Journal of KIIT. Vol. 20, No. 11, pp. 29-38, Nov. 30, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.11.29

단어의 문맥 정보 확장을 위한 지식 임베딩 기법

김예원*, 김장원**

A Study on Knowledge Embedding Method for Extending Contextual Information of Words

Yewon Kim*, Jangwon Gim**

본 연구는 한국과학기술정보연구원 「디지털 기술 기반 국가·사회 현안 문제 해결(K-22-L04-C06-S01)」 사업으로부터 지원받아 수행된 연구임

요 약

사전 학습 언어 모델은 대용량 텍스트 데이터를 학습에 사용하여 다양한 자연어 처리 분야에서 우수한 성능을 보인다. 단어의 문맥 정보를 반영한 임베딩 기법들은 사전 학습 모델의 성능에 중요한 역할을 한다. 그렇지만 문맥에 출현한 명시적 단어들의 문맥 정보만 임베딩에 사용되고 있으므로 단어들이 가질 수 있는 다양한 관계 정보를 동시에 임베딩에 반영하기 위한 임베딩 기법이 필요하다. 본 논문에서는 지식 임베딩 기법을 통해 문장에 출현한 개체가 트리플에서 주어 개체인 경우뿐만 아니라 목적어 개체인 경우까지 포함시켜 출현 개체가 가질 수 있는 지식 정보를 확장하여 임베딩에 반영하는 확장 방법을 제안한다. 제안 모델의 성능 평가 결과 기존 지식 기반 사전 학습 모델인 CoLAKE보다 우수한 성능을 보인다. 따라서 지식 임베딩을 통한 사전 학습 모델의 성능 향상을 통해 응용 분야 문제 해결에 도움을 줄 것으로 기대한다.

Abstract

The pre-training language model shows excellent performance in various natural language processing fields by using large amounts of text data for training. Embedding techniques that reflect contextual information of words play an important role in the performance of the pre-training model. However, since only the context information of explicit words appearing in the context is used for embedding, an embedding technique is needed to simultaneously reflect various relational information that words can have in embedding. In this paper, we propose an extension method that extends the knowledge information that the appearing entity can have and reflects it in embedding by including the case where the object appearing in the sentence is not only the subject object in the triple but also the object object through the knowledge embedding technique. As a result of the performance evaluation of the proposed model, it was confirmed that it showed better performance than CoLAKE, which is the existing knowledge-based pre-training model. Therefore, it is expected that it will help solve application problems by improving the performance of the pre-training model through knowledge embedding.

Keywords

pre-train language model, knowledge embedding, knowledge expansion

* 군산대학교 소프트웨어학과
- ORCID: https://orcid.org/0000-0002-0125-1181
** 군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4480-7944

· Received: Sep. 16, 2022, Revised: Oct. 12, 2022, Accepted: Oct. 15, 2022
· Corresponding Author: Jangwon Gim
Dept. of Software Science & Engineering, Kunsan University, South Korea
Tel.: +82-63-469-8916, Email: jwgim@kunsan.ac.kr

Ⅰ. 서 론

빅데이터 환경에서 대용량의 텍스트 데이터의 급증으로 인하여 텍스트 데이터로부터 보다 가치 있는 정보 추출을 위한 딥러닝 기술이 급속도로 발전하고 있다. 특히 자연어 처리 분야에서는 번역, 기계 독해, 개체명 인식 등의 다양한 문제 해결을 위해 딥러닝 기술을 적용하는 연구가 수행되었다. 특히 사전 학습 언어 모델은 대용량 텍스트 데이터로부터 언어의 의미를 학습하여 다양한 태스크에 범용적으로 활용할 수 있는 기술로 자연어 처리 태스크에서 우수한 성능을 보인다[1][2].

사전 학습 언어 모델은 기존의 워드 임베딩 기법과 달리 대상 단어에 대한 주변 단어들과의 문맥 벡터를 임베딩함으로써 다양한 자연어 처리 태스크에 적용하여 우수한 성능을 보인다. 그렇지만 사전 학습 모델은 대상 단어 주변의 한정된 문맥 정보만을 임베딩하여 대상 단어가 표현할 수 있는 관계 정보가 충분히 반영되지 않는다.

2012년 구글은 검색 엔진의 결과를 향상하기 위해 사용한 기술로 지식 그래프를 발표했다[3]. 지식 그래프는 일종의 기반 지식으로서 다양한 관계 정보를 그래프 형태로 표현할 수 있어 챗봇, 자연어 이해, 검색 고도화 등의 분야에서 지식 그래프를 활용하고 있다[4]. 따라서 본 연구에서는 이러한 사전 학습 모델의 표현력 부족함을 극복하기 위해 사전 학습 시 확장이 가능한 임베딩 모델을 제안하고자 한다. 제안 모델은 대상 단어가 가질 수 있는 다양한 관계 정보를 지식 그래프로 표현하고 사전 학습 모델의 임베딩에 이러한 지식 그래프를 사용한다. 이때 사전 학습 모델의 임베딩에 사용된 단어(텍스트)와 연관된 지식 정보를 함께 임베딩하여 사전 학습 모델의 성능을 높이고자 한다.

본 논문의 2장에서는 관련 연구인 사전 학습 언어 모델, 지식 기반 사전 학습 언어 모델에 대한 연구 동향을 기술한다. 3장에서는 본 논문에서 제안하는 모델의 핵심 내용을 기술한다. 4장에서는 기존 모델과 제안 모델의 비교 평가를 수행한다. 마지막으로 5장에서는 결론 및 향후 연구를 기술한다.

Ⅱ. 관련 연구

2.1 사전 학습 언어 모델

사전 학습 언어 모델은 대용량 텍스트 데이터를 사용하여 단어에 대한 문맥 정보를 학습한 후 다양한 자연어 처리 태스크에 적용하는 기술이다.

대표적으로 BERT(Bidirectional Encoder Representations from Transformer)는 단어에 대한 문맥 정보를 반영하기 위해서 사용하는 임베딩은 그림 1과 같이 3가지 형태로 토큰 임베딩, 세그먼트 임베딩, 위치 임베딩으로 구성된다[5].

그림 1. BERT 모델 임베딩 구조
Fig 1. Embedding structure of BERT model

토큰 임베딩은 각 토큰에 대한 워드 임베딩, 세그먼트 임베딩은 각 토큰 속한 문장 정보를 반영하고 위치 임베딩은 각 토큰의 문장 내 위치를 반영한다. 3가지 임베딩의 합이 입력으로 들어간다. 따라서 BERT는 이러한 3가지 임베딩 계층을 통해 각 단어의 문맥 정보를 표현할 수 있다.

BERT 모델 이후 대용량 텍스트 데이터로부터 언어 모델의 학습 성능을 향상시키기 위해 다양한 모델이 제안되었으며 RoBERTa[6], SpanBERT[7], BART[8], ELECTRA[9] 등이 있다. 사전 학습 언어 모델은 단어의 문장 내 출현 순서, 위치 등의 문맥 정보를 임베딩에 활용하여 대상 단어의 표현력을 높인다. 그렇지만 대상 단어가 가질 수 있는 연관된 지식 정보의 반영은 부족하다.

2.2 지식 기반 사전 학습 언어 모델

지식 정보란 대상 단어의 언어학적 특징뿐 아니라 대상 단어가 다른 단어와 가질 수 있는 다양한 관계를 의미하는 것으로 단어에 대한 지식 정보를 임베딩에 활용하여 기존 사전 학습 언어 모델의 표현력을 보다 높일 수 있다. 이러한 지식 정보를 주어, 서술어, 목적어 형태의 트리플 구조로 표현하고 트리플 간의 관계를 그래프 형태로 표현한 지식 그래프를 임베딩에 적용한 ERNIE, KnowBERT 모델이 등장하였다[10][11].

ERNIE(Enhanced language representation with informative entities)는 말뭉치와 지식 그래프를 퓨전 형태로 인코딩하는 BERT 기반의 사전 학습 언어 모델이다. ERNIE는 텍스트 인코더와 지식 인코더를 통해 대상 단어와 대상 단어와 관련된 개체를 함께 인코딩한다. 이를 위해서 사전 학습 이전에 문장 내의 단어와 관련된 개체가 식별되고 식별된 개체와 관련된 지식 정보가 함께 인코딩에 사용한다. 이를 위해서 지식 임베딩에서 널리 사용되는 TransE[12] 모델을 통해 지식 임베딩을 수행한다. 따라서 ERNIE는 단어 임베딩과 TransE에 의한 임베딩을 통합한 지식 인코더를 통해 대상 단어의 표현력을 높일 수 있다. 그림 2는 ERNIE 모델의 아키텍처이며, 제안 모델과 기존 연구와의 차이점인 지식 인코딩 과정에 대한 차이점을 표현하고자 K-Encoder의 구조를 강조하여 표현하였다. T-Encoder는 텍스트를 위한 선행 표현 학습으로 토큰에서 어휘적, 문맥적 정보를 얻고, K-Encoder는 T-Encoder에서 인코딩된 토큰 임베딩과 엔티티 임베딩을 연결하여 문맥 정보를 통합한다. ERNIE 모델은 개체의 타입 분류와 관계 분류에서 BERT보다 우수한 성능을 보인다. 그렇지만 ERNIE는 단어와 연관된 일부 식별된 개체들이 가지는 개체만을 인코딩에 사용한다. 따라서 개체들이 가질 수 있는 다양한 관계 정보의 표현력이 풍부하지 않다.

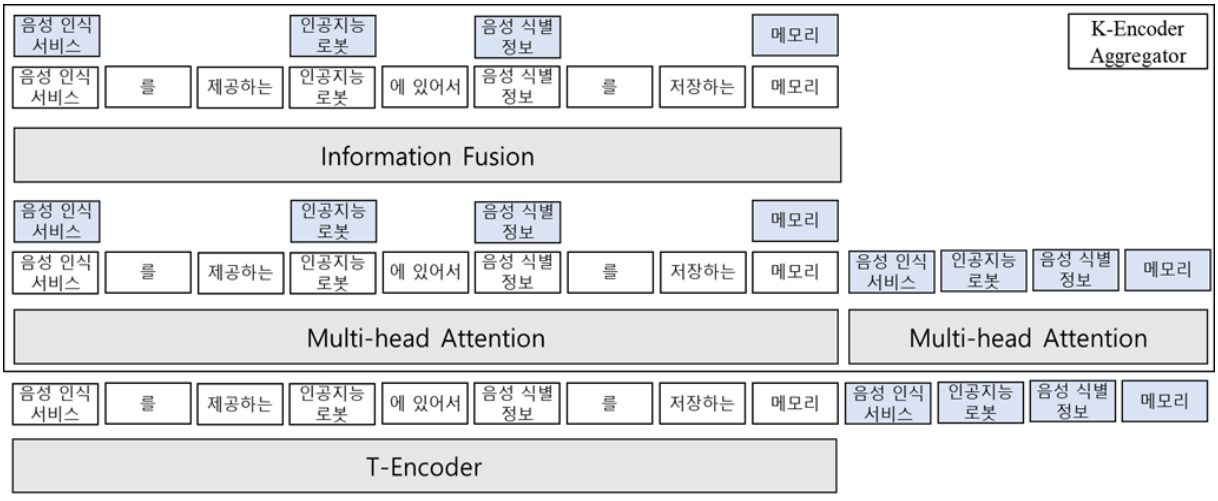

그림 2. ERNIE 모델의 아키텍처
Fig 2. Architecture of ERNIE model

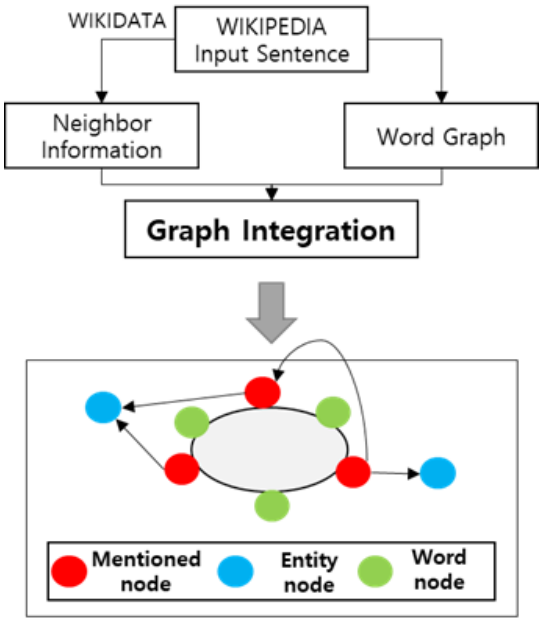

그림 3. CoLAKE 모델의 아키텍처
Fig 3. Architecture of CoLAKE model

이러한 ERNIE가 가지는 지식 정보 표현력의 한계를 보완하기 위한 CoLAKE(Contextualized Language and Knowledge Embedding) 모델이 제안되었다[13]. 그림 3은 CoLAKE의 아키텍처로 문장에 출현한 단어와 관련된 지식 정보를 통합하여 지식 그래프를 생성한다. CoLAKE는 지식 정보 인코딩을 위해 위키 데이터를 지식 인코딩에 활용하여 언급된 개체와 이웃한 개체들의 관계 정보까지 인코딩에 사용하여 확장된 문맥 정보를 모델 학습에 사용한다.

또한 단어 임베딩과 지식 임베딩을 개별로 학습하는 것이 아닌 동시에 학습하여 맥락에 맞는 지식 정보를 함께 인코딩에 사용한다. 이를 위해서 지식 그래프에서 문장에 출현된 개체를 중심으로 트리플을 추출하여 하위 그래프를 생성한다. 더 풍부한 지식 정보를 사용하고 지식 임베딩과 언어 임베딩을 병행으로 학습함으로써 지식 기반 태스크에서 ERNIE보다 높은 성능을 보인다.

그렇지만 ERNIE에서는 대상 단어와 관련된 개체 정보만을 지식 인코딩에 활용하고, CoLAKE 모델에서는 관계 정보도 지식 정보로 활용하지만 언급된 개체가 주어인 경우만을 지식 정보로 사용한다. 따라서 관계 정보도 지식 정보로 활용하며 출현된 개체가 주어인 경우와 목적어인 경우 모두 지식 정보를 활용하여 지식 표현을 향상시킬 수 있는 임베딩 모델이 필요하다.

III. 제안 모델

본 논문에서는 지식 기반 사전 학습 사전 훈련을 위한 지식 확장 임베딩 방법을 적용한 모델을 제안한다. 제안 모델에서는 지식 확장 방법으로 개체 확장을 통한 트리플 구축과 지식 확장 방법을 적용한다. 3장은 개체 확장을 통한 트리플 구축 방법, 지식 확장 방법, 제안 모델의 개요 3단계로 구성된다.

3.1 사전 기반 개체 확장

사전 기반 개체 확장은 지식 확장을 위해 동사를 기준으로 개체 확장을 통해 트리플을 구축한다. 트리플은 개체들 간의 관계를 <주어, 서술어, 목적어> 구조로 표현한다. 주어와 목적어는 개체, 서술어는 관계로 주어와 목적어는 서술어의 관계를 가진다.

과학 기술 문헌에서 동사를 추출하고 동사의 빈도수에 따라 정렬하고 동사의 개수의 임계치를 지정하여 트리플 확장 후보를 선정한다. 본 논문에서는 전체 빈도수의 약 35%를 차지하는 상위 빈도수 15개의 동사를 사용했다. 정규 표현식을 사용하여 선정된 동사의 왼쪽 문장과 오른쪽 문장의 패턴을 추출한다. 문장에서 출현된 단어들이 가질 수 있는 다양한 관계를 추출하기 위해서 본 논문에서는 기존의 지식 베이스(사전)에 출현하는 단어를 개체로 식별하고 해당 개체들이 가질 수 있는 개체들 간의 관계를 임베딩에 활용한다. 그러므로 개체와 개체들 간의 관계를 개체의 지식 정보라고 정의한다.

그림 4는 개체 확장 방법의 예로 확장할 동사로 선정된 '제공', '저장'과 같은 동사를 추출한 후 해당 동사를 기준으로 문장의 왼쪽과 오른쪽에 배치된 <명사, 동사, 명사> 패턴을 추출한다. 추출된 패턴에서 왼쪽에 있는 명사를 주어, 오른쪽에 있는 명사를 목적어의 후보로 사용한다. 후보의 단어가 일반 명사 또는 고유 명사인 경우 '<음성 인식 서비스, 제공, 인공 지능 로봇>, <음성 식별 정보, 저장, 메모리>'처럼 트리플로 확장한다.

주어와 목적어의 후보가 기존 정의되어 있는 엔

티티 집합에 존재하지 않을 경우 신규 개체로 확장한다.

3.2 지식 확장

CoLAKE 모델은 출현한 개체가 주어인 경우의 트리플만 사용하여 이웃 정보로 사용하므로 출현 개체가 목적어인 경우에 대한 개체 관계를 고려하지 않는다. 따라서 제안 모델의 지식 확장 단계에서 목적어와 관련된 관계 정보를 추가 지식으로 확장한다. 그림 5는 지식 확장 방법의 예시로 입력 문장에 '사용자 단말', '안내 로봇', '정보 제공 시스템'이라는 개체명이 언급되었으며 언급된 개체가 주어인 경우의 이웃 정보뿐만 아니라 목적어인 경우의 이웃 정보인 '<카메라, 인식, 사용>, <클라우드, 연동, 사용자>'와 같은 이웃 정보까지 사용한다.

3.3 제안 모델의 개요 및 임베딩 구조

그림 6은 개체 확장 방법과 지식 확장 방법을 적용한 제안 모델의 개요도이다. 그림 6에서의 1 단계는 기존 정의된 트리플인 AMI(Ambient human & Machine Intelligence) 특허 트리플 셋에 정의되어 있지 않은 개체를 신규 개체로 확장하며 개체 확장을 통해 트리플 셋을 구축한다.

1 단계에서 (a)는 AMI 특허 트리플 셋에 정의되어 있는 트리플 중 입력 문장에 언급된 개체를 대상으로 하는 트리플이다. (b)는 개체 확장을 통해 구축한 트리플 중 입력 문장에 언급된 개체를 대상으로 하는 트리플이다. (a)와 (b)의 트리플을 언급된 개체의 지식 정보로 사용한다. 2 단계는 확장된 트리플 셋에서 언급된 개체가 주어인 경우와 목적어인 경우를 함께 반영하여 학습에 사용할 지식 그래프를 확장한다.

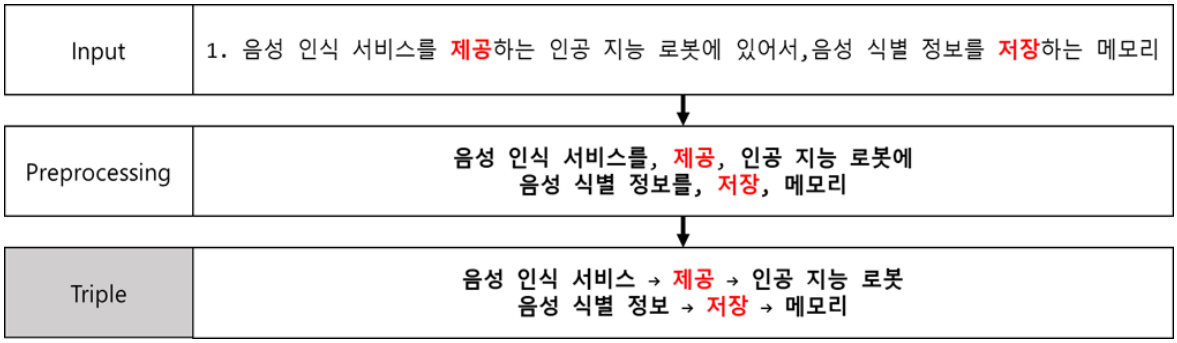

그림 4. 개체 확장을 통한 트리플 구축 예시
Fig 4. Constructing triples considering expanded entities

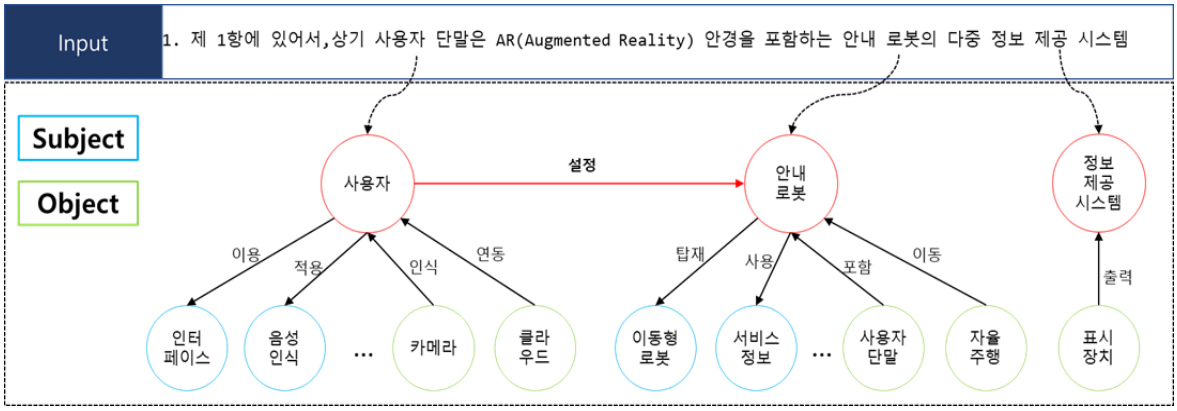

그림 5. 개체 관계 정보를 이용한 지식 확장
Fig 5. Extending knowledge using entity relationships

트리플 셋을 사용하여 입력 문장에 나온 개체들을 식별하고, 식별된 엔티티의 이웃 정보들을 지식으로 반영한다. 식별된 엔티티가 목적어인 경우의 이웃 정보까지 지식으로 확장한다.

그림 7은 제안 모델의 임베딩 구조이다. 입력 문장에 대한 언어 표현과 지식 표현을 함께 병행으로 학습한다. 임베딩 레이어는 토큰 임베딩, 타입 임베딩, 포지션 임베딩으로 구성된다. 토큰 임베딩은 형태소 분석기를 사용하여 입력 문장들을 토큰화하여 토큰 임베딩을 수행한다. 유형 임베딩은 각 토큰들의 유형을 단어, 엔티티 및 관계를 나타낸다. 위치 임베딩은 각 토큰들의 인덱스를 할당한다. 그림 7에서의 빨간색 토큰은 입력 문장에 출현한 개체, 파란색은 주어인 개체, 초록색은 목적어인 개체, 노란색은 관계를 의미한다. 그러므로 개체와 개체 간의 관계에 대한 정보를 포함해 입력 단어의 토큰 뒤에 지식 정보들을 함께 표현하여 임베딩을 수행하므로 모델은 단어에 대한 언어 표현과 개체와 개체 간의 관계에 대한 지식 표현을 동시에 학습할 수 있다.

Ⅳ. 실 험

4.1 실험 환경 및 방법

제안하는 사전 학습 모델 생성 및 학습된 모델을 활용한 관계 예측을 위해 구성된 서버의 환경은 표 1과 같다.

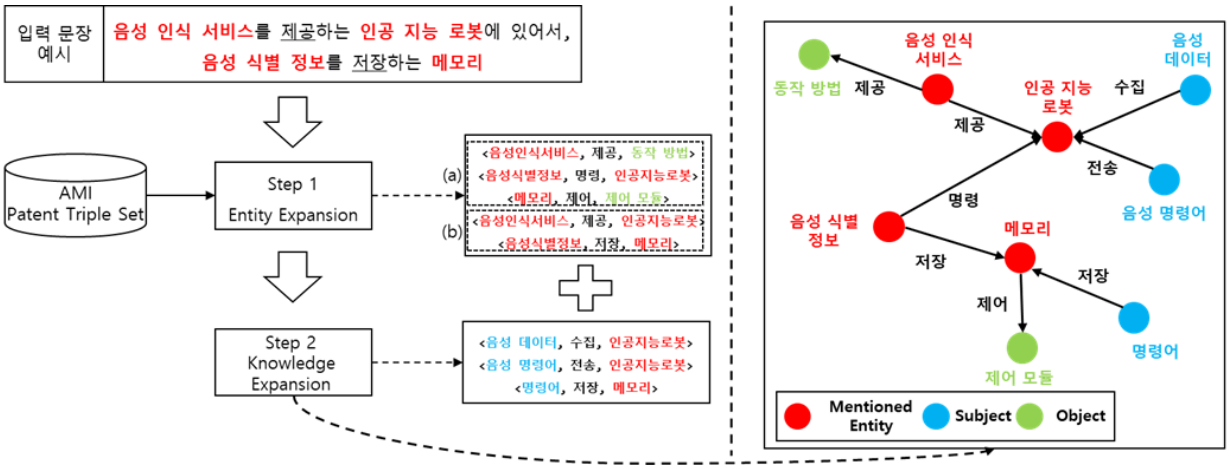

그림 6. 제안 모델 아키텍처
Fig 6. Architecture of the proposed model

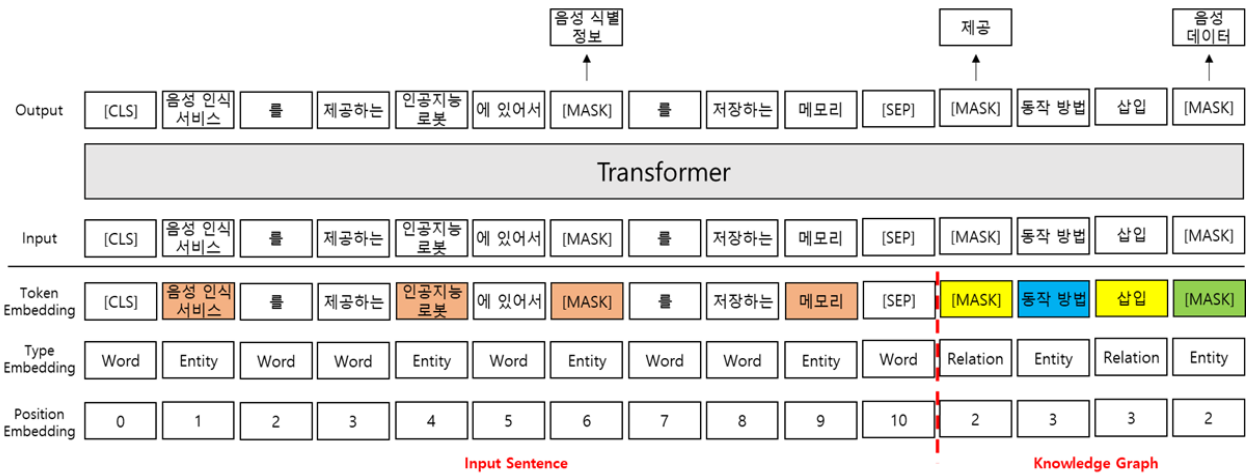

그림 7. 개체 지식 정보를 반영한 임베딩 구조
Fig 7. Embedding structure reflecting entity knowledge

표 1. 실험 환경
Table 1. Experiment environment

OS	Ubuntu 18.04.06 LTS
CPU	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
Memory	64GB
GPU	RTX A6000 (48GB, x2)

실험은 기존 CoLAKE 모델과 제안하는 확장 방법을 적용한 개체 확장(Step 1), 지식 확장(Step 2)의 개체들 간의 관계 예측에 대한 성능을 비교한다.

개체 확장 방법은 문장에 사용된 동사를 기준으로 왼쪽과 오른쪽에 언급된 일반 명사 또는 고유명사를 개체로 확장하고 해당 개체와 동사를 트리플 셋으로 구축한다. 지식 확장 방법은 문장에 출현한 개체가 주어인 경우와 목적어인 경우의 트리플을 지식 정보로 임베딩 한다.

그림 8은 제안 모델의 흐름도로 사전 학습과 미세 조정에 대한 순서를 나타낸다. 먼저 개체 확장을 통해 특허 트리플 셋을 구축하고 구축된 트리플 셋을 활용하여 사전 학습을 진행한다. 학습된 제안 모델을 사용하여 관계 추출 태스크에 대한 실험을 진행하고 기존 모델의 성능과 비교 평가한다.

4.2 실험 데이터

본 논문에서의 실험 데이터는 4차 산업혁명과 관련된 기술 개체들이 가질 수 있는 관계에 대한 지식 확장을 목표로 한다. 그러므로 한국특허정보원(KIPRIS)에서 제공되는 4차 산업과 관련된 기술 분류 코드(B25J, B33Y, G06V, G06W, G06Y, G16Y)들을 포함한 특허 문헌 총 565건을 실험 데이터로 사용한다[14]. 565건 중 80%인 455건은 사전 학습 실험 데이터로 사용하고 20%인 110건은 미세 조정의 실험 데이터로 사용한다. 제안 모델에 대한 관계 예측 태스크 성능 평가를 위해 사용된 트리플 데이터 건수는 표 2와 같다.

표 2. 트리플, 개체, 관계 수
Table 2. Number of triple, entity, relation

Target / Number of Data	Triple	Entity	Relation
AMI patent triple set	14,110	3,035	442

4.3 실험 평가

제안 모델을 통해서 실험 데이터 집합으로부터 총 2,743건의 트리플이 확장되었으며, 개체는 315건 확장되었다. 그러므로 이렇게 확장된 트리플과 개체, 관계에 대한 지식 정보가 사전 학습 모델의 임베딩에 사용된다. 이에 대한 성능 평가를 위해서 제안 모델과 CoLAKE 모델에 대하여 관계 예측 태스크를 수행하고 관계 예측에 대한 F1-Score, 정밀도, 재현율 지표로 모델의 성능을 비교 평가한다.

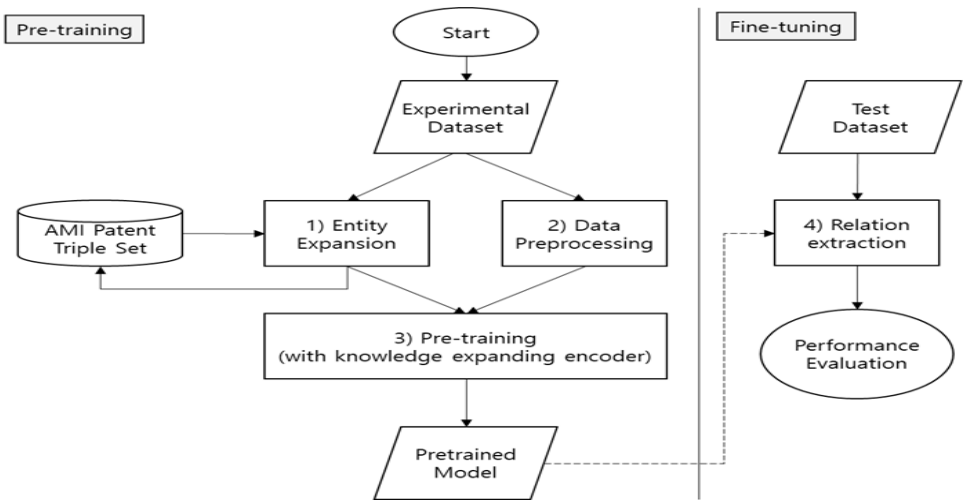

그림 8. 제안 모델의 흐름도
Fig 8. Flowchart of the proposed model

CoLAKE는 개체 정보뿐만 아니라 관계 정보도 지식 정보로 사용하여 ERNIE보다 높은 성능을 보였지만 출현한 개체가 주어 개체인 경우만 지식 정보로 사용한다. 따라서 본 논문은 출현 개체가 주어인 경우와 목적어인 경우 모두 지식 정보로 사용하는 지식 확장 방법과 개체 확장 방법을 제안하며 제안 방법과 기존 모델인 CoLAKE의 관계 예측 성능을 비교한다.

표 3은 기존 모델인 CoLAKE 모델을 사용해 관계 추출 태스크를 실험한 결과로 7 에폭일 때 가장 우수한 성능을 보인다. 표 4는 개체 확장을 통해 구축된 트리플을 적용한 관계 추출 태스크 실험 결과로 10 에폭일 때 가장 우수한 성능을 보인다.

표 3. CoLAKE 실험 결과
Table 3. Experiment results of CoLAKE using the dataset

Accuracy / Epoch	F1-score	Precision	Recall
1 epoch	0.00563	0.00698	0.01895
2 epoch	0.05171	0.7910	0.06709
3 epoch	0.39124	0.45667	0.40682
4 epoch	0.56435	0.60370	0.59608
5 epoch	0.63307	0.65382	0.65815
6 epoch	0.70306	0.72291	0.73172
7 epoch	**0.75157**	**0.77864**	**0.77137**
8 epoch	0.72486	0.75537	0.74994
9 epoch	0.74720	0.76997	0.77111
10 epoch	0.74675	0.76894	0.77248

표 4. 제안 모델의 1 단계 실험 결과
Table 4. Experimental results for the first step of the proposed model

Accuracy / Epoch	F1-score	Precision	Recall
1 epoch	0.02770	0.02778	0.03563
2 epoch	0.28148	0.34230	0.28682
3 epoch	0.48405	0.52313	0.50301
4 epoch	0.60896	0.66758	0.59875
5 epoch	0.67546	0.70157	0.68255
6 epoch	0.69545	0.70240	0.72571
7 epoch	0.76110	0.78305	0.76298
8 epoch	0.74541	0.76028	0.77015
9 epoch	0.79202	0.81581	0.80042
10 epoch	**0.80564**	**0.80329**	**0.83048**

표 5는 개체 확장과 지식 확장을 적용한 관계 추출 태스크 실험 결과로 9 에폭일 때 가장 좋은 성능을 보인다. 에폭별 정확도 결과를 통해 에폭이 증가함에 따라 모델이 수렴되는 것을 보인다.

표 6은 각 모델별 가장 우수한 성능을 보인 비교 결과이다. 이를 통해 제안 모델의 1단계만 수행한 모델이 CoLAKE 보다 F1-Score가 약 5%p 성능이 향상됐다. 이는 제안 모델의 1 단계인 개체 확장을 통해 구축된 트리플을 사용하면 풍부한 지식 표현이 가능한 것을 보인다. 또한 제안 모델은 CoLAKE 보다 약 9%p 성능이 향상했고 제안 모델의 1 단계만 수행한 모델 보다 약 4%p 향상됐다. 이를 통해 언급된 개체가 주어인 경우뿐만 아니라 목적어인 경우의 트리플을 지식으로 활용한 2 단계의 지식 확장 방법이 관계 추출 태스크에서 효과적이라는 것을 확인했다. 또한 개체 확장 방법인 1 단계와 지식 확장 방법인 2 단계의 두 가지 확장 방법을 함께 적용했을 때 관계 추출 태스크에 효과적이라는 것을 확인했다.

표 5. 제안 모델의 2 단계 실험 결과
Table 5. Experimental results for the second step of the proposed model

Accuracy / Epoch	F1-score	Precision	Recall
1 epoch	0.04383	0.05421	0.04647
2 epoch	0.44631	0.48549	0.44369
3 epoch	0.60296	0.64142	0.61991
4 epoch	0.70103	0.74173	0.70914
5 epoch	0.78829	0.80956	0.80403
6 epoch	0.77811	0.79710	0.79787
7 epoch	0.80107	0.82174	0.82469
8 epoch	0.83176	0.85116	0.84427
9 epoch	**0.84974**	**0.87492**	**0.86070**
10 epoch	0.83781	0.86685	0.84600

표 6. 제안 모델과 기존 모델(CoLAKE)의 비교 결과
Table 6. Comparison results between the proposed model and the existing model (CoLAKE)

Accuracy / Epoch	F1-score	Precision	Recall
CoLAKE	0.75157	0.77864	0.77137
Entity expansion (Step 1)	**0.80564**	**0.80329**	**0.83048**
Knowledge expansion(Step 2)	**0.84974**	**0.87492**	**0.86070**

V. 결론 및 향후 과제

대용량 데이터 학습을 통하여 의미 있는 정보를 추출하기 위한 사전 학습 언어 모델이 연구되고 있다. 사전 학습 언어 모델은 주변 단어들과의 문맥 정보를 학습하여 다양한 자연어 처리 태스크에서 우수한 성능을 보인다. 그렇지만 단어 주변의 한정된 문맥 정보만 반영하여 다양한 관계 정보가 충분히 반영되지 않는 문제점이 존재한다. 따라서 개체와 개체들 간의 관계를 포함하는 지식 그래프를 임베딩하는 사전 학습 언어 모델이 연구되었다. 그렇지만 이러한 지식 기반 사전 학습 언어 모델은 개체 정보만 반영하고 개체들 간의 관계 정보를 반영하지 않거나 지식의 일부만 지식으로 활용한다. 그러므로 본 논문에서는 언어 표현 및 지식 표현력을 향상하기 위해 개체 확장 방법과 지식 확장 방법을 적용한 모델을 제안했다.

제안 모델은 지식 확장을 위해서 기존 트리플 데이터를 확장하고 확장된 트리플에 대한 개체 추출을 통한 지식을 확장하였다. 그 결과 개체명 간의 관계를 추출하는 태스크에서 풍부한 이웃 정보를 활용하는 지식 확장의 필요성과 지식 표현이 향상되어 CoLAKE보다 지식 확장 측면에서 성능이 우수함을 확인하였다. 본 논문의 한계는 트리플 확장 시 동사를 기준으로 확장하였으나 동사의 수동형과 능동형을 구분하지 않아 트리플 방향의 통일성이 부족하다. 따라서 동사의 수동형과 능동형 구분을 통해 트리플의 주어와 목적어에 대한 통일성을 높이는 트리플 확장을 향후 연구로 한다.

References

[1] Garg Siddhant, Thuy Vu, and Alessandro Moschitti, "Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 05, pp. 7780-7788, Apr. 2020. https://doi.org/10.1609/aaai.v34i05.6282.

[2] Edunov, Sergey, et al., "Understanding back-translation at scale", arXiv preprint arXiv:1808.09381, Aug. 2018.

[3] Fensel, Dieter, et al., "Introduction: what is a knowledge graph?", Knowledge Graphs. Springer, Cham, pp. 1-10, Feb. 2020. https://doi.org/10.1007/978-3-030-37439-6_1.

[4] Kejriwal Mayank and Pedro Szekely, "Knowledge graphs for social good: An entity-centric search engine for the human trafficking domain", IEEE Transactions on Big Data, Vol. 8, No. 3, pp. 592-606, Oct. 2017. https://doi.org/10.1109/TBDATA.2017.2763164.

[5] Devlin Jacob, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, Oct. 2018. https://doi.org/10.48550/arXiv.1810.04805.

[6] Liu Yinhan, et al., "Roberta: A robustly optimized bert pretraining approach", arXiv preprint arXiv:1907.11692, Jul. 2019. https://doi.org/10.48550/arXiv.1907.11692.

[7] Joshi Mandar, et al., "Spanbert: Improving pre-training by representing and predicting spans", Transactions of the Association for Computational Linguistics, Vol. 8, pp. 64-77, Jan. 2020. https://doi.org/ 10.1162/tacl_a_00300.

[8] Lewis Mike, et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", arXiv preprint arXiv:1910.13461, Oct. 2019. https://doi.org/10.48550/arXiv.1910.13461.

[9] Clark Kevin, et al., "Electra: Pre-training text encoders as discriminators rather than generators", arXiv preprint arXiv:2003.10555, Mar. 2020. https://doi.org/10.48550/arXiv.2003.10555.

[10] Zhang Zhengyan, et al., "ERNIE: Enhanced language representation with informative entities", arXiv preprint arXiv:1905.07129, May. 2019. https://doi.org/10.48550/arXiv.1905.07129.

[11] Peters Matthew E., et al., "Knowledge enhanced

contextual word representations", arXiv preprint arXiv:1909.04164, Sep. 2019. https://doi.org/10.48550/arXiv.1909.04164.

[12] Bordes Antoine, et al., "Translating embeddings for modeling multi-relational data", Advances in neural information processing systems 26, 2013.

[13] Sun Tianxiang, et al., "Colake: Contextualized language and knowledge embedding", arXiv preprint arXiv:2010.00309, Oct. 2020. https://doi.org/10.48550/arXiv.2010.00309.

[14] KIPRIS, http://www.kipris.or.kr/khome/main.jsp [accessed: Oct. 05, 2022]

저자소개

김 예 원 (Yewon Kim)

2019년 3월 ~ 현재 : 군산대학교 소프트웨어학과 학부과정

관심분야 : 자연어 처리, 빅데이터 분석

김 장 원 (Jangwon Gim)

2008년 2월 : 고려대학교 컴퓨터학과(이학석사)

2012년 8월 : 고려대학교 컴퓨터·전파·통신공학과(공학박사)

2013년 3월 ~ 2017년 3월 : 한국과학기술정보연구원 선임연구원

2017년 4월 ~ 현재 : 국립군산대학교 소프트웨어학과 부교수

관심분야 : 자연어 처리, 지식그래프, 지식임베딩, 실시간 빅데이터 분석