에 활용하는지에 대한 연구도 이루어졌다[8].

1.2 Structural Modeling and Interpretability

최근 연구 동향은 단순한 성능 향상뿐만 아니라, LLM이 구문 구조를 어떻게 학습하고 왜 특정 행동을 보이는지에 대한 인과적 설명을 찾는 데 집중하고 있다.

구문 정보의 명시적 통합. LLM에 외부 언어학적 지식을 명시적으로 통합하여 성능 및 해석력을 개선하려는 시도가

이루어졌다. 이에는 Syntax-BERT 및 Syntax-enhanced Pre-trained Model (SEPREM)처럼 구문적 제약을 트랜스포머 아키텍처에 직접 통합하는 방식이 포함된다[9][10].

추론 과정의 분석. LLM의 복잡한 추론 능력이 부상하면서, CoT(Chain-of-Thought)와 같은 단계적 추론 기법이 모델의 내부 작동을 이해하는 도구로도 사용된다[11]. 이러한 추론 과정은 문장 간에 강한 어텐션 연결 패턴을 형성하는 방식으로 확인할 수 있다.

2. Background

KLUE(Korean Language Understanding Evaluation). 한국어 자연어 이해(NLU) 연구 발전을 촉진하기 위해 설계된 종합 벤치마크이다[3]. NER 등 총 8가지 다운스트림 태스크와 개별 데이터셋으로 구성되어 있으며, 본 연구는 이 중 한국어의 형태론적/구문론적 특성을 분석하는 데 적합한 NER 태스크에 초점을 맞춘다.

KLUE RoBERTa Base. BERT의 장점을 계승하며 강력하게 최적화된 사전 학습 접근법인 RoBERTa 아키텍처를 기반으로 한국어 코퍼스에 사전 학습된 인코더 전용 트랜스포머(Encoder-only Transformer) 모델이다. KLUE 벤치마크의 베이스라인 모델로, 12개의 어텐션 레이어와 12개의 어텐션 헤드, 768개의 은닉 차원을 가진다. 이 모델을 포함한 KLUE PLM들은 한국어의 특성을 반영하기 위해 Mecab-ko와 같은 형태소 분석기를 사용하여 원시 텍스트를 형태소 단위로 사전 토큰화한 후 BPE를 적용하는 형태소 기반 서브워드 토크나이징(Morpheme-based Subword Tokenization) 방식을 채택했다.

NER(Named Entity Recognition). 문장에서 개체명의 경계를 탐지하고 인물(PER), 위치(LOC), 기관(OG) 등 6가지 유형으로 분류하는 순차 태깅(Sequence Tagging) 태스크이다. 한국어는 교착어적 특성으로 인해 띄어쓰기 단위(어절)가 내용어와 기능어의 복합체인 경우가 많아, KLUE-NER 데이터셋은 이러한 형태론적 복잡성을 반영하기 위해 기존 영어 데이터셋과 달리 문자 수준 BIO 태깅(Character-level BIO tagging) 방식을 채택하였다. BIO 태깅은 각 토큰(또는 문자)에 대해 개체 경계와 소속을 동시에 표현하는 표기 체계로, 여기서 'B-'는 특정 유형의 개체가 시작(Begin)됨을, 'I-'는 동일 개체의 내부(Inside)에 있음을, 'O'는 비(非)개체(Outside)임을 의미한다. 예컨대 인명(PER) 엔티티 "홍길동"이 연속 문자로 주어지면 "B-PER I-PER I-PER"로 표기된다. 동일 문장 내에서 서로 다른 인명 개체가 연속해 등장할 경우, 새 개체의 첫 글자에는 다시 B-PER가 부여되어 경계를 명확히 구분한다. 평가 지표로는 개체명 전체의 정확도를 측정하는 개체 수준 Entity F1과 더불어, 모델이 한국어의 어간과 접사를 분해하는 능력을 평가하기 위해 부분적인 일치도까지 측정하는 문자 수준 Char F1을 도입했다.

BertViz. BERT 기반 LLM의 멀티헤드 어텐션(Multi-head Attention) 분포를 시각적으로 분석하기 위해 개발된 반응형 도구이다[12]. 입력 텍스트의 각 토큰에 할당하는 가중치 분포를 레이어별, 헤드별 형태로 보여주어, 모델이 어떤 토큰 쌍 사이의 관계에 집중하는지 파악할 수 있도록 돕는다.

III. The Proposed Scheme

1. Research Overview

본 연구에서는 KLUE RoBERTa Base를 백본(backbone)으로 하여, 한국어의 언어적 특성에 최적화된 태스크별 파인튜닝(Fine-tuning)이 적용된 모델을 설계하였다. 한국어 교착어의 분절 특성과 KLUE 벤치마크의 어노테이션 규칙을 반영하여 라벨-토큰 정렬 일관성을 보장하도록 토크나이징 및 손실 계산 전략을 구성하였다.

2. Design of the Fine-Tuning Model

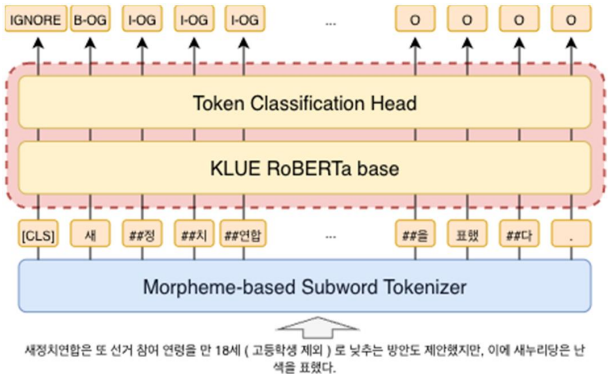

Fig. 1. Model Architecture

NER 태스크 파인튜닝은 토큰 분류 전용 헤드(Token Classification Head)를 결합하여 구현했다(Fig. 1). KLUE-NER 데이터셋은 문자 단위(char-level) 라벨로 구성되어 있다. 이에 토크나이징 과정에서 각 서브워드의 시작과 끝 오프셋 정보를 획득하고, 각 서브워드의 시작 오프셋이 엔티티의 시작 문자 인덱스와 일치할 때에만 해당 토큰에 개체명 라벨을 부여하였다. 즉, 하나의 단어가 여러 개의 서브워드로 분해되는 경우 첫 번째 서브워드에만 라벨을 할당하고, 나머지 서브워드들은 IGNORE_INDEX(-100)으로 마

스킹하여 손실 계산에서 제외하였다. RoBERTa의 마지막 은닉층(768차원)에서 반환된 벡터 임베딩은 NER 태그 집합(총 13개)에 대응하는 단일 선형 분류층(Linear layer)으로 변환된다. 출력은 토큰 단위 확률 분포 형태로 나타나며, 학습 시 IGNORE_INDEX가 아닌 모든 토큰에 대해 교차 엔트로피(Cross-Entropy) 손실이 계산된다.

이러한 서브워드 기반 라벨 정렬 전략은 토큰 단위와 문자 단위 라벨 간의 불일치를 제거하여, NER 모델의 안정적인 학습을 유도한다. 파인튜닝 시 적용한 하이퍼파라미터 정보는 Table 1과 같다.

Table 1. NER Configurations and Evaluation Metrics

Item	Value
Optimizer	AdamW
Learning Rate	2e-5
Weight Decay	0.01
Batch Size(train/eval)	8/8
Epoch	3
Evaluation Metric	Precision, Recall, F1, Accuracy
Loss Function	Cross-Entropy

3. Visualizing Attention Maps

본 연구에서는 파인튜닝된 한국어 LM의 내부 어텐션 구조를 시각적으로 분석하기 위해 BertViz 기반의 어텐션 맵 시각화를 구현하였다. 파인튜닝 모델로부터 레이어별·헤드별 어텐션 가중치를 추출하고, 이를 다양한 관점에서 시각화함으로써 모델 내부의 정보 흐름을 직관적으로 해석할 수 있도록 설계하였다.

3.1 Attention Extraction

입력 문장이 주어지면, 토크나이저를 통해 문장을 토큰 단위로 분할하고, 모델 입력 형식에 맞게 텐서 형태로 변환한다. 이때 특수 토큰([CLS], [SEP], [PAD])을 자동으로 포함하여, 모델의 입력 시퀀스를 완전하게 구성한다. 모델 추론 시 각 레이어별 어텐션 텐서를 추출한다. 이 텐서는 (B,H,S,S)형태의 4차원 구조를 가진다(B: 배치 크기, H: 어텐션 헤드 수, S: 입력 시퀀스의 길이). 모든 레이어에서 추출된 어텐션 행렬은 특수 토큰 마스킹을 통해 의미 있는 토큰 간의 어텐션 정보만 남기며, 각 행의 합이 1이 되도록 정규화 과정을 거친다. 이를 통해 레이어 간 혹은 헤드 간 비교 시 스케일링 차이로 인한 왜곡을 방지한다.

3.2 Visualization Interface

시각화 인터페이스는 Head View와 Model View 두 가지 형태를 동시에 지원한다. Head View는 특정 레이어 내 각 헤드가 개별적으로 어떤 토큰 간 상호작용을 학습했는지를 색상 맵으로 표현하며, Model View는 전체 레이어의 어텐션 변화를 한눈에 파악할 수 있도록 통합적으로 시각화한다. 본 연구는 Head View를 사용하여 구체적인 사례들을 제시하고자 한다.

IV. Experiment Result

1. Qualitative Study

본 절에서는 KLUE-NER로 파인튜닝한 KoBERT 기반 모델의 어텐션 맵을 BertViz로 정성적으로 분석한 결과를 정리한다. 여기서 내부 응집, 경계 정렬, 장거리 의존성 현상을 한국어 언어적 특성과 연관 지어 서술한다. 두 용어는 시각화에서 관찰된 패턴을 기술하기 위한 분석적 개념으로 정의하였다.

1.1 Span Cohesion

KLUE-NER에서 예측의 실제 단위는 연속 토큰 구간인 스팬(span)이며, 토큰 단위 B/I/O 라벨은 스팬의 경계와 유형을 복원하기 위한 수단이다. 한국어는 복합명사·합성어가 발달해 하나의 엔티티가 다수 서브워드로 분절되므로, 모델은 서브워드 간 형태론적 재결합을 수행하고, 이후 스팬 단위로 표상을 집약하는 내부 응집(Span Cohesion)을 필수적으로 형성한다. 이를 통해 스팬 내부 라벨의 연속성과 경계 안정성이 향상되어 엔티티 유형 판별이 견고해진다. 결국 내부 응집은 한국어의 형태론적 복잡성(과도한 서브워드 분절)을 상쇄해 스팬 수준 의미를 회복하는 구조적 적응이며, 모델이 한국어다운 형태론적 현실을 얼마나 내재화했는지를 가늠하는 수단으로 활용할 수 있다. 이와 관련된 사례로 동일 스팬 내부 I-태그 토큰들이 일관되게 스팬의 시작점(B-태그) 또는 말단(마지막 I-태그)으로 주의를 집중하는 경향이 확인되었다.

Fig. 2를 보면 '새정치연합'(B-OG, I-OG: Idx 1-4)에서는 I-OG 토큰이 B-OG 토큰 또는 마지막 I-OG 토큰에 집중하는 패턴을 확인할 수 있다. '새누리당'(B-OG, I-OG: Idx 29-31)에서도 동일한 경향이 나타난다. Fig. 3에서도 '뉴욕타임스'(B-OG, I-OG: Idx 1-3) I-OG 토큰이 B-OG 토큰 또는 마지막 I-OG 토큰에 집중하는 패턴을 확인할 수 있다. 이를 통해 모델이 어텐션 수준에서 형태소 기반

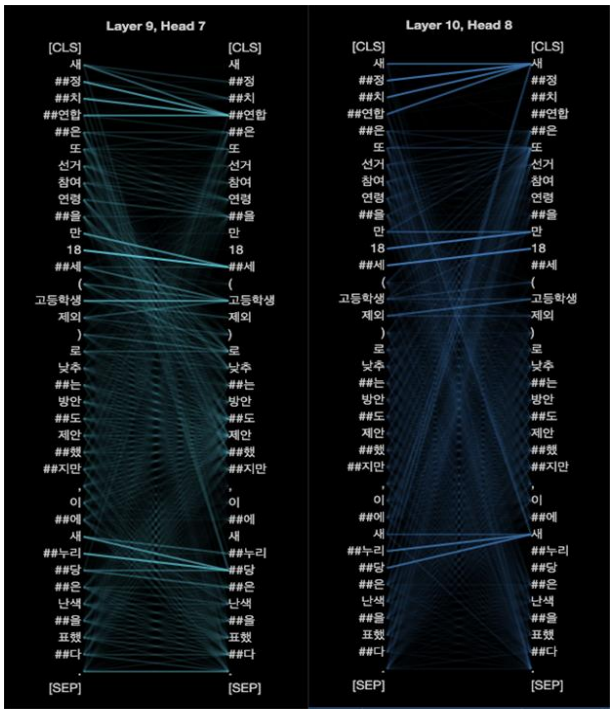

Fig. 2. Span Cohesion (case1)

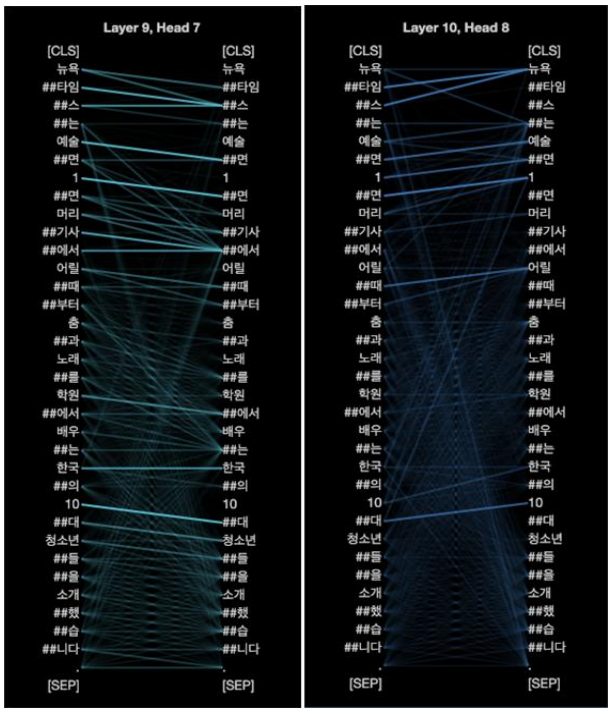

Fig. 3. Span Cohesion (case2)

서브워드 토크나이징으로 인한 분절을 복원하여, 기관/정당명과 같은 고유명사 스팬을 단일 의미 단위로 재결합함을 시사한다.

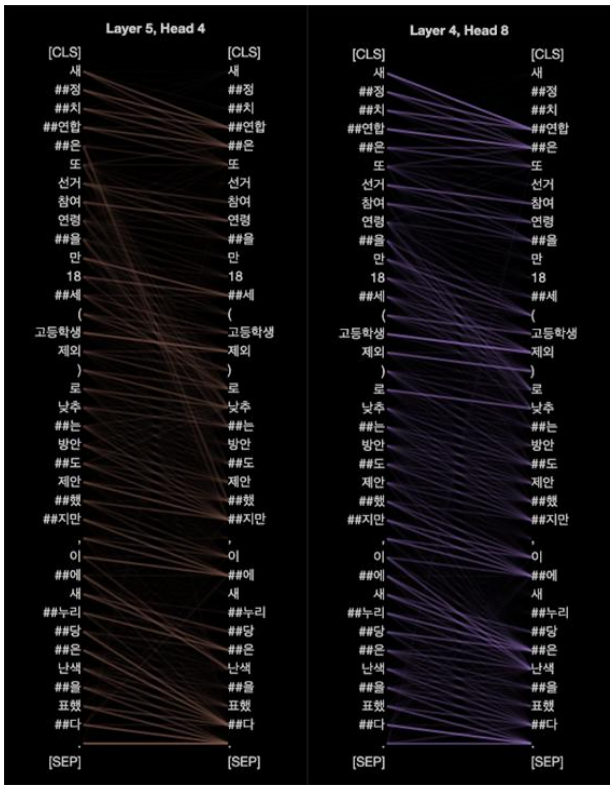

Fig. 4. Boundary Alignment (case1)

1.2 Boundary Alignment

경계 정렬(Boundary Alignment)은 한국어의 교착어적 특성과 밀접하게 연관된다. 한국어는 실질어 뒤에 조사나 어미가 뒤에 붙는 구조로, 고유명사 스팬은 대개 “엔티티 + 조사(O 태그)” 형태로 실현되어 둘 사이에 명확한 경계를 형성한다. ‘에서’, ‘까지’, ‘은/는’ 등 조사는 엔티티 일부가 아닌 문장 기능을 표지하므로, 모델에게는 “여기서부터는 엔티티가 아니다”라는 경계 신호로 작용한다. 엔티티의 B/I 태그 토큰이 직후 조사(O 태그) 토큰으로 집중하는 패턴은 이러한 한국어의 구조를 모델이 경계 판별에 활용하고 있음을 보여준다.

Fig. 4를 보면 ‘새정치연합’, ‘새누리당’에서는 B/I-OG 토큰들이 바로 뒤에 붙은 ‘##은’(O)에 집중하는 패턴을 확인할 수 있다. Fig. 5에서도 ‘뉴욕타임스’의 B/I-OG 토큰들이 뒤에 붙은 ‘##는’(O)에 집중하는 부분이 보인다. 이는 한국어의 조사·어미가 엔티티 경계를 표시하는 구조적 표지로 작동하며, 엔티티 종료 지점에서 즉시 O구간으로 정렬하여 경계 판별을 정밀화한다는 점을 보여준다.

1.3 Long-distance Dependency

장거리 의존성(Long-distance Dependency)은 문장 내에서 서로 멀리 떨어진 요소들 사이에 의미적 연결이 형성되

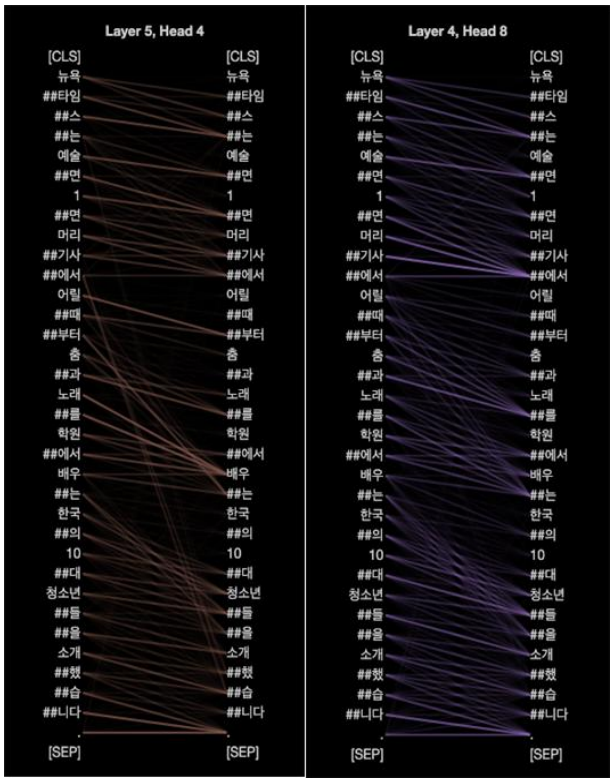

Fig. 5. Boundary Alignment (case2)

어 해석에 활용되는 현상을 말한다. Fig. 6에서 2가지 의미적 결속이 관찰된다.

첫 번째는 정당 명칭 스팬들 사이의 결속이다. 문장 전반의 기관/정당 스팬 '새정치연합'과 문장 후반의 '새누리당'은 표면적으로는 비인접하나, 담화 구조 상 전자의 정책 제안에 대해 후자가 태도를 표명하는 대조적 담화 역할을 수행한다. 시각화에서 양 스팬의 토큰들(스팬 시작 B-토큰부터 말단 I-토큰) 사이에 상호 어텐션 패턴이 확인되며, 이는 모델이 정당 간 관계를 장거리 연결로 확인한다는 점을 시사한다. 다시 말해, 모델은 동일 유형의 엔티티를 담화 내 대응 쌍으로 인식하고, 전·후반 문맥을 가로지르는 관계를 어텐션 그래프로 재현한다.

두 번째는 정책 제안과 반응(난색 표명) 사이의 원거리 논항 결속이다. 수량·단위 스팬 '18세'(B/I-QT: Idx 12-13)를 핵으로 하는 "연령을 낮추는 방안 … 제안" 구간(Idx 9-23)은 후반부의 "난색을 표했다" 구간(Idx 33-36)과 직접 인접하지 않음에도, 해당 정책 제안에 대한 태도 표명으로 기능적 연결을 형성한다. 어텐션 맵에서 '낮추/##는/방안/제안/##했/##지만'과 '난색/표했/##다' 사이의 어텐션이 뚜렷하게 보인다. 이는 모델이 논항 구조를 문장 전역에서 통합하며, 비인접한 정책-평가 구간을 인과·대응 관계로 엮어 해석함을 보여준다.

이와 유사하게 Fig. 7에서도 문장 마지막 부분의 '소개

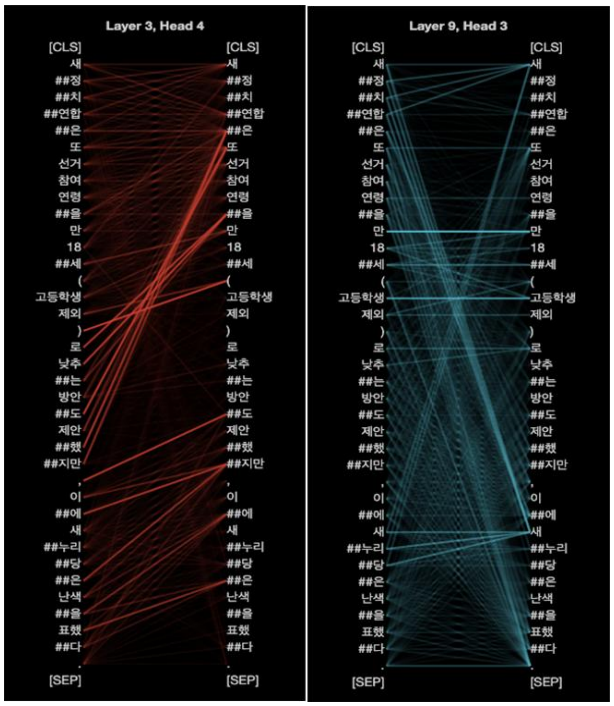

Fig. 6. Long-distance Dependency (case1)

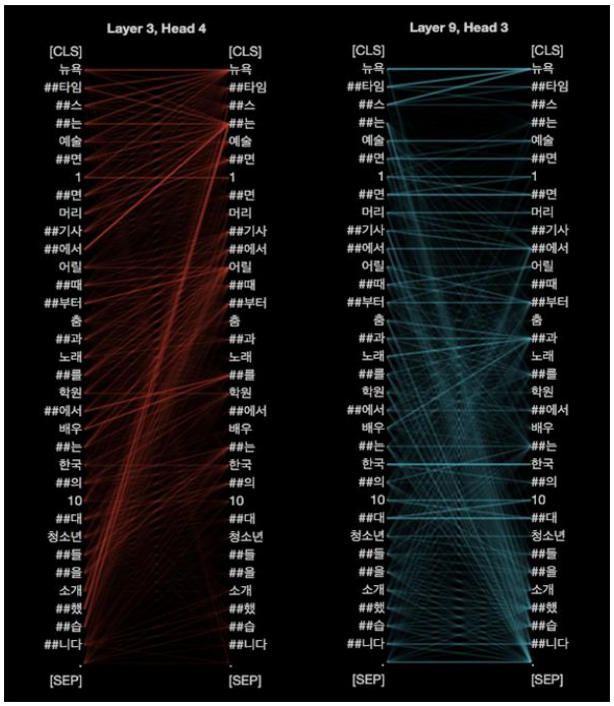

Fig. 7. Long-distance Dependency (case2)

/##했/##습/##니다' 토큰이 멀리 떨어진 '어릴'에서 '##는'까지 토큰들과 연결하여 소개할 대상에 대한 정보를 얻으려는 패턴을 확인할 수 있다.

2. Quantitative Study

본 절에서는 앞선 정성적 분석에서 확인된 세 가지 패턴별 어텐션 강도가 모델 레이어와 어텐션 헤드에 따라 어떻

게 분포하는지 정량적으로 평가하였다. 각 패턴의 어텐션 강도는 모든 레이어와 헤드에서 계산된 어텐션 가중치를 기반으로 산출되었고 이를 통해 생성된 어텐션 분포는 히트맵으로 확인할 수 있다.

2.1 Computing Pattern-Specific Attention Score

내부 응집 강도는 엔티티 내부 토큰들이 스팬의 경계 토큰(시작 토큰 B와 마지막 토큰 I)에 집중하는 정도를 측정한다. 각 엔티티 스팬에 대해 내부 토큰들(B 토큰을 제외한 I 토큰들)이 경계 토큰들에 부여하는 어텐션 가중치의 합을 해당 내부 토큰의 전체 어텐션 가중치 합으로 나눈 비율을 계산한다. 여기서 $A_{t,b}$는 내부 토큰 t가 경계 토큰 b에 부여하는 어텐션 가중치를 나타내며, I_s와 B_s는 각각 스팬 s의 내부 토큰 집합과 경계 토큰 집합을 의미한다. 이 값이 높을수록 엔티티 내부 토큰들이 경계 토큰에 집중하여 스팬의 응집성이 높음을 의미한다.

$$Cohesion(s) = \frac{\sum_{t \in I_s} \sum_{b \in B_s} A_{t,b}}{\sum_{t \in I_s} \sum_{j} A_{t,j}}$$

경계 정렬 강도는 엔티티 직후에 있는 O 태그 토큰(일반적으로 조사나 문법적 마커)이 해당 엔티티의 마지막 토큰(경계 토큰)에 집중하는 정도를 측정한다. 각 엔티티-조사 쌍에 대해, O 태그 토큰이 엔티티의 마지막 토큰에 부여하는 어텐션 가중치를 해당 O 토큰의 전체 어텐션 가중치 합으로 나눈 비율을 계산한다. 여기서 A_{o,b_e}는 O 태그 토큰 o가 엔티티 e의 경계 토큰 b_e에 부여하는 어텐션 가중치를 나타낸다. 이 값이 높을수록 엔티티 직후 조사가 경계 신호로 작동하여 엔티티 경계를 인식하는 데 이바지함을 의미한다.

$$Alignment(e, o) = \frac{A_{o,b_e}}{\sum_{j} A_{o,j}}$$

장거리 의존성 강도는 시퀀스 내에서 서로 멀리 떨어진 엔티티 간의 어텐션 연결 강도를 측정한다. 두 엔티티 스팬의 중심점 간 거리가 임곗값 이상인 경우에만 분석 대상으로 포함하며, 각 엔티티 쌍에 대해 양방향 어텐션 가중치의 평균을 계산한다. 여기서 r_1과 r_2는 각각 엔티티 e_1과 e_2의 대표 토큰(일반적으로 B 토큰)을 나타내며, A_{r_i,r_j}는 토큰 r_i가 토큰 r_j에 부여하는 어텐션 가중치를 의미한다. 이 값이 높을수록 멀리 떨어진 엔티티 간의 의미적 연결이 강함을 의미한다.

$$Dependency(r_1, r_2) = \frac{A_{r_1,r_2} + A_{r_2,r_1}}{2}$$

2.2 Attention Score Heatmap

각 패턴의 점수는 모든 레이어와 헤드에서 개별적으로 계산된 후, 해당 레이어-헤드 조합에서 관찰된 모든 엔티티 또는 엔티티 쌍에 대한 평균값을 산출한다. 히트맵을 사용하면 각 레이어-헤드 조합이 특정 패턴을 얼마나 강하게 나타내는지 시각적으로 확인할 수 있다. 히트맵은 12×12 행렬 형태로 구성되며, 행은 레이어(0-11번), 열은 헤드(0-11번)를 나타낸다. 각 셀의 값은 해당 레이어-헤드 조합에서 계산된 패턴 점수의 평균을 의미하며, 색상 강도(노란색-주황색-빨간색 계열)로 시각화하여 패턴의 분포를 직관적으로 파악할 수 있다.

내부 응집 패턴의 경우, 상위 레이어(8-11번)에서 전반적으로 강한 응집성이 관찰되었다(Fig. 8). 이는 상위 레이어가 엔티티의 의미적 통합을 담당함을 시사한다.

경계 정렬 패턴은 중간 레이어(4-7번)에서 상대적으로 높은 점수를 나타내었으며, 특히 11번 헤드에서 두드러진 경계 인식 능력을 보였다(Fig. 9). 이는 중간 레이어가 문법적 경계 정보를 처리하는 데 중요한 역할을 함을 의미한다.

장거리 의존성 패턴은 전반적으로 낮은 점수를 보였으나, 중간 레이어(3-6번)와 상위 레이어(9-11번)에서 상대적으로 큰 값을 나타내어, 장거리 의미 관계 추론이 주로 상위 레이어에서 수행됨을 시사한다(Fig. 10).

이러한 분포의 차이는 각 패턴이 모델의 어텐션 메커니즘을 통해 구현됨을 보여주며, 이는 모델이 다양한 언어 현상을 처리하고 있음을 보여준다.

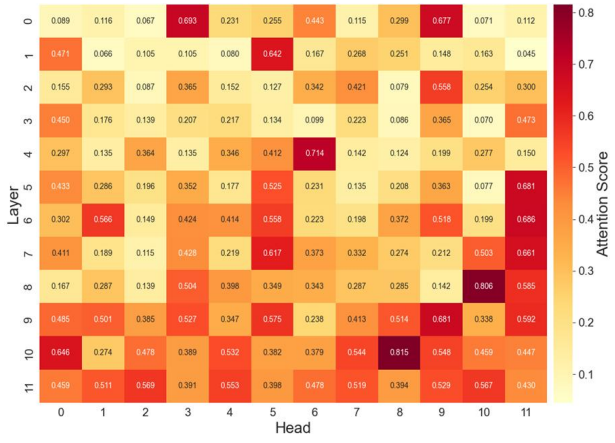

Fig. 8. Heatmap for Span Cohesion

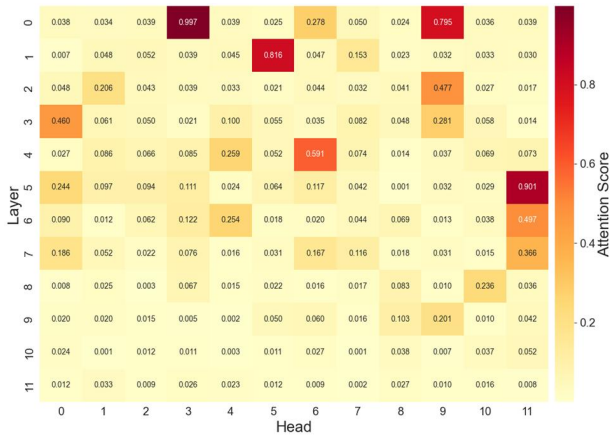

Fig. 9. Heatmap for Boundary Alignment

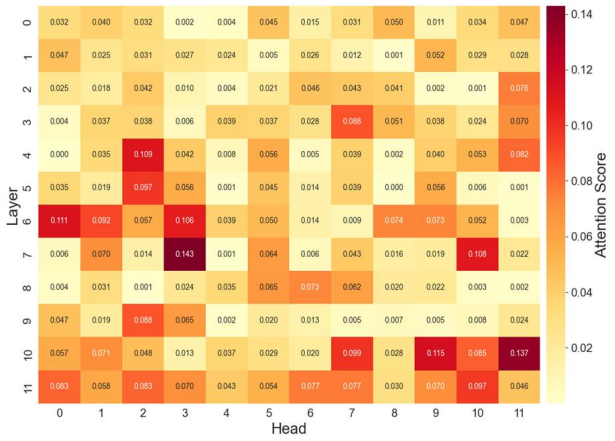

Fig. 10. Heatmap for Long-distance Dependency

V. Conclusions

본 연구는 한국어 LM을 대상으로, 내부 어텐션 구조를 분석하기 위한 방법론을 제시하였다. 한국어는 교착어적 특성과 형태소 결합 규칙을 지닌 언어로, 이러한 구조적 특성이 LM의 내부 표현에 어떠한 방식으로 반영되는지를 밝히는 것은 해석 가능성과 성능 향상 양 측면에서 중요한 의미를 갖는다. 또한 KLUE 벤치마크라는 표준화된 실험 환경 위에서 한국어 LM의 내부 작동 메커니즘을 탐구할 수 있는 방법론을 제안하였다. 이를 통해 모델의 언어 이해 과정에 대한 분석이 구체적인 레이어와 헤드 수준의 분석 절차로 확장될 수 있음을 보였다.

본 연구는 한국어 대규모 언어모델의 내부 어텐션 분포를 중심으로 형태, 의미 수준의 해석 가능성을 탐구하였으나, 이후 연구에서는 다음과 같은 방향으로 확장할 수 있다. 먼저 모델 및 언어 범위의 확장이 필요하다. 본 연구가 KLUE RoBERTa Base라는 단일 아키텍처를 중심으로 이루어졌다면, 향후에 GPT, LLaMA, Gemma, Polyglot 등 다양한 한국어 및 다국어 LLM에 동일한 분석을 적용하여, 언어 구조적 차이가 내부 표현에 미치는 영향을 교차 비교할 수 있을 것이다. 다음으로 본 연구에서 제시한 해석을 다운스트림 태스크에 응용하는 방향으로 발전시킨다면 실질적 성능 개선으로 이어지는 "설명가능한 성능 향상"의 실증적 기반을 마련할 수 있다. 마지막으로 시각화 및 언어 자원화의 측면에서도 확대 가능성이 있다. 어텐션 패턴과 구조 정합성 결과를 반응형 시각화 형태로 제공하거나, 한국어 어텐션 패턴 레포지토리 구축을 통해 향후 학습 효율이 높은 LLM 설계에 기여할 수 있다.

REFERENCES

[1] M. Ormerod, J.M. Rincón, and B. Devereux, "How is a "kitchen chair" like a "farm horse"? Exploring the representation of noun-noun compound semantics in transformer-based language models". Computational Linguistics, Volume 50, Issue 1, March 2024. DOI: 10.1162/coli_a_00495

[2] H. Shin, and S. Trott, "Do language models capture implied discourse meanings? An investigation with exhaustivity implicatures of Korean morphology", In Proceedings of the Society for Computation in Linguistics, pages 150-161, Irvine, CA. Association for Computational Linguistics, June 2024. DOI: 10.48550/arXiv.2405.09293

[3] S. Park, J. Moon, S. Kim, W. Cho, J. Han, and J. Park et al, "KLUE:Korean Language Understanding Evaluation", Advances in Neural Information Processing Systems, October 2021. DOI: 10.48550/arXiv.2105.09680

[4] S. Yu, K. Kim, J. Yun, and Y. Kim, "Making Sense of Korean Sentences: A Comprehensive Evaluation of LLMs through KoSEnd Dataset", In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 455-469, Vienna, Austria. Association for Computational Linguistics, July 2025. DOI: 10.48550/arXiv.2507.03378

[5] H. Zhao, A. Panigrahi, R. Ge, and S. Arora, "Do Transformers Parse while Predicting the Masked Word?", In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16513-16542, Singapore. Association for Computational Linguistics, December 2023. DOI: 10.48550/arXiv.2303.08117

[6] G. Jawahar, B. Sagot, and D. Seddah, "What Does BERT Learn about the Structure of Language?", In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651-3657, Florence, Italy. Association for Computational Linguistics, July 2019. DOI: 10.18653/v1/P19-1356

[7] V. Nastase, G. Samo, C. Jiang, and P. Merlo, "Exploring Syntactic

Information in Sentence Embeddings through Multilingual Subject-verb Agreement", In Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), pages 631–643, Pisa, Italy. CEUR Workshop Proceedings, December 2024. DOI: 10.48550/arXiv.2409.06567

[8] Z. Jin, Y. Chen, F.G. Adauto, J. Liu, J. Zhang, J. Michael, B. Schölkopf, and M. Diab, "Analyzing the Role of Semantic Representations in the Era of Large Language Models", In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3781–3798, Mexico City, Mexico. Association for Computational Linguistics, July 2024. DOI: 10.48550/arXiv.2405.01502

[9] J. Bai, Y. Wang, Y. Chen, Y. Yang, J. Bai, J. Yu, and Y. Tong, "Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees", In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3011–3020, Online. Association for Computational Linguistics, April 2021. DOI: 10.48550/arXiv.2103.04350

[10] Z. Xu, D. Guo, D. Tang, Q. Su, L. Shou, M. Gong, W. Zhong, X. Quan, D. Jiang, and N. Duan, "Syntax-Enhanced Pre-trained Model", In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5412–5422, Online. Association for Computational Linguistics, August 2021. DOI: 10.48550/arXiv.2012.14116

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E.H. Chi, Q.V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models", In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22), Curran Associates Inc., Red Hook, NY, USA, Article 1800, 24824–24837, November 2022. DOI: 10.48550/arXiv.2201.11903

[12] J. Vig, "Visualizing Attention in Transformer-Based Language models.", April 2019. DOI: 10.48550/arXiv.1904.02679

Authors

Jung-Kyu Shin received the B.S. degree in Information Communication Engineering from Dongguk University, Seoul, South Korea, in 2022. He is currently pursuing the M.S. degree in Information Systems at Yonsei University, Seoul, South Korea. He is interested in Natural Language Processing and Efficient AI.

Beak-Cheol Jang received the Ph.D. degrees in Computer Science and Engineering from North Carolina State University, United States, in 2009. Dr. Jang joined the faculty of Graduate School of Information at Yonsei University, Korea, in 2021. He is currently a Professor. His research interests are Wireless Networking and Artificial Intelligence.