JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 7, pp. 45-54, July 2026
https://doi.org/10.9708/jksci.2026.31.07.045

Injecting Hanja Radical Structure into Korean BERT through LoRA-based Contrastive Learning

Euna Song*, Beak-Cheol Jang**

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

[Abstract]

Modern Korean vocabulary comprises approximately 66–70% Sino-Korean words (漢字語)[1] whose sub-character semantic structures encoded in Hanja radicals (部首) are not explicitly modeled by existing Korean pre-trained language models. We propose a parameter-efficient fine-tuning framework applying LoRA to KLUE BERT-base with InfoNCE contrastive loss on 19,396 word pairs from 3,128 Sino-Korean words (0.98% of total parameters updated), using Hanja radicals as supervision signals. The proposed method achieves k-NN Acc@1 of 0.7007 on a hold-out structural generalization task—outperforming all baselines including KoSimCSE (0.5758)—and WSD accuracy of 0.8402 on the auxiliary task, demonstrating that Hanja radical structure serves as an effective supervision signal for semantic clustering in Korean embedding spaces.

▸ **Key words**: Hanja Radical, Contrastive Learning, LoRA, Korean BERT, Word Sense Disambiguation

[요 약]

현대 한국어 어휘의 절반 이상은 한자어(漢字語)로 구성되며, 한자 관련 혼종어를 포함할 경우 약 66~70%에 이른다[1]. 그러나 기존 한국어 사전 학습 언어모델은 서브워드 단위로 텍스트를 처리하여 한자 부수(部首) 기반의 의미 구조를 명시적으로 학습하지 않으며, 의미적으로 연관된 한자어들이 임베딩 공간에서 일관된 군집 구조를 형성하지 못하는 한계가 있다. 이를 해결하기 위해 본 논문에서는 한자 부수를 대조 학습의 지도 신호로 활용하는 경량 파인튜닝 프레임워크를 제안한다. KLUE BERT-base에 LoRA를 적용하고, 19,396쌍의 학습 데이터를 InfoNCE 손실함수로 학습함으로써 전체 파라미터의 0.98%만을 갱신한다. 학습에 사용되지 않은 627개 단어에 대한 hold-out 평가에서 k-NN Acc@1 0.7007을 달성하여 모든 비교 모델을 상회하였으며, 보조 태스크인 WSD에서도 KoSimCSE(0.8554)에 근접한 0.8402의 정확도를 달성하였다. 본 연구는 한자 부수 구조가 한국어 임베딩 공간의 의미 군집화에 효과적인 지도 신호로 활용될 수 있음을 실증한다.

▸ **주제어**: 한자 부수, 대조 학습, LoRA, 한국어 BERT, 한자어 중의성 해소

• First Author: Euna Song, Corresponding Author: Beak-Cheol Jang
*Euna Song (thddbsk0853@yonsei.ac.kr), Graduate School of Information, Yonsei University
**Beak-Cheol Jang (bjang@yonsei.ac.kr), Graduate School of Information, Yonsei University
• Received: 2026. 05. 06, Revised: 2026. 06. 04, Accepted: 2026. 06. 17.

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

I. Introduction

최근 트랜스포머(Transformer)[2] 기반의 사전 학습 언어모델(Pre-trained Language Model, PLM)은 한국어 자연어 처리(NLP) 분야에서 광범위하게 활용되고 있다. 그 대표적 사례로, Park et al.[3]은 한국어 자연어 이해 벤치마크인 KLUE를 제안하고 이를 위한 사전 학습 모델(KLUE BERT-base)을 공개하였다. 그러나 한국어는 어휘적으로 다른 언어와 구별되는 고유한 특성이 있다. 현대 한국어 어휘의 절반 이상은 한자어(漢字語)로 구성되어 있으며, 한자 관련 혼종어를 포함할 경우 그 비율은 약 66~70%에 이른다[1]. 이들 단어는 단순한 표기 단위를 넘어 한자 부수(部首)라는 의미 계층을 공유한다. 예를 들어, '水(물 수)' 부수를 포함하는 단어들—수심(水深), 수량(水量), 담수호(淡水湖)—은 모두 물 또는 액체와 관련된 의미를 지닌다. 이러한 부수 기반의 의미 계층은 형태론적으로 불투명한 한자어들 사이의 잠재적 의미 연관성을 체계적으로 포착할 수 있는 구조적 근거를 제공한다.

그러나 KLUE BERT-base[3]를 비롯한 기존 한국어 PLM은 한국어 텍스트를 서브워드 단위로 처리하며, 한자 부수와 같은 자소 하위 수준(sub-character level)의 의미 구조를 명시적으로 학습하지 않는다. 이로 인해 두 가지 문제가 발생한다. 첫째, 동음이의 한자어—예컨대 '수심(水深, 물의 깊이)'과 '수심(愁心, 근심)'—을 문맥 기반으로 어느 정도 구분할 수 있으나(Zero-shot WSD 정확도 0.7707), 부수라는 상위 의미 계층을 공유하는 단어들이 임베딩 공간에서 일관된 군집 구조를 형성하지 못하는 한계가 있다(k-NN Acc@1 0.4258). 둘째, 동일 부수를 공유하는 단어들이 임베딩 공간에서 체계적으로 군집화되지 않아, 부수 기반의 의미적 연관성을 활용한 검색이나 분류 태스크에서 활용할 수 있는 구조적 표현이 부재하다.

이에 본 논문에서는 한자 부수 정보를 대조 학습(Contrastive Learning)의 지도 신호로 활용하여, 기존 한국어 언어모델의 임베딩 공간에 한자어 의미 구조를 경량 파인튜닝만으로 인코딩하는 프레임워크를 제안한다. 구체적으로, Unicode Unihan DB와 KRDICT 사전을 기반으로 20개의 핵심 부수 그룹에 속하는 3,128개 한자어를 수집하고, 동일 부수 단어 쌍을 양성 쌍(positive pair)으로 구성한 InfoNCE 기반 대조 학습 데이터(총 19,396쌍)를 구축한다. 이후 KLUE BERT-base에 저랭크 적응(Low-Rank Adaptation, LoRA)[4]를 적용하여 전체 파라미터의 0.98%만을 학습함으로써, 사전 학습된 언어 능력을 보존하면서 한자어 의미 구조를 임베딩 공간에 반영한다. 대조 학습 프레임워크의 기본 설계는 SimCSE[5]를 참고하였으며, Projection Head 구조는 SimCLR[6]의 설계를 따른다. 비교 모델로는 대표적인 한국어 문장 임베딩 모델인 KoSimCSE[8]을 활용한다.

제안 방법의 유효성은 두 가지 태스크를 통해 검증된다. Task 1(한자어 중의성 해소, WSD)은 동음이의 한자어가 등장하는 484개 문항에서 올바른 한자 의미를 선택하는 태스크이며, Task 2(구조적 일반화)는 학습에 사용되지 않은 627개 단어가 올바른 부수 군집으로 분류되는지를 k-NN 정확도(Acc@1)로 측정한다. 아울러, 파인튜닝 후 인코더의 일반 언어 표현 능력 보존 여부를 KLUE-NLI 선형 프로빙으로 확인하며, Biderman et al.[7]의 기준을 적용하여 치명적 망각(catastrophic forgetting) 수준을 판정한다.

본 논문의 주요 기여는 다음과 같다.

- 한자 부수를 대조 학습의 지도 신호로 활용하는 한국어 특화 임베딩 파인튜닝 방법론을 제안한다.
- 학습에 사용되지 않은 단어에 대한 구조적 일반화 능력을 k-NN 기반 hold-out 평가로 직접 검증하는 것을 본 연구의 핵심 평가 태스크로 삼는다. WSD는 임베딩 품질의 보조적 검증 수단으로 활용한다.
- 경량 PEFT(LoRA)[4]가 소규모 도메인 데이터(~19K 쌍) 환경에서 전체 파인튜닝을 상회함을 실험적으로 검증한다.
- KLUE-NLI 선형 프로빙을 통해 파인튜닝 후 인코더의 일반 표현 능력이 최소 수준의 변화에 그침을 확인한다.

II. Preliminaries

1. Related works

1.1 Korean Pre-trained Language Models

BERT(Bidirectional Encoder Representations from Transformers)[9]는 대규모 텍스트 코퍼스에서 양방향 문맥을 학습하는 인코더 전용 Transformer 모델로, 다양한 NLP 태스크에서 사전 학습 후 파인튜닝(fine-tuning) 패러다임의 기반을 마련하였다. Park et al.[3]은 이를 한국어로 확장하여 KLUE BERT-base를 제안하였다. KLUE BERT-base는 12개의 어텐션 레이어, 12개의 헤드, 768차원의 은닉 상태를 가지며 약 1억 1천만 개의 파라미터로 구성된다. 한국어의 형태론적 특성을 반영하기 위해 형태소 기반 서브워드 토크나이징을 채택하였으며, 한국어

NLP 태스크의 표준 기준 모델로 널리 활용된다. 그러나 이 모델은 서브워드 단위로 텍스트를 처리하므로, 한자 부수와 같은 자소 하위 수준의 의미 구조는 명시적으로 모델링 되지 않는다.

1.2 Parameter-Efficient Fine-Tuning and LoRA

사전 학습 언어모델의 전체 파라미터를 파인튜닝 하는 방식은 소규모 도메인 데이터 환경에서 과적합(overfitting) 및 치명적 망각(catastrophic forgetting)의 위험을 내포한다. 이를 완화하기 위해 다양한 파라미터 효율적 파인튜닝(Parameter-Efficient Fine-Tuning, PEFT) 기법이 제안되었다. 그 중 저랭크 적응(Low-Rank Adaptation, LoRA)[4]는 Transformer의 어텐션 가중치 행렬에 저랭크 행렬 분해를 통해 추가적인 학습 파라미터를 삽입하는 방법으로, 추론 시 오버헤드 없이 원래 가중치와 병합이 가능하다는 장점이 있다. 구체적으로, 동결된 가중치 행렬 W에 저랭크 행렬 B와 A의 곱 BA를 더함으로써($W_{new} = W + BA$) 소수의 파라미터만으로 적응이 이루어진다. Biderman et al.[7]은 LoRA가 전체 파인튜닝 대비 도메인 외 태스크에서의 성능 저하, 즉 망각이 적음을 실증적으로 보였다.

1.3 Contrastive Learning and InfoNCE

대조 학습(Contrastive Learning)은 양성 쌍(positive pair)은 임베딩 공간에서 가깝게, 음성 쌍(negative pair)은 멀게 만드는 학습 패러다임이다. Chen et al.[6]은 SimCLR을 통해 Projection Head와 in-batch negative 전략을 결합한 InfoNCE(NT-Xent) 손실함수의 유효성을 시각적 표현 학습에서 검증하였다. 이를 자연어 처리로 확장한 SimCSE[5]는 dropout 기반 비감독 증강과 NLI 기반 감독 대조 학습을 통해 문장 임베딩의 품질을 크게 향상시켰으며, KoSimCSE[8]은 이를 한국어로 적용한 모델이다. 본 연구는 부수 그룹을 기반으로 단어 수준의 양성/음성 쌍을 구성하고, InfoNCE 손실함수를 적용한다. 분류(cross-entropy) 기반 접근과 달리, InfoNCE는 클래스 수에 무관하게 의미 유사도 신호를 학습할 수 있어 부수 그룹 수의 변화에도 확장성이 높다.

1.4 Hanja Structure in NLP

한자 구조를 언어모델에 통합하려는 시도는 중국어 NLP 분야에서 활발히 이루어져 왔다. Shi and Liu[10]은 한자의 부수 정보를 문자 임베딩에 결합한 Radical Embedding을 제안하였으며, Yin et al.[11]은 이를 다입도(multi-granularity) 임베딩으로 확장하였다. 한국어의 경우, 기존 연구들은 주로 자모(字母) 단위의 형태론적 특성을 모델에 반영하는 데 초점을 두었으나[3], 한자어의 어원적 구조인 부수 정보를 대조 학습의 지도 신호로 직접 활용하는 연구는 아직 시도된 바 없다. 본 연구는 이 공백을 채우는 것을 목표로 한다.

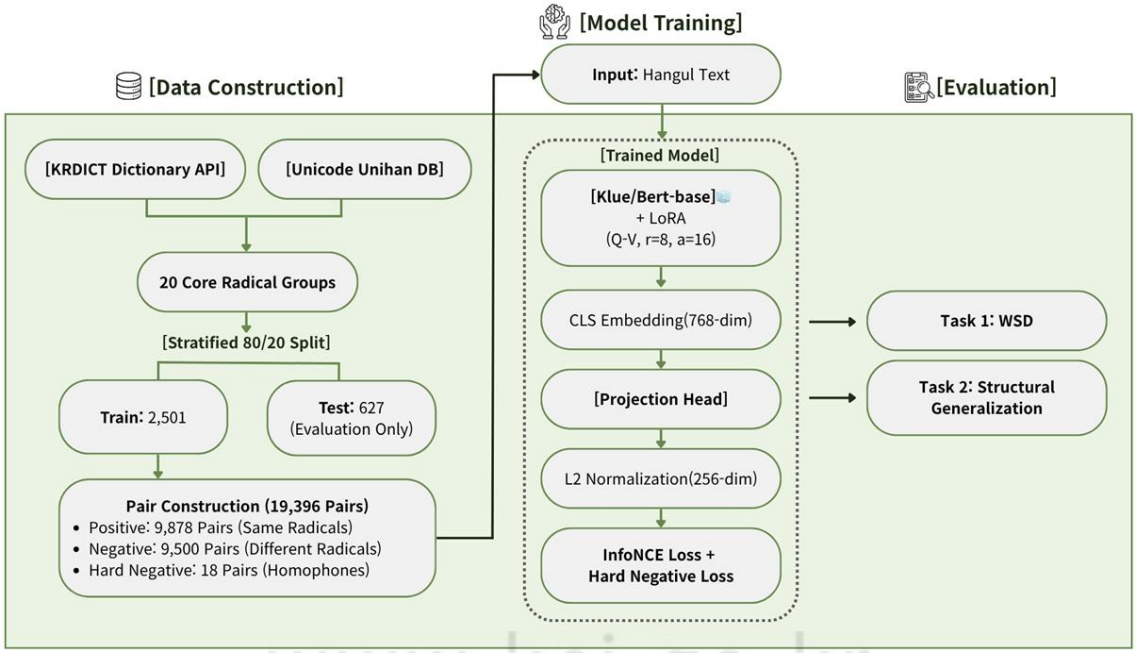

Fig. 1. Overall Framework of the Proposed Method

2. Background

2.1 Hanja Radical System

한자 부수(部首)는 한자를 의미적으로 분류하기 위한 전통적인 체계로, Unicode Unihan DB는 214개의 부수 정보를 포함한다. 현대 한국어에서 사용되는 한자어는 이 부수 체계를 통해 상위 의미 범주로 분류될 수 있다. 예를 들어, '水(수)' 부수를 포함하는 한자어들은 물·액체 관련 의미를, '心(심)' 부수를 포함하는 한자어들은 마음·감정 관련 의미를 공유하는 경향이 있다. 이러한 부수-의미 대응 관계는 한자어 어휘 전반에 걸쳐 일관되게 관찰되며, 본 연구는 이를 임베딩 공간의 군집 구조를 위한 지도 신호로 활용한다.

2.2 Word Sense Disambiguation for Sino-Korean Words

한자어 중의성 해소(Word Sense Disambiguation, WSD)는 동일한 한글 표기를 가지나 서로 다른 한자 표기를 가지는 단어—예컨대 '수심(水深)'과 '수심(愁心)'—를 문맥에 기반하여 올바른 의미로 분류하는 태스크이다. 기존 NLP 연구에서 WSD는 주로 사전 정의(dictionary definition)와 문장 임베딩의 유사도를 활용하여 접근된다[12]. 본 연구에서는 문장의 CLS 임베딩과 각 후보 '한자(뜻풀이)' 임베딩 사이의 코사인 유사도를 비교하여 최적 후보를 선택하는 방식을 채택한다.

III. Proposed Method

1. Overview

본 장에서는 한자 부수 기반 대조 학습으로 한국어 BERT 임베딩을 파인튜닝하는 전체 프레임워크를 기술한다. 학습에 사용되는 입력은 한글 텍스트만이며, 한자 부수 정보는 학습 쌍(pair) 구성 시 레이블로만 활용되고 추론 시에는 사용되지 않는다. Fig. 1은 데이터 구축, 모델 학습, 평가의 세 단계로 구성된 전체 프레임워크를 나타내며, 각 단계의 상세 내용은 이후 절에서 기술한다.

2. Data Construction

2.1 Radical Group Selection and Word Collection

전체 214개의 한자 부수 중, 한국어 어휘 빈도를 고려하여 의미적으로 뚜렷하고 부수당 최소 35개 이상의 한국어 단어를 보유하는 20개의 핵심 부수를 선정하였다. 이 기준은 충분한 대조 신호를 확보하면서 의미적 다양성을 유지하기 위한 균형점이다. Unicode Unihan DB를 통해 한자-부수 매핑을 구성하고, KRDICT 사전 API를 통해 각 부수에 해당하는 한국어 단어를 수집하여 총 3,128개의 단어를 구성하였다. 데이터 구축 과정에서 다음과 같은 기준을 적용하였다. 첫째, 한 단어가 복수의 한자로 구성되는 경우(예: '수심(水深)'은 水와 深 두 한자로 구성), 해당 단어의 의미를 가장 직접적으로 결정하는 한자의 부수를 대표 부수로 선정하였다. 이는 Unicode Unihan DB의 kRSUnicode 필드에 기록된 부수 정보를 우선적으로 참조하였으며, 의미적으로 모호한 경우 KRDICT의 뜻풀이를 기준으로 수동 검토하였다. 둘째, 동일 단어가 복수의 부수 그룹에 속할 수 있는 경우, 해당 단어를 하나의 부수 그룹에만 배정하여 그룹 간 중복을 허용하지 않았다. 셋째, 부수 매핑 오류를 최소화하기 위해 20개 부수 그룹 각각에 대해 무작위 표본 20개를 추출하여 수동 검토를 수행하였으며, 오류율은 전체의 1% 미만으로 확인되었다. Table 1은 대표 부수 및 전체 Train/Test 분포를 요약하여 나타낸다.

Table 1. Core Radical Groups and Word Counts

Radical	Meaning	Total	Train	Test
刀(18)	Cutting	317	254	63
木(75)	Wood	298	238	60
人(9)	Human	250	200	50
水(85)	Water	244	195	49
⋮	⋮	⋮	⋮	⋮
山(46)	Mountain	35	28	7
Total (20개)		**3,128**	**2,501**	**627**

2.2 Train/Test Split and Pair Construction

데이터 누수를 방지하기 위해 각 부수 그룹 내에서 계층적 80/20 분할(stratified split, seed=42)을 수행하여 학습 단어 2,501개와 평가 단어 627개를 분리하였다. 평가 단어는 학습 쌍 구성에 일절 포함되지 않는다.

학습 단어를 기반으로 세 종류의 쌍을 구성하였다. 양성 쌍(positive pair)은 동일 부수 그룹에 속하는 두 단어의 조합이며, 음성 쌍(negative pair)은 서로 다른 부수 그룹의 단어 조합이다. 추가로, 동음이의어 쌍(hard negative)을 명시적으로 포함하여 발음이 동일하지만 한자가 다른 단어들 사이의 의미 차이를 직접 학습하도록 하였다. Table 2는 학습 쌍의 구성을 나타낸다.

Table 2. Training Pair Composition

Pair Type	Count
Positive	9,878
Negative	9,500
Hard Negative	18
Total	**19,396**

3. Model Architecture

3.1 Backbone and LoRA

기반 모델로 KLUE BERT-base(110.6M 파라미터, 12레이어, 12헤드, 768차원 은닉 상태)를 사용한다[3]. BERT 인코더 전체를 동결(frozen)하고, 각 레이어의 어텐션 Query(Q) 및 Value(V) 가중치 행렬에 LoRA[4]를 적용한다. LoRA는 동결된 가중치 행렬 $W \in \mathbb{R}^{d \times d}$에 저랭크 행렬 $B \in \mathbb{R}^{d \times r}$와 $A \in \mathbb{R}^{r \times d}$를 추가하여 다음과 같이 표현된다.

$$W_{new} = W + BA$$

rank r=8, α=16으로 설정하며, 12개 레이어의 Q·V 행렬에 적용한 LoRA 학습 파라미터는 294,912개이다. BERT 인코더를 동결하는 것은 사전 학습된 한국어 언어 능력을 보존하고, 소규모 데이터(~19K쌍)에서의 과적합을 방지하기 위함이다.

3.2 Projection Head

LoRA가 적용된 BERT의 CLS 임베딩(768차원)을 대조 학습 전용 공간으로 사영하기 위해 SimCLR[6]의 설계를 참고하여 Projection Head를 구성한다. 구조는 다음과 같다.

CLS (768) → Linear (768→768) → LayerNorm → GELU → Linear (768→256) → L2 normalization

학습 시에는 Projection Head의 출력(256차원)을 InfoNCE 손실 계산에 활용한다. 평가 Task 1(WSD)에서는 문장 의미 비교를 위해 CLS 임베딩(768차원)을 직접 사용하며, Task 2(구조적 일반화)에서는 Projection Head 출력(256차원)을 활용한다. Projection Head의 학습 파라미터는 788,992개이며, LoRA 파라미터를 합산한 총 학습 파라미터는 1,083,904개로 전체의 0.98%에 해당한다. Table 3은 모델 파라미터 구성을 나타낸다.

Table 3. Model Parameter Configuration

Component	Parameters	Trainable
KLUE BERT-base	110,617,344	Frozen
LoRA	294,912	✓
Projection Head	788,992	✓
Trainable Total	**1,083,904 (0.98%)**	

4. Training Objective

4.1 InfoNCE Loss

배치 내 B개의 양성 쌍$(e_1^{(i)}, e_2^{(i)})$에 대해 양방향 cross-entropy 기반의 InfoNCE 손실을 적용한다.

$$L_{InfoNCE} = \frac{1}{2}\left[CE\left(\frac{e_1 \cdot e_2^T}{\tau}, I\right) + CE\left(\frac{e_2 \cdot e_1^T}{\tau}, I\right)\right]$$

$$s_{ij} = \frac{\cos(e_1^{(i)}, e_2^{(j)})}{\tau}$$

배치 내 다른 양성 쌍의 단어들이 자동으로 음성 예시(in-batch negative)로 작용한다. 온도 파라미터 τ=0.15는 하이퍼파라미터 탐색을 통해 결정되었다. 분류 기반 손실함수와 달리 InfoNCE는 부수 클래스 수에 무관하게 의미 유사도 신호를 학습할 수 있어 확장성이 높다.

4.2 Hard Negative Loss

$$L_{hard} = \max\left(0,\ m - \left(1 - \frac{e_a \cdot e_b}{|e_a||e_b|}\right)\right)$$

동음이의어 쌍에 대해 힌지 마진 손실(hinge margin loss)을 추가 적용하여 동음이의어 쌍이 임베딩 공간에서 명시적으로 멀어지도록 유도한다. 최종 학습 손실은 $L = L_{InfoNCE} + \lambda \cdot L_{hard}$이며, λ=1.0, margin=0.5로 설정하였다. 단, 본 연구에서 수집된 동음이의어 쌍은 18쌍으로 규모가 제한적이므로, Hard Negative Loss는 WSD 성능을 직접 최적화하기 위한 목적보다는 임베딩 공간에서 동음이의어 간의 명시적 분리를 보조하는 정규화 역할로 해석함이 적절하다. WSD에 특화된 성능 향상을 목표로 할 경우 hard negative 데이터의 대규모 확장이 필요하며, 이는 향후 연구 과제로 남긴다.

4.3 Hyperparameter Search

최적 하이퍼파라미터는 rank → temperature → dropout의 3단계 순차 탐색(seed=42)을 통해 결정하였다. 각 단계에서 이전 단계의 최적값을 고정한 채 다음 파라미터를 탐색하는 방식을 취하며, 평가 지표로 사용된 Train R@1은 학습 단어 집합 내에서 수행한 자기 검색

(self-retrieval) Recall@1이며, 최종 hold-out 평가에 사용되는 627개 평가 단어는 하이퍼파라미터 탐색에 사용하지 않았다.

Table 4. LoRA Rank Ablation (τ=0.10, dropout=0.1, seed=42)

rank (r)	α	Trainable Params	Train R@1	MRR	Best Epoch
2	4	862,720	0.7018	0.7517	95
4	8	936,448	0.6986	0.7514	90
8	**16**	**1,083,904**	**0.7113**	**0.7618**	**45**
16	32	1,378,816	0.7033	0.7511	40
32	64	1,968,640	0.7097	0.7619	105

Table 4에서 rank r=8이 Train R@1 0.7113으로 가장 높은 성능을 보이며, Best Epoch 45로 빠르게 수렴한다. r=2, 4는 표현력 부족으로 성능이 낮고, r=16, 32는 학습 파라미터가 늘어남에도 r=8 대비 성능 향상이 없어 소규모 데이터 환경에서 과도한 랭크가 불필요함을 시사한다. 이에 r=8, α=16을 고정하고 다음 단계로 진행한다.

Table 5. Temperature Ablation (r=8, dropout=0.1, seed=42)

τ	Train R@1	Train R@5	Train R@10	MRR
0.05	0.6842	0.8102	0.8692	0.7449
0.07	0.6858	0.8102	0.8596	0.7501
0.1	0.7113	0.8134	0.8437	0.7618
0.15	**0.7289**	**0.7943**	**0.823**	**0.7643**
0.2	0.7129	0.7879	0.8262	0.7539

Table 5에서 τ=0.15가 Train R@1 0.7289, MRR 0.7643으로 가장 우수한 성능을 보인다. τ가 낮을수록 (0.05, 0.07) R@1은 낮으나 R@10은 높아지는 경향이 있는데, 이는 낮은 온도가 임베딩 공간을 과도하게 집중시켜 근거리 검색 정밀도는 떨어지되 광범위한 군집화에는 유리하게 작용하기 때문으로 해석된다. τ=0.15는 R@1과 MRR을 동시에 최대화하는 균형점으로, 이를 고정하고 다음 단계로 진행한다.

Table 6. Dropout Ablation (r=8, τ=0.15, seed=42)

Dropout	Train R@1	Train R@5	MRR
0	0.7193	0.7959	0.7604
0.05	0.7241	0.799	0.7621
0.1	**0.7289**	**0.7943**	**0.7643**

Table 6에서 dropout=0.1이 Train R@1 0.7289, MRR 0.7643으로 가장 높은 성능을 보인다. dropout=0.0 대비 일관된 성능 향상이 관찰되어, 소규모 데이터 환경에서 dropout이 정규화 효과를 제공함을 확인한다.

위 탐색 결과로부터 최적 설정을 r=8, α=16, τ=0.15, dropout=0.1로 확정하였다. 최종 모델은 이 설정으로 seed=42, 1, 2의 3-seed 실험을 수행하여 결과의 통계적 신뢰성을 확보하였다

IV. Experimental Results

1. Experimental Setup

1.1 Evaluation Tasks

Task 1(한자어 중의성 해소, WSD)은 보조 평가 태스크로, 동음이의 한자어가 등장하는 484개 문항에서 올바른 한자 의미를 선택하는 태스크이다. 평가 데이터는 다음과 같이 구축하였다. KRDICT 사전 API를 통해 동일한 한글 표기를 가지나 서로 다른 한자 표기를 지닌 동음이의어 목록을 수집하고, 각 동음이의어에 대해 KRDICT가 제공하는 용례 문장을 평가 문장으로 활용하였다. 각 문항은 하나의 평가 문장과 두 개 이상의 후보 한자 표기(뜻풀이 포함)로 구성되며, KRDICT 용례의 표제어 정보를 정답 레이블로 사용하였다. 정답 레이블은 KRDICT 용례의 표제어 정보를 기준으로 자동 부여하였으며, 품질 검증을 위해 전체 484개 문항 중 무작위로 추출한 214개 문항(95% 신뢰수준, 오차범위 ±5%)에 대해 연구자가 직접 용례 문장과 정답 레이블의 일치 여부를 수동 검수하였다. 검수 결과 오류율은 1% 미만으로 확인되어 데이터 품질이 양호함을 검증하였다. 학습 단어(2,501개)가 포함된 문항은 사전에 제거하여 데이터 누수를 방지하였으며, 최종 484개 문항이 평가에 사용되었다. 평가 방식은 문장 내 대상 단어를 [MASK]로 치환한 후, 문장 CLS 임베딩(768차원)과 각 후보 '한자(뜻풀이)' 임베딩 간의 코사인 유사도를 계산하여 최고 유사도 후보를 선택한다.

1.2 Baselines

비교 모델은 메인 비교와 어블레이션으로 구분한다. 메인 비교에서는 네 가지 모델을 사용한다. Zero-shot BERT는 파인튜닝 없는 KLUE BERT-base로 성능 하한을 나타내며, BM25는 어휘적 중복(lexical overlap) 기반의 비신경망 기준선이다. KoSimCSE[8]은 SimCSE 방법론을 한국어로 확장한 공개 문장 임베딩 모델로, 별도의 한자어

특화 학습 없이 범용 한국어 코퍼스만으로 학습된 모델이다. 본 연구에서는 이를 강력한 범용 한국어 임베딩 기준선으로 활용한다.

어블레이션에서는 다섯 가지 변형 모델을 비교한다. B3(random grouping)는 동일한 LoRA 아키텍처를 사용하되 단어를 무작위로 20개 그룹에 배정하여 부수 기반 그루핑의 효과를 검증한다. B2(flat grouping)는 모든 한자어를 단일 그룹으로 취급한다. Proj-only는 BERT를 완전히 동결하고 Projection Head만 학습하여 LoRA의 기여도를 측정한다. BERT Full FT는 전체 12개 레이어를 언프리즈(unfreeze)하여 동일한 학습 데이터와 목적함수(InfoNCE)로 학습한 모델로, 소규모 데이터에서 LoRA 대비 전체 파인튜닝의 효과를 비교한다. 공정한 비교를 위해 B2, B3, Ours는 모두 동일한 LoRA 아키텍처(r=8, τ=0.15, dropout=0.1)를 사용하며, 차이는 그루핑 레이블뿐이다.

2. Main Results

Table 7은 Task 1(WSD)과 Task 2(구조적 일반화)에 대한 메인 비교 결과를 나타낸다. Ours는 3-seed(seed=42, 1, 2) 평균 및 표준편차로 표기된다.

Table 7. Main Comparison Results

Model	Task 1 WSD Acc.	Task 2 k-NN Acc@1
Zero-shot BERT	0.7707	0.4258
BM25	0.4917	0.6443
KoSimCSE	**0.8554**	0.5758
Ours	0.8402 ± 0.0059	**0.7007 ± 0.0112**

Task 1(WSD)에서 Ours(0.8402)는 KoSimCSE(0.8554)에 1.5%p 차이로 근접하며, Zero-shot BERT(0.7707) 대비 7.0%p 향상을 보인다. Ours가 KoSimCSE에 소폭 미치지 못하는 것은, KoSimCSE가 대규모 일반 코퍼스에서 문장 수준의 의미 유사도를 집중적으로 학습한 반면, 본 연구는 단어 수준의 부수 그룹 신호(~19K쌍)만으로 학습하였기 때문으로 해석된다. 그럼에도 소규모 한자어 특화 데이터만으로 대규모 범용 모델에 준하는 WSD 성능에 도달하였다는 점은 유의미하다. BM25(0.4917)의 낮은 성능은 어휘적 접근만으로는 한자어 의미 구분에 불충분함을 시사한다.

Task 2(구조적 일반화)에서 Ours(0.7007)는 Zero-shot BERT(0.4258) 대비 27.5%p, KoSimCSE(0.5758) 대비 12.5%p 향상을 달성한다. 특히 Task 1에서 강세를 보인 KoSimCSE(0.5758)가 Task 2에서는 BM25(0.6443)에도 미치지 못하는 결과는, 문장 수준 의미 유사도에 특화된 범용 임베딩이 부수 기반 의미 군집 구조를 포착하지 못함을 보여준다.

두 태스크의 결과를 종합하면, KoSimCSE는 WSD에는 강하지만 구조적 일반화에는 한계를 지니는 반면, Ours는 두 태스크 모두에서 균형 있는 성능을 달성한다. 이는 소규모 도메인 데이터(~19K쌍)만으로도 한자어 특화 태스크에서 범용 모델을 웃돌 수 있다는 본 연구의 핵심 주장을 지지한다. 특히 Task 1과 Task 2에서 모델 간 순위가 역전되는 양상은, 두 태스크가 서로 다른 임베딩 능력을 측정하고 있음을 보여주며, 단일 태스크 평가만으로는 임베딩 모델의 한자어 표현 능력을 온전히 평가하기 어렵다는 점을 시사한다.

Table 7의 정량적 결과를 보완하기 위해 Fig. 2에 test

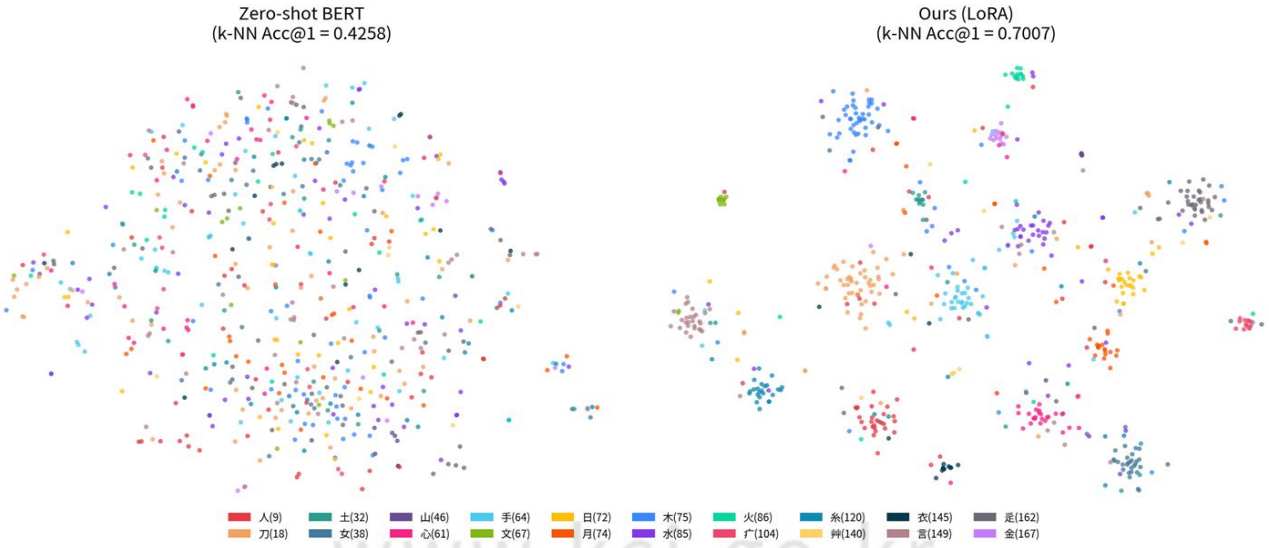

Fig. 2. t-SNE Visualization of Test Word Embeddings (627 words, 20 radicals)

단어 627개의 t-SNE 시각화를 제시한다. Zero-shot BERT(k-NN Acc@1=0.4258)의 경우 20개 부수에 해당하는 단어들이 임베딩 공간에서 뚜렷한 군집 구조 없이 혼재되어 있음을 확인할 수 있다. 반면 Ours(k-NN Acc@1=0.7007)는 동일 부수에 속하는 단어들이 공간적으로 응집된 군집을 형성하며, 부수 간 경계가 명확하게 나타난다. 이는 제안 방법이 한자 부수 기반의 의미 구조를 임베딩 공간에 효과적으로 인코딩함을 시각적으로 뒷받침한다.

3. Ablation Study

Table 8은 어블레이션 결과를 나타낸다.

Table 8. Ablation Study Results

Model	Task 1 WSD Acc.	Task 2 k-NN Acc@1
Zero-shot BERT	0.7707	0.4258
B2 (flat grouping)	0.8037	0.5460
B3 (random grouping)	0.8182	0.5375
Proj-only	0.7707	0.5534
BERT Full FT	0.7542 ± 0.0029	0.6858 ± 0.0065
Ours	**0.8402 ± 0.0059**	**0.7007 ± 0.0112**

각 단계의 결과는 다음과 같이 해석된다. Zero-shot에서 B2로의 변화(Task 2: 0.4258 → 0.5460)는 한자어 전체를 하나의 의미 범주로 취급하는 것만으로도 대조 학습이 부분적인 구조 신호를 제공함을 보여준다. B2에서 B3로의 변화(0.5460 → 0.5375)는 무작위 그루핑이 오히려 flat 그루핑보다 소폭 낮은 성능을 보임을 나타내며, 이는 의미 없는 무작위 레이블이 대조 학습의 신호 품질을 저하시킬 수 있음을 시사한다. Proj-only는 Task 2에서 0.5534를 기록하여 Projection Head 학습만으로도 일정 수준의 개선이 가능함을 보여준다. 그러나 Ours(0.7007)와 비교하면 LoRA를 통한 인코더 적응 없이는 부수 기반 구조를 충분히 반영하는 데 한계가 있음을 시사한다. BERT Full FT는 Task 2에서 0.6858로 큰 향상을 보이나, Task 1 WSD에서는 Zero-shot(0.7707) 대비 오히려 성능이 저하(0.7542)되어 소규모 데이터 환경에서의 과적합 또는 치명적 망각을 시사한다. 최종적으로 Ours는 부수 기반 그루핑과 LoRA를 결합하여 두 태스크 모두에서 최고 성능을 달성하며, 소규모 데이터 환경에서 LoRA가 전체 파인튜닝보다 효과적임을 검증한다.

4. Per-Radical Performance Analysis

Table 9는 20개 부수 각각에 대한 Ours와 B2의 Task 2 hold-out k-NN 성능 비교를 나타낸다(3-seed 평균). Ours는 모든 20개 부수에서 B2를 웃돌며, Wilcoxon 부호순위 검정(W=210.0, $p<0.0001$)으로 차이의 통계적 유의성이 확인된다.

Table 9. Per-Radical Task 2 Hold-out k-NN Acc@1 Comparison (Ours vs. B2, 3-seed Average)

Radical	Test Words	Ours	B2	Δ
金(167)	15	0.9333	0.6222	+0.3111
疒(104)	14	0.9048	0.7857	+0.1191
女(38)	32	0.8750	0.6979	+0.1771
⋮	⋮	⋮	⋮	⋮
人(9)	50	0.5267	0.3467	+0.1800
艸(140)	9	0.4444	0.4074	+0.0370
Average (20)		**0.7161**	**0.5634**	**+0.1527**

성능 상위 부수는 金(0.9333), 疒(0.9048)으로 의미적 일관성이 높은 그룹에서 뚜렷한 군집화가 이루어짐을 확인할 수 있다. 단어 수와 성능 간의 Spearman 상관계수는 $r=-0.260$($p=0.269$)로 통계적으로 유의하지 않아, 성능 결정 요인이 단어 수가 아닌 부수의 의미적 일관성임을 시사한다.

5. Catastrophic Forgetting Verification

파인튜닝 후 BERT 인코더의 일반 언어 이해 능력이 손상되었는지 확인하기 위해 KLUE-NLI 선형 프로빙을 수행하였다. BERT 인코더를 완전히 동결한 채 선형 분류 헤드만 학습(30 epoch)하여 함의/중립/모순 3분류 정확도를 측정하였다. Table 10은 Zero-shot BERT와 LoRA 적용 모델의 동결 인코더 기반 선형 프로빙 결과를 나타낸다.

Table 10. KLUE-NLI Linear Probing Results

Model	NLI Accuracy
Zero-shot BERT	0.5470
Ours (LoRA, 3-seed Average)	0.5138 ± 0.0064
Difference (Ours − Zero-shot)	−0.0332

Biderman et al.[7]의 기준에 따르면, 차이가 −0.05 이상 −0.01 미만인 경우 '최소 표현 변화(minimal drift)'로 판정된다. 본 실험에서 −0.0332의 차이는 이 범주에 해당하며, LoRA 파인튜닝이 한자어 특화 태스크에 적응하는 과정에서 불가피하게 발생하는 미세한 표현 변화임을

확인한다. 이는 Biderman et al.[7]의 "LoRA forgets less than full fine-tuning" 결론과 일치하며, 제안 방법이 일반 언어 이해 능력을 유의미하게 손상시키지 않음을 시사한다.

V. Conclusions

본 논문에서는 한자 부수를 대조 학습의 지도 신호로 활용하여 한국어 BERT 임베딩을 경량 파인튜닝 하는 방법론을 제안하였다. 20개의 핵심 부수 그룹으로부터 구성된 19,396쌍의 학습 데이터와 InfoNCE 손실함수, 그리고 LoRA를 결합하여 전체 파라미터의 0.98%만을 학습함으로써 한자어 의미 구조를 임베딩 공간에 효과적으로 인코딩하였다.

실험 결과, 본 연구의 핵심 평가 태스크인 구조적 일반화(Task 2)에서 k-NN Acc@1 0.7007을 달성하여 KoSimCSE(0.5758)를 포함한 모든 비교 모델을 상회하였다. 보조 평가인 한자어 중의성 해소(Task 1)에서는 소규모 도메인 데이터만으로 KoSimCSE(0.8554)에 근접한 0.8402의 정확도를 달성하였으며, 이는 제안 방법이 WSD에도 유효한 임베딩 표현을 학습함을 시사한다. 어블레이션 분석은 부수 기반 그루핑과 LoRA가 각각 독립적인 기여를 하며, 소규모 데이터(~19K쌍) 환경에서 LoRA가 전체 파인튜닝보다 효과적임을 확인한다. 아울러, KLUE-NLI 선형 프로빙을 통해 파인튜닝 후 인코더 표현의 변화가 최소 수준(−0.0332)에 그침을 검증하였다.

본 연구의 한계는 다음과 같다. 제안 방법은 한자 어원을 가지는 한자어(한국어 어휘의 약 66~70%[1]) 범위 내에서만 유효하며, 순수 고유어('하늘', '바람' 등)에는 직접 적용이 불가하다. 또한 학습 데이터는 KRDICT 기반 20개 부수, 3,128개 단어로 제한되어 있어, 부수 범위 확장 및 더 많은 어휘를 포함한 검증이 향후 과제로 남는다.

향후 연구로는 부수 범위를 20개에서 확장하여 더 넓은 한자어 어휘를 포괄하는 방향, 그리고 제안된 임베딩을 한자어 기반 정보 검색 및 어휘 의미망 구축에 직접 적용하는 방향을 고려할 수 있다.

REFERENCES

[1] C. Heo, "Analysis of Korean Vocabulary Composition and Investigation of Chinese Character Word Formation Ability Based on Korean Dictionary Entries," Dongbang Hanmunhak, 2008. https://kiss.kstudy.com/Detail/Ar?key=2803855

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems, Vol. 30, pp. 5998-6008, Long Beach, USA, Dec. 2017. DOI: 10.48550/arXiv.1706.03762

[3] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, S. Kim, K. Lim, J. Lee, K. Park, J. Shin, S. Kim, L. Park, A. Oh, J. Ha, and K. Cho, "KLUE: Korean Language Understanding Evaluation," Advances in Neural Information Processing Systems, Vol. 34, pp. 1-52, Dec. 2021. DOI: 10.48550/arXiv.2105.09680

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations, Apr. 2022. DOI: 10.48550/arXiv.2106.09685

[5] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6894-6910, Punta Cana, Dominican Republic, Nov. 2021. DOI: 10.18653/v1/2021.emnlp-main.552

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," Proceedings of the 37th International Conference on Machine Learning, pp. 1597-1607, Virtual, USA, Jul. 2020. DOI: 10.48550/arXiv.2002.05709

[7] D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham, "LoRA Learns Less and Forgets Less," Transactions on Machine Learning Research, Aug. 2024. DOI: 10.48550/arXiv.2405.09673

[8] BM-K, "Sentence-Embedding-Is-All-You-Need," https://github.com/BM-K/Sentence-Embedding-Is-All-You-Need

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, Minneapolis, USA, Jun. 2019. DOI: 10.18653/v1/N19-1423

[10] T. Shi and C. Liu, "Radical Embedding: Delving Deeper to Chinese Radicals," Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pp. 594-598, Beijing, China, Jul. 2015. DOI: 10.18653/v1/P15-2098

[11] Y. Yin, F. Wei, L. Dong, K. Xu, M. Zhang, and M. Zhou, "Multi-Granularity Chinese Word Embedding," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 981-986, Austin, USA, Nov. 2016. DOI:

10.18653/v1/D16-1100

[12] R. Navigli, "Word Sense Disambiguation: A Survey," ACM Computing Surveys, Vol. 41, No. 2, pp. 1-69, Feb. 2009. DOI: 10.1145/1459352.1459355

Authors

Euna Song received the B.S. degree in Chinese Language and Literature from Kyung Hee University, South Korea, in 2025. She is currently pursuing the M.S. degree in Information Systems at Yonsei University, Seoul, South Korea. She is interested in Natural Language Processing and Data Science.

Beak-Cheol Jang received the Ph.D. degree 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.