BERT 모델과 지식 그래프를 활용한 지능형 챗봇

An Intelligent Chatbot Utilizing BERT Model and Knowledge Graph

저자 (Authors)
유소엽, 정옥란
SoYeop Yoo, OkRan Jeong

출처 (Source)
한국전자거래학회지 24(3), 2019.8, 87-98(12 pages)
The Jounal of Society for e-Business Studies 24(3), 2019.8, 87-98(12 pages)

발행처 (Publisher)
한국전자거래학회
Society for e-Business Studies

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08769539

APA Style
유소엽, 정옥란 (2019). BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 한국전자거래학회지, 24(3), 87-98

http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.24, No.3, August 2019, pp.87-98
https://doi.org/10.7838/jsebs.2019.24.3.087

BERT 모델과 지식 그래프를 활용한 지능형 챗봇

An Intelligent Chatbot Utilizing BERT Model and Knowledge Graph

유소엽(SoYeop Yoo)*, 정옥란(OkRan Jeong)**

초 록

인공지능이 활발하게 연구되면서 이미지, 영상, 자연어 처리와 같은 다양한 분야에 적용되고 있다. 특히 자연어 처리 분야는 사람이 말하고 쓰는 언어들을 컴퓨터가 이해할 수 있도록 하기 위한 연구들이 진행되고 있고 인공지능 기술에서 매우 중요한 영역 중 하나로 여겨진다. 자연어 처리에서 컴퓨터에게 사람의 상식을 이해할 수 있도록 학습시키고 사람의 상식을 기반으로 결과를 생성하도록 하는 것은 복잡하지만 중요한 기술이다. 단어들의 관계를 이용해 연결한 지식 그래프는 컴퓨터에게 쉽게 상식을 학습시킬 수 있다는 장점이 있다. 하지만 기존에 고안된 지식 그래프들은 특정 언어나 분야에만 집중해 구성되어 있거나 신조어 등에는 대응하지 못하는 한계점을 갖고 있다. 본 논문에서는 실시간으로 데이터를 수집 및 분석하여 자동으로 확장 가능한 지식 그래프를 구축하고, 이를 기반 데이터로 활용하는 챗봇 시스템을 제안하고자 한다. 특히 자동 확장 그래프에 BERT 기반의 관계 추출 모델을 적용시켜 성능을 향상시키고자 한다. 자동 확장 지식 그래프를 이용해 상식이 학습되어 있는 챗봇을 구축하여 지식 그래프의 활용 가능성과 성능을 검증한다.

ABSTRACT

As artificial intelligence is actively studied, it is being applied to various fields such as image, video and natural language processing. The natural language processing, in particular, is being studied to enable computers to understand the languages spoken and spoken by people and is considered one of the most important areas in artificial intelligence technology. In natural language processing, it is a complex, but important to make computers learn to understand a person's common sense and generate results based on the person's common sense. Knowledge graphs, which are linked using the relationship of words, have the advantage of being able to learn common sense easily from computers. However, the existing knowledge graphs are organized only by focusing on specific languages and fields and have limitations that cannot respond to neologisms. In this paper, we propose an intelligent chatbotsystem that collects and analyzed data in real time to build an

본 논문은 2019년 과학기술정보통신부의 재원으로 한국연구재단의 기초연구사업 지원과 문화체육관광부 및 한국저작권위원회의 저작권기술개발사업, 과학기술정보통신부의 SW중심대학 사업의 연구결과로 수행되었음. (Nos. NRF-2019R1A2C1008412, 2019-CONTEXT-9500, 2015-0-00932)

* First Author, Department of Software, Gachon University(bbusso@gc.gachon.ac.kr),
** Corresponding Author, Department of Software, Gachon University(orjeong@gachon.ac.kr)

Received: 2019-06-28, Review completed: 2019-07-23, Accepted: 2019-08-09

automatically scalable knowledge graph and utilizes it as the base data. In particular, the fine-tuned BERT-based for relation extraction is to be applied to auto-growing graph to improve performance. And, we have developed a chatbot that can learn human common sense using auto-growing knowledge graph, it verifies the availability and performance of the knowledge graph.

키워드 : 지식 그래프, BERT, 챗봇, 관계 추출
Knowledge Graph, BERT, Chatbot, Relation Extraction

1. 서 론

수없이 많이 축적된 빅데이터와 복잡한 연산들을 빠르게 처리할 수 있도록 하는 하드웨어의 성능이 향상됨에 따라 인공지능 기술에 대한 관심이 높아지고 다양한 연구들이 진행되고 있다. 특히 4차 산업혁명은 빅데이터, IoT 등의 정보기술과 인공지능 소프트웨어의 지능기술의 융합을 핵심으로 하고 있다. 많은 연구들이 수집된 데이터를 분석하거나 예측하기 위한 인공지능 기술에 집중되어 있고, 최근에는 인공지능 기술을 다른 분야, 기술과 융합하기 위한 연구가 이루어지고 있다[6, 8, 10].

인공지능 기술이 적용된 융합 기술을 가장 쉽게 접할 수 있는 시스템은 챗봇이다. 인간과 대화를 주고받기 위해 자연어 처리(NLP: Natural Language Processing), 인식(NLU: Natural Language Understanding), 생성 (NLG: Natural Language Generation) 등에 인공지능 기술을 적용하여 챗봇의 성능을 높이기 위한 연구가 진행되고 있다[6, 8, 10]. 챗봇에 필요한 다양한 인공지능 기술 중 컴퓨터가 인간의 상식(common sense)을 이해할 수 있도록 하는 것은 매우 중요한 기술이다. 인간들에게 언어는 사회적, 문화적으로 자연스럽게 학습되는 일종의 상호간의 규약이지만, 컴퓨터는 인간의 상식을 이해하기 위해 별도의 학습이 필요하다. 이를 가능하게 하는 많은 기술들 중 지식 그래프는 단어와 단어를 관계로 연결하여 그래프로 나타내기 때문에 챗봇 시스템에서 핵심적인 기반 데이터로 활용이 가능하다[2, 14].

컴퓨터가 마치 인간처럼 단어의 의미, 단어 사이의 관계를 파악하기 위해서는 지식 그래프는 핵심 기술이 된다. 하지만 기존의 지식 그래프들은 특정 언어나 분야에만 집중되어 있거나 신조어에 대응하지 못하는 한계점을 갖고 있다. 본 논문에서는 실시간으로 데이터를 수집, 분석하여 새로운 관계들을 추출하고 기존 그래프에 연결함으로써 자동으로 확장 가능한 지식 그래프를 제안한다. 또한 제안하는 자동 확장 지식 그래프를 기반으로 챗봇 시스템을 설계 및 구현하여 지식 그래프의 활용 가능성을 검증하고자 한다.

2. 관련 연구

2.1 지식 그래프

최근 인간의 문장을 인간이 인지하는 것처럼 컴퓨터가 이해할 수 있도록 돕기 위한 연구가 활발하게 이루어지고 있다. 하지만 여전히 인

간의 인지 능력까지 끌어올리는 데에는 한계가 존재한다. 보다 인간처럼 문장을 이해하기 위해서 단어와 단어 사이를 관계로 연결한 지식 그래프 기술이 필요하다. WordNet, YAGO, Probase, ConceptNet 등 지식 그래프에 대한 다양한 연구들이 진행되고 있다[2, 4, 9, 13-15].

WordNet[3]은 영어에 대한 단어 데이터베이스로, 단어의 의미를 기반으로 명사, 동사, 형용사 등으로 분류되어 있다. 동의어 집합(synset)으로 분류하고 개념적인 의미와 관계로 서로 연결한다. 약 20만 7천 개의 단어-의미 쌍으로 구성되어 있다. YAGO[9]는 Wikipedia, WordNet, GeoNames 데이터를 기반으로 구축된 지식 그래프이다. 약 1천만 개 이상의 개체들이 연결되어 있다. Probase[15]는 Microsoft에서 구축한 지식 그래프로 약 16억 개 이상의 웹 페이지로부터 데이터를 수집했다. 약 2백만 개의 개체들과 2억 개 이상의 페어들로 구성되어 있다. ConceptNet[13]은 오픈소스 지식 그래프로 영어, 프랑스어 등 10개의 핵심 언어들에 대해 대량의 데이터를 갖고 있고, 한국어 등 약 68개의 언어들에 대해서는 소량의 데이터를 보유하고 있다. 8백만 개 이상의 개체들이 서로 40개의 관계를 기반으로 연결되어 있다.

이외에도 다양한 지식 그래프들이 연구되고 있지만 현존하는 지식 그래프들은 데이터가 쌓인 후 일정 기간 이상 지나면 한꺼번에 지식 그래프를 확장하는 등 여전히 인간의 개입이 많이 필요하다. 또한 특정 언어, 분야를 기반으로 데이터를 사용하기 때문에 언어 종속성이 존재하고 신조어에 대한 대응이 어렵다.

2.2 BERT

딥러닝 기술이 빠르게 발전하면서 이미지, 영상 등 다양한 분야에서 이미 높은 성능을 보였고 다양하게 활용되고 있다. 하지만 사람의 말을 처리하는 자연어 처리 분야에서는 최근 딥러닝 기술을 활발하게 적용하고 있다. 자연어 처리 분야에서 중요하게 여겨지는 기술은 이미지 처리의 전이 학습처럼 사전 학습된 언어 모델을 이용하는 것이다[5, 7, 16].

대표적인 사전 학습된 언어 모델에는 ELMo[11], OpenAI GPT[12], BERT[4] 등이 있다. 특히 BERT는 최근 다양한 자연어 처리 분야의 챌린지에서 가장 좋은 성능을 내면서 여러 가지 일들을 수행하기 위해 사용되고 있다. 대용량의 데이터를 직접 학습시키기 위해서는 매우 많은 자원과 시간이 필요하다. 하지만 BERT 모델은 기본적으로 대량의 단어 임베딩 등에 대해 사전 학습이 되어있는 모델을 제공하기 때문에 상대적으로 적은 자원만으로도 충분히 자연어 처리의 여러 일들의 수행이 가능하다.

본 논문에서는 인간과 대화하는 챗봇에 인간의 상식을 이해하는데 도움을 줄 수 있는 지식 그래프를 적용하여 컴퓨터에게 일반적인 지식, 상식을 학습시키고자 한다. 기존 지식 그래프의 한계점을 개선하고 효율적으로 데이터를 처리하고 분석하기 위해 실시간 빅데이터 분석 및 예측 시스템인 Polaris[17]를 기반으로 한다. Polaris는 실시간으로 데이터 수집, 이벤트 감지, 경로 분석, 감정 분석 및 예측이 가능한 시스템으로 기반 데이터로 자동 확장 지식 그래프인 PolarisX를 사용한다. PolarisX는 새로운 데이터 소스를 실시간으로 수집, 분석해 자동으로 지식 그래프의 확장이 가능하다. 특히 최근 자연어 처리 분야에서 높은 성능을 보여주는 Google의 BERT를 이용해 관계 추출 모델을 적용하여 보다 높은 성능의 지식 그래프를

구축하고자 한다. 이러한 PolarisX를 기반으로 관련된 지식에 대한 관계를 보여주는 챗봇인 PolarisX-bot을 설계 및 구현하였다.

3. 지능형 챗봇 시스템

본 논문에서 제안하는 자동 확장 지식 그래프 기반 지능형 챗봇인 PolarisX-bot은 <Figure 1>과 같이 설계되었다. 시스템은 지식 그래프 레이어와 챗봇 레이어로 구성된다. 지식 그래프 레이어에서는 기존 지식 그래프를 자동으로 확장하기 위해 트위터와 뉴스를 통해 새로운 데이터를 계속해서 수집하고 BERT 모델을 기반으로 새로운 관계를 추출한다. 추출된 관계를 지식 그래프에 연결함으로써 자동으로 확장 가능한 지식 그래프를 구축한다. 챗봇 레이어에서는 사용자의 질문 의도를 파악하고 그에 따른 결과를 지식 그래프를 기반으로 보여줄 수 있도록 구성된다.

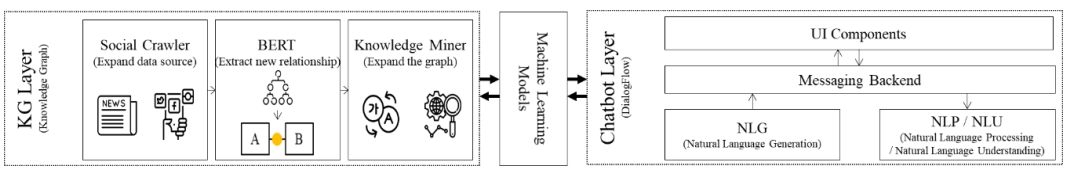
3.1 BERT를 이용한 자동 확장 지식 그래프

지식 그래프는 단어, 문장 등의 관계를 연결하여 지식을 그래프로 나타낸다. 지식의 관계를 포함하고 있어 컴퓨터에게 인간의 상식을 학습시키는데 유용하게 활용 가능하다. 하지만 기존의 많은 지식 그래프들은 영어, 프랑스어와 같은 특정 언어에만 집중되어 있고, Wikipedia처럼 기존에 수집된 데이터를 기반으로 하기 때문에 다양한 언어나 신조어에 대응하지 못한다. 이러한 한계를 개선하고 신조어에 대응 가능하고 언어 종속성이 없는 지식 그래프의 구축이 필요하다.

자동 확장 지식 그래프는 기존 지식 그래프의 한계를 개선하기 위해 분석 대상이 되는 데이터 소스를 확장한다. 실시간으로 뉴스, 소셜 미디어 등의 데이터를 수집하고, 수집된 데이터를 분석해 단어와 단어 사이의 관계를 추출한다. 추출된 단어, 관계 쌍은 기존의 지식 그래프에 존재하는지 확인하여 없을 경우 추가함으로써 그래프를 자동으로 확장해 나간다.

우리는 자동 확장 지식그래프의 구축에서 새로운 관계를 추출하기 위해 BERT[3] 모델을 사용한다. BERT는 구글에서 발표한 사전 학습된(pre-trained) 언어 모델로 NLP 분야의 11개 실험에서 State-of-the-art를 차지했다. 기존 데이터를 사전에 학습시켜 일반적인 언어 모델로 공개된 BERT는 수행하고자 하는 일에 따라 학습 데이터를 이용해 미세 조정(fine-tuning)이 가능하다. 우리는 BERT를 이용해 키워드(entity) 사이의 관계를 추출할 수 있도록 모델을 미세 조정한다.

BERT 모델이 자동 확장 지식 그래프에서 사용될 수 있도록 새로운 관계 추출 모델 구축



〈Figure 1〉 System Architecture

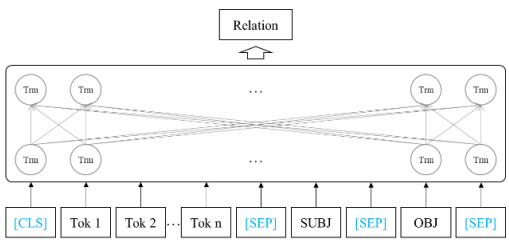
을 위해 관계 기반 데이터인 TACRED[19]를 사용한다. TACRED 데이터는 뉴스나 웹 텍스트로 만들어진 관계 추출을 위한 데이터셋이다. 41개의 관계 종류가 존재하고, 관계가 없는 경우는 'no_relation'으로 라벨링 된다.

관계의 추출을 위해서는 두 개의 객체(entity)와 관계를 나타내는 'relation'을 찾아야 한다. <Figure 2>는 제안하는 관계 추출 모델의 입력과 출력을 보여준다. 관계 추출 모델의 학습을 위해서 BERT 모델에 입력값으로 문장을 입력할 때 TACRED에서 제공해 주는 주어와 목적어를 전체 문장 뒤에 연결한다. 하나의 문장으로 연결하여 BERT 모델에 인풋으로 하고, 최종적으로 나오는 아웃풋, 즉 라벨을 관계로 한다. 이렇게 학습과 검증을 거친 관계 추출 모델은 새로운 문장에 대해 관계를 추출하기 위해 사용된다.

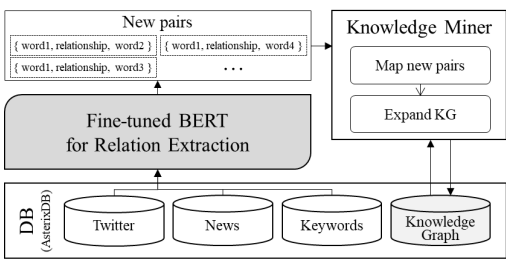
<Figure 2>에서 [CLS]와 [SEP]는 BERT 모델에서 사용하는 별도의 토큰으로 각각 문장의 시작과 끝을 의미한다. Tok 1, Tok 2, …, Tok n은 BERT 모델에 학습을 위해 인풋을 토큰화(tokenization)한 후의 각 토큰을 의미한다. 또한 SUBJ는 입력 문장 내에서 주어 즉, 관계의 주체가 되는 단어를 뜻하고, OBJ는 목적어 즉, 관계의 대상이 되는 단어를 뜻한다. 양방향 사전 학습 언어 모델인 BERT에서 아웃풋은 SUBJ와 OBJ 사이의 관계를 나타낸다.

소셜 미디어인 트위터와 뉴스 데이터를 실시간으로 수집하여 많이 언급되는 키워드를 추출한다. 추출된 키워드가 포함된 문장들을 트위터와 뉴스 데이터에서 확인하고 해당되는 문장을 제안하는 BERT 기반의 관계 추출 모델에 입력값으로 입력한다. 키워드와 다른 단어 사이의 관계까지 추출해 {키워드, 관계, 단어}와 같은 쌍을 만든다. 만들어진 쌍은 기존의 지식 그래프와 매핑 함으로써 그래프를 확장한다.

<Figure 3>은 BERT 모델을 이용해 새로운 관계를 추출하고 추출된 관계를 기존의 지식 그래프에 확장하는 방법을 보여준다. 실시간으



〈Figure 2〉 Relation Extraction Using BERT



〈Figure 3〉 Flow of Extracting New Relations and Expanding Knowledge Graph

로 수집한 트위터, 뉴스 데이터로부터 키워드를 추출하고 키워드가 포함된 문장을 관계 추출을 위해 TACRED 데이터셋을 기반으로 미세 조정된 BERT 모델에 입력한다. 모델은 입력된 문장에 대해 새로운 관계를 예측해서 보여주고, 이 결과를 기반으로 기존 지식 그래프와 매칭되는 노드를 찾아 연결한다.

PolarisX는 기존의 지식 그래프 중 오픈소스 지식 그래프인 ConceptNet[13]을 기반으로 확장해 나간다. 새로운 관계 지식을 얻기 위한 데이터 소스는 트위터와 뉴스를 기반으로 한다. 오픈소스 빅데이터 관리 시스템인 AsterixDB[1]는 FeedAdapter 기능을 통해 간단한 설정과 쿼리를 입력하면 실시간으로 트위터 데이터의 수집 및 저장을 가능하게 한다. 또한 다양한 인덱싱 기법을 통해 빅데이터 처리의 효율성을 높여줄 수 있어 PolarisX의 데이터베이스 시스템으로 사용한다.

실시간으로 수집되는 트위터와 뉴스 데이터에서 핵심 키워드들을 수집하고 키워드가 포함되어 있는 문장들을 추출해 관계를 추출한다. 관계 추출을 위해서 BERT를 기반으로 TACRED 데이터를 이용해 조정된 딥러닝 모델을 이용한다. 모델을 기반으로 키워드와 다른 단어 사이의 연결되어 있는 관계 지식을 추출해 새로운 관계 지식 쌍을 만든다. 지식 쌍은 각각 {키워드, 관계, 단어}의 형태로 연결된다. 최종적으로 기존 지식 그래프와 비교 매핑을 통해 계속해서 확장 가능한 지식 그래프를 구축한다.

3.2 자동 확장 지식 그래프를 이용한 지능형 챗봇

인간과의 대화를 수행하는 챗봇은 인간의 자연어를 처리해 의도를 파악하고 의도에 해당하는 답변을 하는 것이 중요하다. 챗봇 시스템은 주로 인간의 말, 텍스트 등 자연어를 인식하기 위한 모듈과 의도를 파악하는 모듈, 의도에 따른 답변 내용을 다시 자연어로 만드는 모듈 등으로 구성된다. 실제로 챗봇 시스템을 처음부터 끝까지 구축하기 위해서는 수많은 모듈과 모델들이 필요하다. 본 논문에서는 쉽게 챗봇 시스템을 구현하기 위해 구글의 대화형 챗봇 서비스인 DialogFlow를 활용한다.

구글의 DialogFlow는 몇 가지 설정을 통해 필요한 챗봇을 구축할 수 있도록 도움을 주는 서비스이다. 기본적인 인사 등을 직접 입력해서 활용할 수도 있지만 우리는 자동 확장 지식 그래프인 PolarisX를 기반으로 하는 PolarisX-bot을 구축하기 위해 기반 데이터로 PolarisX를 활용한다. DialogFlow는 기본적으로 자연어 인식, 처리 모듈을 갖고 있다. 이 모듈을 이용해 사용자가 입력한 문장을 분석해서 질문의 의도를 파악하는 설정을 한다.

기본적으로 사용자의 입력 문장을 분석했을 때 'is a', 'has a'와 같은 질문의 형태가 들어있을 경우 PolarisX에 대한 질문으로 인식할 수 있도록 의도 설정을 한다. 질문에 해당하는 'what'이 입력되면 PolarisX에 대한 질문으로 인식하도록 객체를 설정하고, 'is a', 'has a'와 같은 단어들은 관계로 설정한다. 문장 내에서 질문 객체와 관계 객체가 아닌 단어는 PolarisX에서 검색할 키워드로 설정한다.

DialogFlow의 모듈에 의해 분석된 문장 결과를 이용해 PolarisX에서 해당하는 키워드와 관계를 검색하고 결과값을 사용자에게 보여준다. 하나의 키워드와 관계에 대해 다양한 결과가 나올 수 있기 때문에 단순하게 문장으로 결과값을 보여주는 것이 아닌 상위 5개의 결과를

그래프로 표현하여 보여줌으로써 보다 직관적으로 지식을 얻을 수 있도록 한다.

PolarisX-bot은 자동 확장 지식 그래프인 PolarisX와 DialogFlow를 활용해 구축한다. 인간의 일반적인 지식과 상식을 지식 그래프를 통해 챗봇을 학습시키고 그에 따른 대화가 가능하기 때문에 자동 확장 지식 그래프의 활용성을 지능형 챗봇을 통해 검증할 수 있다.

4. 실 험

우리는 BERT 모델을 이용해 자동으로 확장하는 지식 그래프를 구축하고, 이를 기반으로 대화를 주고받을 수 있는 지능형 챗봇을 구현한다. 제안하는 시스템의 검증을 위해 BERT 기반 자동 확장 지식 그래프의 정확도를 측정하고, 지능형 챗봇의 구현 결과를 보여준다.

4.1 데이터셋

제안하는 시스템의 실험을 위해서는 크게 2가지 분류의 데이터 종류를 활용한다. 첫 번째는 BERT 기반 자동 확장 지식 그래프가 계속해서 확장해 나가기 위해 실시간으로 수집하는 소셜 미디어와 뉴스 데이터이다. 두번째로 수집된 데이터로부터 새로운 관계를 추출할 수 있는 딥러닝 모델의 학습을 위한 데이터이다.

〈Table 1〉 Dataset

Data	Size
Twitter	About 15,000,000 tweets
News	About 102,000 articles
TACRED	About 106,000 sentences

<Table 1>은 실험에 사용한 데이터셋을 요약해서 보여준다. 트위터와 뉴스 데이터는 지식 그래프의 데이터 소스 확장을 위해 사용된다. 실제로는 실시간으로 데이터를 수집해서 자동으로 확장이 가능하지만 실험에서는 2018년 11월 한 달 간의 데이터를 활용했다. 트위터 데이터는 Apache AsterixDB의 FeedAdapter 기능을 이용해 실시간으로 트위터 스트리밍 데이터를 수집하고, 뉴스 데이터는 NewsAPI를 통해 수집했다.

TACRED 데이터는 지식 그래프에서 새로운 관계를 추출하기 위한 딥러닝 모델의 학습 데이터이다. 또한 실험을 통해 지식 그래프의 관계 추출 모델의 검증을 위해서도 활용된다. TACRED 데이터는 LDC(Linguistic Data Consortium)을 통해 얻을 수 있고, 총 106,264개의 문장들로 이루어져 있다. TAC Knowledge Base Population challenge에서 사용되는 코퍼스로부터 만들어진 관계 추출 데이터셋으로 'no_relation'을 포함해 총 42개의 관계를 포함하고 있다.

4.2 BERT를 이용한 지식 그래프에 대한 실험

제안하는 자동 확장 지식 그래프는 새로운 데이터로부터 새로운 관계를 찾아내 그래프로 만드는 것이 매우 중요하다. 우리는 사전 학습된 BERT 모델을 관계 추출에 적합하도록 TACRED 데이터셋을 이용해 학습시키고 관계 추출을 위한 딥러닝 모델을 구축한다. 모델 구축을 위해 Google colab의 TPU 환경을 이용한다.

BERT는 대소문자 구분 여부, 계층(layer)의 수, 히든 유닛의 수에 따라 다른 사전 학습된 모델들이 존재한다. 우리는 자동 확장 지식 그

래프의 검증을 위해 12개의 계층으로 구성되어 있고, 768개의 히든 유닛, 그리고 12개의 헤드로 구성된 대소문자 구별이 가능하고 104개의 언어를 지원하는 'BERT-Base, Multilingual Cased' 모델을 실험에 이용한다. 또한 지식 그래프에 활용하기 위해 TACRED 데이터셋을 이용해 모델을 학습시킨다. TACRED를 각각 약 65%, 20%, 15%로 학습(train), 검증(dev), 테스트(test) 셋을 분리하여 실험한다.

〈Table 2〉 Comparison Results on Relation Extraction

Model	F1 score
Logistic Regression[18]	59.4
PA-LSTM[19]	65.1
C-GCN+PA-LSTM[18]	68.2
BERT-based model (our)	75.7

<Table 2>는 제안하는 BERT 기반 관계 추출 모델과 기존 연구와의 비교 실험 결과를 보여준다. 전통적인 기법인 로지스틱 회귀(Logistic Regression)를 모델과 딥러닝을 기반으로 하는 모델들과 비교 실험을 했다. 단어의 위치를 활용한 LSTM 모델인 PA-LSTM[19]과 그래프 합성곱 신경망을 적용한 Zhang et al. [19]의 모델을 이용했다. TACRED 데이터셋을 이용한 모델들 중 가장 높은 실험 결과를 보여준 모델과 비교했을 때, BERT를 기반으로 한 관계 추출 모델이 76.7의 F1 점수로 좋은 성능을 보여줬다.

관계 추출은 관계를 추출하기 위해 그 대상이 되는 두 개의 개체를 선택하고 또 관계를 추출해야 하기 때문에 기존의 다른 NLP 과제들보다 상대적으로 복잡하다. BERT-Base 모델을 TACRED 데이터셋으로 실험했을 때 기존의 모델들보다 좋은 결과를 보여줬다. BERT base 모델보다 계층의 수, 히든 유닛의 수 등이 더 많이 존재하고 더 큰 데이터셋으로 학습된 BERT-Large 모델을 적용할 경우 보다 높은 성능을 보일 수 있을 것으로 기대된다.

4.3 지능형 챗봇 시스템 구현

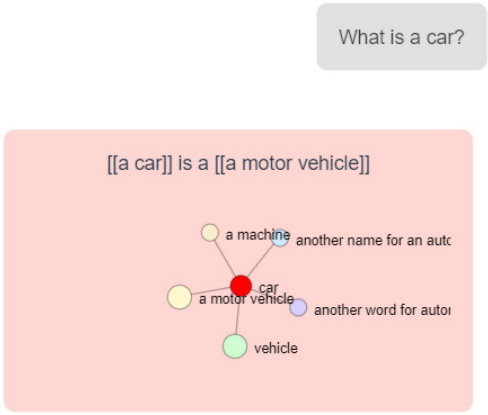
본 논문에서 자동 확장 지식 그래프의 기반 지식 그래프로 활용한 ConceptNet은 IsA, HasA, PartOf 등 40개의 관계를 갖고 있다. A라는 개체와 B라는 개체 사이에 IsA 관계가 존재한다면 A가 B의 하위 종류이거나 예시임을 나타낸다. 따라서 모든 A는 B에 속하게 된다. HasA 관계라면 B가 A에 속한다는 것을 의미한다. 보통 HasA는 PartOf 관계의 반대이기도 하다[13].

<Figure 4>~<Figure 6>은 구현된 자동 확장 지식 그래프 기반 챗봇을 이용해 대화하는 예시를 보여준다. <Figure 4>는 'What is a car?(자동차는 무엇입니까?)'라는 질문에 대해 지식 그래프를 기반으로 검색하고 해당되는 답을 그래프로 나타내고 있고, <Figure 5>는 'What car has?(자동차는 무엇을 갖고 있습니까?)'라는 질문에 대한 답을 보여준다. <Figure 6>은 'What is selfie?(셀피는 무엇입니까?)'라는 질문에 대한 답을 보여준다.

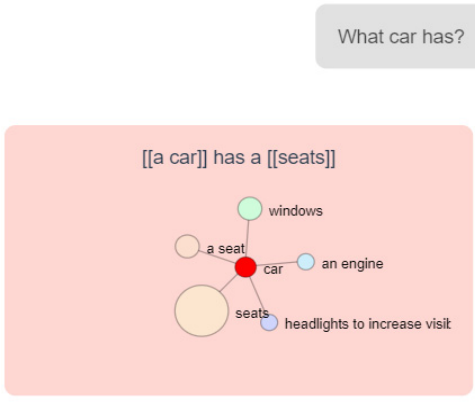
구현 결과로 보여준 <Figure 4>와 <Figure 5>에서는 각각 'car(자동차)'라는 키워드에 대해 IsA 관계에 해당되는 단어들과 HasA 관계에 해당되는 단어들을 보여주고 있다. <Figure 4>는 자동차와 IsA 관계에 있는 단어들에 대해 질문했기 때문에 motor vehicle(자동차), machine(기계), vehicle(차량) 등의 결과를 가중치

에 따라 노드의 크기를 다르게 하여 보여준다. 가장 높은 가중치를 갖고 있는 답인 motor vehicle은 문장으로도 답변한다. <Figure 5>는 자동차와 HasA 관계에 있는 단어들인 seats(좌석), windows(창문), engine(엔진) 등을 결과로 보여준다.

<Figure 6>은 실제로 신조어에 잘 대응했는지 검증하기 위해 신조어를 이용해 질문하고

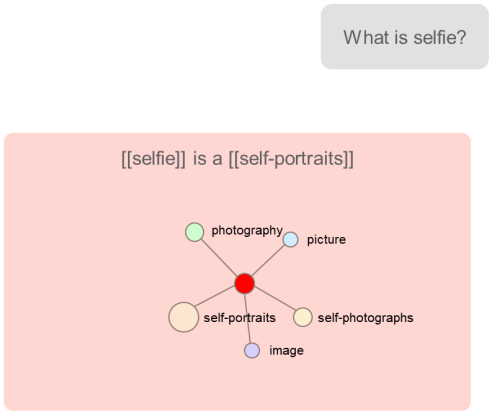


〈Figure 4〉 Conversation Example using Chatbot (Question: What is a car?)



〈Figure 5〉 Conversation Example using Chatbot (Question: What car has?)

답변 결과를 확인한 결과이다. 'selfie(셀피)'라는 키워드에 대해 IsA 관계에 해당되는 단어들을 보여준다. 셀피와 IsA 관계에 있는 단어들은 self-portraits(자화상), self-photographs(본인 사진, 셀카), photography 등의 결과를 가중치에 따라 보여주고, 가장 높은 결과인 self-portraits를 문장으로 답변하게 한다.



〈Figure 6〉 Conversation Example using Chatbot (Question: What is selfie?)

자동 확장 지식 그래프 기반 지능형 챗봇은 지식 그래프를 기반으로 하고 있음을 보여주고, 사용자에게 하나의 답보다 여러 개의 답을 보여주기 위해 결과를 그래프 형태로 나타낸다. 문장을 통해 결과를 보여주는 것뿐만 아니라 그래프로 시각화함으로써 보다 직관적으로 결과 확인을 가능하게 한다.

5. 결 론

인공지능 기술이 활발하게 연구되고 활용되고 있는 지금, 컴퓨터에게 인간의 상식을 학습

시키기 위한 기술은 매우 중요한 기술이다. 인간의 상식이 반영된 단어의 관계를 그래프로 표현하고 이를 기반 데이터로 활용하게 되면 컴퓨터에게 쉽게 인간의 상식을 학습시킬 수 있다.

본 논문에서는 자동으로 확장되는 BERT 기반의 지식 그래프를 구축하고 적용한 지능형 챗봇을 제안한다. 자동 확장 지식 그래프가 기존 지식 그래프와 달리 언어 종속성이 없고 신조어에도 대응 가능하며 다양한 활용성이 있다는 것을 실험과 구현을 통해 검증했다.

References

[1] Alsubaiee, S., Altowim, Y., Altwaijry, H., Behm, A., Borkar, V., Bu, Y., and Gabrielova, E., "AsterixDB: A scalable, open source BDMS," Proceedings of the VLDB Endowment, Vol.7, No.14, pp. 1905-1916, 2014.

[2] Athreya, R. G., Ngonga Ngomo, A. C., and Usbeck, R., "Enhancing Community Interactions with Data-Driven Chatbots-The DBpedia Chatbot," In Companion of the The Web Conference 2018, pp. 143-146, 2018.

[3] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[4] Fellbaum, C., "WordNet: An Electronic Lexical Database," Cambridge, MA: MIT Press, 1998.

[5] Hyun, Y. J. and Kim, N. G., "Text Mining-based Fake News Detection Using News And Social Media Data," The Journal of Society for e-Business Studies, Vol.23, No.4, pp. 19-39, 2018.

[6] Ji, G., He, S., Xu, L., Liu, K., and Zhao, J., "Knowledge graph embedding via dynamic mapping matrix," In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol.1, pp. 687-696, 2015.

[7] Lee, D. H. and Kim, K. H., "Web Site Keyword Selection Method by Considering Semantic Similarity Based on Word2Vec," The Journal of Society for e-Business Studies, Vol.23, No.2, pp. 83-96, 2018.

[8] Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X., "Learning entity and relation embeddings for knowledge graph completion," In Twenty-ninth AAAI conference on artificial intelligence, 2015.

[9] Mahdisoltani, F., Biega, J., Suchanek, F. M., "YAGO3: A Knowledge Base from Multilingual Wikipedias," Conference on Innovative Data Systems Research (CIDR), 2015.

[10] Paulheim, H., "Knowledge graph refinement: A survey of approaches and evaluation method," Semantic web, Vol.8, No.3, pp. 489-508, 2017.

[11] Peters, M. E., Neumann, M., Iyyer, M.,

Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., "Deep contextualized word representations," arXiv preprint arXiv:1802.05365, 2018.

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I., "Language models are unsupervised multitask learners," OpenAI Blog, Vol.1, No.8, 2019.

[13] Speer, R., Chin, J., and Havasi, C., "Conceptnet 5.5: An open multilingual graph of general knowledge," In Thirty-First AAAI Conference on Artificial Intelligence, Feb. 2017.

[14] Tarau, P. and Figa, E., "Knowledge-based conversational agents and virtual storytelling," In Proceedings of the 2004 ACM symposium on Applied computing, pp. 39-44, 2004.

[15] Wu, W., Li, H., Wang, H., and Zhu, K. Q., "Probase: A probabilistic taxonomy for text understanding," In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 481-492, 2012.

[16] Yang, Y. J., Lee, B. H., Kim, J. S., and Lee, K. Y., "Development of An Automatic Classification System for Game Reviews Based on Word Embedding," The Journal of Society for e-Business Studies, Vol.24, No.2, pp. 1-14, 2019.

[17] Yoo, S., Song, J., and Jeong, O., "Social media contents based sentiment analysis and prediction system," Expert Systems with Applications, Vol.105, pp. 102-111, 2018.

[18] Zhang, Y., Peng, Q., and Christopher D. M., "Graph Convolution over Pruned Dependency Trees Improves Relation Extraction," In Proceeding of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2205-2215, 2018.

[19] Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D., "Position-aware attention and supervised data improve slot filling," In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 35-45, 2017.

저 자 소 개

유소엽 (E-mail: bbusso@gc.gachon.ac.kr)
2014년 가천대학교 소프트웨어설계경영학과 (학사)
2016년 가천대학교 일반대학원 소프트웨어설계경영학과 (석사)
2018년~현재 가천대학교 일반대학원 IT융합공학과 소프트웨어학 전공 박사과정
관심분야 빅데이터, 소셜미디어 마이닝, 머신러닝, 딥러닝

정옥란 (E-mail: orjeong@gachon.ac.kr)
2005년 이화여자대학교 컴퓨터공학과 (공학박사)
2005년~2006년 서울대학교 컴퓨터공학부 박사후 연구원
2007년 Univ. of Illinois of Urban Champaign (visiting scholar)
2008년~2009년 성균관대학교 정보통신공학부 연구교수
2009년~현재 가천대학교 소프트웨어학과 부교수
관심분야 빅데이터, 소셜미디어 마이닝, 머신러닝