J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr


# BERT와 지식 그래프를 이용한 한국어 문맥 정보 추출 시스템☆

# Korean Contextual Information Extraction System using BERT and Knowledge Graph

유 소 엽[1] 정 옥 란[1*]
SoYeop Yoo OkRan Jeong

**요 약**

인공지능 기술의 비약적 발전과 함께 사람의 언어를 다루는 자연어 처리 분야 역시 활발하게 연구가 진행되고 있다. 특히 최근에는 구글에서 공개한 언어 모델인 BERT는 대량의 코퍼스를 활용해 미리 학습시킨 모델을 제공함으로써 자연어 처리의 여러 분야에서 좋은 성능을 보이고 있다. BERT에서 다국어 모델을 지원하고 있지만 한국어에 바로 적용했을 때는 한계점이 존재하기 때문에 대량의 한국어 코퍼스를 이용해 학습시킨 모델을 사용해야 한다. 또한 텍스트는 어휘, 문법적인 의미만 담고 있는 것이 아니라 전후 관계, 상황과 같은 문맥적인 의미도 담고 있다. 기존의 자연어 처리 분야에서는 어휘나 문법적인 의미를 중심으로 연구가 주로 이루어졌다. 텍스트에 내재되어 있는 문맥 정보의 정확한 파악은 맥락을 이해하는 데에 있어 중요한 역할을 한다. 단어들의 관계를 이용해 연결한 지식그래프는 컴퓨터에게 쉽게 문맥을 학습시킬 수 있는 장점이 있다. 본 논문에서는 한국어 코퍼스를 이용해 사전 학습된 BERT 모델과 지식 그래프를 이용해 한국어 문맥 정보를 추출하는 시스템을 제안하고자 한다. 텍스트에서 중요한 요소가 되는 인물, 관계, 감정, 공간, 시간 정보를 추출할 수 있는 모델을 구축하고 제안한 시스템을 실험을 통해 검증한다.

☞ 주제어 : 문맥 정보 추출, 인물 추출, 관계 추출, 감정 추출, BERT, 지식 그래프

**ABSTRACT**

Along with the rapid development of artificial intelligence technology, natural language processing, which deals with human language, is also actively studied. In particular, BERT, a language model recently proposed by Google, has been performing well in many areas of natural language processing by providing pre-trained model using a large number of corpus. Although BERT supports multilingual model, we should use the pre-trained model using large amounts of Korean corpus because there are limitations when we apply the original pre-trained BERT model directly to Korean. Also, text contains not only vocabulary, grammar, but contextual meanings such as the relation between the front and the rear, and situation. In the existing natural language processing field, research has been conducted mainly on vocabulary or grammatical meaning. Accurate identification of contextual information embedded in text plays an important role in understanding context. Knowledge graphs, which are linked using the relationship of words, have the advantage of being able to learn context easily from computer. In this paper, we propose a system to extract Korean contextual information using pre-trained BERT model with Korean language corpus and knowledge graph. We build models that can extract person, relationship, emotion, space, and time information that is important in the text and validate the proposed system through experiments.

☞ keyword : contextual information extraction, person extraction, relation extraction, sentiment extraction, BERT, knowledge graph

## 1. 서 론

하드웨어, 소프트웨어 등 다양한 기술의 발전으로 실시간으로 쌓이는 대용량 데이터에 대한 분산 처리와 빠른 연산 처리가 가능해졌다. 빅데이터의 활용과 컴퓨팅 능력의 향상은 복잡한 연산을 요구하는 인공지능 기술 관련 분야의 연구 활성화로 이어졌다. 특히 이미지, 영상 처리 분야에서는 비약적인 발전이 이루어졌다[1,2].

사람의 언어를 다루는 자연어 처리(Natural Language Processing; NLP) 분야 역시 활발한 연구가 진행되고 있다. 특히 최근 ELMO(Embeddings from Language Models)[3],

1 Dept. of AI·Software, Gachon University, Seongnam, 13120, Korea
* Corresponding author (orjeong@gachon.ac.kr)
[Received 11 March 2020, Reviewed 19 March 2020, Accepted 13 April 2020]
☆ This research was supported by Basic Science Research Program through the NRF(National Research Foundation of Korea), and the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute for Information & communications Technology Promotion) (Nos. NRF2019R1A2C1008412, 2015-0-00932).

BERT(Bidirectional Encoder Representation from Transformers)[4], XLNET[5] 등과 같은 언어 모델이 등장하면서 빠른 속도로 NLP 분야 관련 기술의 성능이 향상되고 있다. 다양한 언어에 대해 연구가 진행되고 있고, 한국어에 대한 연구 또한 다방면에서 이루어지고 있지만 여전히 한계점이 존재한다[1-5].

한국어는 이미 많은 발전을 이룬 영어와 다른 문법, 형태적 특성을 갖고 있다. 한국어와 영어의 가장 큰 차이점은 언어의 최소 단위이다. 영어는 한 글자씩 쪼갤 수 있는 알파벳이 최소 단위인 반면 한국어는 초성, 중성, 종성으로 분리되는 자모가 가장 작은 단위이다. 따라서 한국어 모델은 영어 모델과 다른 특징을 갖고 있기 때문에 현재 NLP 분야에서 활발하게 사용되고 있는 BERT 모델을 그대로 적용할 경우 성능에 한계가 있다[6,7]. 이는 구글이 대용량의 코퍼스를 사전 학습시켜 다국어에 적용할 수 있는 모델을 제공했음에도 한국어의 코퍼스가 상대적으로 부족하고 영어와 동일한 방법으로 학습시켰기 때문에 발생하는 한계점이다.

또한 텍스트는 기존의 자연어 처리 분야에서 주목한 어휘, 문법적 의미만 담고 있지 않다. 텍스트의 앞뒤 상황, 전체적인 상황과 관계 등의 문맥을 내포하고 있다. 텍스트에 내재되어 있는 문맥 정보의 파악은 전후 관계와 상황을 이해하는데 매우 중요한 부분으로 이를 통해 문서 전체의 요약, 의도 추론, 상황 판단 등 다양한 활용 가능성이 있다.

겉으로 드러나는 텍스트에 대한 정보를 수집하는 것과 달리 다양한 정보를 내포하고 있는 문맥 정보는 여러 분야에서 중요한 정보로 활용 가능하다. 하지만 오랜 사회, 문화, 언어적 학습을 통해 문맥 정보를 쉽게 파악할 수 있는 사람과 달리 컴퓨터는 정확한 문맥 정보의 파악이 어렵다.

본 논문에서는 한국어 텍스트에서 문맥 정보를 컴퓨터가 파악할 수 있는 시스템을 제안한다. 언어 모델을 통해 문맥 정보의 추출도 가능하지만 정확도를 향상시키기 위해 지식 그래프를 활용한다. 지식 그래프는 단어들을 관계를 이용해 연결하여 나타내기 때문에 컴퓨터가 쉽게 문맥을 이해할 수 있도록 한다.

한국어 텍스트의 문맥정보 추출을 위해 한국어의 특성이 반영될 수 있도록, 기존의 BERT 모델을 대량의 한국어 코퍼스를 이용해 학습시킨 한국어 BERT 모델을 활용한다. 또한 데이터를 관계 기반으로 연결하여 표현해주는 지식 그래프를 활용하여 텍스트 내의 문맥 정보를 추출할 수 있는 시스템을 제안하고 실험을 통해 활용 가능성을 검증하고자 한다.

## 2. 관련 연구

### 2.1 문맥 정보 (Contextual Information)

문맥이라는 단어는 개념의 폭이 매우 넓어서 다양한 관점에서 다르게 해석되고 사용된다. 언어 내적 요소인 텍스트 그 자체를 나타내기도 하고, 언어 외적 요소인 심리적, 사회적 맥락을 의미하기도 한다. 특히 텍스트를 사용하는 사람과 관련된 언어적 행동과 지식, 언어활동이 이루어지는 사회적, 시간적, 공간적 배경까지 모두 포함한다[8-10]. 문맥의 다양한 해석 중 우리는 언어 외적 요소, 즉 언어를 구성하고 있는 상황적 맥락으로 문맥을 정의한다.

텍스트의 문맥은 다양한 상황 정보를 통해 파악할 수 있다[9]. 사람은 오랜 사회, 문화적 학습을 통해 텍스트의 앞뒤, 전체적인 흐름 등을 인지하고 자연스럽게 텍스트가 내포하고 있는 문맥을 인지할 수 있다. 하지만 컴퓨터는 문맥을 이해하기 위해서는 별도의 학습 과정을 필요로 한다.

문맥 정보는 텍스트의 의미를 정의하는, 더 나아가 텍스트에 내재되어 있는 의미를 파악할 수 있는 주변의 여러 상황적 맥락 정보라고 정의 할 수 있다[9]. 상황적 맥락을 이해할 수 있는 정보들에는 사회적, 시간적, 공간적 배경까지 모두 포함되기 때문에 본 논문에서는 다양한 문맥 정보 중 인물, 시간, 공간, 관계, 감정, 5가지의 문맥 정보를 활용하고자 한다.

### 2.2 BERT 언어 모델

2018년 10월에는 논문으로 발표되고, 11월에는 오픈소스로 공개된 BERT[4]는 구글의 언어 표현 모델이다. 공개 당시 자연어 처리의 11개 태스크에서 State-of-the-art를 기록하며 자연어 처리 분야의 ImageNet으로 평가받고 있다.

위키 데이터와 같은 대형 코퍼스를 이용해 라벨링되지 않은 데이터로 미리 학습시킨 후, 특정 자연어 처리 태스크에 따라 별도의 아키텍처 없이 하나의 태스크 처리를 위한 레이어만을 추가함으로써 지도 학습을 수행하도록 하는 전이 학습 모델이다. BERT 이전에도 이러한 방법을 이용하는 모델로 ELMo[3], OpenAI GPT[11] 등이 있지만 이전 모델들은 접근 방식이 단방향이거나 얕은

양방향으로는 언어 표현에 한계가 있었다[4].

BERT의 구조는 Transformer[12]를 사용하지만, 사전 학습(pre-training)과 미세 조정(fine-tuning) 시의 아키텍처를 다르게 하여 전이 학습을 용이하게 만드는 것이 핵심이다. Transformer 중에서도 인코더 부분만 사용해 적용한다.

BERT는 모델의 크기에 따라 base와 large 모델을 제공한다. 레이어의 개수, 히든 유닛의 크기 등의 차이가 있다. 또한 대소문자 구별 여부에 따라 uncased와 cased 모델로 구별 가능하고, 최근에는 104개 언어들을 지원하는 'BERT-Base, Multilingual Cased' 모델을 제공하고 있다.

다국어 모델을 적용할 경우 쉽게 BERT 모델을 한국어에 적용해 일정 성능을 달성하는 모델을 구축할 수 있다. 하지만 다국어 모델은 대량의 코퍼스로 사전 학습되어 있지만 한국어 코퍼스가 상대적으로 적을 뿐만 아니라 기존 영어의 특성에 따라 적용된 임베딩 모델이 동일하게 사용되어 한국어 데이터에 최적화되어 있지 않다.

따라서 한국어 데이터에 최적화시키는 과정이 필요하다. 대량의 한국어 코퍼스를 통해 기존 모델을 다시 학습시킨 모델을 사용해야 더 좋은 성능을 기대할 수 있다. 본 논문에서는 한국어 위키와 뉴스 데이터로 기존 구글의 다국어 모델을 재학습시킨 SKTBrain의 KoBERT[13] 모델을 활용한다.

### 2.3 지식 그래프

지식 그래프는 단어와 단어를 관계로 연결하여 단어의 관계적 의미를 보여주는 그래프이다. 지식 그래프는 관계를 기반으로 단어가 표현되어 있기 때문에 컴퓨터가 사람이 단어의 여러 맥락적 의미를 인지하는 것처럼 추론할 수 있도록 도와주는 기술로 활용 가능하다. 지식 그래프는 WordNet[14], YAGO[15], ConceptNet[16] 등 다양한 종류를 갖고 있다.

특히 ConceptNet[16]은 오픈소스 지식 그래프이다. 영어, 프랑스어, 한국어 등 다양한 언어들에 대한 데이터를 보유하고 있으며, OMCS(Open Mind Common Sense) 프로젝트의 일환으로 위키피디아, 전문가가 만든 자료 등 크라우드소싱 데이터를 기반으로 한다. ConceptNet은 8백만 개 이상의 개체들이 서로 40개의 관계를 기반으로 연결되어 있는 지식 그래프이다. 우리는 ConceptNet을 활용해 텍스트 내의 관계적 문맥 정보를 추출하고자 한다.

또한 SenticNet[17]은 기존의 지식 그래프 기술을 활용해 감정의 개념을 그래프로 표현한다. 텍스트 내에서 개념 수준의 감정 분석을 가능하게 하는 SenticNet은 약 10만 개 이상의 단어에 관련된 감성, 감정 등을 포함하고 있다. 감성(sentic)은 즐거움(pleasantness), 흥미(attention), 민감도(sensitivity), 적성(aptitude)의 4개 차원 내의 감정을 표현하는 단어들로 이루어진다. 극성(polarity)은 부정과 긍정을 -1부터 +1 사이의 실수로 표현한다. 본 논문에서는 텍스트 내의 감정적 문맥 정보를 추출하기 위해 SenticNet[17]을 활용한다.

텍스트에 내재되어 있는 의미나 맥락을 이해하는데 도움을 주는 문맥 정보를 추출하기 위해서는 기존의 모델을 각 문맥 정보 추출에 최적화 하는 과정을 거쳐 모델을 구축해야 한다. 본 논문에서는 기존 BERT 모델을 한국어 코퍼스를 이용하여 다시 학습시킨 한국어 BERT 모델과 관계 정보를 나타낼 수 있는 지식 그래프를 이용한 한국어 문맥 정보 추출 시스템을 제안한다.

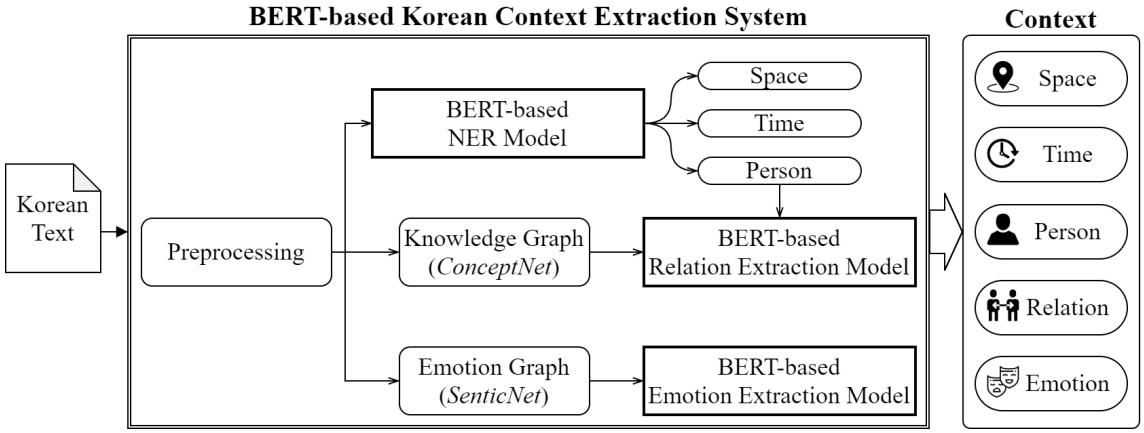
## 3. 한국어 문맥 정보 추출 시스템

문맥 정보(context)는 글의 의미나 형식을 뜻하기도 하지만 주로 해당 텍스트가 존재하는 전후 관계, 혹은 연관되어 나타나는 상황을 의미한다. 이러한 문맥 정보는 주어진 텍스트에 내재되어 있는 의미 또는 맥락을 이해하는데 매우 중요한 요소이다. 문맥 정보는 매우 다양하지만 본 논문에서는 텍스트에서 나타나는 전후 관계, 연관된 상황에 초점을 맞추어 이를 표현할 수 있는 여러 정보 중 '인물, 시간, 공간, 관계, 감정'의 5가지 정보로 제한한다.

컴퓨터는 사람의 말을 바로 이해할 수 없기 때문에 문맥 정보 추출을 통해 텍스트의 의미와 맥락을 파악하고 가치 있는 정보 산출이 가능하다. 문맥 정보 추출 시스템을 통해 문맥의 전후 관계와 상황 등의 판단을 가능하게 하는 유용한 정보를 추출 할 수 있으므로, 텍스트의 요약이나 의도 추론 등의 분야에서 다양하게 활용이 가능하다.

본 논문에서는 대표적인 언어 모델인 BERT[4]와 지식 그래프 ConceptNet[16]과 감정 기반 지식 그래프 SenticNet[17]을 이용한 한국어 문맥 정보 추출 시스템을 제안하고자 한다. 한국어 텍스트에서 인물, 관계, 감정, 시간, 공간 총 5가지의 문맥 정보를 추출한다.

그림1은 제안하는 BERT 기반 한국어 문맥 정보 추출 시스템의 구조를 나타낸다. BERT 기반의 모델은 인물, 시간, 공간을 추출하기 위한 개체명 인식 모델과 관계 추출을 위한 모델, 감정 추출을 위한 모델로 구분된다. 딥러닝 모델이기 때문에 각 모델은 학습 단계와 추론 단계로 분류 가능하다. 학습 단계에서는 각기 다른 데이터셋




(그림 1) 한국어 문맥 정보 추출 시스템
(Figure 1) Korean context extraction system

과 학습 파라미터를 이용해 학습된다. 학습된 모델은 실제 한국어 문맥 정보 추출 시스템에 추론 단계의 모델로 사용된다.

한국어 문맥 정보 추출 시스템은 입력으로 주어지는 한국어 텍스트를 3가지 추출 모델 별로 설정된 입력 형태에 따라 다른 전처리 과정을 거친다. 전처리 된 문장들은 각 모델의 추론을 통해 최종적으로 인물, 관계, 감정, 시간, 공간 정보를 추출한다.

본 논문에서는 SKTBrain의 KoBERT[13] 모델을 사용한다. 구글의 다국어 모델에 한국어 위키 5백만 문장과 한국어 뉴스 2천만 문장을 추가로 학습시켜 한국어 데이터에서 보다 좋은 성능을 낼 수 있는 모델이다.

우리는 KoBERT 모델을 인물, 시간, 공간 정보 추출을 위한 개체명 인식과 관계 정보 추출, 감정 정보 추출까지 총 3가지의 모델로 학습시키고 추출 시스템에 활용한다. 특히 관계 정보와 감정 정보 추출 모델은 문맥을 고려한 추출이 필요하기 때문에 지식 그래프와 감정 기반 지식 그래프를 적용해 학습시킨다. 우리는 다양한 지식 그래프 중 오픈 소스로 공개되어 있는 ConceptNet과 SenticNet을 활용한다.

제안하는 한국어 문맥 정보 시스템의 각 추출 모델에 대한 설명은 다음 3.1, 3.2, 3.3에서 자세하게 기술한다.

## 3.1 인물, 시간, 공간 정보 추출

인물, 시간, 공간 정보의 추출을 위해 개체명 인식 기술을 활용한다. 개체명 인식 기술은 텍스트에서 인명, 지명, 기관명 등의 개체명을 인식하는 기술이다. 자연어 처리 분야에서 활발한 연구가 진행되고 있는 기술 중 하나이다. 개체명 인식 기술을 활용해 한국어 텍스트 내에서 인물, 시간, 공간 정보의 추출이 가능하다.

(표 1) NER 데이터셋의 태그 리스트
(Table 1) Tag list of NER dataset

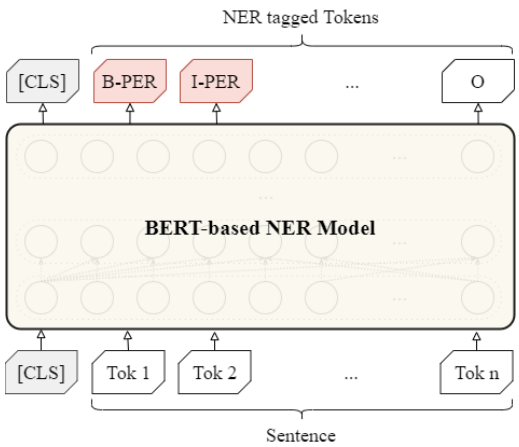
| Category | Tag | Definition |
|---|---|---|
| Person | PER | Person name |
| Space | LOC | Location |
| | ORG | Organization |
| Time | DAT | Date |
| | TIM | Time |
| Others | DUR | Duration |
| | MNY | Money |
| | PNT | Proportion |
| | NOH | Others in number representation |
| | POH | Others |

한국어 개체명 인식, 특히 인물, 시간, 공간 정보의 추출을 위한 BERT 모델의 학습을 위해 한국해양대학교의 개체명 인식 데이터셋[18]을 활용한다. 표 1과 같이 총 8개의 태그로 이루어진 데이터셋이다. 8개의 태그 정보 중 사람 이름은 인물 정보, 지명, 기관명은 공간 정보로 분류하고, 날짜와 시간은 시간 정보로 분류한다. 이외의 태그들은 기타로 분류하여 인물, 공간, 시간 정보 추출을 위한 모델의 학습에는 사용하지 않는다.

인물, 시간, 공간 정보 추출을 위한 개체명 인식 모델의 학습을 위해서는 텍스트 전처리 과정이 필요하다. 전

처리 과정은 모델의 학습 단계와 추론 단계에서 동일하게 적용된다. 모델의 입력으로 들어가는 문장은 BERT에서 제공되는 토크나이저를 이용해 토큰화되어 토큰의 배열 형태로 전처리된다.

그림 2는 제안하는 개체명 인식 모델의 입력과 출력 예제를 보여준다. 토크나이저로 토큰화 된 입력 문장을 모델에 입력하면 각 토큰에 해당하는 태그를 결과로 추출한다. 동일한 형태의 입력과 출력으로 BERT 모델을 학습시키고 인물, 시간, 공간 정보 추출을 위해 사용한다.




(그림 2) NER 모델의 입력값과 출력값
(Figure 2) Input and output of NER model

## 3.2 관계 정보 추출

개체명 인식 모델을 통해 추출된 인물, 시간, 공간 정보 중 인물 정보와 공간 정보는 관계 정보 추출을 위해 활용된다. 시간 정보 또한 활용 가능하나 현재 관계 추출을 위해 사용하는 데이터셋은 '인물-인물', '인물-공간'의 관계 정보만 존재하기 때문에 별도로 시간 정보는 활용하지 않는다. 추후 시간 정보를 활용할 수 있는 데이터셋을 확보해 인물, 공간, 시간 정보를 모두 활용할 계획이다.

관계 정보를 추출하기 위해서는 주체가 되는 인물/공간과 대상이 되는 인물/공간이 필요하다. 추출된 인물과 공간이 포함된 문장을 필터링하여 사용한다.

관계 정보 추출 모델의 학습을 위해서 카이스트에서 공개한 한국어 관계 라벨링 데이터셋[19]을 사용한다. 총 89개의 다양한 관계 태그가 존재하지만, 본 논문에서는 인물 사이, 혹은 인물과 공간 사이의 관계를 나타낼 수 있는 태그 10개를 선정해 사용한다.

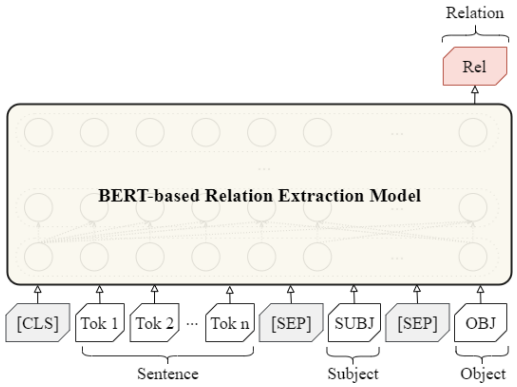
표 2는 논문에서 사용하는 관계 데이터 태그이다. 각 태그들은 주체와 객체 사이의 관계를 표현하는 태그이다. 우리는 주체를 문장 내에서 관계의 주가 되는 인물로 정의하고 [[SUBJ]]로 표현한다. 객체는 주체와 관계 대상이 되는 인물 혹은 공간으로 정의하고 [[OBJ]]로 표현한다.

(표 2) 관계 추출 데이터셋의 태그 리스트
(Table 2) Tag list of relation extraction dataset

| Tag | Definition |
|---|---|
| child | {{OBJ}} is child of {{SUBJ}} |
| spouse | {{OBJ}} is spouse of {{SUBJ}} |
| relative | {{OBJ}} is relative of {{SUBJ}} |
| opponent | {{OBJ}} is opponent of {{SUBJ}} |
| parent | {{OBJ}} is parent of {{SUBJ}} |
| commander | {{OBJ}} is commander of {{SUBJ}} |
| birthPlace | {{OBJ}} is birth place of {{SUBJ}} |
| country | {{OBJ}} is country of {{SUBJ}} |
| foundedBy | {{SUBJ}} is founded by {{OBJ}} |
| influenced | {{SUBJ}} is influenced by {{OBJ}} |

관계 정보 추출 모델을 위한 입력 텍스트의 전처리는 개체명 인식과 달리 주체와 대상이 필요하다. 이때 주체와 대상은 인물 혹은 공간이 된다. 3.1에서 추출된 인물과 공간 정보는 모든 쌍으로 표현해서 관계 추출 모델의 입력으로 사용된다. 예를 들어 인물 A, 인물 B, 인물 C가 추출됐다면, (인물 A, 인물 B), (인물 A, 인물 C), (인물 B, 인물 A)와 같은 형태로 가능한 모든 쌍의 조합을 만들고, 쌍의 첫 번째 요소를 주체, 두 번째 요소를 객체로 구분해 관계 추출 모델의 입력으로 사용한다.

개체명 인식 모델에서 추출된 인물과 공간 정보를 기반으로 입력 문장들을 필터링하여 사용한다. 필터링 된 문장들과 주체, 객체의 쌍을 그대로 사용할 경우 데이터셋으로 활용하는 관계 태그와 관련된 문장이 상대적으로 적기 때문에 정확도 측면에서 낮은 결과를 보여준다. 우리는 이러한 문제점을 개선하기 위해 지식 그래프를 활용해 관계 정보를 확장한다. 지식 그래프는 단어와 단어 사이의 동의어, 관계된 단어, 하위 단어 등 다양한 관계를 기반으로 정보를 표현한다. 오픈소스로 제공되는 ConceptNet[16]은 여러 관계 정보를 제공할 뿐만 아니라 한국어의 관계 정보 또한 제공하고 있기 때문에 ConceptNet을 활용해 데이터셋의 관계 태그와 관련 있는 단어들을 이용해 입력에 사용되는 문장을 확장한다. 그림 3은 제안하는 관계 추출 모델의 입력과 출력의 형태를 보여준다.




(그림 3) 관계 추출 모델의 입력값과 출력값
(Figure 3) Input and output of relation extraction model

추출된 인물/공간이 포함된 문장, 주체가 되는 인물/공간, 그리고 대상이 되는 인물/공간을 하나의 배열로 연결하여 관계 정보 추출 모델의 학습과 추론의 입력값으로 사용한다. 입력된 토큰들을 학습 혹은 추론하여 주체와 대상 사이의 관계를 출력값으로 보여준다.

### 3.3 감정 정보 추출

인간의 감정은 수치화하기 어려울 정도로 다양한 감정을 갖고 있다. 감정을 수치화하고 분류하는 기준이 모호하기 때문에 텍스트에서 감정 정보를 추출하는 것은 매우 어렵다. 여러 감정 모델과 방법들이 존재하고 있다.

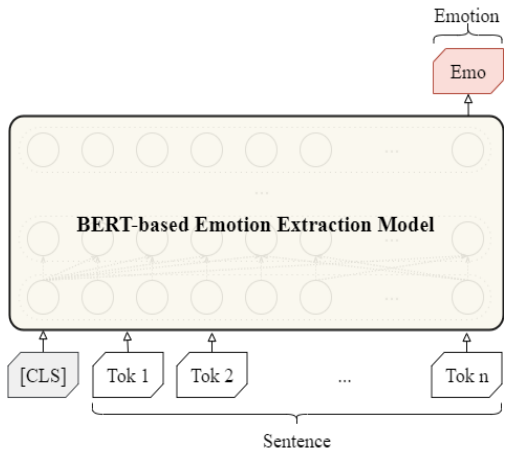
표 3은 동서양의 기본적인 감정의 유형들을 보여준다. 동양권에서는 기본적으로 사람의 감정을 '희(喜), 기쁨', '노(怒), 분노', '애(哀), 슬픔', '락(樂), 즐거움'의 4가지로 분류한다[20]. 반면 서양은 상대적으로 다양한 감정 모델을 갖고 있고 가장 대표적인 Paul Ekman[21]은 감정을 6가지로 분류한다.

(표 3) 동서양의 기본 감정
(Table 3) Basic emotions of East and West

| East | West (Ekman) |
|---|---|
| 희(喜), happiness | happiness |
| 노(怒), anger | anger |
| 애(哀), sadness | sadness |
| 락(樂), joy | - |
| - | disgust |
| - | surprise |
| - | fear |

동양과 서양에서 공통적으로 나타나는 인간의 기본 감정은 '기쁨/행복', '분노', '슬픔'이다. 동서양에서 공통적으로 나타나는 감정을 기본 감정 모델로 정의가 가능하다. 따라서 우리는 기쁨, 분노, 슬픔을 인간의 가장 중요한 핵심적인 감정 정보로 지정하고 기본 감정 모델을 감정 정보 추출 모델에 이용한다.

또한 본 논문에서는 감정 정보 추출을 위해 감정 기반 지식 그래프인 SenticNet[17]을 활용한다. SenticNet은 단어에 내포되어 있는 감정들을 그래프로 연결하고 있기 때문에 감정과 관련된 문장들을 1차적으로 선별하는데 한다. 1차적으로 감정이 내포된 단어를 포함한 문장들을 선정한 후 BERT 모델을 이용한 추론 모델에 활용하기 때문에 정확도 향상에 도움을 준다.




(그림 4) 감정 추출 모델의 입력값과 출력값
(Figure 4) Input and output of emotion extraction model

BERT 기반 감정 정보 추출 모델의 학습을 위해 마이크로소프트의 EmoContext 데이터셋[22]을 활용한다. EmoContext는 감정을 'happy', 'angry', 'sad', 'others' 총 4개로 분류한다. 영어로 된 데이터를 한국어로 활용하기 위해 Naver NMT API[23]를 사용하여 한국어로 번역한 후 BERT 모델로 학습시킨다.

그림 4는 감정 추출 모델의 입력과 출력 형태로 보여주고 있다. 개체명 인식 모델과 마찬가지로 입력값은 토큰화 된 문장이다. 출력값은 'happy', 'angry', 'sad', 'others' 중 1개의 감정이다. EmoContext 데이터셋을 이용해 감정 추출 모델을 학습시키고 추론에 사용한다.

## 4. 실 험

본 논문에서는 한국어 텍스트에서 인물, 공간, 시간, 관계, 감정 총 5가지의 문맥 정보를 추출할 수 있는 KoBERT와 지식 그래프 기반 시스템을 제안한다. 개체명 인식, 관계 정보 추출, 감정 정보 추출 모델을 실험을 통해 검증한다.

### 4.1 실험 환경

BERT 기반 한국어 문맥 정보 추출 시스템의 실험은 구글 colab[24]의 TPU 환경을 이용한다. 모델은 PyTorch를 통해 구현되고, 기반이 되는 BERT 한국어 모델은 SKTBrain에서 제공하는 KoBERT[13]를 사용한다.

모델 학습을 위한 데이터셋은 개체명 인식, 관계, 감정 추출 모델 각각 다른 데이터셋을 사용한다. 인물, 공간, 시간 정보 추출을 위한 개체명 인식 모델의 학습을 위해서는 한국해양대학교의 말뭉치 데이터셋[18]을 활용한다. 관계 정보 추출 모델의 학습에는 카이스트의 관계 데이터셋[19]을 활용하고, 감정 정보 추출 모델의 학습에는 한국어 데이터셋이 없기 때문에 영어 데이터셋인 EmoContext[22]를 Naver NMT API[18]를 이용해 한국어로 번역하여 활용한다.

EmoContext는 학습과 테스트 데이터셋이 분류되어 있기 때문에 제공되는 데이터셋 그대로 실험을 진행한다. 이외의 데이터셋은 80%와 20%의 비율로 랜덤하게 분류해서 학습과 테스트 데이터셋으로 사용한다.

### 4.2 실험 결과

제안하는 한국어 문맥 정보 추출 시스템의 검증을 위해 각 모델을 통해 추출되는 문맥 정보에 대한 정확도를 실험한다. 개체명 인식 모델의 인물, 공간, 시간 정보와 관계 추출 모델의 관계 정보, 그리고 감정 추출 모델의 감정 정보를 실험의 대상으로 한다.

(표 4) 실험 결과
(Table 4) Experiment result

| Context | Precision | Recall | F1-score |
|---|---|---|---|
| Person | 0.9183 | 0.9017 | 0.9099 |
| Space | 0.8936 | 0.9024 | 0.8980 |
| Time | 0.8409 | 0.8675 | 0.8539 |
| Relation | 0.8726 | 0.8507 | 0.8615 |
| Emotion | 0.9375 | 0.9163 | 0.9267 |

표 4는 각 데이터셋에 대한 정확도 실험 결과이다. 실험은 정확도(precision), 재현율(recall), 그리고 F1 점수로 측정한다. 각 문맥 정보에 대해 측정 결과를 보여준다.

개체명 인식 모델을 통해 추출된 인물, 공간 시간 정보는 각각 약 0.90, 0.89, 0.85의 결과를 보여준다. 인물과 공간 정보는 높은 성능을 보여주고 있지만, 시간 정보의 경우 숫자가 포함되어 있는 경우가 많기 때문에 별도로 숫자에 대한 처리를 필요로 한다.

관계 추출 모델을 통해 추출된 관계 정보는 약 0.86의 결과를 보여준다. 관계 정보는 주체와 대상 사이의 관계를 추출하는 다소 복잡한 연산이 필요하기 때문에 상대적으로 낮은 성능으로 보일 수 있다. 이는 앞서 개체명 인식 모델을 통해 추출된 인물, 공간 정보를 입력값으로 활용하기 때문에 이때 발생한 오류가 누적되어 결과값에 영향을 주었기 때문이다. 지식 그래프를 활용하면서 관계 추출 모델의 성능이 일정 부분 향상됐기 때문에 보다 확장된 지식 그래프를 사용한다면 추가적인 성능 향상이 가능하다.

감정 정보를 추출하는 모델의 성능은 약 0.92로 좋은 성능을 보여주고 있다. 한국어로 이루어진 데이터셋이 아닌 영어 데이터셋을 한국어로 번역한 데이터로 학습했음에도 높은 성능을 보여 제안하는 모델의 가능성을 보여준다.

표 5는 기존 구글에서 제공하는 'BERT -Base, Multilingual Cased' 모델만 사용했을 때의 결과와 SKTBrain의 KoBERT 모델만 사용한 결과, 그리고 우리가 제안하는 지식 그래프 기반 모델의 F1 점수를 비교한 결과이다. KoBERT는 기존 BERT 모델을 한국어 코퍼스를 이용해 다시 학습시켰기 때문에 모든 문맥 정보에 대해서 성능이 향상된 것을 확인할 수 있다. 또한 우리가 제안하는 지식 그래프를 적용한 관계와 감정 추출 모델은 KoBERT만 적용했을 때보다 성능이 향상되어 지식 그래프 사용의 가능성을 보여준다.

(표 5) 비교 결과
(Table 5) Comparison result

| Context | F1-score | | |
|---|---|---|---|
| | BERT-Base | KoBERT | Ours |
| Person | 0.8705 | 0.9099 | 0.9099 |
| Space | 0.8416 | 0.8980 | 0.8980 |
| Time | 0.8024 | 0.8539 | 0.8539 |
| Relation | 0.8371 | 0.8504 | 0.8615 |
| Emotion | 0.8638 | 0.9149 | 0.9267 |

각 모델들을 서로 다른 모델이지만 성능 향상의 정도를 평가하기 위해 성능 향상의 정도를 확인할 때, 전체적으로 약 4% 포인트의 성능이 향상됐다. 인물 추출이나 관계 추출에서는 다소 낮은 정도의 성능 향상이 이루어졌지만, 공간, 시간, 감정 추출에서는 약 5% 포인트의 성능 향상을 보여주었다.

## 5. 결 론

본 논문에서는 언어 모델인 BERT과 지식그래프를 이용해 한국어 문맥 정보를 추출하는 시스템을 제안하고 검증한다. BERT 모델은 하나의 모델로 다양한 자연어 처리 분야의 일들을 수행할 수 있지만, 한국어 데이터 처리에서는 한계를 갖고 있다. 한국어 위키 데이터와 뉴스 데이터로 다시 학습된 모델을 사용해 한국어 데이터에 잘 대응할 수 있는 KoBERT 모델을 사용한다.

또한 컴퓨터에게 문맥 정보를 쉽게 학습시키기 위해 관계를 기반으로 단어를 연결한 지식 그래프를 관계와 감정 추출 모델에 활용한다. 한국어 BERT 모델과 지식 그래프를 기반으로 중요한 문맥 정보인 인물, 관계, 감정, 공간, 시간의 5가지 문맥 정보를 추출하기 위한 개체명 인식 모델, 관계 추출 모델, 감정 추출 모델을 학습시켜 각 정보를 추출할 수 있는 시스템을 구축한다.

한국어 문맥 정보 추출 시스템을 통해 추출된 문맥 정보들은 분석과 알고리즘 적용을 통해 검색, 추천 등의 성능 향상을 위해 활용 가능하다. 특히 대화형 인공지능 시스템에서 인간의 언어를 컴퓨터에게 이해시키고 의미 있는 정보를 제공하기 위해 유용하게 활용할 수 있다.

## 참고문헌(Reference)


[1] Deng, L., and Liu, Y. (Eds.), Deep Learning in Natural Language Processing, Springer, 2018. http://dx.doi.org/10.1007/978-981-10-5209-5

[2] Palash, G., Sumit, P., and Karan, J., Deep Learning for Natural Language Processing, Apress, 2018. http://dx.doi.org/10.1007/978-1-4842-3685-7

[3] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., “Deep contextualized word representations,” in Proc. of NAACL, 2018. http://dx.doi.org/10.18653/v1/N18-1202

[4] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. of NAACL, 2019. http://dx.doi.org/10.18653/v1/N19-1423

[5] Zhilin, Y., Zihng, D., Yiming, Y., Jaime, C., Ruslan, S., and Quoc V. L., “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” arXiv preprint, 2019. https://arxiv.org/abs/1906.08237

[6] K. H. Park., S. H. Na., J. H. Shin., and Y. K. Kim., “BERT for Korean Natural Language Processing: Named Entity Tagging, Sentiment Analysis, Dependency Parsing and Semantic Role Labeling,” Korea Computer Congress 2019, 2019, pp. 584-586. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08763261

[7] S. Kwon., Y. Ko., and J. Seo, “Effective vector representation for the Korean named-entity recognition,” Pattern Recognition Letters, Vol. 117, pp. 52-57, 2019. http://dx.doi.org/10.1016/j.patrec.2018.11.019

[8] Sung-Il, Lee., “Contextualism and a Reflection on the Notions of ‘Context’,” Journal of Language Sciences, Vol. 17, No. 3, pp. 67-86, 2010. http://dx.doi.org/G704-001077.2010.17.3.003

[9] Min-Woo, Lee., “Semantic Relations from the Contextual Perspective,” Korean Semantics, Vol. 66, pp. 101-120, 2019. http://dx.doi.org/10.19033/sks.2019.12.66.101

[10] M. S. Shin., “The Characteristics of the Contextual Meaning Evaluation Items of Words – Focusing on the Korean Language Subject of the College Scholastic Ability Text,” KOED, No. 116, pp. 143-185, 2018. http://dx.doi.org/10.15734/koed..116.201809.143

[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, “Attention is

all you need," In Proc. of the 31st Internationcal Conference on Neural Information Processing Systems, pp. 6000-6010, 2017.
http://dx.doi.org/10.5555/3295222.3295349

[13] SKTBrain, "Korean BERT pre-trained cased (KoBERT)," https://github.com/SKTBrain/KoBERT

[14] Fellbaum, C., "WordNet: An Electronic Lexical Database," Cambridge, MA: MIT Press, 1998.
http://dx.doi.org/10.1017/S0142716401221079

[15] Thomas, R., Fabian M. S., Johannes, H., Joanna, B., and Gerhard, W., "YAGO: A Mulitilingual Knowledge Base from Wikipedia, Wordnet, and Geonames," in Proc. of 15th International Semantic Web Conference, pp. 177-185, 2016.
https://doi.org/10.1007/978-3-319-46547-0_19

[16] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge," In Thirty-First AAAI Conference on Artificial Intelligence, 2017.
https://dl.acm.org/doi/10.5555/3298023.3298212

[17] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings," In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16839

[18] Kmounlp, "Definition of Korean Named-Entity Task," https://github.com/kmounlp/NER

[19] KAIST, "Korean Relation Extraction Gold Standard," https://github.com/machinereading/kor-re-gold

[20] S. S. Lee., "A Study on the Analysis of Semantic Relation and Category of the Korean Emotion Words," Journal of Korean Library and Information Science Society, Vol. 47, No. 2, pp. 51-70, 2016.
http://dx.doi.org/10.16981/kliss.47.201606.51

[21] P. Ekman, "Are there basic emotions?" Psychological Review, Vol. 99, No. 3, pp. 550-553, 1992.
http://dx.doi.org/10.1037/0033-295X.99.3.550

[22] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, "SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text," in Proc. of the 13th International Workshop on Semantic Evaluation, pp. 39-48, 2019.
http://dx.doi.org/10.18653/v1/S19-2005

[23] Naver Developers, "Papago NMT API Reference," https://developers.naver.com/docs/nmt/reference/

[24] Google, "Google Colab," https://colab.research.google.com

## ◐ 저 자 소 개 ◑

**유 소 엽(SoYeop Yoo)**
2014년 가천대학교 소프트웨어설계·경영학과 (학사)
2016년 가천대학교 일반대학원 소프트웨어설계·경영학과 (석사)
2018년~현재 가천대학교 일반대학원 IT융합공학과 소프트웨어학 전공 (박사과정)
관심분야 : Machine learning, Deep learning, Big data, Data mining, Social computing, etc.
E-mail : bbusso@gc.gachon.ac.kr

**정 옥 란(OkRan Jeong)**
2005년 이화여자대학교 대학원 컴퓨터공학과 (공학박사)
2006년 서울대학교 컴퓨터공학부 (박사 후 연구원)
2007년 Univ. of Illinois at Urban-Champaign (박사 후 연구원)
2008년 성균관대학교 정보통신학부 (연구교수)
2009년~현재 가천대학교 AI·소프트웨어학부 (부교수)
관심분야 : Big data, Social media mining, Machine learning, etc.
E-mail : orjeong@gachon.ac.kr