CNN과 LSTM 네트워크를 활용한 한국어 QA봇

Korean QAbot using CNN and LSTM Networks

저자 (Authors)
장성은, 이나영, 나병현, 김동호
Sung-eun Jang, NaYoung Lee, ByungHyun Na, Dongho Kim

출처 (Source)
한국정보과학회 학술발표논문집 , 2019.6, 913-915(3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08763371

APA Style
장성은, 이나영, 나병현, 김동호 (2019). CNN과 LSTM 네트워크를 활용한 한국어 QA봇. 한국정보과학회 학술발표논문집, 913-915

CNN과 LSTM 네트워크를 활용한 한국어 QA봇

장성은[O1] 이나영[1] 나병현[1] 김동호[2]
[1]동국대학교 컴퓨터공학과
[2]동국대학교 융합소프트웨어교육원
jse9512@dongguk.edu, lsyy201@naver.com, loversugar@hanmail.net, dongho.kim@dgu.edu

Korean QAbot using CNN and LSTM Networks

Sung-eun Jang[O1], NaYoung Lee[1], ByungHyun Na[1], Dongho Kim[2]
[1]Department of Computer Science and Engineering, Dongguk University
[2]Convergence Institute, Dongguk University

요 약

본 논문에서는 대학생들을 위한 학과 관련 질의응답에 활용이 가능한 한국어 QA봇을 설계하고 SNS 서비스의 일종인 LINE과 연동하여 제공될 수 있도록 구현하였다. 시스템은 CNN 모델을 사용한 한국어 임베딩, LSTM 모델을 통한 문장 의도 분류, 키워드 추출의 단계를 거쳐 사용자의 질문에 대한 답변을 제공하며 각 모델은 자체적으로 제작한 데이터셋을 사용하여 훈련되었다. 구현된 시스템에 대하여 질문의 의도 및 키워드가 올바르게 추출되는 정도를 실험을 통해 확인하였으며 각각 98.5%, 73.3%의 정확도를 보였으며, 의도와 키워드가 모두 올바르게 추출되어 적합한 대답을 하는 정도는 71.3%로 확인되었다.

1. 서론

상담 시스템은 직접 방문에서 전화, 인터넷 게시판과 모바일 앱과 같은 다양한 플랫폼으로 이동하고 있다. 이 변화는 기존의 상담 시스템의 문제를 해결하기 위해 계속 새로운 시스템을 고안한 결과이다. 직접 방문과 전화의 경우 질문자, 답변자 모두에게 높은 시간적 비용과 적절한 응대 방법에 대한 문제가 있다. 인터넷 게시판과 모바일 앱은 상담을 위해 별도의 게시판을 만들어야 하고, 질문자는 모바일 앱에 접근하기 위해 사용법 숙지가 필요하다.

위의 문제들을 해결하기 위해 최근에는 인공지능 분야의 다양한 자연어처리기법을 활용한 질의응답(question answering) 시스템, 일명 QA봇이 등장하고 있다. QA봇은 사용법이 친숙하고 간단하며 시공간적 제약 없이 상담을 제공할 수 있는 시스템이기 때문이다. 이미 페이스북, 텐센트, 구글, 네이버와 같이 고객과 접촉이 많은 국내외의 대다수 IT 기업들은 상담 시스템에 적용되고 있다.[1]

본 논문에서는 딥러닝을 중심으로 다양한 자연어 처리 기법을 활용하여 대학생들을 위한 학과 관련 질의응답에 활용이 가능한 한국어 QA봇을 구현하고 접근성 높은 환경을 구성하기 위해 SNS 서비스의 일종인 LINE과 연동하여 제공될 수 있도록 하였다.[2] 구현한 QA 봇은 자체적으로 제작한 데이터셋을 사용하여 훈련되었으며 문장 의도 분류와 키워드 추출이 성공적으로 이루어지는 것을 실험을 통해 확인하였다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었음 (2016-0-00171)

2. 관련 연구

질의응답 시스템은 사람의 질문에 자동으로 응답하는 시스템을 말하며, 정보 검색 및 개체명 추출의 분야에서도 사용된다.[3] 이는 자연어 처리 연구의 초기 단계부터 진행되었으며, 인간 지능을 모방한 자연어 이해 분야에서 주로 연구되었다.[4]

초기에는 상호참조 분석, 파싱, 품사 태깅 등 다양한 자연어처리 기법을 함께 사용하여 그 구조가 복잡하였으나, 최근에는 많은 양의 데이터를 통해 학습시킨 딥러닝 모델이 적용되어 보다 간단한 구조의 시스템을 만드는 것이 가능해졌다. 이러한 시스템은 보다 길고 일반적인 질문에도 대응할 수 있다는 장점을 지닌다.[3]

이때 사용될 수 있는 딥러닝 모델은 다양하다. CNN(Convolutional Neural Networks) 모델은 일반적으로 이미지 데이터의 특징을 추출하여 처리하는 데에 주로 사용되었다. 이를 응용하여 최근에는 문장을 구성하는 단어나 말뭉치로부터 특징을 추출하여 분산 표상(distributed representation)을 구성할 때 사용될 수 있다. 자연어 문장과 같이 연속적인 데이터를 분석하기 위해 고안된 RNN(Recurrent Neural Networks)의 경우 언어 모델링, 기계 번역, 음성인식 등 다양한 자연어 처리분야에 사용되어 뛰어난 성능을 보이고 있다.[5]

3. 모델 구조

구현된 QA봇은 그림 1에서 볼 수 있듯이 한국어 임베딩, 문장 의도 분류, 키워드 추출의 단계를 거쳐 사용자의 질문에 대한 답변을 제공한다. 이때 단어 임베딩과 문장 의도 분류 해당하는 모델은 최근 자연어처리 분야에서 최첨단의 성능을 내고 있는 다양한 딥러닝 모델을 활용하였다.[5]

또한 사용자의 질문에 존재할 수 있는 철자 및 띄어쓰기 오타에 대처할 수 있도록 전처리 과정을 거치게 되며, 사전에 학습되지 않은 단어(OOV:Out of Vocabulary)를 처리할 수 있도록 설계되었다.

그림 1 모델 구조

3.1 한국어 임베딩

자연어 처리를 위해서는 먼저 문장을 일정한 기준으로 토크나이징하고 각각을 벡터와 같은 연속된 차원으로 임베딩 할 필요가 있다. 한국어 토크나이징의 경우 교착어인 한국어의 특성에 알맞은 다양한 토크나이저(tokenizer)가 존재한다. 그중 비지도 학습 기반 토크나이저인 max score tokenizer의 경우 띄어쓰기가 제대로 이루어지지 않은 문장에 대한 토크나이징이 가능하다는 장점을 가진다. 이를 모델의 토크나이저로 설정하여 사용자의 띄어쓰기 오타를 직접 수정할 수 있도록 설계하였다.

임베딩 모델의 경우, 딥러닝 모델의 일종인 CNN을 기반으로 하는 char-word 임베딩 모델을 사용하여 사전에 학습되지 않은 단어를 처리할 수 있도록 설계하였다. 이렇게 구현된 임베딩 모델은 사용자의 질문을 입력으로 받아 그에 해당하는 고정 길이의 임베딩 벡터를 반환하게 된다.[6]

레이블	질문 유형	키워드
0	수강 학기	강의명
1	강의 정보	강의명
2	교수님의 강의 정보	교수명, 강의명
3	교수님 진행 강의	교수명
4	강의 진행 교수님	강의명
5	교수님 분야	교수명
6	교수님 특성	교수명

표 1. 레이블 구성

3.2 문장 의도 분류

다양한 사용자가 이용하는 QA봇은 다양한 유형과 길이의 질문에 대해서도 분류가 가능한 모델이 필요하다. 이를 위해 문장과 같이 다양한 길이를 가지는 순차적인 데이터를 처리하기 위해 고안된 RNN 모델의 사용을 고려하였으며, 그중에서도 LSTM(Long Short-Term Memory) 네트워크 모델을 사용하였다.[5]

다만 목표하는 QA봇은 소수의 레이블에 대한 답변을 제공한다. 이렇게 레이블의 수가 적은 경우 단순한 모델로도 목표 달성이 가능하다고 판단하여 단순하게 모델을 구성하였다. 구현된 분류 모델은 임베딩 모델을 통해 얻어진 임베딩 벡터를 입력으로 받아 문장이 속한 레이블을 출력하게 된다.

3.3 키워드 추출

분류기를 통해 획득한 레이블을 기반으로 키워드를 추출하여 실제 사용자에게 제공할 답변을 구성하게 된다. 이때 레이블에 따른 키워드의 종류는 표 1의 내용과 같다. 강의명과 교수명 키워드의 수는 각각 88개, 48개로 총 136개의 키워드에 대한 추출이 가능하다.

키워드 추출 모델은 사전에 정의된 키워드의 목록과 사용자 질문 내의 단어를 비교하여 문장에 포함된 키워드를 검색한다. 이때 검색에 사용되는 문장은 원본 문장과 철자 및 띄어쓰기 오타를 수정한 문장이 포함된다. 이 둘을 함께 확인함으로써 사용자가 입력한 문장에 오타가 있을 경우에도 올바른 답변을 제공할 수 있도록 하였다

띄어쓰기 오타의 경우 학습된 토크나이저를 통해 수정이 가능하도록 설계되었다. 철자 오타의 경우 문장에 존재하는 OOV와 사전에 구성한 단어 목록에 있는 단어의 Levenshtein distance를 계산하고, 일정 이상의 유사도를 보이는 단어로 OOV를 수정하는 방식으로 수정된다.

3.4 답변 제공

문장 의도 분류 모델과 키워드 추출 모델을 통해 획득한 레이블과 키워드를 기반으로 사전에 작성된 답변 중 올바른 답변을 골라 반환하게 된다.

4. 실험 및 결과

4.1 데이터셋

문장 의도 분류 모델과 임베딩 모델의 성공적인 학습과 테스트를 위해선 양질의 한국어 질의응답 데이터가 필요하다. 그러나 공개 데이터셋 중에는 적합한 한국어 질의응답 데이터를 구하기 어려우며, 설계된 QA봇은 좁고 특정한 분야를 다루도록 계획되었기 때문에 자체적으로 데이터셋을 제작하여 사용하였다.

데이터셋은 동국대학교의 교육과정과 교내 커뮤니티에 학생들이 작성한 내용을 기반으로 컴퓨터공학과, 멀티미디어공학과, 정보통신공학과의 커리큘럼에 포함된 강의들과 해당 강의를 진행하는 교수님에 대한 질문과 답변 데이터를 구성하였다.

항목	전체	학습	테스트
문장 수	63,000	44,100	18,900
평균문장길이 (글자 수)	34	34	34
최대문장길이 (글자 수)	128	128	108
최소문장길이 (글자 수)	8	8	8

표 2. 데이터셋 구성

4.2 학습

학습은 한국어 임베딩 모델과 문장 의도 분류 모델의 두 가지에 대해 이루어지며, 44,100개의 질문 데이터를 사용하여 이루어 졌다. 구현언어는 파이썬(Python)을 사용하였다.

한국어 임베딩 모델을 구성하는 모델의 경우 깃헙에 공개되어 있는 Kor2Vec와 soyNLP 패키지를 사용하여 구현되었으며 사용자의 질문을 128 x 1 차원을 가지는 20개의 벡터로 반환하도록 구성하였다.

문장 의도 분류 모델을 구성하는 LSTM 모델의 경우 PyTorch 라이브러리를 사용하여 구현되었으며 히든 레이어의 크기는 64로 설정되었다. 미니 배치의 크기는 315로 설정하였으며 10번의 에포크(epoch)를 거쳐 훈련되었다.

4.3 실험

'문장 의도 분류'와 '키워드 추출'이 올바르게 이루어지는지를 실험을 통해 확인하였다. 8,900개의 질문 데이터를 사용하였으며 각각 문장에 해당하는 올바른 레이블이 반환되었는지, 주어진 질문에서 두 가지의 키워드(강의명, 교수명)을 올바르게 추출해 내는지를 확인하였다. 또한 의도와 키워드가 모두 올바르게 추출되어 적절한 답변을 제공할 수 있는지 또한 실험을 통해 확인하였다.

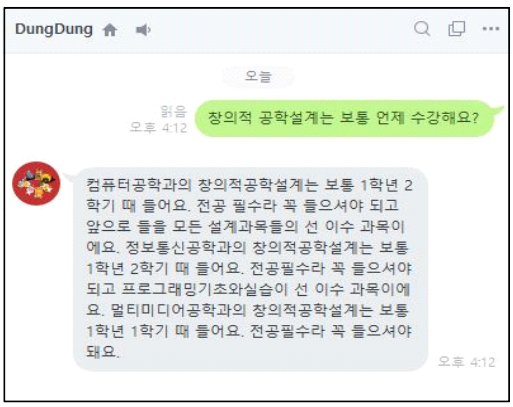

그림 2 실행화면

4.4 결과

실험 결과는 표3의 내용과 같으며 실제 LINE과 연동한 실행화면은 그림 2와 같다.

항목	정확도 (%)
키워드	98.5%
레이블	73.3%
키워드 & 레이블	71.3%

표 3. 실험 결과

5. 결론

본 논문에서는 대학생들을 위한 학과 관련 질의응답에 활용이 가능한 한국어 QA봇을 설계하고 SNS 서비스의 일종인 LINE과 연동하여 제공될 수 있도록 구현하였다. 시스템은 CNN 모델을 사용한 한국어 임베딩, LSTM 모델을 통한 문장 의도 분류, 키워드 추출의 단계를 거쳐 사용자의 질문에 대한 답변을 제공하며 각 모델은 자체적으로 제작한 데이터셋을 사용하여 훈련되었다. 구현된 시스템은 7가지 질문유형과 136개의 키워드에 대응할 수 있다. 이에 대해 질문의 의도 및 키워드가 올바르게 추출되는 정도를 실험을 통해 확인한 결과 각각 98.5%, 73.3%의 정확도를 보였으며, 의도와 키워드가 모두 올바르게 추출되어 적합한 대답을 하는 정도는 71.3%로 확인되었다.

전체적인 시스템의 정확도의 향상을 위해서는 문장 의도 분류 모델의 개선이 필요할 것으로 예상된다. 또한 양질의 데이터셋을 학습에 사용함으로써 보다 다양한 유형의 질문에 대응이 가능할 것으로 예상된다.

참 고 문 헌

[1] 윤상오. "인공지능 기반 공공서비스의 주요 쟁점에 관한 연구." 한국공공관리학보 32.2, 83-104, 2018.

[2] http://aidev.co.kr/chatbotdev/1733

[3] Deng, Li, and Yang Liu, eds. "Deep Learning in Natural Language Processing." 930-933 Springer, 2018.

[4] Andrenucci, Andrea, and Eriks Sneiders. "Automated question answering: Review of the main approaches." Third International Conference on Information Technology and Applications (ICITA'05). IEEE, 514-519, 2005.

[5] Young, Tom, et al. "Recent trends in deep learning based natural language processing." IEEE Computational Intelligence Magazine 13.3, 55-75, 2018

[6] Kim, Yoon, et al. "Character-aware neural language models." Thirtieth AAAI Conference on Artificial Intelligence, 2016