ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2024 March: 30(1): 179~199 http://dx.doi.org/10.13088/jiis.2024.30.1.179

KoSentEval: 한국어 문장 임베딩 평가 연구*

정민화
연세대학교 디지털애널리틱스융합협동과정
(minalang@yonsei.ac.kr)

송민
연세대학교 문헌정보학과
(min.song@yonsei.ac.kr)

최근 다량의 텍스트 데이터를 하나의 밀집 벡터로 압축하는 벡터 데이터베이스가 주목받으며 언어의 의미적/통사적 정보를 통합한 전역적인 문장 임베딩을 개발하고 이를 평가할 수 있는 체계를 구축하려는 노력이 이어지고 있다. 영어연구에서는 문장 임베딩을 평가할 수 있는 SentEval라이브러리가 개발되어 폴란드어, 러시아어 등으로 확산되고 있지만 한국어 연구에서는 데이터 공개 관련 문제로 인해 연구 및 상업적인 활용이 어렵다는 한계가 있다. 본 논문에서는 이러한 한계를 극복하기 위해 KoSentEval을 공개 및 배포한다. KoSentEval은 한국어 문장 임베딩을 평가하기 위한 오픈소스 라이브러리로, 2개의 하위 태스크와 8개의 프로빙 태스크로 구성되어 임베딩의 표층적, 통사적, 의미적 속성을 검증할 수 있다. 또한, 다국어, 한국어, 대조학습 기반 문장 인코더를 적용한 결과를 비교 및 분석하였다. 본 연구는 한국어 문장 임베딩 품질의 측정뿐만 아니라, 다양한 문장 인코더가 한국어의 언어적 특성을 얼마나 잘 담고 있는지를 검증하는 데 의의가 있다.

주제어 : 문장 임베딩, 임베딩 평가, 한국어 문장 임베딩, 프로빙 태스크, 대조학습

논문접수일 : 2024년 2월 15일 논문수정일 : 2024년 3월 8일 게재확정일 : 2024년 3월 11일
원고유형 : Fast Track 교신저자 : 송민

1. 개요

문장 임베딩(sentence embedding)은 그 필요성과 유용함 덕분에 최근 자연어처리의 중요 과제로 인식되고 있다. 기존의 단어 임베딩은 동음이의어나 문맥을 파악하는 것에 단점이 있어(석주리, 2022) 의미와 문맥을 살릴 수 있는 문장 임베딩의 효과가 주목받고 있기 때문이다. 최근 Pinecone[1], Qdrant[2], Weaviate[3] 등의 벡터 데이터베이스를 통해 다량의 텍스트 데이터를 하나의 밀집 벡터(dense vector)로 압축하여 효율적으로 정보를 저장 및 검색하려는 움직임이 일어나고 있고(Morris et al., 2023), 이러한 관점에서 언어정보를 잘 표현한 문장 임베딩을 만들어내는 것이 필요하다. 또한, 문장 임베딩은 문장의 의미 파악, 문장 간 유사성 판별, 문장 분류, 질의응답 시스템 등 다양한 영역에서 유용하게 활용될 수 있다(지인영, 2023).

문장 임베딩은 문장을 하나의 벡터로 투영하는 것을 뜻한다. 하지만 단순히 단어에 대한 임베딩을 종합하여 문장을 표현하는 것은 충분치 않기 때문에(Li et al., 2022), 좋은 품질의 문장 임베딩을 개발하고 이를 평가할 수 있는 체계가 필요하다.

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1A2B5B02002359).
1) https://www.pinecone.io/
2) https://qdrant.tech/
3) https://weaviate.io/

고품질의 문장 임베딩은 단어의 지역적(local)인 정보뿐만 아니라 문장 기저의 의미적/통사적 정보를 통합한 전역적(global)인 정보를 포괄하는 것으로, 다양한 언어권에서 문장 임베딩의 품질과 속성을 평가하기 위한 태스크가 개발 및 공개되었다.

영어권에서는 문장 임베딩의 성능을 평가할 수 있는 하위 태스크(downstream task) 18개와 임베딩의 언어적 속성을 평가할 수 있는 10개의 프로빙 태스크(probing task)로 구성된 SentEval 라이브러리가 배포되어 다양한 연구에서 활용되었다(Tsukagoshi, 2021; Ni, 2022; Xu, 2023). 이후, 폴란드어(Krasnowska-Kieraś & Wróblewska, 2019), 러시아어(Mikhailov et al., 2021) 등 다양한 언어로 SentEval 라이브러리의 체계가 확장되어 각 언어의 표층적, 통사적, 의미적 특성을 반영한 새로운 프로빙 태스크가 개발되었다. 한국어에서는 안애림 등(2021)의 연구에서 한국어의 언어적 특성인 주어 생략, 부정, 경어법을 반영한 한국어 고유의 프로빙 태스크를 개발하였으나, 데이터 공개 관련 문제로 인해 연구 및 상업적 활용이 어렵다는 한계가 있었다.

따라서, 본 연구에서는 선행연구(안애림 등, 2021)의 체계를 따르되 오픈소스 데이터를 활용하여 연구 및 상업적으로 활용이 가능한 8가지의 한국어 프로빙 태스크를 개발하여 공개한다. 또한, 하위 태스크로 문장 임베딩의 품질을 검사할 수 있는 문장 유사도 태스크(Sentence Textual Similarity, STS), 대화형 인공지능 개발 시 적합 여부를 검증할 수 있는 의미 검색 태스크(Semantic Search)를 추가하여 총 10가지의 태스크로 구성된 KoSentEval 라이브러리를 배포한다.

본 연구는 아래와 같은 의의가 있다. 1) 다양한 한국어 문장 인코더의 성능 비교. 2) 각 문장 인코더가 한국어의 언어적 특성을 잘 반영하는지 검증. 3) 한국어 문장 임베딩의 품질을 평가할 수 있는 태스크 및 라이브러리 공개.

본 연구는 다음과 같이 구성되어있다. 2장에서는 관련 연구로 다양한 언어에서 개발한 문장 임베딩 평가 방법론을 비교 및 분석한다. 3장에서는 본 연구에서 개발 및 제안하는 문장 임베딩 평가 태스크를 소개한다. 4장에서는 문장 인코더를 본 연구에서 제안하는 태스크로 실험한 결과에 관해 서술하고, 그 결과를 분석한다. 5장은 연구 결과 및 결론을 논한다. 본 연구에서 개발한 라이브러리는 다음의 링크에서 사용할 수 있다. https://github.com/minalang/KoSentEval

2. 관련 연구

2.1 언어별 문장 임베딩 평가 연구

문장 임베딩에 관한 평가 연구는 영어권에서 시작되어 다양한 언어권으로 확대되었다. 특히, 문장분석, 문장추론 등 하위 태스크와 문장 길이 및 단어 예측 연구(Adi et al., 2017), 그리고 통사적 속성 연구(Shi et al., 2016)와 같이 파편적으로 존재하던 평가 방법을 종합한 SentEval 라이브러리가 공개되면서 그 구조를 모방하되 각 언어에 맞게 태스크를 변형한 연구들이 공개되었다.

먼저, 영어 SentEval 라이브러리는 18개의 하위 태스크와 10개의 프로빙 태스크로 구성되어있다. 하위 태스크의 경우 감성분석, 문장추론, 문장 유사도 판별 등 기존 언어 모델 평가에 활용되던 태스크를 종합하였다. 10가지의 프로빙 태스크는 표면적 속성, 통사적 속성, 의미적 속성을 검증하기 위한 태스크로 이루어졌다. 먼저, 문장의 표면적 속성은 문장 길이(SentLen)와 어휘 분류

(WC)태스크로 검증할 수 있다. 문장의 구조나 기능, 문장 구성요소의 속성인 통사적 속성을 평가하기 위해서는 어휘 순서변화(Bshift)와 문장을 그래프로 구조화하였을 때의 통사구조 깊이(TreeDepth), 구성요소의 종류(TopConst)에 대한 태스크를 포함한다. 마지막으로, 문장의 의미적 속성을 평가하기 위해서는 문장의 시제(Tense), 주어의 개수(SubjNum), 목적어의 개수(ObjNum), 문장 내 어휘 대치 여부(SOMO), 문장의 순서 교체 여부(CoodInv)에 대한 태스크로 구성되어있다.

폴란드어 연구(Krasnowska-Kieraś & Wróblewska, 2019)의 경우 2개의 하위 태스크와 9개의 프로빙 태스크로 구성되어있다. 하위 태스크로는 두 문장의 연관도(Relatedness)와 함의 관계(Entailment)에 대해 상관 계수를 통해 예측하는 과제를 새로이 개발하였다. 프로빙 태스크에서는 영어 연구에 없었던 문장의 태(능동, 수동)를 분류하는 태스크(Passive)와 문장 유형(의문문, 명령문, 평서문)에 대한 태스크(SentType)를 추가하고, 몇몇 태스크를 삭제하여 9개의 과제를 공개하였다.

이와 달리, 러시아어 연구(Mikhailov et al., 2021)는 7개의 프로빙 태스크로 이루어져있다. 언어에 성별이 존재한다는 러시아어의 특징을 태스크에 활용하여 주어와 목적어의 성별을 판별하는 태스크(SubjNumber, ObjNumber)를 개발하였다. 또한, 통사적 속성을 검증하기 위해 접속사의 유형(ConjType)과 문장의 비인칭 주어 활용 여부(Impersonal Sent) 태스크를, 의미적 속성으로는 문장의 상(완료/진행)을 분류하는 과제(PV)를 개발하였다.

마지막으로, 한국어는 생략이 빈번한 언어이기 때문에 주어 혹은 목적어의 수는 중요치 않으며 경어법이 중요하게 사용된다(안애림 등, 2021). 이에, 통사적 속성 검증으로 주어의 생략 여부(SubjOmission)와 의미적 속성으로 경어법(Honorifics) 분류 과제를 개발하였다. 언어별 문장 임베딩 평가 연구를 정리한 내용은 <표 1>과 같다.

〈표 1〉 언어별 문장 임베딩 태스크

	영어	폴란드어	러시아어	한국어
하위 태스크	18개	2개	-	-
문장의 표층적 속성 평가				
문장 길이	SentLen	SentLen	SentLen	**SentLen**
어휘 구분	WC	WC	WC	**WC**
문장의 통사적 속성 평가				
통사적 속성	TreeDepth	TreeDepth	TreeDepth	TreeDepth
서술어의 논항	TopConst	TopDeps	-	**TopDeps**
문장성분생략	-	-	Gapping	**SubjOmission**
어휘 순서 변화	Bshift	-	Nshift	-
접속사 유형	-	-	ConjType	-
비인칭 주어	-	-	Impersonal Sent	-
문장의 의미적 속성 평가				
시제	Tense	Tense	PT	**Tense**
주어의 수	SubjNum	SubjNum	SubjNum	-
목적어의 수	ObjNum	ObjNum	ObjNum	-
주어의 성	-	-	SubjGender	-
목적어의 성	-	-	ObjGener	-
어휘 대치	SOMO	-	-	-
문장 순서교체	CoordInv	-	-	-
태(능동/수동)	-	Passive	PA	-
상(완료/진행)	-	-	PV	-
문장 유형	-	SentType	-	**SentType**
극성(긍정/부정)	-	-	-	**Negation**
경어법	-	-	-	**Honorifics**

영어, 폴란드어, 러시아어의 태스크들은 데이터와 평가 코드가 모두 깃허브(Github)를 통해 공개되어 있지만, 한국어 연구의 경우 데이터 공개 문제로 인해 연구에 활용하기 어렵다는 한계를 가진다. 이에, 본 연구에서는 오픈소스 데이터를 활용하여 데이터 공개 문제를 해결한 태스크와 코드를 공개하였다. 안애림 등(2021)의 구조를 참고한 경우 <표 1>에서 태스크 명을 강조체로 표기하였다.

3. 문장 임베딩 평가 태스크

문장 임베딩을 평가하기 위해 임베딩의 품질을 평가할 수 있는 태스크 2개와 각 임베딩이 언어적 속성을 내포하고 있는지 알아볼 수 있는 프로빙 태스크 8가지를 종합하여 총 10가지 태스크를 개발하였다. 평가를 위한 코드는 2018년 페이스북에서 개발한 SentEval 라이브러리를 활용하였다. 태스크 구성을 위해 사용한 데이터는 태스크에 따라 상이하다. 프로빙 태스크의 체계는 안애림 등(2021)을 참고하되, 사용신청 및 약정 기간 설정이 필요하지 않은 오픈소스 데이터를 활용하였다. 각 태스크의 명칭과 평가 속성, 활용 데이터, 그리고 간단한 설명은 <표 2>와 같다.

〈표 2〉 평가 태스크 개괄

태스크 명	평가 속성	활용 데이터	설명
문장 유사도	임베딩 품질	korSTS	두 문장 임베딩의 코사인 유사도와 0-5점 사이로 라벨링 된 점수 간의 상관관계를 평가
의미 검색	임베딩 품질	rlhf korean dataset	응답후보 중 질문 문장에 대해 가장 적합한 문장을 검색. 가장 유사도가 높은 문장을 정답으로 산출
문장 길이 분류	표층적	open korpos	어절을 기준으로 문장의 길이를 계산하여 분류
어휘 분류	표층적	open korpos	데이터의 중빈도 어휘 목록 1000개를 선정하여 주어진 문장을 어휘에 따라 분류
주어 생략 여부	통사적	KLUE-DP	문장의 주어 유무 판별
서술어의 논항 예측	통사적	KLUE-DP	루트 노드의 하위 노드인 서술어의 논항이 되는 문장 구성성분 예측
시제	의미적	open korpos	문장의 시제(과거, 비과거) 분류
극성분류	의미적	NSMC	문장의 극성(부정, 긍정) 분류
문장 유형 분류	의미적	styleKQC+ paraKQC	문장의 유형(선택의문문, 설명의문문, 요구, 금지) 분류
경어법 분류	의미적	StyleKQC +Korean Smile Style Dataset	존댓말, 반말 문장 분류

3.1 임베딩의 품질 측정

3.1.1 문장 유사도 태스크

문장 유사도 태스크는 두 문장의 의미적인 유사도를 0에서 5 사이의 점수로 라벨링 한 데이터를 활용하여 두 문장 임베딩의 코사인 유사도와 점수 간의 상관관계를 평가하는 태스크(Conneau et al., 2018)로 문장 임베딩의 품질을 평가하기 위해 주요하게 사용되어왔다.

본 연구에서는 영어의 STS-B 데이터를 번역한 KorSTS데이터(Ham et al., 2020)에서 train, dev, test 데이터를 합해 총 8,583개를 활용하였다. KorSTS 데이터는 유영현 외(2022)등 다수의 한국어 임베딩 연구에서 평가에 활용되었다. 또한, 그림에 대한 설명을 담은 캡션(main-captions) 3,234개, 뉴스 헤드라인 및 기사(main-news) 4,297개, 온라인 사용자들의 대화를 모은 유저포럼(main-forums) 1,052개로 이루어져있어 문어체(캡션, 뉴스 헤드라인 및 기사)와 구어체(유저포럼)의 구분이 있고 분야별 성능을 비교할 수 있다는 장점이 있어 태스크 개발에 선정하였다. 데이터의 구성과 예시는 <표 3>과 같다.

〈표 3〉 korSTS데이터 구성 및 예시

장르	문장1	문장2	유사도 점수
main-captions	남자가 담배를 피우고 있다	남자가 스케이트를 타고 있다	0.5
main-news	올림픽 개막식 전 세계에 히트	개막식은 올림픽을 흔들어놓는 출발을 선사한다.	4.2
main-forums	그러니 질문에 대답해.	질문을 피하지 마세요.	3.4

3.1.2 의미 검색 태스크

의미 검색 태스크는 대조학습 기반의 문장 임베딩 방법론인 DSE(Zhou et al., 2022)의 성능평가 중

‘utterance level response selection’을 벤치마크하여 개발하였다. 위 태스크는 응답 후보 중 질문 문장에 대한 가장 적합한 문장을 찾는 태스크이며, 가장 유사도가 높은 문장을 정답으로 산출한다.

위 태스크를 위해 성균관대학교 산학협력단(2023)에서 구축한 rlhf_korean_dataset[4]을 활용하였다. 위 데이터셋은 거대언어모델(Large Language Model) 학습을 위해 구축된 영어 지침 데이터셋(instruction dataset)을 챗지피티(chatGPT)를 통해 번역한 뒤, 자연스럽게 전처리하는 과정을 거친 데이터셋이다. 기존의 다양한 영어지침 데이터셋(alpaca[5], dolly[6] 등)을 결합하여 총 107,158개의 많은 양을 확보하였다는 점과 번역 과정에서 존댓말, 대화체 사용 등의 지침을 통해 가독성을 고려하였다는 장점이 있어 태스크에 활용하였다.

본 데이터셋은 일상대화 챗봇 개발을 목적으로 활용할 수 있으며 지침(instruction), 입력값(input), 출력값(output)으로 구성되어있다. 본 데이터셋의 예시는 <표 4>와 같다.

〈표 4〉 rlhf korean dataset 데이터 구성 및 예시

지침 (instruction)	입력값 (input)	출력값 (output)
고양이와 개의 유사점은 무엇인가요?	-	고양이와 개의 유사점은 두 동물 모두 털이 있고 네 다리를 가지며 애완동물이며, 또한 주인에 대한 애정을 표현할 수 있다는 점입니다.
“행복”에 가장 유사한 5개의 단어를 찾아보세요.	-	쾌활한, 유쾌한, 환한, 기뻐하는, 만족하는.
다음 문장을 동사, 명사, 형용사 등으로 분류하세요.	자연은 아름답습니다.	명사 (자연), 형용사 (아름다운).

총 107,158개의 데이터셋에 대해 지침과 입력값을 합친 문자열을 질문(question)으로, 출력값(output)을 답변(answer)으로 하여 질의응답을 구성하였다. 그 중 5,000개의 데이터를 랜덤하게 추출하여 과제에 사용하였다. 본 과제는 각 질문에 대해 정답을 포함하여 랜덤하게 산출한 100개의 답변 후보 중 문장 유사도가 가장 높은 문장을 반환한다. 과제에 대한 평가로는 가장 유사도가 높은 문장이 정답일 확률(top-1 정확도)과 유사도가 높은 문장 3개, 5개 중에 정답이 있을 확률(top-3 / top-5 정확도)를 계산하였다.

3.2 언어의 표층적 속성

문장 임베딩이 문장의 길이, 어휘와 같은 언어의 표면적인 정보를 잘 반영하고 있는지 검증하기 위해 두 가지 태스크를 제작하였다. 언어의 표층적 속성을 평가하는 태스크는 openkorpos[7]데이터를 활용했다. Openkorpos는 한국어 품사 정보가 함께 표기된 오픈소스 데이터로, 데이터 공개가 어려운 세종 코퍼스[8]와 모두의 말뭉치[9]의 대안으로 개발되었다(Moon et al., 2022). 한국어 위키피디아의 문서에 공개된 형태소 분석기를 활용하여 문장을 분석하고, 각 분석기의 결과로부터 문장 구성성분과 품사 정보를 투표하여 최종적으로 가장 적합한 후보를 선출하는 투표 기반 오픈소스 말뭉치 태깅(voting-based open corpus annotation) 방법을 활용하였다. 각각의 문장은 문장 구성성분으로 분해되어 형태소 정보와 함께 병기되어 있다. 한 문장 구성성분에 여러 개의 형태소가

4) https://github.com/JoJo0217/rlhf_korean_dataset/
5) https://huggingface.co/datasets/tatsu-lab/alpaca
6) https://huggingface.co/datasets/nlpai-lab/databricks-dolly-15k-ko
7) https://github.com/openkorpos/openkorpos

혼합되어 있는 경우 '+'기호를 사용하였다. 한 문서 내에는 10,000개의 문장이 있으며, aa-pn까지의 분류로 구분되어 총 4,031,704개의 문장으로 구성되었다. 데이터에 대한 원문장과 데이터 구축 과정을 통해 구성된 데이터의 예시는 <표 5>와 같다.

〈표 5〉 openkorpos 데이터 구성 및 예시

원문장	문장구성요소와 품사정보 예시
사람들은 환호했다.	["사람","NNG"],["들","XSN"],["은","JX"], [" ","SB"],["환호","NNG"],["했","XSV+EP"], ["다","EF"],[".","SF"]
녹색을 상징색으로 하고 있다.	[["녹색","NNG"],["을","JKO"],[" ","SB"], ["상징","NNG"],["색","NNG"],["으로","JKB"], [" ","SB"],["하","VV"],["고","EC"], [" ","SB"],["있","VX"],["다","EF"],[".","SF"]],
제3대, 제4대 국회의원을 지냈다.	[["제","XPN"],["3","SN"],["대","NNBC"], [",","SC"],[" ","SB"],["제","XPN"], ["4","SN"],["대","NNBC"],[" ","SB"], ["국회의원","NNG"],["을","JKO"],[" ","SB"], ["지냈","VV+EP"],["다","EF"],[".","SF"]]

이 중, a부터 c 분류에 있는 문장 780,000개를 대상으로 전처리하여 태스크별 120,000개의 데이터(학습 데이터셋 100,000개, 검증 데이터셋 10,000개, 테스트 데이터셋 10,000개)를 사용하였다. 이는 영어 프로빙 태스크의 데이터 수와 같다.

3.2.1 문장 길이 분류 태스크 (Sentence Length, SentLen)

문장의 길이를 분류하는 태스크로, <표 5>의 원문장과 같이 문장을 복원한 뒤, 공백을 기준으로 문장을 분리한 어절을 기준으로 문장의 길이를 추출한 후 라벨(0-5)을 부여하여 데이터셋을 구성하였다. 라벨 기준은 선행연구(안애림 등, 2021)를 참고하였다. 각 라벨에 따른 어절 수와 데이터의 개수는 <표 6>과 같다.

〈표 6〉 문장 길이 분류 태스크 라벨 및 어절 수

라벨	0	1	2	3	4	5
어절	5-8	9-12	13-16	17-20	21-15	26-31
데이터 수	20,000	20,000	20,000	20,000	20,000	20,000

3.2.2 어휘 분류 태스크(Word Content, WC)

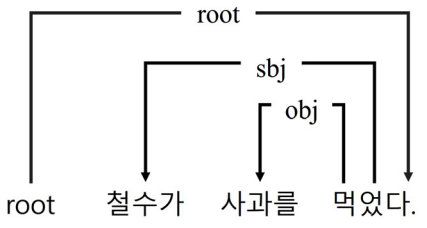
어휘 분류 태스크는 말뭉치에서 중빈도 어휘를 미리 선정하고, 주어진 문장을 어휘에 따라 분류하는 과제로, 임베딩이 자연어 문장에 있는 어휘를 기억하고 있는지 평가하는 태스크이다(Conneau et al., 2018). 위 과제는 openkorpos 데이터를 그대로 활용하게 되면 시제 등의 문제로 인해 정확한 어휘의 빈도를 세기 어렵다는 문제가 있다고 판단하여 kiwi 형태소 분석기(이민철, 2022)를 통해 문장을 분리한 뒤, 빈도를 계산하여 중빈도 어휘를 선정하였다. 예를 들어, <표 5>의 세 번째 문장에서 ["지냈","VV+EP"]의 경우 '지내다'라는 동사의 과거형이기 때문에 기본형인 '지내-'와 별도로 개수를 계산하는 문제가 생긴다. 이에, 형태소 분석기를 통해 문장을 형태소 단위로 분리한 뒤, 일반명사, 고유명사, 동사, 형용사, 어근, 일반 부사, 관형사의 품사를 선정하되, 형용사와 같이 과거시제 등으로 활용이 가능한 어휘들은 기본형으로 복원하였다. 총 120,000개의 데이터가 구축되었으며 학습 데이터셋 100,000개, 검증 데이터셋 20,000개, 테스트 데이터셋 20,000개로

8) https://www.korean.go.kr/front/reportData/reportDataView.do?mn_id=45&report_seq=197&pageIndex=1
9) https://kli.korean.go.kr/

구성되었다. 데이터셋의 라벨이 되는 어휘는 평균적으로 120의 빈도를 가지며 구체적인 목록과 각 어휘의 빈도는 링크[10]를 통해 확인할 수 있다.

3.3 언어의 통사적 속성

문장 임베딩이 문장의 구조나 기능, 문장 구성 요소의 속성인 언어의 통사적 속성을 잘 반영하고 있는지 검증하기 위한 태스크를 개발하였다. 문장의 통사적인 구조를 나타내는 방법 중 하나인 의존 구문 분석은 문장 성분 간의 지배소-의존소 관계를 파악함으로써 문장의 구조를 파악한다. 때문에 문장을 구성하는 요소의 위치에 제약이 적고 생략에도 유연하게 대처할 수 있어 한국어 구문분석에 적합하다(남궁영, 2019). 의존 구문 분석을 통해 분석한 한국어 문장의 예시는 <그림 1>과 같다.



〈그림 1〉 의존 구문 분석을 통한 한국어 구문 분석 예시

<그림 1>은 한국어 문장에서 나타나는 지배소-의존소 관계에 대해 화살표의 방향으로 나타내고 있으며 각 어절은 노드(node)라고 부른다. 즉, '철수가', '사과를'은 지배소인 '먹었다'의 의존소이며 '철수가'는 주어(Subject, SBJ), '사과를'은 목적어(Object, OBJ)의 관계이다. 서술어인 '먹었다'는 지배소-의존소 관계를 나타내기 위해 별도의 루트 노드(root node)를 추가하였다.

본 논문에서는 한국어 문장에 대한 의존 구문 분석 정보를 가진 KLUE-DP 데이터셋(Park et al., 2021)을 활용하였다. KLUE-DP 데이터셋은 한국어 문장의 통사적 구조를 의존 구문 분석을 통해 분석한 정보를 담은 데이터셋이다. 어절의 인덱스, 어절, 품사, 지배소의 인덱스 번호, 어절의 통사/기능적 정보를 표기하고 있다. <그림 1>의 문장에 대한 정보를 KLUE-DP 형식으로 나타내면 <표 7>과 같다. 어절 중 '먹었다'의 경우 루트 노드를 지배소로 가지고 있기 때문에 인덱스를 0으로 표기하였다.

〈표 7〉 KLUE-DP 데이터 구성 및 예시

인덱스 번호	어절	형태소 분해	품사 정보	지배소의 인덱스	통사/기능 정보
1	철수가	철수 가	NNP+JKO	3	NP_SBJ
2	사과를	사과 를	NNG+JKO	3	NP_OBJ
3	먹었다	먹 었 다 .	VV+EP+EF+SF	0	VP

3.3.1 주어 생략 여부 태스크 (SubjectOmission, SubjOmission)

한국어 문장의 통사적 특징 중 하나는 주어의 생략 현상으로, 한국어는 주어가 생략되더라도 그 의미를 회복할 수 있으면 주어가 쉽게 생략된다(박청희, 2013). 본 연구에서는 이러한 특성을 활용하여 주어 생략 여부 태스크를 개발하였다. KLUE-DP의 통사적/기능적 정보에서 주어를 나타내는 SBJ가 있는 경우 주어가 있는 문장(14,790개)으로, 없는 경우에는 주어가 없는 문장(3,346개)

10) https://github.com/minalang/KoSentEval/blob/main/word_counts.txt

으로 하여 총 18,136개의 데이터셋을 구축하였다. 학습 데이터셋은 14,508개, 검증 데이터셋 1,814개, 테스트 데이터셋 1,814개로 구성되어있다. 주어가 있는 문장과 생략된 문장의 예시는 (1)과 같다.

(1) (주어있음) 덕분에 세비야에서의 여행이 참 좋은 기억으로 남았네요.
(주어생략) 한국 분들 위해서 한글로 후기 남깁니다

3.3.2 서술어의 논항 예측 태스크 (Top Dependency Sequence, TopDeps)

서술어의 논항 예측 과제는 최상위 구성 성분인 루트 노드의 하위 노드를 예측하는 것이다(안애림 등, 2021). 루트 노드의 하위노드란 보통 문장의 서술어를 지배소로 가지는 의존소를 의미하며 이를 논항이라고 한다. <그림 1>의 경우를 예로 들면 서술어 '먹었다'를 지배소로 하는 '철수가'와 '사과를'이 논항이 되며 해당 어절의 기능적 정보를 언더바(_)를 통해 이어 'SBJ_OBJ'가 정답 라벨이 된다. 기능적 정보에는 주어(Subject, SBJ), 목적어(Object, OBJ), 명사 수식어(Noun Modifier, MOD), 서술어 수식어(Predicate Modifier, AJT), 보어(Complement, CMP), 접속사(Conjunction, CNJ)가 있으며 빈도를 계산하여 상위 라벨 20가지를 선별하여 총 10,761개의 데이터셋으로 구성하였다. 학습 데이터셋은 8,608개, 검증 데이터셋 1,076개, 테스트 데이터셋 1,076개이다. 상위 20개 라벨의 정보는 <표 8>와 같다.

〈표 8〉 상위 20개 라벨 정보

SBJ, AJT, MOD, AJT_SBJ, SBJ_AJT, OBJ, SBJ_MOD, AJT_AJT, AJT_OBJ, SBJ_OBJ, SBJ_AJT_OBJ, CMP, SBJ_AJT_AJT, AJT_AJT_SBJ, SBJ_SBJ, SBJ_CMP, AJT_SBJ_AJT, SBJ_AJT_AJT_OBJ, MOD_MOD, OTHER (그 외의 태그)

3.4 언어의 의미적 속성

문장 임베딩이 한국어의 다양한 의미 차이를 반영하고 있는지 검증하기 위해 아래와 같은 태스크를 구성하였다. 앞서 3.3절에서 다룬 통사적인 구조가 완벽하다는 것이 문장이 의미적으로 모순이 없다는 것을 보장하지 않기 때문에(Fromkin et al., 2014), 문장의 임베딩이 언어의 의미적인 차이를 잘 구분할 수 있는지 검증하는 것은 중요하다. 본 논문에서는 의미적 속성을 시제, 극성, 문장 유형, 경어법으로 세분화하여 태스크를 구성하였다.

3.4.1 시제 분류 태스크(Tense)

한국어는 시제 언어로서, 과거-현재-미래의 3분 체계를 갖는다. 시간적 위치를 나타내는 문법 요소로써 과거시제의 경우 -었-/-았-, 현재시제로는 -느-, 미래시제로는 -겠-/-을 것이- 등을 사용한다(박진호, 2011). 본 연구에서는 3.2절과 마찬가지로 openkorpos 데이터를 활용하여 문장 서술어의 시제를 기반으로 과거/비과거로 데이터를 구성하였다. 즉, 서술어의 선어말어미가 -었-/-았-인 경우를 선별하여 과거시제로 라벨링 하였다. 과거시제와 달리, 미래시제를 나타내는 선어말어미 -겠-의 경우 시제에 대한 정보 뿐만 아니라 추측의 의미 등을 담기 위해 사용된다. 때문에 형태소 분석만을 통한 선별은 어렵다고 판단하여 과거시제 외의 문장을 비과거로 라벨링하였다.

과거 문장 60,000개, 비과거 문장 60,000개로 총 120,000개의 문장을 무작위로 추출하여 선별하였으며, 학습 데이터셋 100,000개, 검증 데이터셋 10,000개, 테스트 데이터셋 10,000개로 구성하였다. 과거와 비과거 문장의 예시는 (2)와 같다.

(2) (과거) 2024년 3월 31일에 게재되었다.
(비과거) 문장 임베딩은 다음과 같은 뜻이 있다.

3.4.2 극성 분류 태스크(Negation)

문장의 의미를 부정문과 긍정문으로 구분하는 극성 분류 태스크를 수행하였다. 선행연구(안애림 외, 2021)에서는 '안, 못'과 같은 부정부사와 '않다, 말다, 못하다'와 같은 부정적 의미를 나타내는 보조용언과 서술어를 참고하여 부정문을 판단하였지만, 부정문은 그 유형과 사용 동기, 출현 양상에 따라 다양할 수 있다. 예를 들어, '나는 그 친구가 밥을 안 먹은 줄 몰랐다'라는 문장의 경우 부정부사 '안'과 부정을 나타내는 동사 '모르다'가 함께 나타나는 이중부정문이다(김송희, 2021). 이처럼 다양한 경우의 수가 존재하기 때문에 단순 구성요소만을 보고 문장의 극성을 판단하는 것이 충분치 않다고 판단하였다. 이러한 한계점을 극복하기 위해 사전에 라벨링이 되어 있고 감성분석에 주로 사용되는 데이터셋인 네이버 영화리뷰 데이터[11]를 활용하였다. 위 데이터는 영화 감상에 대해 0(부정), 1(긍정)으로 라벨링 한 데이터로 총 200,000개의 문장으로 구성되어있다. 그 중 무작위로 긍정 문장 60,000개, 부정 문장 60,000개로 총 120,000개를 선별하여 실험에 사용하였다. 120,000개의 데이터셋 중 학습 데이터셋 100,000개, 검증 데이터셋 10,000개, 테스트 데이터셋 10,000개를 활용하여 과제를 진행하였다. 긍정과 부정 문장의 예시는 (3)과 같다. 이 중 긍정 문장의 경우 보조용언인 '없다'를 활용하였지만 전체적인 의미는 긍정인 예시를 보여주며 부정 문장은 부정부사나 보조용언을 사용하지 않고도 부정의 감성을 표현하는 문장의 예시이다.

(3) (긍정) 말이 필요없는 영화~ 너무 행복합니다.
(부정) 별 한 개도 아깝다.

3.4.3 문장 유형 분류 태스크 (Sentence Type, SentType)

일반적으로 한국어에서는 문장 유형을 문장의 종결어미에 따라 평서문, 의문문, 명령문, 청유문의 네 가지 형태로 유형을 구분한다(임동훈, 2011). 이에 선행연구에서는 문장 유형 정보를 가진 데이터를 활용하여 데이터셋을 구성하였지만, 위 데이터는 오픈소스로 공개되어 있지 않다는 한계를 가진다. 따라서, 본 태스크에서는 두 가지 중에 하나를 선택하는 선택의문문, 누가, 언제, 어디서 등의 의문사가 포함되는 설명 의문문, 요구, 금지 4가지 유형으로 구성된 paraKQC[12]와 styleKQC[13]를 함께 활용하였다. paraKQC는 자유주제, 메일, 스케줄, 기사, 날씨에 대한 주제에 대해 질문 및 요구 문장 10,000개로 이루어져 있다. StyleKQC는 주제를 메신저, 집안일, 캘린더 일정, 자연 현상, 엔터테인먼트, 쇼핑으로 확장한 질문 및 요구 문장 30,000개로 구성되어있다. 두 데이터를 합친 뒤 결측치와 중복값을 제거하여 총 39,970개의 데이터를 구성하였으며, 학습 데이터

11) https://github.com/e9t/nsmc
12) https://github.com/warnikchow/paraKQC
13) https://github.com/cynthia/stylekqc

셋 31,976개, 검증 데이터셋 3,997개, 테스트 데이터셋 3,997개로 이루어져 있다. 문장의 각 유형에 대한 예시는 (4)와 같다.

(4) (선택의문문) 과제랑 시험공부 중 뭐 먼저 할까?
(설명의문문) 식기 건조대에 어떤 그릇이 담겨 있는지 확인 해주시겠어요.
(요구) 마른 컵은 정리 좀 해주세요.
(금지) 안된다고. 눈 올 때 운전하면.

3.4.4 경어법 분류 태스크(Honorifics)

한국어의 경어법은 다른 언어와 비교하였을 때 특징적으로 정교하게 발달되어 있다(박성일, 2013). 이러한 경어체계는 존대의 대상에 따라 주체경어법과 객체경어법, 상대경어법으로 분류되고, 문법 형태소나 조사, 혹은 특수 높임말에 의해 실현된다. 예를 들어, (5)의 존댓말 문장은 문장의 주어인 '당신'을 높이는 주체높임법을 실현하기 위해 특수 높임말인 '계시다'를 사용하였다. 그리고, 문장을 듣는 청자가 문장의 주어와 같은 '당신'이며 높임의 대상이기 때문에 상대높임법을 실현하기 위해 문장의 종결어미에 격식체인 '하십시오'체를 활용하였다. 이와 같이 한 문장을 말할 때마다 상대방과의 상대적인 지위나 친밀도를 고려하여 적절한 높임의 정도를 결정해야 할 정도로 한국어의 경어체계는 우세하기 때문에(Yeon & Brown, 2017), 본 연구에서는 경어의 유무를 판별하는 경어법 분류 태스크를 고안하였다. 데이터는 styleKQC와 smilegate AI에서 구축한 korean smile style dataset[14]을 활용하였다. styleKQC 데이터셋의 경우 3,000개의 쿼리에 대한 반말 문장 5개, 존댓말 문장 5개로 총 10개의 문장이 존재하여 전체 데이터의 수는 30,000개이다. 이를 분류하여 존댓말(15,000개), 반말(15,000)개의 데이터를 확보하였다. korean smile style dataset은 동일한 대화에 대해 7개의 서로 다른 스타일로 문체를 변환시킨 데이터셋이다. 이 중 존댓말과 반말 스타일을 적용한 문장을 추출하여 본 과제에 활용하였다. 존댓말(3,470개)과 반말(3,470개)로 이루어져 있으며 styleKQC와 종합한 뒤 중복값과 결측치를 제거해 총 36,917개의 데이터를 구성하였다. 학습 데이터셋 29,533개, 검증 데이터셋 3,692개, 테스트 데이터셋 3,692개로 설정하였다. 경어법 분류 태스크에 활용한 존댓말 문장과 반말 문장의 예시는 (5)와 같다.

(5) (존댓말) 뇌우가 있을 때 당신은 실내에 계십시오.
(반말) 뇌우가 있을 때 너는 실내에 있어

4. 실험

4.1 활용 모델

본 논문에서는 문장 임베딩을 생성하는 모델들을 활용하였으며, 그 중에서도 기계독해, 문서분류, 언어분석, 검색 결과 순위화 등 언어이해 유형의 다양한 태스크에서 우수한 성능을 보이는 (임준호, 2021) BERT계열의 인코더 모델 BERT, RoBERTa, ELECTRA모델을 활용하였다. 모델의 다양성을 위해 허깅페이스(huggingface)에 공개된 모델을 활용하였으며 모델의 분류와 개수는 다음과 같다. 1) 다국어를 학습하여 다양한 언어의 임베딩을 생성할 수 있는 다국어 문장 인코더 2개, 2) 대량의 한국어 텍스트를 사전 학습한 한국어

14) https://github.com/smilegate-ai/korean_smile_style_dataset

문장 인코더 3개, 3) 트랜스포머(Transformer) 계열의 사전학습 모델에 대조학습을 활용하여 추가 학습한 대조학습 기반의 문장 인코더 7개를 활용하였다. 또한 형평성을 위해 모두 base모델을 사용하였다.

4.1.1 다국어 문장 인코더

LaBSE(Feng et al., 2020)는 BERT기반의 다국어 문장 인코더로 듀얼 인코더 구조에 BERT의 사전학습 방식인 MLM(Masked Language Modeling)과 TLM(Translation Language Modeling)을 적용한 모델이다. 109개의 언어에 대해 학습되어있다.

XLM-R(Conneau et al., 2020)모델은 RoBERTa 기반의 다국어 문장 인코더로 동적 마스킹(Dynamic Masking)방식으로 사전학습 되었다. 100가지 언어 이상으로 구성된 CommonCrawl corpus를 통해 학습되었다.

4.1.2 한국어 문장 인코더

한국어 문장 인코더의 경우 **KLUE-BERT**, **KLUE-RoBERTa**모델(Moon et al., 2021)을 선정하였다. 이는 비지도 대조학습 방법을 다양한 한국어 사전학습 모델(KoBERT, KR-BERT, KLUE-BERT)에 적용하여 문장 유사도 태스크에 평가하였을 때 KLUE-BERT가 안정적이고 우수한 성능을 보인다는 선행연구(유영현, 2022)의 결과를 참고하였다. 이후 4.1.3에서 설명하는 대조학습 기반 한국어 문장 인코더는 모두 KLUE-BERT, KLUE-RoBERTa에 추가적인 학습을 통해 구축된 모델을 활용하여 성능의 개선이 있는지 분석하였다. 두 모델은 모두의 말뭉치, 나무위키 등 약 62G의 대용량 한국어 코퍼스를 통해 사전학습되었다.

KoELECTRA(박장원, 2020)는 생성기(Generator)에서 나온 토큰(token)을 보고 판별기(Discriminator)에서 실제 문장의 토큰인지, 대체된 토큰인지를 판별하는 방식(Replaced Token Detection)으로 학습하는 ELECTRA모델의 한국어 버전으로, 약 34G의 한국어 텍스트를 통해 사전학습 되었다.

4.1.3 대조학습 기반 한국어 문장 인코더

대조학습(Contrastive Learning)은 학습대상이 되는 앵커(anchor)문장과 유사한 데이터인 정례(positive pair)는 잠재공간에서 서로 가깝게 위치하는 비슷한 벡터로 표현되고, 상이한 데이터인 부례(negative pair)는 잠재공간에서 멀리 위치하는 벡터로 표현되도록 학습하는 방법이다(지인영, 2023). 트랜스포머 계열의 사전학습 모델에 대조학습을 활용하여 문장 임베딩을 생성하는 것은 좋은 품질의 임베딩을 만드는 학습방법으로 밝혀졌다(Giorgi et al., 2021). 대조학습에서 정례와 부례를 구성하는 방법은 각 방법론에 따라 다양하다. 본 논문에서는 허깅페이스(huggingface)에 공개된 대조학습 기반의 한국어 문장 인코더 모델을 활용하며, 모든 모델은 한국어 모델임을 표기하기 위해 접두어로 'Ko'를 표기한다. 모든 대조학습 기반의 모델은 KLUE-BERT, KLUE-RoBERTa 모델을 사전학습모델로 활용하여 추가 학습되었다.

KoSimCSE-BERT, **KoSimCSE-RoBERTa**모델(김봉민, 2022)은 정례로 자연어추론(Natural Language Inference)데이터셋에서 하나의 전제(premise)문장을 앵커문장이라고 가정할 때, 위 문장을 함의(entailment)하는 문장을 활용하고, 부례로 앵커문장과 모순(contradiction)된 문장을 활용하여 데이터셋을 구성하는 지도적인 simCSE(Gao et al., 2021)방법론에 의해 학습되었다. 학습데이터셋은 카카오브레인의 korNLI데이터(Ham et al., 2020)를 활용하였다.

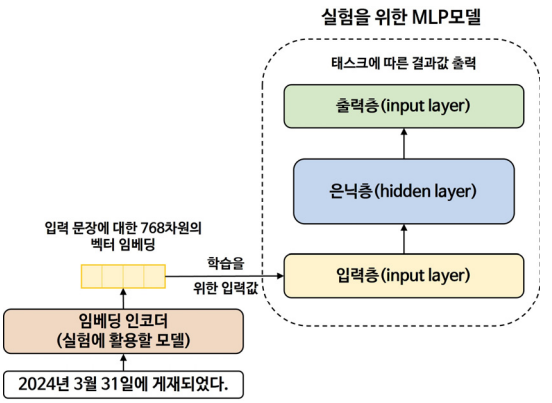
KosimCSE-Unsup-BERT, KosimCSE-Unsup-RoBERTa모델(김봉민, 2022)은 앵커가 되는 문장을 단순히 같은 인코더에 한 번 더 넣어 랜덤한 dropout mask를 적용한 문장의 임베딩을 정례로 사용하고, 배치 내의 다른 문장을 부례로 활용하는 비지도적인 simCSE방법론을 통해 학습되었다. 학습데이터는 위키문서를 종합한 텍스트이다.

KoDSE-BERT, **KoDSE-RoBERTa**모델(정민화, 2023)은 정례를 앵커문장에 연속된 대화문장으로, 부례를 배치 안의 문장 각각을 앵커와의 유사도에 따라 가중치를 부여하여 활용하는 hard negative sampling 전략을 통해 학습하는 DSE(Zhou et al., 2022)방법론을 따른다. AI허브에 게시된 용도별 목적 대화 데이터셋[15]과 주제별 텍스트 일상 대화 데이터셋[16]에서 speaker_type이 1:1인 데이터만을 추출 후 활용하고, 한국어 멀티세션 대화[17]를 추가하여 총 4,030,024개의 문장 쌍 데이터셋을 학습 데이터로 활용하였다.

KoDiffCSE-RoBERTa는 비지도적인 simCSE방법론에서 dropout mask를 활용한 정례 생성을 민감하지 않은 변형으로, 문장을 마스킹한 뒤 생성기(generator, G)를 통해 복원한 문장을 민감한 변형으로 두 가지를 혼합하여 학습하는 DiffCSE방법론(Chuang et al., 2022)을 사용하였다. 비지도적인 방법론이기 때문에 KoSimCSE-Unsup모델들과 마찬가지로 위키문서를 종합한 코퍼스를 통해 학습되었다. KodiffCSE-BERT의 경우 KLUE모델에 BERT-small 또는 distillBERT모델이 존재하지 않아 논문과 같은 조건으로 구현하기 어렵기 때문에 본 실험에서 제외하였다.

4.2 실험 방법

본 논문의 실험은 SentEval 라이브러리의 코드를 활용하였다. SentEval은 각 태스크에 대해 임베딩 벡터를 평가할 수 있는 라이브러리로 각 태스크를 구성하는 데이터셋의 문장을 인코더(실험에 활용되는 모델)를 통해 벡터로 변환한 값을 입력값으로, 각 문장의 라벨 값을 정답 데이터로 사용하여 MLP(Multi Layer Perceptron) 분류기(classifier)에 학습한다. 이후 각 태스크에 따라 학습된 분류기의 정확도를 측정하여 결과값으로 출력하는 구조이다. 각 태스크에 맞는 학습/검증 데이터셋을 통해 학습이 이루어지며 분류 정확도 결과는 테스트 데이터 예측값의 정확도를 평가한다. 본 논문에서 이루어진 실험을 도식화하면 <그림 2>와 같다.



〈그림 2〉 실험과정 도식화

15) https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=544
16) https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=543
17) https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71630

위 라이브러리에서는 데이터 경로(task_path), pytorch 사용유무(usepytorch), kfold 교차검증을 사용할 경우 k의 수(kfold), MLP 분류기 은닉층의 수(nhid), 옵티마이저(optimizer), 배치 크기(batch_size), 훈련 중단 기준(tenacity), 학습 횟수(epoch_size)를 설정할 수 있다. 본 실험에서 설정한 구체적인 하이퍼파라미터 세팅은 <표 9>와 같다.

〈표 9〉 실험의 하이퍼파라미터

```
params_senteval = {'task_path': PATH_TO_DATA,
                   'usepytorch': True, 'kfold': 10}
params_senteval['classifier'] = {'nhid': 1, 'optim': 'adam',
                 'batch_size': 128, 'tenacity': 5, 'epoch_size': 5}
```

4.3 실험 결과

하위 태스크 2개, 프로빙 태스크 8개에 대한 다국어 문장 인코더 2개, 한국어 문장 인코더 3개, 대조학습 기반 문장 인코더 7개의 실험 결과는 다음과 같다. 표에 기술된 수치의 경우 소수점 셋째 자리에서 반올림하였다.

4.3.1 문장 임베딩의 품질 평가

4.3.1.1 문장 유사도 태스크 평가

문장 유사도 태스크에서는 스피어만 상관관계(spearman correlation)를 계산하여 모델이 예측한 두 문장의 코사인 유사도와 실제 사람이 라벨링한 점수의 상관관계를 비교하였다. 총계의 경우, 데이터 수에 따라 가중평균한 값을 기록하였다. 모델별 실험 결과는 <표 10>과 같다.

〈표 10〉 문장 유사도 태스크 결과

모델	그림 캡션	뉴스 헤드라인 및 기사	유저포럼	총계 (가중치 평균)
다국어 문장 인코더				
LaBSE	0.80	0.69	0.62	0.71
XLM-R	0.16	0.10	0.15	0.13
한국어 문장 인코더				
KLUE-BERT	0.47	0.45	0.47	0.46
KLUE -RoBERTa	0.14	0.23	0.20	0.19
KoELECTRA	0.16	0.21	0.22	0.20
대조학습 기반 문장 인코더				
KosimCSE -BERT	**0.91**	**0.71**	**0.70**	**0.79**
KosimCSE -RoBERTa	0.89	0.68	0.64	0.76
KosimCSE -Unsup-BERT	0.80	0.68	0.62	0.72
KosimCSE -Unsup -RoBERTa	0.80	0.67	0.60	0.71
KoDSE-BERT	0.77	0.53	0.60	0.63
KoDSE -RoBERTa	0.72	0.49	0.56	0.59
KodiffCSE -RoBERTa	0.81	0.70	0.60	0.73

문장 유사도 태스크의 경우 대조학습 기반의 인코더 중 simCSE-BERT의 성능이 월등하였다. 특히 DSE방법론과 diffCSE방법론의 경우 simCSE의 성능을 개선한 것으로 제안되었으나, 한국어의 경우 simCSE 기반의 모델(KosimCSE-BERT, KosimCSE-RoBERTa)의 성능이 더 좋았다.

모델적으로는 RoBERTa, ELECTRA 기반의 모델보다 BERT 기반의 모델(LaBSE, KLUE-BERT 등)이 더 성능이 높은 것으로 확인되었다.

4.3.1.2 의미 검색 태스크 결과

모델별 의미 검색 태스크에 대한 결과는 <표 11>과 같다.

〈표 11〉 의미 검색 태스크 결과

모델	top1 정확도	top3 정확도	top5 정확도
다국어 문장 인코더			
LaBSE	0.69	0.79	0.82
XLM-R	0.09	0.12	0.14
한국어 문장 인코더			
KLUE-BERT	0.30	0.37	0.41
KLUE -RoBERTa	0.13	0.18	0.19
KoELECTRA	0.05	0.09	0.11
대조학습 기반 문장 인코더			
KosimCSE -BERT	**0.79**	**0.89**	**0.91**
KosimCSE -RoBERTa	0.74	0.86	0.89
KosimCSE -Unsup-BERT	**0.79**	0.88	**0.91**
KosimCSE -Unsup -RoBERTa	**0.79**	0.88	0.90
KoDSE -BERT	0.77	0.86	0.89
KoDSE -RoBERTa	0.64	0.76	0.81
KodiffCSE -RoBERTa	0.80	0.88	0.90

의미 검색 태스크에서도 인코더의 종류로 구분하였을 때 다국어, 한국어 문장 인코더에 비해 대조학습 기반 문장 인코더가 전반적으로 우수한 성능을 달성하였다. 대조학습 기반의 문장 인코더 중에서는 top1, top3, top5정확도 모두에서 simCSE기반의 방법론을 활용한 모델이 높은 성능을 보였다. 대화 데이터셋을 활용한 태스크이기 때문에 정례로 연속된 대화를 추가 학습한 DSE 방법론이 높은 성능을 달성할 것이라고 기대하였으나 그보다는 정례로 앵커문장을 함의하는 문장을 학습하는 형태인 simCSE방법론이 보다 풍부한 의미관계를 학습하여 좋은 성능을 보였다고 해석할 수 있다. 또한, 모델에 대한 비교에서는 RoBERTa, ELECTRA기반의 모델보다는 BERT기반의 모델이 더욱 높은 성능을 달성하였다. 의미 검색 태스크에서 높은 성능을 보여준 모델은 대화형 인공지능 개발과 검색 쿼리에 대한 유사한 문서 등을 찾는 포털 검색 등에 활용할 수 있다.

4.3.2 언어의 표층적 속성 / 통사적 속성 평가

각 모델에 대해 표층적/통사적 속성을 평가하는 태스크의 실험 결과는 <표 12>와 같다.

〈표 12〉 표층적/통사적 속성 실험 결과

모델	문장 길이	어휘 분류	주어 생략 여부	서술어의 논항 예측
다국어 문장 인코더				
LaBSE	67.21	41.09	90.96	27.21
XLM-R	79.14	31.43	88.53	25.81
한국어 문장 인코더				
KLUE-BERT	77.78	47.15	**91.01**	26.74
KLUE -RoBERTa	79.36	46.80	89.47	24.61
KoELECTRA	**80.61**	17.44	89.86	25.35
대조학습 기반 문장 인코더				
KosimCSE -BERT	73.19	**51.29**	89.31	**27.30**
KosimCSE -RoBERTa	69.52	48.82	88.92	22.93
KosimCSE -Unsup-BERT	73.19	47.83	89.03	26.09
KosimCSE -Unsup -RoBERTa	68.25	50.02	88.48	23.49
KoDSE-BERT	77.19	46.46	89.69	26.00
KoDSE -RoBERTa	75.71	44.50	88.59	23.96
KodiffCSE -RoBERTa	67.00	49.90	88.20	24.30

언어의 표층적/통사적 속성에 대한 실험결과 인코더의 종류로는 다국어 문장 인코더에 비해 한국어와 대조학습 기반의 문장 인코더가 더 높은 성능을 보여주었다. 표층적 속성에서는 중빈도 어휘 1,000개를 분류하는 어휘 분류가, 통사적 속성에서는 서술어의 논항이 되는 문장구성요소 상위 20종류를 분류하는 서술어의 논항 예측이 전반적으로 낮은 성능을 보여주어 다른 태스크에 비해 상대적으로 복잡한 과제였음을 알 수 있었으며 이러한 결과는 영어를 대상으로 한 실험(Conneau et al., 2018)에서도 비슷한 경향성이 나타났다. 두 과제에 대해서는 대조학습 기반의 KosimCSE-BERT모델이 가장 높은 성능을 보여 대조학습 기반의 학습법이 언어의 속성을 임베딩에 반영하는 효율적인 방법임을 증명하였다.

4.3.3 언어의 의미적 속성 평가

언어의 의미적 속성을 평가할 수 있는 태스크에 대한 평가 결과는 <표 13>과 같다.

〈표 13〉 의미적 속성 실험 결과

모델	시제	극성	문장유형	경어
다국어 문장 인코더				
LaBSE	95.42	81.79	87.44	90.85
XLM-R	93.63	82.01	81.81	98.65
한국어 문장 인코더				
KLUE-BERT	**96.09**	87.34	85.62	99.49
KLUE-RoBERTa	95.42	86.26	87.39	99.54
KoELECTRA	94.81	83.85	86.24	99.08
대조학습 기반 문장 인코더				
KosimCSE-BERT	94.37	**88.47**	84.83	93.53
KosimCSE-RoBERTa	93.42	85.85	**88.62**	93.20
KosimCSE-Unsup-BERT	94.07	84.61	83.06	98.43
KosimCSE-Unsup -RoBERTa	91.83	82.16	84.69	98.02
KoDSE-BERT	95.84	85.69	84.01	99.67
KoDSE-RoBERTa	94.62	86.97	83.69	**99.78**
KodiffCSE-RoBERTa	91.88	82.20	84.74	97.94

표층적, 통사적 속성에 비해 의미적 속성에 대한 프로빙 태스크의 성능이 전반적으로 높았으며 이러한 경향성은 한국어 선행연구(안애림 등, 2021)에서도 동일하게 나타났다. 이러한 결과는 대부분의 인코더가 의미적인 문제 해결을 목적으로 학습되기 때문이라고 예측된다. 인코더의 종류에 따라서는 다국어 문장 인코더에 비해 한국어와 대조학습 기반의 문장 인코더가 좋은 성능을 보여주었다.

4.4 결과 분석

본 논문의 실험결과는 다음과 같은 특징을 보인다. 먼저, 인코더의 종류에서 다국어 인코더에 비해 한국어와 대조학습 기반 문장 인코더가 전반적으로 좋은 성능을 보여주었다. 이는 한국어 텍스트로 학습한 모델이 한국어의 속성을 잘 반영하고 있다고 해석할 수 있다. 특히, 대조학습 기반의 문장 인코더의 경우 문장 유사도와 의미검색 태스크에서 다른 인코더에 비해 월등한 성능을 보이며 그 중에서도 KosimCSE-BERT모델은 프로빙 태스크에서도 높은 성적을 보여 한국어의 특징을 잘 반영한 임베딩 모델이면서도 대화형 인공지능, 검색 서비스 등에서의 활용도가 높을 것으로 기대된다.

모델별 비교에서는 RoBERTa, ELECTRA기반의 모델보다 BERT기반의 모델 성능이 우수하였다. RoBERTa와 ELECTRA모델이 BERT모델의 성능을 개선한 모델임에도, BERT모델의 성능이 우수하였던 것은 다음과 같은 이유라고 추측된다. 먼저, BERT모델의 사전학습 방식인 NSP(Next Sentence Prediction)가 문장 임베딩 분야에 보다 적합하였을 수 있다. 또한, 실험 문장의 최대길이가 64로 지정되었기 때문에, 긴 문장으로 학습한

ELETCRA와 RoBERTa모델의 장점이 잘 활용되지 않았을 수 있다. 보다 긴 문장으로 모델을 학습 및 평가하는 것은 본 연구의 향후 과제가 될 것이다. 최근 긴 문서를 하나의 벡터로 축약하여 데이터베이스로 구축하는 벡터 DB가 대세로 떠오르고 있기 때문에(Morris et al., 2023), 이러한 연구의 확장은 유의미할 것이다.

마지막으로, 대부분의 인코더에서 표층적, 통사적 속성을 평가한 태스크보다 의미적 속성을 평가한 태스크에서 높은 성능을 보였다. 이러한 경향성은 인코더가 표층적/통사적 속성 보다는 문장의 의미적인 부분을 주요하게 학습한다는 것을 시사한다.

5. 결론

최근 문장 기저의 통사적/의미적 정보까지 내포하는 문장 임베딩에 대한 수요가 높아지며 다양한 언어에서 임베딩의 품질을 평가할 수 있는 체계가 개발되고 있지만, 한국어에서 연구 및 상업적으로 활용할 수 있는 방법론은 부재하였다. 이에 본 논문에서는 완전 공개가 가능한 오픈소스 데이터셋을 활용하여 평가 태스크와 이를 적용할 수 있는 라이브러리를 공개하였다.

평가 태스크는 임베딩 평가에 보편적으로 활용되며 다양한 언어에서 활용한 SentEval라이브러리의 체계를 따랐으며, 각 태스크는 재배포를 허용하지 않는 데이터셋을 활용하여 공개가 불가하였던 선행연구(안애림 등, 2021)의 한계점을 극복하기 위해 오픈소스 데이터셋으로 구성하였다. 임베딩의 언어학적 속성(표층적, 통사적, 의미적)을 검증할 수 있는 8개의 태스크와 임베딩의 품질과 활용성을 판단할 수 있는 문장 유사도, 의미 검색 태스크를 추가하여 총 10가지의 태스크로 다각적인 분석을 시도하였다. 언어학적 속성 평가의 경우, 주어 생략 여부, 경어법과 같이 한국어의 특성을 반영한 태스크를 개발하였다.

실험 결과, 다국어 인코더에 비해 한국어와 대조학습 기반의 인코더가 한국어의 언어적 특성을 잘 반영한 것으로 나타났다. 특히, 대조학습 기반의 문장 인코더 중 KosimCSE-BERT는 한국어의 언어적 속성을 잘 반영한 임베딩을 생성하고, 임베딩의 품질과 활용성을 검증하는 하위 태스크에서 월등한 성능을 보여 향후 챗봇, 검색 등 다양한 분야에서 활용 가능할 것으로 보인다.

또한, 모델에서는 RoBERTa, ELECTRA모델에 비해 BERT기반의 모델의 성능이 높았다. 이는 문장 임베딩이라는 도메인의 특징으로 보이며 향후 RoBERTa와 ELECTRA모델의 장점인 긴 길이의 문장 임베딩에 대한 태스크를 개발하여 여전히 BERT기반의 모델의 성능이 우수한지 검증할 필요가 있다. 마지막으로, 모든 인코더에서 표층적/통사적 속성에 비해 의미적 속성에 대한 태스크에서 높은 성능을 보였다. 이는 인코더의 사전학습 방식이 의미적인 문제를 해결하는 것이기 때문으로 추측된다.

본 연구의 의의는 다음과 같다. 먼저, 다양한 한국어 문장 인코더 모델을 태스크에 적용하고, 그 결과를 비교·분석 하였다. 본 논문을 통해 그동안 성능검증이 어려웠던 다양한 한국어 문장 임베딩 모델의 비교가 이루어졌다. 성능 검증은 문장 임베딩이 한국어의 언어적 특성(표층적, 통사적, 의미적)을 잘 내포하고 있는지 뿐만 아니라 임베딩의 품질과 활용성을 판단할 수 있는 10가지의 태스크를 통해 가능하다. 또한, 본 논문에서 설명한 임베딩 평가 태스크와 라이브러리는 깃허브를 통해 공개되어 연구 및 상업적인 활용이

가능하다. 공개한 태스크와 라이브러리를 활용하여 사용자는 새롭게 개발한 문장 임베딩 모델을 검증하거나, 개발하고자 하는 목적에 맞는 문장 임베딩 모델을 선정하여 자연어 기술개발에 활용할 수 있을 것이다.

한편, 본 연구의 한계와 이를 극복하기 위한 향후 연구방향은 아래와 같다. 먼저, 하위 태스크의 수가 2개로, 18개인 영어 라이브러리에 비해 개수가 적다는 한계가 있다. 이는 기존 자연어처리의 다양한 하위태스크(기계 독해, 문장추론, 혐오탐지)를 추가하거나, 향후 임베딩의 품질을 검증할 수 있는 새로운 태스크가 연구되어 추가되어야 할 것이다. 또한, 문장 최대길이가 상대적으로 짧게(64) 설정되었다는 실험 설정에서의 한계가 있다. 주요 실험 결과 중 BERT모델이 다른 모델에 비해 전반적으로 성능이 높게 나온 것이 문장 길이 설정의 영향일 수 있다고 판단된다. 이러한 한계를 극복하기 위해 문장 최대길이에 대한 다양한 설정값으로 실험하거나, 보다 긴 길이의 문장을 다루는 태스크가 추가적으로 개발되어 실험에 포함되는 등 다양한 조건에서의 추가 실험이 이루어질 필요가 있다. 마지막으로, 대조학습 기반의 문장 인코더 중 KodiffCSE-BERT모델이 부재하였다는 한계가 있다. 이는 베이스모델로 활용한 KLUE모델에 BERT-small이나 distilBERT 모델이 부재하여 diffCSE방법론을 적용한 BERT 모델을 학습하기 어려웠기 때문이다. 이러한 한계는 향후 연구에서 KLUE모델 뿐만 아니라 다양한 사전학습 모델을 활용하여 추가학습한 모델이 개발된다면 해소할 수 있을 것이다. 본 논문에서 발생한 한계를 극복한 추가 활발히 이루어져 한국어 문장 임베딩을 평가할 수 있는 SentEval리더보드를 개발되어 문장 임베딩 평가에 대한 객관적인 기준이 세워지기를 기대한다.

참고문헌(References)

[국내 문헌]

김봉민. (2022). Sentence-Embedding-Is-All-You-Need. Github. Retrieved Jan 02, 2024, from https://github.com/BM-K/Sentence-Embedding-Is-All-You-Need/tree/main

김송희. (2021). 한국어 이중부정문의 사용 동기와 출현 양상. *國語學,* 100, 243-273. https://doi.org/10.15811/kjl.2021..100.008

남궁영, 김창현, 천민아, 박호민, 윤호, 최민석, 김재균, 김재훈. (2019). 구묶음을 반영한 한국어 의존 구조 말뭉치 생성. *제31회 한글 및 한국어 정보처리 학술대회(HCLT 2019),* 한국과학기술원, 대전.

박성일. (2013). 인칭 범주에 기반한 한국어 경어법의 교육 내용 연구. *국어교육연구,* 32, 41-67.

박장원. (2022). KoELECTRA: Pretrained ELECTRA Model for Korean. Github. Retrieved Jan 02, 2024, from https://github.com/monologg/KoELECTRA

박진호. (2011). 시제, 상, 양태. *國語學,* 60, 289-322. https://doi.org/10.15811/jkl.2011..60.011

석주리. (2022). 단어 기반 임베딩과 문장 기반 임베딩의 한국어 문장 임베딩 성능 비교 연구. 고려대학교.

성균관대학교 산학협력단. (2023). rlhf_korean_dataset [Online]. Github. https://github.com/JoJo0217/rlhf_korean_dataset

스마일게이트(Smilegate-AI). (2022). korean SmileStyle Dataset[Online]. Github. https://github.com/smilegate-ai/korean_smile_style_dataset

안애림, 고병일, 이다니엘, 한경은, 신명철, 남지순. (2021). *한글 및 한국어 정보처리 학술대회 (HCLT 2021),* 온라인.

유영현, 이규민. (2022). 한국어 문장 표현을 위한 비지도 대조 학습 방법론의 비교 및 분석,

제34회 한글 및 한국어 정보처리 학술대회 (HCLT), 경주 교원드림센터, 경상북도.

이민철. (2022). Kiwi, Korean Intelligent Word Identifier. Github. from https://github.com/bab2min/Kiwi

임동훈. (2011). 한국어의 문장 유형과 용법. *國語學,* 60, 323-359. https://doi.org/10.15811/jkl.2011..60.012

임준호. (2021). 딥러닝 사전학습 언어모델 연구 동향. *정보과학회지,* 39(4), 9-14.

정민화. (2023). 대조학습 기반 한국어 문장 임베딩 방법론 평가 및 비교분석. 연세대학교.

지인영, 김희동, 김태혁, 배홍식. (2023). 대조학습을 이용한 문장 임베딩 연구. *언어와언어학,* 99, 95-124.

[국외 문헌]

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:*1608.04207. https://doi.org/10.48550/arXiv.1608.04207

Cho, W. I., Moon, S., Kim, J. I., Kim, S. M., & Kim, N. S. (2021). StyleKQC: A style-variant paraphrase corpus for Korean questions and commands. *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2021),* 7122-7128.

Chuang, Y. S., Dangovski, R., Luo, H., Zhang, Y., Chang, S., Soljačić, M., ... & Glass, J. (2022). DiffCSE: Difference-based contrastive learning for sentence embeddings. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 4207-4218. https://doi.org/10.18653/v1/2022.naacl-main.311

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics(ACL 2021),* 8440-8451. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018),* 2126-2136. https://doi.org/10.18653/v1/P18-1198

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022),* 878-891. https://doi.org/10.18653/v1/2022.acl-long.62

Fromkin, V.,Rodman, R., Hyams, N. (2014). *An Introduction to Language (10th ed.).* Wadsworth: Cengage Learning

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 6894-6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

Giorgi, J., Nitski, O., Wang, B., & Bader, G. (2020). Declutr: Deep contrastive learning for unsupervised textual representations. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021),* 879-895. https://doi.org/10.18653/v1/2021.acl-long.72

Ham, J., Choe, Y., Park, K., Choi, I. & Soh, H. (2020). KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding. *Findings of the Association for Computational Linguistics(EMNLP 2020),* 422-430. https://doi.org/10.18653/v1/2020.findings-emnlp.39

Krasnowska-Kieraś, K., & Wróblewska, A.. (2019). Empirical Linguistic Study of Sentence Embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics(ACL 2019),* 5729-5739. https://doi.org/10.18653/v1/P19-1573

Li, R., Zhao, X., & Moens, M. (2022). A Brief Overview of Universal Sentence Representation Methods: A Linguistic View. *ACM Computing Surveys.* 1-42. https://doi.org/10.1145/3482853

Mikhailov, V., Taktasheva, E., Sigdel, E., & Artemova, E. (2021). RuSentEval: Linguistic Source, Encoder Force!. *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing(BSNLP 2021),* 43-65. https://aclanthology.org/2021.bsnlp-1.6

Moon, S., Cho, W. I., Han, H. J., Okazaki, N., & Kim, N. S. (2022). OpenKorPOS: Democratizing Korean Tokenization with Voting-Based Open Corpus Annotation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference(LREC 2022).* 4975-4983. https://aclanthology.org/2022.lrec-1.531

Morris, J.X., Kuleshov, V., Shmatikov, V., & Rush, A.M. (2023). Text Embeddings Reveal (Almost) As Much As Text. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2023),* 12448-12460. https://doi.org/10.18653/v1/2023.emnlp-main.765

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., & Yang, Y. (2021). Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *Findings of the Association for Computational Linguistics(ACL 2021),* 1864-1874. https://doi.org/10.18653/v1/2022.findings-acl.146

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. *arXiv preprint arXiv*:2105.09680. https://doi.org/10.48550/arXiv.2105.09680

Shi, X., Padhi, I., & Knight, K. (2016). Does string-based neural MT learn source syntax?. *Proceedings of the 2016 conference on Empirical Methods in Natural Language Processing (EMNLP 2016),* 1526-1534.

Tsukagoshi, H., Sasano, R., & Takeda, K. (2021). DefSent: Sentence embeddings using definition sentences. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021),* 411-418. https://doi.org/10.18653/v1/2021.acl-short.52

Xu, L., Xie, H., Li, Z., Wang, F. L., Wang, W., & Li, Q. (2023). Contrastive learning models for sentence representations. *ACM Transactions on Intelligent Systems and Technology,* 14(4), 1-34. https://doi.org/10.1145/3593590

Yeon, J. & Brown Lucien. (2017). *Korean: a comprehensive grammar.* United Kingdom: Routledge.

Zhou, Z., Zhang, D., Xiao, W., Dingwall, N., Ma, X., Arnold, A. O., & Xiang, B. (2022). Learning dialogue representations from consecutive utterances. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 754-768. https://doi.org/10.18653/v1/2022.naacl-main.55

Abstract

KoSentEval: A Study of Korean Sentence Embedding Evaluation

Minhwa Jung* · Min Song**

Recently, efforts have focused on developing vector databases to compress large volumes of text data into a single dense vector, aiming to create global sentence embeddings that integrate semantic and syntactic information. In English research, the development of the SentEval library has enabled the evaluation of sentence embeddings, leading to its widespread adoption in languages such as Polish and Russian. However, in Korean research, there is a limitation due to challenges related to data availability, making it difficult for both academic research and commercial applications. To overcome this limitation, this paper introduces and releases KoSentEval. KoSentEval is an open-source library designed for evaluating Korean sentence embeddings, comprising two sub-tasks and eight probing tasks that validate the surface, syntactic, and semantic attributes of embeddings. Additionally, we conduct a comparative analysis of results obtained using multilingual, Korean-specific, and contrastive learning-based sentence encoders. This research not only contributes to measuring the quality of Korean sentence embeddings but also validates how well various sentence encoders capture the linguistic characteristics of the Korean language.

Key Words : Sentence Embedding, Embedding Evaluation, Korean Sentence Embedding, Probing Task, Contrastive Learning

Received : February 15, 2024 Revised : March 8, 2024 Accepted : March 11, 2024
Corresponding Author : Min Song

* Digital Analytics, Yonsei University
** Corresponding author: Min Song
Professor of Library and Information Science Department of Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: +82-2-2123-2405, Fax: +82-2-393-834, E-mail: min.song@yonsei.ac.kr

저 자 소 개

정민화

한국외국어대학교 언어인지과학과/language&AI에서 학사 학위를 취득하였으며 연세대학교 디지털애널리틱스융합협동과정에서 석사 학위를 취득하였다. 관심 연구 분야는 자연어처리, 언어 임베딩, 대화형 인공지능 등이다. 좋은 데이터와 언어기술의 힘을 믿는다.

송민

연세대 문헌정보학과 정교수, 디지털 애널리틱스, 인공지능 대학원 겸임교수이며 2014년에 언더우드 특훈교수로 임명되었다. 전공 분야는 텍스트 마이닝이며 2012년 연세대 부임 이후 SCI급 논문 120여 편과 단행본 3편을 저술하였다. Frontiers in Research Metrics and Analytics 저널에 텍스트 마이닝 분야 Specialty Chief Editor이고 Scientometrics와 Journal of Informetrics 저널의 편집위원으로 활동 중이다.