DoAjou(도아주) : 아주대학교 학생들을 위한 질의응답용 챗봇 개발

DoAjou : Developing Q&A Chatbot For Ajou Univ. Student

저자 (Authors)
김치헌, 박승현, 최순원, 손경아
Chiheon Kim, Seunghyun Park, Soonwon Choi, Gyeonga Sohn

출처 (Source)
한국정보과학회 학술발표논문집 , 2018.6, 2098-2100(3 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07503623

APA Style
김치헌, 박승현, 최순원, 손경아 (2018). DoAjou(도아주) : 아주대학교 학생들을 위한 질의응답용 챗봇 개발. 한국정보과학회 학술발표논문집, 2098-2100

DoAjou(도아주) : 아주대학교 학생들을 위한 질의응답용 챗봇 개발

김치헌, 박승현, 최순원, 손경아
kch21026@naver.com, shp9408@gmail.com, wony6731@naver.com, kasohn@ajou.ac.kr
아주대학교 소프트웨어학과

DoAjou : Developing Q&A Chatbot For Ajou Univ. Student.

Chiheon Kim, Seunghyun Park, Soonwon Choi, Gyeonga Sohn
Department of Software and Computer Engineering, Ajou University

요약

기존의 검색형 모델 기반 챗봇에서는 대부분 정보 제공을 위하여 사용자가 질문의 유형을 선택받도록 되어있다. 이는 사용자 입장에서 필요한 정보를 얻기 위해 매번 질문의 유형을 선택해야하는 번거로움이 있다. DoAjou 챗봇은 이러한 번거로움을 개선하고 아주대학교 학생들에게 학교 내 궁금한 정보를 제공한다. 본 논문은 DoAjou가 제공하는 정보 범위 내에서 사용자의 질문에 대한 답변을 출력하는 알고리즘을 제공한다. 우선 사용자의 입력을 Soynlp[1]를 통해 한국어 자연어 처리를 한 후 Sentence2Vec[2]으로 벡터화한다. 그 후, 저장되어 있는 데이터 셋(의도 질문 문장)과 비교하여 유사도(Cosine Similarity)가 가장 높은 의도 질문을 찾고 이에 대한 답변을 출력한다. 추가로 대화로그와 Slot filling[3] 기술을 구현하여 사용자의 질문을 기억할 수 있도록 하였다. 만약 유사도가 높은 의도 질문을 찾지 못하면 Seq2Seq[4] 기반의 생성모델을 통해 스스로 답변을 생성하도록 하였다. 이 후 구성된 모델을 Django[5]를 통해 카카오톡 플러스친구로 구축하여 사용자가 쉽게 사용할 수 있도록 하였다.

1. 서 론

본 논문에서 개발한 DoAjou는 아주대학교 학생들이 학교 생활을 하며 손쉽게 궁금증을 가지는 질문에 대해 답변을 해주는 챗봇이다. 사용자로부터 질문을 입력 받으면 자연어 처리 과정을 거친 후 만들어 놓은 데이터 셋에서 가장 유사도가 높은 의도 질문을 찾아 간다. 그리고 나서 찾아 간 의도 질문에 해당하는 답변을 출력하여 사용자에게 정보를 제공한다.

본 논문의 구성은 다음과 같다. 2장에서 시스템 구성을 살펴보고, 3장에서는 실험을 통해 시스템의 성능을 분석하고, 4장에서 결론 및 향후 연구에 대해서 논의한다.

2. 시스템 구성

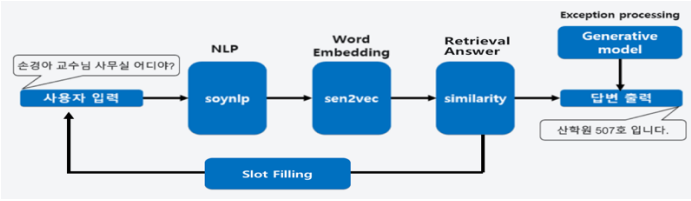

Figure 1. 시스템 구성도

DoAjou 는 검색 기반 모델(Retrieval-based model)과 생성 모델(Generative model) 2 종류의 모델을 사용한 챗봇이다. Figure 1 은 전체 시스템 구성을 간략히 나타낸 것이다. 사용자는 아주대학교 내 궁금한 정보를 챗봇에게 질문한다. 질문을 받은 챗봇은 입력문장을 Soynlp 토크나이저를 사용하여 토큰화하고, Sentence2Vec 을 사용하여 문장을 벡터화 한다. 벡터화를 마친 문장을 가지고 있는 의도 질문들과 유사도를 비교하여 유사도가 가장 높은 의도 질문을 찾는다. 그 후 유사도가 가장 높은 의도 질문과 쌍으로 이루어진 응답 문장을 답변으로 출력한다. 만약 질문에 필요한 정보들이 모두 담겨져 있지 않다면 Slot filling 을 통해 사용자로부터 필요한 정보를 얻는다. DoAjou 가 제공하는 정보 범위 밖의 질문에 대해서는 생성 모델을 통하여 답변한다.

2.1 데이터 준비

검색 기반 모델에서 사용되는 데이터는 Key-Value를 가지는 csv파일에 저장된다. Key에 저장된 데이터는 벡터화된 사용자의 질문과 유사도를 비교하기 위한 의도 문장들이다. Value에 저장된 데이터는 각각의 Key에 해당하는 응답 문장들이다. 이

[1] 한국어 정보처리를 위한 Python Package, https://github.com/lovit/soynlp
[2] Word2Vec을 기반으로 하는 Word Embedding 기법.
[3] Dialogflow, https://dialogflow.com/docs/how-tos/slot-filling
[4] RNN Encoder–Decoder, consists of two recurrent neural networks (RNN) that act as an encoder and a decoder pair.
[5] Opensource Web Application Framework.

데이터를 통해 사용자의 질의와 가장 유사도가 높은 의도 문장(key)을 찾고, 해당하는 응답 문장(Value)을 답변으로 내놓는다. 또한 하루마다 바뀌는 데이터의 경우는 crontab[6]을 이용하여 매일 Web Crawling[7]을 할 수 있게 했다. 즉, 매일 데이터베이스를 갱신하여 해당 문제를 해결하였다.

예외 처리를 담당하는 생성 모델에 사용되는 데이터는 질문-대답 134쌍의 데이터다. 이 데이터들은 아주대학교 학생들의 입장에서 할 수 있는 범위 밖 질문-대답으로 seq2seq 모델을 학습시키는데 사용되었다.

2.2 한국어 자연어 처리

Konlpy는 한국어 문장들을 단어로 토큰화하고, 각 단어의 형태소를 알려준다. Konlpy 내부에는 Kkma, Mecab, Twitter 등의 토크나이저가 있지만 그 중에 Twitter를 기반으로 한 Twitter 토크나이저가 자주 사용된다. 하지만 Konlpy에서는 고유명사의 토큰화에 문제가 생긴다. 아주대학교 정보로 교수 성함이 있는데, 성함은 고유명사로 Konlpy 토크나이저가 토큰화를 잘하지 못한다. 예를 들어, '이정태'에 대한 정보를 묻는 질의 문장이 들어오면 Konlpy는 '이정태'를 한 단어로써 인식하지 못하여 '이정', '태'로 토큰화한다. 이를 해결하기 위해 서울대학교에서 개발한 Soynlp를 사용하였다.

```
[('이정', 'Noun'),
 ('태', 'Noun'),
 ('교수', 'Noun'),
 ('님', 'Suffix'),
 ('연구실', 'Noun'),
 ('어디', 'Noun'),
 ('야', 'Josa'),
 ('?', 'Punctuation')]
```

['이정태', '교수님', '연구실', '어디야?']

<Konlpy> <Soynlp>

Figure 2. Konlpy와 Soynlp 비교

Soynlp는 기본적으로 띄어쓰기 기반으로 문장을 토큰화하지만 score라는 Argument를 받아 지정한 단어에 대해서는 띄어쓰기 없이 이어진 문장 내에서도 토큰화를 할 수 있다는 장점이 있다. Figure 2는 "이정태교수님 연구실 어디야?"라는 문장을 Konlpy와 Soynlp를 통해 토큰화한 결과다. Konlpy는 '이정태'를 '이정', '태'로 토큰화 하지만, Soynlp는 '이정태'라는 단어에 대해 score를 주면 '이정태'를 우선적으로 토큰화하여 결과적으로 '이정태'와 '교수님'으로 토큰화하게된다. 이러한 Soynlp의 장점을 통해 DoAjou 챗봇은 고유명사의 토큰화 문제를 해결 하였다.

2.3 Sentence2Vec

Sentence2Vec은 사용자의 질문을 벡터화한다. Soynlp로 토큰화된 문장의 단어들을 Gensim[8]의 Word2Vec[9]을 통해 벡터화하고 벡터들의 평균을 내어 전체 문장에 대한 벡터를 만든다.

Sentence2Vec을 통해 만들어진 벡터들은 해당하는 의도 질문의 벡터와 높은 유사도를 가진다. 이는 사용자의 질문이 해당하는 의도 질문을 찾아가 사용자가 원하는 답변을 따라 갈 수 있도록 해준다.

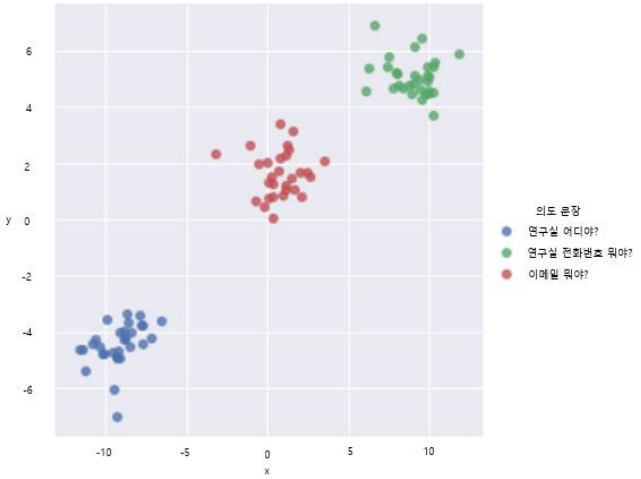

Figure 3. Sentence2Vec를 이용한 3가지 의도 시각화

Figure 3은 Sentence2vec을 통해 사용자의 질문들의 유사도를 시각화한 것이다. 사용자의 질문 중 교수의 연구실 위치를 물어보는 질문들은 그림에 파란 점들이다. 이는 같은 의도를 가진 질문들이 서로 높은 유사도를 가진다는 것을 알 수 있다.

2.4 답변 출력

Sentence2Vec을 통해 벡터화된 사용자의 질의는 기존에 저장된 모든 Key(의도 질문) 데이터와 유사도를 비교한다. 가장 높은 유사도를 가진 Key 데이터는 사용자의 질의에 형식화된 의도 질문이라 할 수 있다. 예를 들어, 사용자가 'OOO 교수님 연구실 전화번호 좀 가르쳐줘!'를 질문하면 Key 데이터들 중 최고의 벡터 유사도를 가진 'OOO 교수님 연구실 전화번호 뭐야?'를 찾게 된다. 유사도 비교를 통해 Key 데이터를 찾으면 Key에 해당하는 Value(응답 질문) 'OOO 교수님 연구실 전화번호는 xxx-xxx-xxxx입니다.'를 사용자에게 답변한다.

2.5 예외 처리

예외처리는 DoAjou가 다루지 않는 범위 밖 질문들에 대하여 진행하였다. 예를 들어, 사용자가 날씨를 묻거나, 안부를 묻는 질문을 입력할 시 기존에 저장해 놓은 의도 질문 중 유사

[6] 해당 시간에 자동으로 특정 명령어를 실행시키는 Linux 명령어
[7] 특정 월드 와이드 웹을 탐색하는 기법
[8] Word2vec Python package,
[9] 단어의 의미와 맥락을 고려하여 단어를 벡터로 표현하는 Word Embedding 기법.

도가 높은 질문을 찾기 힘들다. 이러한 경우 유사도에 기준 값을 두어 보다 낮은 유사도가 나올 시 생성 모델로 답변을 하도록 하였다.

DoAjou의 생성 모델은 Seq2Seq를 사용하여 아주대학교 학생이 할 수 있는 범위 밖 질문과 대답을 134쌍을 학습시켰다. 이 생성 모델은 범위 밖 정보들에 대해서 학습 데이터를 바탕으로 직접 답변을 생성하여 출력한다.

2.6 기억 기능

챗봇 개발 과정에서 흔히 겪는 문제는 챗봇이 이전의 질문을 기억하지 못한다는 점이다. 이 부분에 대해서 DoAjou는 Slot filling의 개념을 도입하였다. Slot filling이란 정보 제공을 위한 slot들을 두어 모든 slot에 값들이 채워지지 않으면 빈 slot을 채우기 위해 사용자에게 정보를 요구하는 개념이다. DoAjou는 사용자의 대화 로그들을 기억하여 slot이 모두 채워질 때까지 질문을 반복하고 slot이 다 채워지면 답변을 하게 된다.

3. 실험

3.1 실험 환경

본 논문에서 구현된 챗봇은 AWS EC2 ubuntu 16.04 환경에서 Django와 Python을 통해 카카오톡 플러스친구로 구현했다.

3.2 실험 결과 및 분석

Figure 4. 정보 제공

Figure 5. 예외 처리

본 논문의 챗봇은 아주대학교 학생들이 학교 생활을 하며 궁금증을 가지는 질문에 대해 답변을 해주는 챗봇이다. Figure 4와 Figure5를 통해 DoAjou의 우수한 성능을 볼 수 있다.

Figure 4는 DoAjou의 세 가지 핵심 기능을 보여준다. 정보 제공, slot filling을 통한 필요한 정보 되묻기, 대화로그 기억이다. 교수님 연구실 전화번호에 대해서는 2가지 slot(교수님 성함, 의도)으로 구성되어 있다. 질문을 통해 의도 slot은 채워졌으나 교수님 성함 slot은 채워지지 않아 되묻는 과정을 볼 수 있다. 한국어 자연어 처리에 대한 어려움은 Soynlp의 score기능을 이용함으로써 완벽하게 해결하였다. 또한, 대화 로그를 기억함으로써 좀 더 유연한 대화를 할 수 있도록 하였다.

Figure 5는 범위 밖 질문이 들어왔을 경우 생성 모델을 통해 답변을 출력한 모습이다. 비록 문법과 문맥이 맞지 않지만 말을 직접 생성하여 대답을 하고 있다.

제공하는 정보 범위 내에서는 정확한 정보를 전달하고, 범위 밖 정보에 대해서는 검색 기반 모델의 단점을 보완하기 위해 생성 모델을 사용하여 대화의 자유도를 높였다.

4. 결론 및 향후 연구

본 논문에서는 아주대학교 학생들을 위한 챗봇에 대하여 다루었다. 제공하는 정보 범위 내의 질문에 대하여 꽤나 정확한 답변을 보일 뿐만 아니라 범위 밖의 질문에 대해서는 생성 모델을 통하여 비록 문법과 문맥이 맞지 않더라도 답변을 생성하여 출력하는 것을 확인할 수 있다. 향후 연구는 생성 모델에 중점적으로 투자하여 좀 더 정확한 답변을 생성하여 출력하도록 할 것이며 정보 제공 범위 또한 확장할 계획이다.

DoAjou는 아주대학교 학생들을 대상으로 개발되었다. 하지만 사용된 시스템 구조와 알고리즘은 질의에 대해 응답을 하는 모든 상황에 적용될 수 있다. 또한 정보 제공 범위도 쉽게 추가가 가능하여 다방면에서 유연하게 사용되어 질 수 있다.

Acknowledgement

This research was supported by the MIST(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute for Information & communications Technology Promotion)" (2015-0-00908)

참 고 문 헌

[1] "Sentence2Vec : Evaluation of popular theories – Part 1(Simple average of word vectors)". sentence2Vec. https://medium.com/@premrajnarkhede/sentence2vec-evaluation-of-popular-theories-part-i-simple-average-of-word-vectors-3399f1183afe

[2] "파이썬 한국어 NLP". Konlpy. http://konlpy-ko.readthedocs.io/ko/v0.4.3/

[3] "한국어 자연어 처리를 위한 파이썬 라이브러리". Soynlp. https://github.com/lovit/soynlp

[4] "챗봇". Seq2Seq. https://github.com/golbin/TensorFlow-Tutorials/tree/master/10%20-%20RNN/ChatBot

[5] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[6] Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).