

챗봇의 의도 예문 자동 입력을 위한 Text-CNN 기반 의도 분류 방법

박 필 원*

Text-CNN Based Intent Classification Method for Automatic Input of Intent Sentences in Chatbot

Pill-Won Park*

요 약

본 논문에서는 기존 챗봇 프레임워크의 추론율을 높이기 위해 주어진 인텐트에 대한 예문을 Text-CNN을 이용하여 자동으로 분류해서 생성하는 방법을 제안하였다. 의도 분류시스템은 전처리 과정을 거쳐서 준비된 각 문장들의 단어들에 대한 워드벡터, 위치벡터들로 구성된 데이터를 Text-CNN을 이용하여 학습한다. 제안된 Text-CNN 구조는 컨볼루션 층, 맥스 풀링 층을 가지며 완전연결 소프트 맥스 층을 출력으로 갖는다. 또한, 드롭아웃(dropout)을 적용하여 정규화를 수행한다.

본 논문은 국립휴양림 사이트의 질의응답 및 휴양림 후기를 웹스크래핑을 이용하여 총 9000건의 문장을 수집하여 실험하였다. 취사안내 의도에 대한 Text-CNN 학습 결과로 얻은 정확도는 약 94%였다. Text-CNN 학습 결과로 생성한 모델로 레이블링되지 않은 나머지 문장을 분류하였으며 총 63건의 취사 관련 예문을 추출하였으며 이를 챗봇의 의도 예문으로 추가하였다.

Abstract

In this paper, we propose how to automatically categorize and generate examples of given intents using Text-CNN in order to increase the inference rate of the existing chatbot framework. The Intent Classification System uses Text-CNN to learn data consisting of word vectors and position vectors for each prepared sentence through the preprocessing process. The proposed Text-CNN structure has a construction layer, a max pooling layer, and a fully connected soft max layer as its output. In addition, dropout is applied to perform regularization.

For the experiment, a total of 9,000 sentences were collected using web scraping. An experiment showed that the accuracy obtained from Text-CNN's learning of kitchen intents was about 94%. The rest of the sentences, not labeled with the model produced by Text-CNN, were grouped and 63 cases of cooking were extracted in total and the sentences were input in Chatbot.

Keywords

chatbot framework, natural language processing, intent classification, CNN, increase the inference rate, text

* 동국대학교 컴퓨터공학과 전문연구원
- ORCID: <https://orcid.org/0000-0003-4693-6042>

• Received: Dec. 23, 2019, Revised: Jan. 21, 2020, Accepted: Jan. 24, 2020
• Corresponding Author: Pill-won Park
Dept. of Computer Science & Engineering, Dongguk University, 30,
Pildong-ro 1-gil, Jung-gu, Seoul, Republic of Korea
Tel.: +82-2-3899-3203, Email: pillwon79@gmail.com

1. 서 론

챗봇(Chatbot)은 채팅과 로봇의 합성어로 로봇의 인공지능을 대화형 인터페이스에 접목한 기술로 인공지능을 기반으로 사람과 상호작용하는 대화형 시스템을 지칭한다. 챗봇은 1960년부터 개발된 기술이나 인공지능 기술의 발전으로 머신러닝(Machine learning) 기술을 기반으로 사용자의 질문을 이해하고 의도를 파악하여 답변을 할 수 있게 되면서 복잡한 서비스를 제공하는 것이 가능해졌다. 또한, 스마트폰의 보급과 모바일 메신저 사용자가 급증하면서 메신저 앱을 사용하여 챗봇 서비스를 제공하는 사례도 증가하였다[1].

챗봇에서 이용되는 중요한 개념으로 자연어 이해(NLU, Natural Language Understanding), 의도(Intent), 엔티티(Entity, 개체) 등이 있다. NLU는 챗봇이 사람이 사용하는 자연어를 이해하는 것을 말한다. 복잡한 사람의 언어를 챗봇이 이해하기 위해서는 문장을 단어로 분해해서 분류하고 의도를 파악해야 한다. 챗봇은 자연어 이해 기술을 이용해 입력된 문장의 문법 구조를 파악하고 사용자의 의도와 개체를 분류한다.

란 사용자가 챗봇에게 전달하는 문장의 목적을 말한다. 챗봇은 사용자의 목적을 파악하고 이에 적합한 답변을 해야 한다. 예를 들어서, 호텔 예약 대행 챗봇에서 “제주도 호텔에는 뭐가 있니?”, “8월에 강원도 숙박 가능한 리조트 알려줘.”, “8월에 가족에게 가장 인기 있는 제주도 호텔은?”과 같은 사용자 문장들의 공통적인 목적인 “예약”이다. 즉, 이 문장들의 의도는 ‘예약’이며 챗봇은 예약에 관한 답변을 해야 한다. 엔티티는 문장에 포함된 핵심 단어를 말한다. 일반적으로 사용자의 문장에는 여러 핵심 단어들이 존재한다. 앞서 예로 들었던 사용자 문장 “8월에 강원도 숙박 가능한 리조트 알려줘.”에서 시기를 알 수 있는 “8월”, 위치를 알 수 있는 “강원도” 등이 있다. 의도가 예약이라면 “8월”, “강원도”는 시기와 위치를 알 수 있는 엔티티이다. 챗봇은 의도와 엔티티를 통해 사용자의 문장을 이해하고 이에 답변한다.

최근 개발된 챗봇 프레임워크들(단비[2], 클로버[3], 기가지니[4], 에이챗[5] 등)은 대부분 의도와 엔

티티를 기반으로 문장을 이해한다. 챗봇 프레임워크들에서는 의도 예문에 따라 대화 추론율이 달라질 수 있다. 추론율을 높이기 위해서는 의도에 관련된 예문이 충분히 주어지는 것이 중요하다. 이 논문에서는 Text-CNN(Convolution Neural Network)을 기반으로 한 의도 분류를 수행하여 챗봇 프레임워크에 의도 예문을 추가할 수 있는 방법을 제안한다. 이를 위해서 본 논문에서 제안하는 방법은 웹 스크래핑(Web scrapping)을 통해서 챗봇을 구축하는 영역에 관련된 질의응답, 후기 등의 문장들을 수집한다. 또한, 수집한 문장들을 Text-CNN 기반의 문장분류기를 통해서 의도별로 분류하여 챗봇 프레임워크에 입력한다.

논문의 구성은 다음과 같다. 2장에서는 본 논문에서 제안하는 방법과 관련된 기존 연구들을 조사 분석한다. 3장에서는 제안하는 방법에 대해서 기술하고 4장에서는 실험결과를 보여준다. 마지막으로 5장에서 결론을 맺는다.

II. CNN 기반 의도 자동분류를 이용한 챗봇의 추론율 개선기법

2.1 관련 연구

텍스트 분류는 자연어처리 분야에서 지속적으로 연구되어온 문제이며 기사 주제 분류, 스팸 필터링, 감정 분석 등에 활용된다.

챗봇 시스템은 음성(SIRI)이나 문자(페이스북 메신저 플랫폼에서 개발된 챗봇처럼)로 인간과 통신한다. 챗봇은 어떤 소통의 의미든 입력 텍스트를 이해해야 고객들에게 올바른 답을 제공할 수 있다. 챗봇 시스템에서 이 작업을 담당하는 구성 요소는 NLU라고 불리며, 여기에는 여러 가지 자연어 처리(NLP, Natural Language Processing) 기법이 통합되어 있다. 그림 1은 챗봇시스템의 가장 중요한 요소인 의도 분류시스템의 구조를 보여준다[6].

최근, 텍스트 분류를 위해서 심층학습(Deep learning) 방법을 적용하는 연구가 진행되고 있다[6]-[9]. 이 연구들은 RNN(Recurrent Neural Network), CNN, GCN(Graph Convolutional Network) 등의 심층 신경망(Deep neural network) 구조가 사용되어왔다.

그림 1은 [8]에서 제안한 Word2Vec을 활용한 CNN 구조이다. [8]에서 제안한 모델은 다양한 크기의 필터들을 여러 개 적용한 컨볼루션 층(Convolution layer)과 하나의 맥스 풀링 층(Max pooling layer)으로 이루어져 있으며 완전연결 층(Fully connected layer)이나 은닉 층(Hidden layer)이 없고 소프트맥스 출력 층(Softmax output layer)만을 가진다.

Word2Vec[10]은 신경망을 이용하여 단어를 벡터로 표현하는 방법을 제안하고 있다. 단어 벡터는 단어를 숫자로 이루어진 수십에서 수백 차원의 벡터로 변환한 것이다. 단어 벡터는 단순히 단어가 숫자로 변환된 것이 아니라 단어와 단어간의 벡터 공간 상에서의 거리를 표현한다. Word2Vec은 내부적으로 CBOW(Continuous Bag-Of-Words)와 스킵그램(Skip-Gram)방식을 이용하여 단어들의 벡터 표현을 학습한다.

CBOW 모델은 주어진 단어에 대해 앞 뒤로 단어를 검색하여 주어진 단어를 정확히 추론하기 위한 방식이며, Skip-gram 모델은 CBOW와는 반대 방향

으로 접근하는 모델로서 현재 주어진 단어 하나를 가지고 주위에 등장하는 나머지 몇 가지의 단어들의 등장 여부를 유추하는 것이다. 이 때 예측하는 단어들의 경우 현재 단어 주위에서 샘플링하는데, ‘가까이 위치해있는 단어일수록 현재 단어와 관련이 더 많은 단어일 것이다’라는 이론을 적용하기 위해 멀리 떨어져있는 단어일수록 낮은 확률로 택하는 방법을 사용한다.

[9][12][13]에서는 기존 Word2Vec를 이용한 CNN 텍스트 분류 방법에 Doc2Vec을 이용한 문서의 벡터 표현을 추가하여 분류 성능을 향상하는 방법을 제안하였다. 그림 2는 [9]에서 제안한 Word2Vec 및 Doc2Vec을 이용한 Text-CNN 텍스트 분류 방법의 구조도이다.

[6]에서는 RNN과 CNN을 이용한 텍스트 분류 방법의 단점을 다음과 같이 지적하고 있다. RNN은 시퀀스 데이터를 처리에 특화되어 있어 텍스트 분류를 할 때 텍스트의 단어들이 문장 내에서 거리가 먼 경우 관계추출이 어렵다.

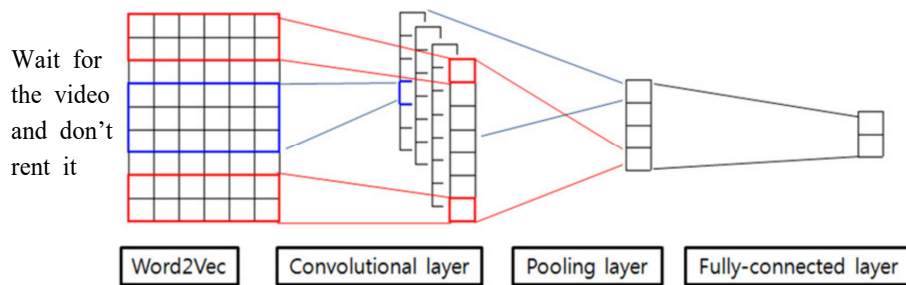


그림 1. Word2Vec 기반의 텍스트 분류를 위한 CNN구조[8]
Fig. 1. CNN architecture for text classification based on Word2Vec[8]

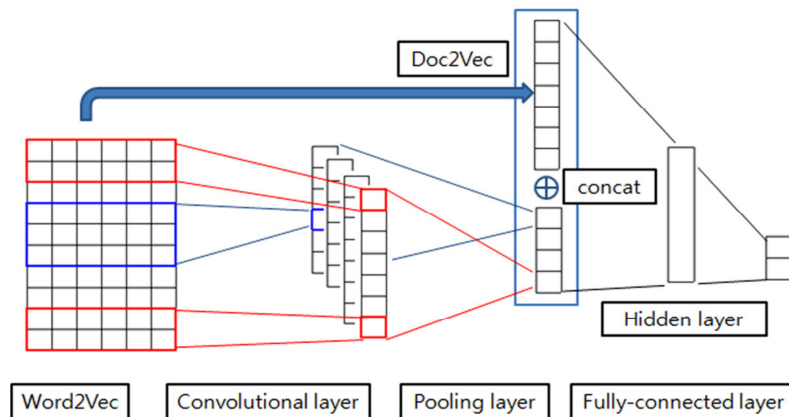


그림 2. Word2Vec 및 Doc2Vec을 이용한 텍스트 분류를 위한 CNN 구조[9]
Fig. 2. CNN architecture for text classification based on Word2Vec and Doc2Vec[9]

CNN 역시 그림 1이나 2에서처럼 문장 내의 단어 순서대로 입력을 만든다는 점에서 RNN과 유사한 문제를 가진다. [6]에서는 GCN을 이용하여 단일 문단을 해당 문단 내 단어만 포함하는 그래프로 인코딩한 후 GCN을 통해 텍스트를 분류하는 기법을 제시한다. 그래프 표현방법은 모든 단어 간의 관계를 표현할 수 있으므로 RNN 및 CNN의 단점을 해결할 수 있다고 주장한다.

이 논문에서는 기존 제안된 텍스트 분류 방법을 이용하여 챗봇의 추론율을 향상시키기 위한 의도 예문을 분류한다. 앞에서 기술한 방법들 외에도 다양한 심층 학습기법을 기반으로 한 다양한 텍스트 분류 방법이 존재한다. 이 논문에서는 기존 방법들 중에서 Word2Vec 기반의 CNN 분류기를 이용한다. 분류 대상이 되는 문장들이 문서형태로 존재하는 것이 아니고 간단한 질의/응답 형태의 문장이므로 Doc2Vec을 이용한 특징 추출은 필요가 없다. 또한, [6]에서 지적한 것처럼 문단내의 단어 관계를 표현하지 못하는 단점은 [11]에서 제안하는 위치 정보를 추가하여 해결한다.

2.2 챗봇의 추론율 개선기법

그림 3은 이 논문에서 개발하는 휴양림안내 챗봇을 위한 의도 분류기의 전체 구조이다. 휴양림안내 챗봇 의도 예문을 만들기 위해서 데이터 수집기는 먼저 휴양림과 관련된 기존의 질의응답, 사용자 후기 등의 데이터를 수집한다.

전처리기는 수집한 문장들 중 일부에 대하여 의문문으로 표현되는 문장들만 따로 추출하여 그림 4에서와 같이 형태소 분석을 수행하고 각 형태소에

대한 워드 임베딩을 수행한다. 워드 임베딩은 Word2Vec을 이용하여 수행한다. 그림 4에서 상단은 예문을 형태소 분석을 수행한 결과이고 하단은 형태소 분석된 결과를 워드 임베딩을 통해서 워드 벡터로 변환한 것이다. 또한 워드벡터의 특징값은 형태소들 간의 관계에 따라 부여된 숫자이다.

이어서 각 단어의 위치벡터를 추출한다. 각 단어 중 주어와 술어를 추출하고 모든 단어가 주어, 목적어, 술어로부터 얼마나 떨어져 있는지 계산한다. 그림 5에서처럼 “신분증”은 목적어, “보내주”는 술어이다. 주어는 생략되었다고 가정한다. 이때 “신분증”, “보내주”의 위치 벡터는 (주어에서의 위치, 목적어에서의 위치, 술어에서의 위치)로 각각 (-1,0,5), (-1,5,0)이 된다.

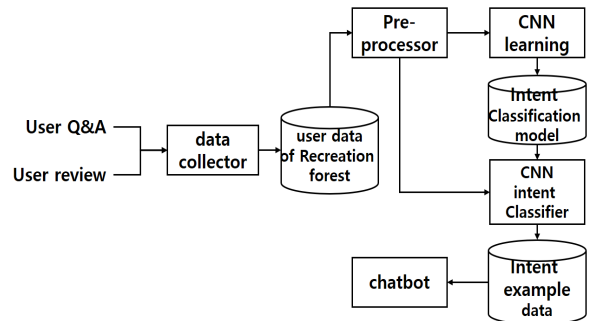


그림 3. Text-CNN 기반 의도 분류기 전체 구조
Fig. 3. Overall architecture of Text-CNN based intent classifier

Word	Word vector	Position vector
신분증	0.079, -0.071, ...	-1, 0, 5
...		
보내주	0.381, 0.1284, ...	-1, 5, 0

그림 5. 전처리 과정 - 워드 벡터, 위치 벡터
Fig. 5. Preprocessing - word vector, position vector

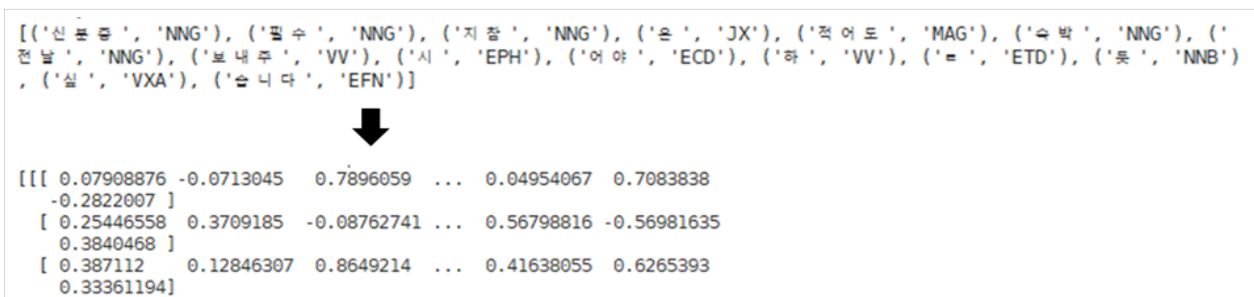


그림 4. 전처리 과정 - 형태소 분석 및 워드 임베딩
Fig. 4. Preprocessing - morpheme analysis and word embedding

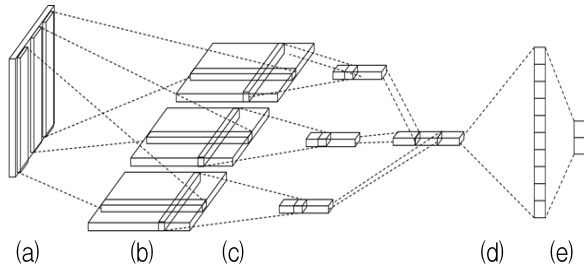


그림 6. 본 논문에서 사용한 Text-CNN구조 [9]
 Fig. 6. Text-CNN architecture used in the proposed method [9], (a) Input, (b) Convolution, (c) Mapping, (d), (e) Fully connected

CNN 학습기는 전처리 과정을 거쳐서 준비된 각 문장들의 단어들에 대한 워드벡터, 위치벡터들로 구성된 데이터를 [9]에서 제안한 Text-CNN을 이용하여 학습하고 분류한다. 그림 6에서 [9]의 Text-CNN 구조를 보여주고 있다. 그림에서처럼 제안하는 Text-CNN 구조는 컨볼루션 층, 맥스 풀링 층을 가지며 완전연결 소프트 맥스 층을 출력으로 갖는다. 또한, 드롭아웃(Dropout)을 적용하여 정규화를 수행한다.

Text-CNN을 이용한 의도 분류기는 학습한 의도 분류모델을 이용하여 수집한 전체 데이터에 대해서 분류를 수행하고 분류한 의도 별 예문을 휴양림 안내 챗봇에 업로드한다.

III. 실험 및 고찰

이 논문에서는 사이트의 질의응답 및 휴양림 후기를 웹스크래핑을 이용하여 총 9,000건의 문장을 수집하였다. 수집한 전체 문장 중 사용자의 질문이 드러나는 의문문으로 되어 있는 문장 3,249건을 추출하였다. 의문문은 형태소 분석을 통해 ‘?’ 부호가 있는 문장들을 대상으로 하였다. 그 중 중복되거나 무의미한 내용을 제외한 500건에 대해서는 직접 취사 안내와 그렇지 않은 나머지로 레이블링을 하였다. 레이블링된 문장들을 400건의 학습데이터와 100건의 테스트 데이터로 나누고 학습 및 테스트를 진행하였다. 레이블링된 각 문장들에 대해 3장에서 기술한 바와 같이 전처리 과정을 통해 단어별 (워드벡터, 위치벡터)를 생성하여 Text-CNN 학습을 수행하여 모델을 생성하였고 생성한 모델에 대해서 테스트를 수행하였다.

실험에 사용한 서버는 Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz, RAM 16G, TITAN Xp GPU, Samsung SSD 960 EVO 250GB의 사양을 가지며 Ubuntu 18.04.3 LTS 운영체제를 설치하여 사용하였다. 의도 분류 시스템은 Python 3.7.4, Keras 2.3.0, Tensorflow 2.0.0을 이용하여 개발하였다. 또한, 챗봇은 단비를 이용하여 개발하였으며 숙박안내, 취사안내, 액티비티안내 의도를 갖도록 하였다.

학습 및 분류에 사용한 Text-CNN구조는 128차원의 워드벡터와 3차원의 위치벡터를 합쳐서 입력으로 받는다. 또한, 히든(Hidden) 층에 3개의 컨볼루션 층이 Relu를 활성화함수로 하는 3개의 맥스 풀링 층과 연결된다. 출력 층은 완전연결 소프트 맥스 층으로 구성된다. 드롭아웃 비율은 0.5로 하였다.

취사안내 의도에 대한 Text-CNN 학습 결과로 얻은 모델의 정확도는 약 94%였다. Text-CNN 학습 결과로 생성한 모델을 이용하여 9,000건 중 학습에 사용된 3,249건을 제외한 나머지 5,751건에 대해서 분류를 수행하였으며 총 63건의 취사 관련 예문을 추출할 수 있었다. 그림 7은 2개의 취사안내 의도 예문들과 그렇지 않은 의도 예문을 보여준다.

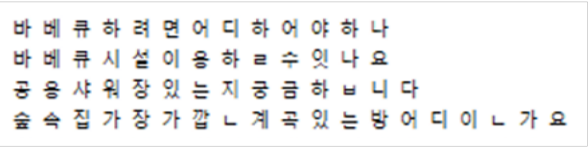


그림 7. 분류 결과 문장 예
 Fig. 7. Classified example sentences

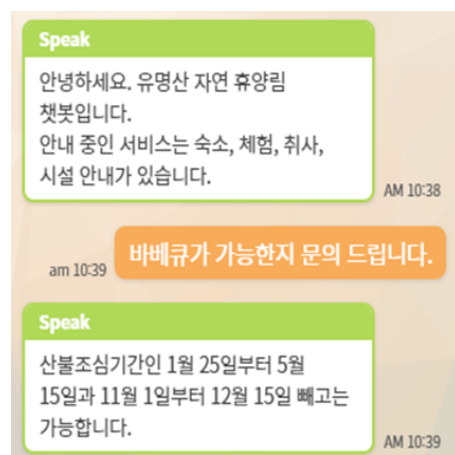


그림 8. 취사안내 의도에 대한 대화
 Fig. 8. Conversation about cooking guide intent

추출한 취사안내 의도 예문을 단비에 추가한 후 챗봇과 취사 관련 대화를 진행한 결과가 그림 8에 나타나 있다. 그림에서처럼 바비큐에 대한 질문에 대해 취사안내로 인지하고 취사 안내 대화를 이어 가는 모습을 볼 수 있다.

IV. 결 론

본 논문에서는 기존 챗봇 프레임워크의 추론율을 높이기 위해 주어진 의도에 대한 예문을 자동으로 분류해서 생성하는 방법을 제안하였다. 이를 위해서 본 논문에서는 휴양림 안내를 대상 영역으로 설정하고 웹 스크래핑을 통해서 데이터를 수집하였다. 수집한 문장들에 대해서는 워드 벡터와 위치 벡터를 추출하여 입력데이터를 구성하고 Text-CNN을 이용하여 학습시켰다. 이를 통해 생성한 모델을 이용하여 문장들을 분류하고 챗봇에 반영하였다.

챗봇은 단비를 이용하여 개발하였으며 숙박안내, 취사안내, 액티비티 안내 의도를 갖도록 하였다. 캠핑이 가능한 휴양림과 같은 상황에 대한 실험결과, 취사안내 의도에 대한 Text-CNN 학습 결과로 얻은 정확률은 약 94%로 나타났다. Text-CNN 학습 결과로 생성한 모델로 레이블링되지 않은 나머지 문장을 분류하였으며 총 63건의 취사 관련 예문을 추출하고 챗봇에 의도 예문으로 추가하였다. 추가된 예문을 기반으로 챗봇에 대화를 시도한 경우 이를 정확히 인식하는 것을 볼수 있었다.

References

[1] Byungrak Seo, "Introducing ChatBot Technology and Service Cases", AIRI Report, RInstitute of Intelligence and Information, 2017.7
 [2] <https://danbee.ai/>. [accessed: Dec. 31, 2019]
 [3] <https://clova.ai/ko>. [accessed: Dec. 31, 2019]
 [4] <https://apilink.kt.co.kr/api/menu/apiSvcDetail.do?sysId=GIGAGENIE>. [accessed: Dec. 31, 2019]
 [5] <https://www.aibril.com/web/solution/getMarketDetail.do?solSeq=2>. [accessed: Dec. 31, 2019]
 [6] Pham Quang Nhat Minh, "Three basic NLP

problems when one develops a chatbot system and some typical approaches", FTRI Report, 13/04/2017.
 [7] Minwoo Lee, Yanghoon Kim, and Kyomin Jung, "Text Classification via Sentence-level Graph Convolutional Networks", Journal of KIISE, Vol 25, No. 8, pp. 677-679, Aug. 2019.
 [8] Hyun-jung Park, Min-chaee Song, and Kyung-Shik Shin, "Sentiment Analysis of Korean Reviews Using CNN - Focusing on Morpheme Embedding", Journal of Intelligence and Information Systems, Vol. 24, No. 2, pp. 59-83, Jun. 2018.
 [9] Yoon Kim, "Convolutional Neural Network for Sentence Classification", Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), pp. 1746-1751, Oct. 2014.
 [10] Dowoo Kim and Myoung-Wan Koo, "Categorization of Korean News Articles Based on Convolutional Neural Network Using Doc2Vec and Word2Vec", Journal of KIISE, Vol. 44, No. 7, pp. 742-747, Jul. 2017.
 [11] Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification", arXiv preprint arXiv, 2016.
 [12] Y. Bengio, R. Ducharme, and P. Vincent, "Neural Probabilistic Language Model", Journal of Machine Learning Research, Vol. 3, pp. 1137-1155, Feb. 2013.
 [13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch", Journal of Machine Learning Research, Vol. 12, pp. 2493-2537, Aug. 2011.

저자소개

박 필 원 (Pill-Won Park)



2008년 2월 : 충남대학교

컴퓨터공학과 졸업(공학사)

2010년 9월 : 고려대학교 컴퓨터

전파통신공학과 졸업 (공학석사)

2017년 8월 : 고려대학교 컴퓨터

전파통신공학과 졸업 (공학박사)

2020년 1월 현재 : 동국대학교

컴퓨터공학과 전문연구원

관심분야 : 이동통신 차량통신