JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 6, pp. 21-29, June 2025
https://doi.org/10.9708/jksci.2025.30.06.021

Hi-BERT : A Semantic Relation Learning Model for Recruitment Ontology and Academic Papers

Sung-Kwang Song*, Hyeon-Jeong Mun**, Young-Ji Kim**, Yong-Tae Woo***

*Student, Dept. of Computer Engineering, Changwon National University, Gyeongsangnamdo, Korea
**Principal Research Engineer, Hibrainnet Co., Gyeongsangnamdo, Korea
***Professor, Dept. of Computer Engineering, Changwon National University, Gyeongsangnamdo, Korea

[Abstract]

In this study, we propose the Hi-BERT model to identify effectively and to learn semantic associations between recruitment ontology generated automatically from recruitment information and academic papers using generative AI. The proposed model can learn semantic associations between ontology-transformed sentences, which are structural information of recruitment ontology converted into natural language, and academic papers. In addition, domain-specific characteristics are learned from recruitment information and academic papers, and optimized by the contrastive learning technique. The proposed model can improve the limitations of knowledge transfer between different domains of existing sentence embedding models such as KLUE and KoSimCSE-BERT. In particular, the learning method utilizing ontology-transformed sentences proposed in this study can provide richer semantic understanding than existing models trained by simple text corpus. In the future, the proposed model is expected to be used as a base model for developing expert search systems.

▸ **Key words**: BERT, Pre-trained Language Model, Ontology, Generative AI, Contrastive Learning

[요 약]

본 연구에서는 생성형 AI를 이용하여 채용공고에서 자동 생성된 채용 온톨로지와 학술논문 간의 의미적 연관성을 효과적으로 학습하기 위한 Hi-BERT 언어모델을 제안한다. 제안 모델에서는 채용 온톨로지의 구조적인 정보를 자연어 형태로 변환한 온톨로지 변환 문장에 의해 채용공고와 학술논문 간의 의미적인 연결성을 학습한다. 또한, 채용공고와 학술논문에서 도메인별로 고유한 특성을 학습하여 대조 학습 기법으로 최적화한다. 제안 모델은 KLUE와 KoSimCSE-BERT 같은 기존의 문장 임베딩 모델이 가지는 서로 다른 도메인 간의 지식 전이의 한계를 개선할 수 있다. 특히, 본 연구에서 제안한 온톨로지 변환 문장을 활용한 학습 방법은 단순 텍스트 코퍼스에 의해 학습된 기존 모델보다 의미론적 이해가 가능하다. 향후, 본 제안 모델은 전문가 검색 시스템 개발을 위한 기반 언어모델로 사용할 수 있으리라 기대된다.

▸ **주제어**: BERT, 사전학습 언어모델, 온톨로지, 생성형 AI, 대조학습

• First Author: Sung-Kwang Song, Corresponding Author: Yong-Tae Woo
*Sung-Kwang Song (sksong@hibrain.net), Dept. of Computer Engineering, Changwon National University
**Hyeon-Jeong Mun (hjmun@hibrain.net), Hibrainnet Co.
**Young-Ji Kim (yjkim@hibrain.net), Hibrainnet Co.
***Yong-Tae Woo (ytwoo@changwon.ac.kr), Dept. of Computer Engineering, Changwon National University
• Received: 2025. 05. 08, Revised: 2025. 05. 29, Accepted: 2025. 05. 29.

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

I. Introduction

현대 사회는 인공지능이나 반도체와 같은 핵심 기술이 산업을 선도하는 4차 산업혁명 시대를 맞이하고 있다. 기관에서 필요로 하는 전문가를 채용하는 일은 조직의 경쟁력 확보와 미래 성장을 위한 중요한 이슈 중의 하나이다. 하지만 채용기관과 전문가 간의 전공 일치도를 판단하기 어려운 관계로 후보자의 프로필을 수동적으로 추천하는 방식에 의존하고 있다. 또한 전문가들은 하이브레인넷 같은 전문 인력 채용사이트에서 직접 채용공고를 검색하는 방식을 사용하고 있다[1-5]. 이러한 방식은 채용기관이나 지원자 모두 시간과 비용이 많이 소요되어 비효율적이다.

최근에는 IT 기술을 이용하여 전문가를 효과적으로 검색하기 위한 기법에 대한 연구가 활발하게 진행되고 있다. 전문가 검색을 위하여 제시된 방식은 키워드 기반 검색, 벡터 기반 검색, 온톨로지 기반 검색 그리고 인공지능 기술을 활용한 기법 등이 있다.

먼저, 키워드 기반 검색 기법은 채용분야에 대한 단순 키워드 매칭에 의해 검색하는 방식이다. 이 방식은 구현은 쉽지만 문맥상 의미를 충분히 반영하기 어렵다. 벡터 기반 검색 기법은 문서의 의미를 수치로 표현하여 문맥적인 정보를 부분적으로 반영할 수 있다. 하지만 이 방법은 전문 분야에서 요구하는 세부적 의미를 충분하게 반영하기 어려운 문제점이 있다[6].

온톨로지 기반의 검색 기법은 특정 도메인의 개념과 개념 간의 관계를 명시적으로 구조화하여 의미적으로 검색하기 위한 기법이다[7]. 이 기법은 전문 분야의 지식을 의미적으로 표현하여 의미 기반의 전문가 검색이 가능하다. 하지만 온톨로지 구축과 유지 보수를 위한 전문 인력이 필요하고, 많은 시간과 비용이 소모되는 단점이 있다. 또한 전문 지식이 급속도로 발전하는 첨단 분야에서는 기 구축된 온톨로지를 최신 상태로 유지하기 어렵다[8].

최근에는 인공지능 기술에 기반한 BERT (Bidirectional Encoder Representations from Transformers) 모델 같은 사전 학습 언어모델을 이용한 검색 기법에 대한 연구가 진행되고 있다[9-11]. 이 모델은 방대한 텍스트로부터 사전 학습된 언어모델로써 자연어 처리 분야에서 의미적 관계나 문맥 정보를 이해하는 데 우수한 성능을 보인다[12]. 하지만 이 모델은 일반 도메인에 기반하여 학습된 모델로 특화된 분야의 전문 지식을 효과적으로 반영하기 어렵다[13]. 이에 따라 특정 도메인에 대한 전문적인 지식을 반영한 새로운 검색 기법에 대한 연구가 필요하다.

본 연구에서는 채용공고 데이터와 학술논문 데이터 간의 의미적인 연관성을 학습한 채용-학술 도메인에 특화된 Hi-BERT 언어모델을 제안한다. 제안 모델에서는 생성형 AI를 이용하여 채용기관에서 제시한 직무와 관련된 의미를 추출하여 채용공고 온톨로지를 자동으로 생성하는 방법을 제안한다. 채용공고 온톨로지는 학습용 데이터셋을 구축하는데 이용된다. 그리고 논문에서 구성한 학술논문 데이터셋과 함께 사전 언어모델을 학습한 도메인 특화형 언어모델을 제안한다.

제안 모델의 효율성을 검증하기 위하여 실험용 데이터셋을 수집하여 성능 평가를 진행하였다. 먼저, 채용공고 사이트에서 석박사급 전문가에 대한 채용공고를 수집하였고, KCI(Korean Citation Index)에서 학술논문을 수집하였다. 실험 결과, 본 연구에서 제안한 모델이 기존의 사전 학습 언어모델보다 의미기반 검색에서 우수한 성능을 보였다. 제안 모델은 채용공고와 학술논문에서 추출한 정보를 생성형 AI 기법의 언어모델로 결합하여, 기관에서 요구하는 전문가를 효과적으로 검색하기 위한 전문가 검색 시스템 개발에 적용될 수 있으리라 기대한다.

II. Preliminaries

1. Related works

1.1 Expert Search Method based on Keywords and Vectors

먼저, 키워드 기반의 전문가 검색 기법은 전문가에 대한 프로필 정보로부터 사전에 정의된 키워드의 존재 여부를 비교하여 검색하는 기법이다[14]. 하지만 이 기법은 '인공지능'과 'AI'처럼 동일한 개념을 다른 용어로 표현한 경우에 유사성을 식별하기 어려운 단점이 있다. 벡터 기반 검색 기법은 텍스트 데이터를 벡터 형태로 변환하고, 벡터 간의 유사도에 따라 검색하는 방식이다. 이 기법은 문서 내에서 단어의 중요도를 수치화하여 검색하는 TF-IDF(Term Frequency-Inverse Document Frequency) 기법이나 문맥을 고려한 단어 임베딩을 통해 의미적인 유사성을 표현하는 Word2Vec 기법이 있다[15]. 하지만 이 기법은 문서의 의미를 수치 벡터로 압축하는 과정에서 세부적인 의미상의 차이가 무시될 수 있다.

1.2 Ontology-based Expert Search Method

온톨로지 기반의 검색 기법은 도메인 내의 개념과 관계를 명시적으로 정의하여 구조화된 온톨로지를 구축하고, 이를 기반으로 검색하는 기법이다[16]. 이 기법은 해당 도

메인에서 개념 간의 계층적인 관계와 의미적인 연결성을 표현할 수 있다. 하지만 온톨로지를 구축하는 과정에서 전문 분야에 대한 이해와 인적·시간적 비용이 많이 소요되고, 전문가의 지속적인 참여가 필요하다. 또한 첨단 분야에서 온톨로지를 최신 상태로 유지하기 어렵다. 최근에는 생성형 AI를 활용하여 온톨로지 구축을 자동화하기 위한 연구가 진행되고 있다[17,18]. 하지만 이들 연구에서 LLM(Large Language Model)은 온톨로지의 자동 생성보다는 온톨로지의 개별 요소를 예측하기 위해 사용하는 관계로 온톨로지를 자동 생성하는데 한계가 있다.

1.3 BERT based Expert Search Method

BERT 모델은 자연어 처리 분야에서 획기적인 성과를 가져온 사전학습 언어모델이다[19]. Devlin et al.에 의해 제안된 이 모델은 양방향 트랜스포머 구조를 기반으로 텍스트의 문맥적 표현을 효과적으로 학습할 수 있다. 이 모델의 핵심 학습 기법인 MLM(Masked Language Model) 방식은 입력 텍스트에서 일정 비율의 토큰을 무작위로 마스킹한 후, 주변 문맥을 활용하여 마스킹된 토큰의 원래 단어를 예측하도록 학습하는 방식이다. 이러한 MLM 기법은 문맥을 통한 언어 이해 능력을 학습하는데 중요한 역할을 한다. 또한 전통적인 언어 모델에서는 한 방향으로 문맥을 처리하는 반면, MLM은 마스킹된 단어의 좌우 문맥을 동시에 고려하여 보다 정확한 의미적 표현을 학습할 수 있다. 그러나, 일반 도메인에서 사전 학습된 BERT 기반의 검색 기법은 특정 전문 분야의 전문가를 효과적으로 검색하는데 한계가 있다.

1.4 Contrastive Learning

대조 학습(Contrastive Learning) 기법은 유사 데이터 쌍의 임베딩 간의 거리는 최소화하고, 비 유사 데이터 쌍 간의 거리는 최대화하는 자기지도학습 기법이다[20]. 이 기법은 InfoNCE(Information Noise-Contrastive Estimation) 손실 함수를 통해 상호 정보를 최적화하며, 레이블 없이도 데이터의 의미적인 구조를 효과적으로 학습할 수 있다. 이 기법은 약한 감독 신호(Weak Supervision Signal)만으로도 데이터 간의 관계를 모델링할 수 있다. 따라서 레이블이 없는 데이터셋에서도 학습이 가능한 장점이 있다. 최근에 이 기법은 컴퓨터 비전이나 자연어 처리 분야에서 우수한 성과를 보이고 있다.

III. The Proposed Scheme

기존의 BERT 모델은 서로 다른 도메인 간의 의미적인 연관성을 찾기 어려운 문제점이 있다. 본 연구에서는 석박사급 전문가 채용공고의 모집분야와 학술논문을 결합하여 서로 다른 도메인 간의 의미적인 연관성을 효과적으로 학습하는 Hi-BERT 언어모델을 제안한다. 제안 모델에서는 생성형 AI에 의해 자동 생성된 온톨로지를 이용하여 사전 언어모델을 학습하는 새로운 방식을 제안한다. 그리고 기존의 BERT 모델을 확장하여 채용공고와 학술논문처럼 서로 다른 도메인 간의 의미 관계를 효과적으로 학습하는 새로운 형태의 도메인 특화형 언어모델을 제안한다.

먼저, 제안 모델에서는 수집한 채용공고에 대해 생성형 AI를 사용하여 채용공고 온톨로지를 자동으로 생성한다. 그리고 KCI에서 수집한 논문 데이터에 대해 전처리 과정을 거쳐 학술 데이터셋으로 구성한다. 학습용 데이터셋에 의해 기존의 BERT 모델을 학습시키기 위하여 도메인 적응 기법과 대조 학습 방법을 순차적으로 적용한다. 이를 통해 채용공고와 학술논문 간의 의미적 연관성을 이해할 수 있는 Hi-BERT 모델을 구축한다. 그림 1은 본 연구에서 제안한 Hi-BERT 모델의 전체적인 개념도이다.

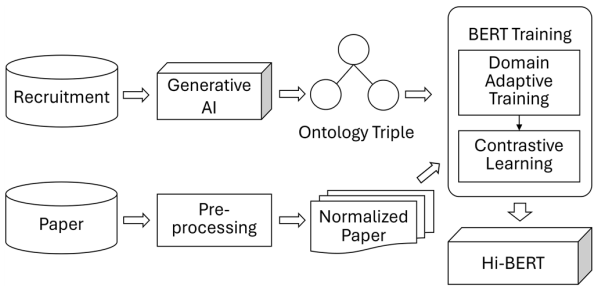

Fig. 1. Concept Diagram for Hi-BERT Model

1. Generation of Training Dataset

본 과정은 채용공고와 학술논문 도메인 간의 의미 관계를 효과적으로 학습하기 위하여 채용공고와 학술논문에 대한 데이터셋을 생성하기 위한 과정이다.

1.1 Generation of Recruitment Dataset based on Recruitment Ontology

본 연구에서는 생성형 AI를 사용하여 도메인 전문가의 도움없이 채용공고에서 전문분야에 대한 의미를 추출하고, 온톨로지를 자동으로 생성하는 방법을 제안한다. 본 연구에서 채용 온톨로지는 학술논문과의 의미적 관계를 연결하기 위한 목적으로 구성된다. 채용공고에는 기관소개, 직

무, 전공, 필요기술, 응시자격, 제출서류, 전형방법 등과 같은 다양한 항목을 포함하고 있다. 채용 온톨로지는 이러한 항목 중에서 채용 분야와 관련된 직무, 전공, 전공 키워드, 기술, 자격요건에 대한 관계로 설계된다.

본 연구에서는 채용 온톨로지를 기반으로 채용공고를 분석하고, 자동으로 트리플(주어-술어-목적어)을 생성하는 방법을 제안한다. 예를 들어, '제목: 인공지능 전문가 채용, 기술: 딥러닝 기술보유'와 같은 채용공고는 채용 온톨로지의 정의에 따라, '인공지능-hasSkill-딥러닝' 과 같은 형태의 트리플로 생성된다. 제안 방법은 전문 지식이 빠르게 발전하는 최신 분야에서 온톨로지를 최신 상태로 유지할 수 있는 장점을 제공한다. 그림 2는 생성형 AI를 사용하여 표나 텍스트 형태의 채용공고를 채용 온톨로지 기반의 트리플로 자동 생성하는 예이다.

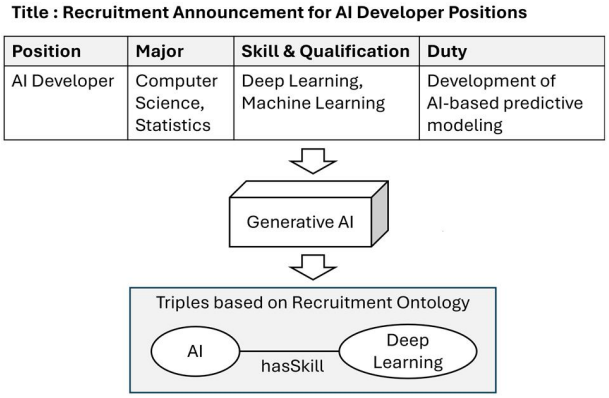

Fig. 2. Automatic Generation of Triples based on Recruitment Ontology using Generative AI

그림 2에서 자동 생성된 트리플은 채용공고 도메인 학습과 학술논문 간의 의미적 관계와 관계 보정을 위하여 도메인 적응 기법과 대조 학습 과정에서 사용된다. 이를 위하여 온톨로지의 직무와 전공, 직무와 기술, 전공과 전공 키워드, 직무와 자격요건 간의 관계를 자연어 문장으로 변환한다. 예를 들면, '{AI 개발자}는 {컴퓨터공학} 분야의 전공자입니다.', '{AI 개발자}는 {딥러닝, 머신러닝} 기술이 요구됩니다.' '{AI} 분야는 {딥러닝, 머신러닝} 전공 키워드를 포함합니다.'와 같은 자연어 문장으로 생성된다.

1.2 Generation of Dataset based on Academic Papers

본 과정은 학술논문 데이터셋을 생성하기 위한 과정이다. KCI에서 수집한 학술논문에서 추출한 제목, 초록, 키워드, 분류 체계, 저자 정보, 출판 연도와 같은 정보를 분리하여 메타 데이터를 구성한다. 그리고 구조화된 CSV 포맷으로 변환하여 제안 모델을 학습하기 위한 학술논문 데이터셋을 생성한다. 그림 3은 학술논문에서 메타 데이터를 추출하여 학술 도메인을 학습하기 위하여 데이터셋을 생성하는 과정이다.

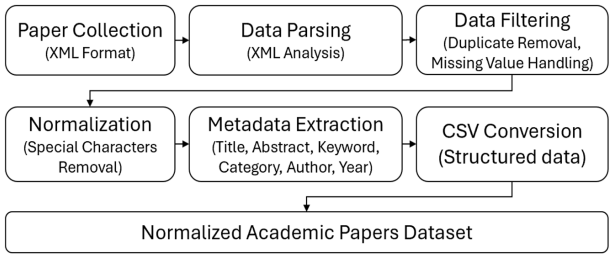

Fig. 3. Process of Generation Academic Papers Dataset

2. Hi-BERT Model

본 연구에서는 채용공고와 학술논문을 결합하여 서로 다른 도메인 간의 의미적인 연관성을 효과적으로 학습한 Hi-BERT 모델을 제안한다. Hi-BERT 모델은 도메인 적응 학습 단계와 대조 학습 단계를 통해 구축된다. 먼저, 도메인 적응 학습 단계는 MLM 기법을 이용하여 해당 도메인에서 고유한 용어와 맥락을 이해하기 위한 과정이다. 그리고 대조 학습 단계는 의미적으로 관련 있는 정보는 서로 가깝게, 관련 없는 정보는 서로 멀게 임베딩하는 교차 도메인 통합 학습 과정이다.

2.1 MLM-based Domain Adaptive Training

본 과정은 Hi-BERT 모델이 채용공고와 학술논문의 문맥적 특성을 이해할 수 있도록 MLM 학습 방법을 사용하여 적응 학습하는 과정이다. 이 과정에서 채용공고에는 [JOB], 학술논문에는 [PAPER], 그리고 채용 온톨로지에서 생성한 문장에는 [TRIPLE]을 추가하여 각 도메인을 구별한다. 예를 들어, 채용공고는 '[JOB] 직무: 인공지능 전문가 채용, 전공 : 인공지능, 기술 : 딥러닝' 형태로, 학술논문은 '[PAPER] 제목 : 딥러닝 기반 스팸처리 모델, 분류 : 보안, 키워드 : 딥러닝, 스팸처리' 형태로, 그리고 온톨로지 변환 문장은 '[TRIPLE] 인공지능 분야는 딥러닝 기술이 필요하다'와 같은 형태로 학습 데이터를 구성한다.

MLM 학습 방법은 입력 텍스트의 일부를 무작위로 선택하여 [MASK] 토큰으로 대체하고, 주변 맥락을 기반으로 마스킹된 토큰에 대한 원래 단어를 예측하는 방식이다. 예를 들어, '인공지능 전문가는 [MASK] 기술이 필요하다'와 같은 문장에서 마스킹된 부분을 '딥러닝', '파이썬' 등과 같이 적절한 단어를 예측하도록 훈련된다. 이 과정을 통하여 제안 모델은 문맥 속에서 단어의 의미와 관계를 이해하

게 된다. 그림 4는 채용공고 데이터셋을 대상으로 MLM 학습 방법에 의해 학습하는 과정이다.

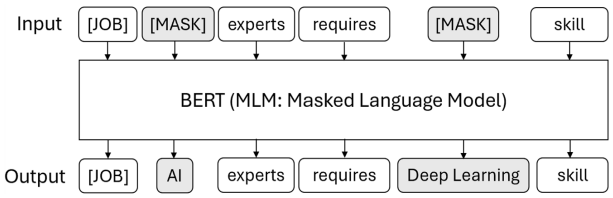

Fig. 4. The Process of Training by MLM on the Recruitment Dataset

제안 모델에서 온톨로지 변환 문장은 서로 다른 도메인 간의 연결성을 강화하는 역할을 한다. 즉, 온톨로지 변환 문장을 통해 채용공고와 학술논문 간의 연관성을 학습함으로써 채용공고와 연관성이 약한 학술논문도 효과적으로 연결할 수 있는 교량 역할을 한다. 예를 들어, '인공지능 전문가 채용' 채용공고와 '딥러닝 기반 스팸처리 모델 연구' 학술논문은 문맥상 연관성이 약하다. 하지만, '인공지능 엔지니어는 딥러닝 기술이 필요하다'라는 온톨로지 변환 문장과 채용공고와의 연관성에 의해, 채용공고와 학술논문의 문맥적인 연관성을 찾을 수 있다. 이러한 과정을 통해 채용공고와 학술논문에 특화된 채용공고 도메인 인코더와 학술논문 도메인 인코더가 생성된다.

2.2 Cross-domain Integration through Contrastive Learning

본 과정은 MLM에 의해 사전 학습된 두 도메인 인코더를 대조 학습에 의해 통합하는 과정이다. 이를 통해 단일 BERT 구조를 기반으로 채용공고와 학술논문의 도메인별 특성을 반영하여 의미적 관계를 학습할 수 있다. 대조 학습은 채용공고와 학술논문 간에 연관성이 높으면 긍정 쌍으로, 낮으면 부정 쌍으로 학습하는 기법이다. 본 연구에서는 긍정 쌍과 부정 쌍에 대한 효과적인 분류를 위하여 약한 감독 신호를 사용한다. 그리고 채용공고와 학술논문 간의 키워드, 전공분야 매칭 관계, 온톨로지 변환 문장과의 유사도를 약한 관계 신호로 사용한다.

대조 학습에서 온톨로지 변환 문장은 채용공고와 학술논문 간의 의미적인 관계를 보정하는 역할을 한다. 예를 들어, '인공지능 시대에 농업 기술의 발전 방향 연구' 학술논문과 '인공지능 전문가 채용' 채용공고는 키워드 매칭에 의해 긍정 값을 가진다. 하지만, '인공지능은 컴퓨터공학 분야이다.' 와 같은 온톨로지 변환 문장에 의해 '인공지능'은 문맥상 전공이나 전공 분류를 나타내는 용어임을 이해한다. 그러나, 학술논문에서 '인공지능'은 전공이나 기술적인 의미가 아니라 문맥상 산업 패러다임을 의미한다. 따라서 온톨로지 변환 문장과 유사도 비교에 의해 부정 값으로 보정된다. 결론적으로, 온톨로지 변환 문장에 의해 긍정 쌍의 거리는 더욱 가깝게 되고, 부정 쌍의 거리는 더욱 멀어진다. 그림 5는 대조 학습 후에 채용공고와 학술논문 간의 임베딩 거리 변화를 보여주는 예이다.

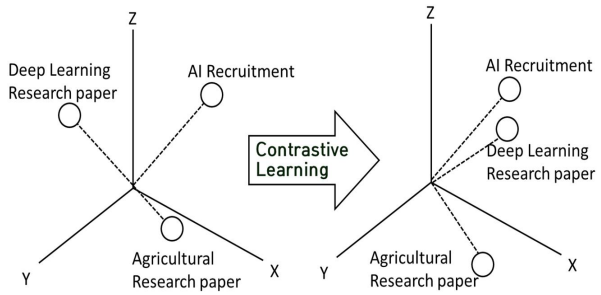

Fig. 5. Changes in Embedding Distance between Recruitment and Academic Papers

IV. Experimental Results

본 연구에서 제안한 Hi-BERT 모델을 구축하고 성능을 평가하기 위하여, 석박사급 전문가 채용사이트인 하이브레인넷에서 채용공고 1,014건과 한국연구재단에서 제공하는 한국학술지색인(KCI)에서 학술논문 21,911건을 수집하여 실험하였다. 수집된 데이터는 무작위로 분할하여 70%는 학습용, 15%는 검증용, 15%는 테스트용으로 사용하였다. 제안 모델의 성능을 비교하기 위하여 기존의 한국어 사전학습 언어모델인 KLUE(Korean Language Understanding Evaluation) 모델과 KoSimCSE-BERT (Korean Simple Contrastive Learning Sentence Embeddings) 모델과 성능 비교 실험을 수행하였다.

1. Generating Datatset Collection

본 단계에서는 생성형 AI를 사용하여 1,014건의 채용공고를 분석하여 채용 온톨로지 트리플을 생성하였다. 생성된 트리플은 총 131,184건이며, 전공 키워드(hasMajorKeyword) 관계에 대한 출현 빈도가 가장 높았다. 관계 유형의 출현 빈도는 채용공고에서 관계별 중요도를 나타내는 주요 지표로 사용되며, 대조 학습에서 약한 감독 신호에 대한 가중치 값으로 사용된다. BERT 모델에서는 문장에 대해 학습하는 관계로 트리플을 자연어 문장으로 변환하였다. 그림 6은 온톨로지 트리플을 자연어 문장으로 변환하는 과정이다.

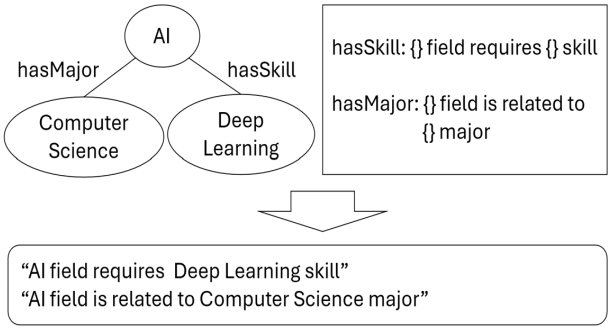

Fig. 6. Converting Ontology-based Triples to Natural Language Sentences

2. Experiments for Hi-BERT Model Training

본 실험에서는 Hi-BERT 모델을 구축하기 위하여 MLM 학습과 대조 학습을 수행하였다. 실험에서는 한국어 채용공고와 학술논문에 특화된 언어 표현을 효과적으로 학습하기 위하여 KoSimCSE-BERT를 사전학습을 위한 기본 모델로 사용하였다.

2.1 Experimental Process for MLM-based Domain Adaptive Training

본 실험에서는 Hi-BERT 모델이 채용공고와 학술논문 간의 문맥적 특성을 효과적으로 이해할 수 있도록 MLM 기반의 도메인 적응 학습 실험을 순차적으로 수행하였다. 채용공고와 학술논문이라는 서로 다른 도메인의 문장 특성과 맥락을 명확하게 인식할 수 있도록 '[JOB]', '[PAPER]', '[TRIPLE]'이라는 도메인에 특화된 특수 토큰을 문장에 추가하여 학습하였다. 그림 7은 채용공고, 학술논문, 온톨로지 변환 문장의 텍스트에 도메인에 특화된 토큰을 추가하여 구성한 예이다.

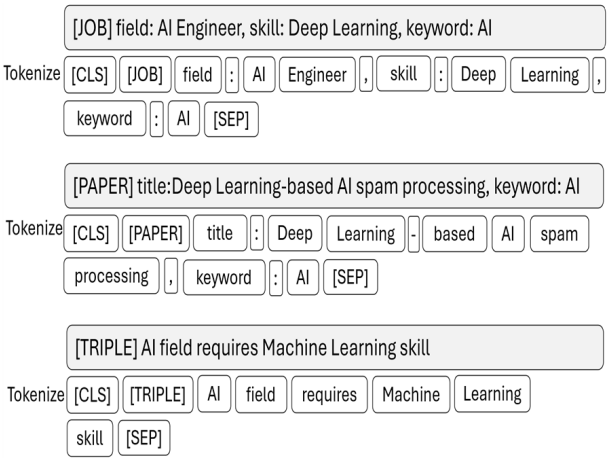

Fig. 7. Tokenized Results by Adding Domain-specific Tokens to the Input Text

제안 모델에서 MLM 학습 실험은 토큰 값 중에서 15%를 무작위로 선택하여 [MASK] 토큰으로 대체하였다. MLM 학습의 정확도는 교차 엔트로피 손실 함수(Cross Entropy Loss)로 평가하였다. 이 값이 작을수록 [MASK]된 토큰들을 원래 단어로 예측한 분포가 높다는 것을 의미한다. MLM 학습은 손실 함수의 값이 더 이상 감소하지 않고, 수렴하는 시점에 종료한다. 실험에서는 3회 반복 학습(Epoch)을 통해 손실 함수가 수렴되어 학습을 종료하였고, 채용공고 도메인 인코더와 학술논문 도메인 인코더가 구축되었다. 실험에서 도메인 최적화를 위해 하이퍼파라미터를 휴리스틱하게 구성하여 실험하였다. 표 1은 MLM 학습 실험 과정에 사용된 하이퍼파라미터의 종류와 값이다.

Table 1. Hyperparameters used in MLM-based Domain Adaptive Training

Hyperparameter	Value
Base Model	KoSimCSE-BERT
Batch Size	32
Max Sequence Length	256
Epoch	3
Learning Rate	1e-5
Embedding Dimension	768
Dropout Rate	0.1

2.2 Experimental Process of Cross-domain Integration through Contrastive Learning

본 실험은 MLM 학습 과정을 통해 구축된 채용공고와 학술논문 도메인 인코더를 대조 학습으로 통합하여, 최종적으로 Hi-BERT 모델을 구축하기 위한 과정이다. 먼저, 대조 학습에 필요한 긍정 쌍, 부정 쌍의 데이터셋을 구성하기 위하여 채용공고와 학술논문을 대상으로 약한 감독 신호 생성 방법을 사용하여 후보를 분류하였다. 그리고, 약한 신호에 대한 신뢰도를 판단하기 위하여 임계값 기반의 접근법을 적용하였다. 실험에서는 임계값이 0.155 이상이면 긍정 쌍, 0.145 이하이면 부정 쌍으로 분류하고, 불확실성이 높은 0.145 ~ 0.155 사이 값은 제외하였다. 또한, 데이터셋의 균형을 위하여 긍정 쌍과 부정 쌍은 약 1:1의 비율로 구성하였다. 총 4,537개의 학습 쌍에 대하여 52.4%는 긍정 쌍, 47.6%는 부정 쌍으로 구성하였다.

대조 학습은 도메인에 관계없이 긍정 쌍은 가깝게, 부정 쌍은 멀어지도록 임베딩 공간을 재정렬하기 위하여 대조 손실 함수(Contrastive Loss)를 사용한다. 이 함수는 각 쌍의 두 임베딩 벡터 간의 유클리드 거리를 계산한 값으로, 긍정 쌍은 거리가 멀수록 그리고 부정 쌍은 거리가 가까울수록 손실이 커진다. 이 함수의 성능은 각 쌍에 대한

비교 강도를 조절하는 온도(Temperature)와 최소 거리를 나타내는 마진(Margin) 하이퍼파라미터의 영향을 받는다.

본 실험에서는 그리드 서치(Grid Search) 기법을 통해 하이퍼파라미터를 탐색하였다. 그리드 서치 기법은 모든 하이퍼파라미터 조합을 대상으로 모델을 학습하고 평가하여 가장 우수한 성능을 보이는 조합을 찾아내는 방법이다. 온도는 [0.05, 0.07, 0.1]의 범위에서, 마진 파라미터는 [0.1, 0.2, 0.3] 범위에서 총 9가지 조합을 교차 검증하였다. 각 조합에 대해 검증 데이터셋의 MRR(Mean Reciprocal Rank)와 NDCG(Normalized Discounted Cumulative Gain) 지표로 평가하였고, 온도는 0.07과 0.2 조합일 때 가장 우수한 성능을 보였다. 표 2는 대조 학습 실험에 사용된 하이퍼파라미터의 종류와 값이다.

Table 2. Hyperparameters used in Contrastive Learning

Hyperparameter	Value
Base Model	KoSimCSE-BERT
Batch Size	32
Max Sequence Length	256
Epoch	3
Learning Rate	5e-6
Embedding Dimension	768
Dropout Rate	0.1
Temperature	0.07
Margin	0.2

2.3 Comparative Experimental Results between Existing Model and Hi-BERT Model

Hi-BERT 모델의 성능을 검증하기 위하여 기존의 KLUE 모델 및 KoSimCSE-BERT 모델과 비교 실험을 수행하였다. 실험은 채용공고 테스트 데이터셋의 각 샘플에 대해 학술논문 데이터셋으로부터 연관성이 높은 학술논문을 검색하는 방식으로 수행하였다. 성능 평가는 채용공고 테스트 데이터셋 300건과 학술논문 테스트 데이터셋 1,000건에 대하여 모델별로 채용공고와 유사한 학술논문을 검색한 결과로 평가하였다.

성능 평가에서 사용한 지표는 언어모델 평가에서 일반적으로 사용하는 MRR(Mean Reciprocal Rank), NDCG(Normalized Discounted Cumulative Gain), Precision, MAP(Mean Average Precision) 지표를 사용하였다. MRR 지표는 질문에 대해 모델이 첫 번째 정답을 얼마나 빠르게 제공하는지를 평가하는 지표이다. NDCG는 정답의 순서를 고려하여 중요한 정답이 먼저 제공되는지를 평가하는 지표이다. Precision은 하나의 질문에 대한 상위 N개의 검색 결과에서 정답이 많이 포함되었는지 여부를 평가하는 지표이다.

그림 8은 Hi-BERT 모델과 기존 BERT 모델 간의 성능을 비교 실험한 결과이다. 그림 8에서 Precision@5 지표의 결과는 제안 모델이 채용공고와 학술논문 도메인에서 높은 관련성을 제공할 수 있음을 의미한다. 또한 검색 결과에서 전반적인 관련성을 평가하는 MAP 지표의 결과는 온톨로지 트리플 문장 기반의 학습이 단순 유사성 학습을 넘어 온톨로지에 명시한 관계까지 효과적으로 학습할 수 있음을 보여준다. 특히, 제안 모델은 채용공고와 학술논문 도메인에서 기존의 BERT 모델보다 대부분 지표에서 우수한 성능을 보였다.

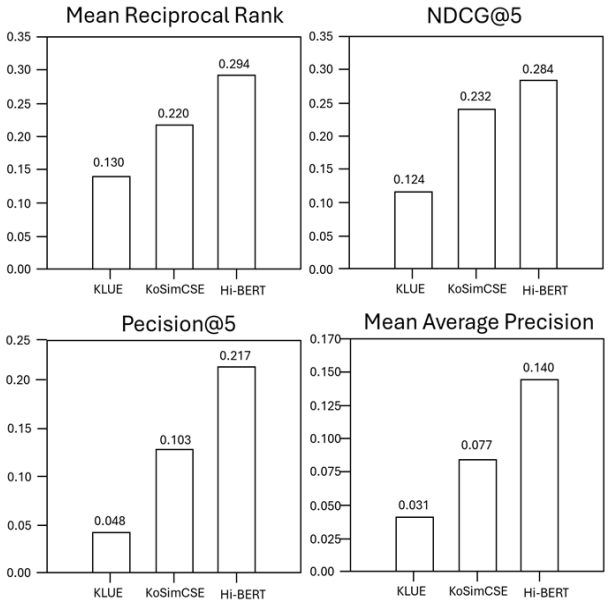

Fig. 8. Comparative Experimental Results between Existing Model and Hi-BERT Model

V. Conclusions

본 연구에서는 생성형 AI에 의해 자동 생성된 온톨로지를 이용하여 사전언어모델을 학습하는 Hi-BERT 모델을 제안하였다. 제안 모델은 기존의 BERT 모델을 확장하여 채용공고와 학술논문처럼 서로 다른 도메인 간의 의미 관계를 효과적으로 연결하는 새로운 형태의 도메인 특화형 모델이다. 먼저, 본 연구에서 설계한 채용 온톨로지를 기반으로 생성형 AI를 사용하여 자동으로 트리플을 생성하였다. KCI에서 수집한 논문 데이터는 전처리 과정을 거쳐 학술 데이터셋으로 구성하였다. BERT 모델에 대한 학습은 도메인 적응 방법과 대조 학습 방법을 순차적으로 적용하였다. 이를 통해 채용공고와 학술논문 간의 의미적 연관성을 이해할 수 있는 Hi-BERT 모델을 구축하였다.

Hi-BERT 모델의 성능을 검증하기 위하여 기존의 KLUE 모델 및 KoSimCSE-BERT 모델과의 비교 실험을 진행하였다. 성능은 각 모델별로 채용 데이터셋을 기반으로 유사한 학술논문 데이터셋을 검색한 결과를 비교하여 평가하였다. 실험 결과, 제안 모델이 기존의 사전학습 언어모델보다 의미기반 검색에서 우수한 성능을 보였다. 특히, 채용공고를 기반으로 학술논문을 검색하는 경우에는 다른 모델보다 검색 정확도가 더 우수하였다.

결과적으로, 본 연구에서 제안한 채용 온톨로지를 이용한 방법은 채용공고와 학술논문 도메인 간의 의미적인 연결성을 강화하는 효과를 보였다. 즉, 온톨로지 기법을 통해 채용공고와 학술논문 간의 연관성을 학습함으로써 채용공고와 연관성이 약한 학술논문도 의미적으로 연결될 수 있음을 보였다. 본 모델은 연구기관에서 요구하는 채용 분야와 관련된 논문을 효과적으로 검색할 수 있는 방법을 제공하여, 전문가 검색 시스템 개발을 위한 언어모델로 사용할 수 있으리라 기대한다. 앞으로 본 연구에서 제안한 채용공고 온톨로지 기법과 다양한 분야에 대한 학술 온톨로지를 자동으로 생성하여, 채용공고와 학술 데이터 간의 의미적 관계를 더욱 강화한 검색 모델을 개발할 예정이다.

REFERENCES

[1] S. W. Son and J. Y. Oh, “A Study on the Performances of AI Recruitment System: A Case Study on Domestic and Abroad Companies”, The Journal of Korean Career·Entrepreneurship & Business Association, Vol. 7, No. 2, pp.137-155, March 2023. DOI: 10.48206/kceba.2023.7.2.137

[2] S. H. Son, “A Study on the Exploration of the Core Capabilities of Design Future Talent against the Fourth Industrial Revolution”, pp. 305-315, Jun. 2019. DOI: 10.18208/ksdc.2019.25.2.305

[3] R&D Trend Report, https://www.kird.re.kr/newsletter/html/vol125/images/sub/KIRD_R&D_HDR_vol.4.pdf

[4] HRD Trend Report, https://www.kpc.or.kr/download/pt/KPC_2025_HRD_Trend_Report.pdf

[5] G. H. Han, H. Y. Kim, T. T. Lim, D. J. Choi, H. B. Lee, Y. H. Ho, D. W. Pyun, M. J. Bang, J. W. Jeon, K. S. Bok, and J. S. Yoo, “Expert Search System Through Analysis of Multiple Academic Information”, Proceedings of the Korean Institute of Communication Sciences Conference, Vol. 2020, No. 8, pp. 1331-1332, 2020.

[6] H. S. Kang and J. H. Yang, “Performance Comparison of Word2vec and fastText Embedding Models”, Journal of Digital Contents Society, Vol. 21, No. 7, pp. 1335-1343, Jul. 2020. DOI: 10.9728/dcs.2020.21.7.1335

[7] H. J. Mun, I. H. Jun, and Y. T. Woo, “A Researcher Model based on Ontology and a Social Network Construction Technique”, Journal of Korea Multimedia Society, Vol. 12, No. 7, pp. 1022-1031, Jul. 2009.

[8] F. Freitas, H. Stuckenschmidt, and N. F. Noy, “Ontology Issues and Applications Guest Editor’s Introduction”, Journal of the Brazilian Computer Society, Vol. 11, No. 2, pp. 5-16, Nov. 2005. DOI: 10.1590/S0104-65002005000300001

[9] I. A. Mannix and E. Yulianti, “Academic expert finding using BERT pre-trained language model”, International Journal of Advances in Intelligent Informatics, Vol. 10, No. 2, pp. 280-295, May 2024. DOI: doi.org/10.26555/ijain.v10i2.1497

[10] Y. J. Kim, J. H. Kim, J. M. Lee, M. J. Jang, Y. J. Yum, S. T. Kim, U. S. Shin, Y. M. Kim, and H. J. Joo, “A pre-trained BERT for Korean medical natural language processing”, Scientific Reports, Vol. 12, No. 1, Aug. 2022. DOI: 10.1038/s41598-022-17806-8

[11] G. S. Park, and J. T. Kim, “Legal search method using S-BERT”, Journal of the Korea Society of Computer and Information, Vol. 27, No. 11, pp. 57-66, Nov. 2022. DOI: 10.9708/jksci.2022.27.11.057

[12] M. Belocif and C. Bieman, “Probing Pre-trained Language Models for Semantic Attributes and their Values”, In Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2554-2559, Nov. 2021. DOI: 10.18653/v1/2021.findings-emnlp.218

[13] A. Nayak, H. Timmapathini, K. Ponnalagu, and V. G.Venkoparao, “Domain adaptation challenges of BERT in tokenization and sub-word representations of Out-of-Vocabulary words”, Proceedings of the First Workshop on Insights from Negative Results in NLP, pp. 1-5, Nov. 2020. DOI: 10.18653/v1/2020.insights-1.1

[14] C. K. Park and G. C. Yang, “Design and Implemantation of Information Retrieval System based on Semantic Information”, Proceedings of the Korea Contents Association Conference, Vol. 2, No. 2, pp. 265-268, 2004.

[15] D. S. Park and H. J. Kim, “ A Proposal of Join Vector for Semantic Factor Reflection in TF-IDF Based Keyword Extraction”, Journal of Korean institute of information technology, Vol 16. No. 2, pp. 1-16, Feb. 2018. DOI: 10.14801/jkiit.2018.16.2.1

[16] J. X. Huang, J. A. Shin, and K. S. Choi, “Building Domain Ontology through Concept and Relation Classification”, Korean Institute of Information Scientists and Engineers, Vol. 35, No. 9, pp. 562-571, Sep. 2008.

[17] H. B. Giglou, J. D’Souza, and S. Auer, “LLMs4OL: Large Language Models for Ontology Learning”, The Semantic Web – ISWC 2023, pp. 408-427, Oct. 2023. DOI: 10.1007/978-3-031-47240-4_22

[18] H. B. Giglou, J. D'Souza, F. Engel, and S. Auer, "LLMs4OM: Matching Ontologies with Large Language Models", The Semantic Web: ESWC 2024 Satellite Events, pp. 25-35, Jan. 2025. DOI: 10.1007/978-3-031-78952-6_3

[19] 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, Vol. 1, pp. 4171-4186, Jun. 2019. DOI: 10.18653/v1/N19-1423

[20] D. N. Rim, D. N. Heo, and H. Y. Choi, "Adversarial Training with Contrastive Learning in NLP", Korean Institute of Information Scientists and Engineers, Vol. 52, No. 1, pp. 52-61, Jan. 2025. DOI: 10.5626/JOK.2025.52.1.52

Authors

Sung-Kwang Song received the B.S. and M.S. degrees in Computer Engineering from Changwon National University, Korea, in 2009 and 2012, respectively. Mr. Song is a Senior Research Engineer in Hibrainnet Co. He is interested in Generative AI application services, Big data analysis, and Ontology.

Hyeon-Jeong Mun received the M.S. and Ph.D. degrees in Computer Engineering from Changwon National University, Korea, in 1996 and 2003, respectively. Dr. Mun is a Principal Research Engineer at Hibrainnet Co. She is interested in Ontology, Data Modeling, and Data Analysis.

Young-Ji Kim received the M.S. and Ph.D. degrees in Computer Engineering from Changwon National University, Korea, in 1997 and 2004, respectively. Dr. Kim is a Principal Research Engineer at Hibrainnet Co. She is interested in Generative AI application services, Big data analysis, and hyper-personalization.

Yong-Tae Woo received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Kyungpook National University, Korea, in 1982, 1984 and 1995, respectively. Dr. Woo is a Professor in the Department of Computer Engineering, Changwon National University since 1987. He is also CEO of Hibrain.net Co. He is interested in Data Modeling, Internet Business, and Big Data Analysis areas.