

교육환경에서의 기계독해 기반 질의응답 시스템

주준하⁰, 박상현, 남승완, 임경태

한밭대학교 컴퓨터공학과

{20197132, 20171612, 20171768}@edu.hanbat.ac.kr, ktlim@hanbat.ac.kr

Machine Reading Comprehension-based Q&A System in Educational Environment

Jun-Ha Ju⁰, Sang-Hyun Park, Seung-Wan Nam, Kyung-Tae Lim
Dept. of Computer Engineering, Hanbat National University

요 약

코로나19 이후로 교육의 형태가 오프라인에서 온라인으로 변화되었다. 하지만 온라인 강의 교육 서비스는 실시간 소통의 한계를 가지고 있다. 이러한 단점을 해결하기 위해 본 논문에서는 기계독해 기반 실시간 강의 질의응답 시스템을 제안한다. 본 논문연구에서는 질의응답 시스템을 만들기 위해 KorQuAD 1.0 학습 데이터를 활용해 BERT를 fine-tuning 했고 그 결과를 이용해 기계독해 기반 질의응답 시스템을 구축했다. 하지만 이렇게 구축된 챗봇은 강의 내용에 대한 질의응답에 최적화되어있지 않기 때문에 강의 내용 질의응답에 관한 문장형 데이터 셋을 구축하고 추가 학습을 수행하여 문제를 해결했다. 실험 결과 질의응답 표를 통해 문장형 답변에 대한 성능이 개선된 것을 확인할 수 있다.

주제어: 질의응답, 기계독해(MRC), BERT, 교육

1. 서론

코로나19 이후 교육의 형태가 변화되고 있다. 기존의 K-MOOC, KOCW 등 여러 대학의 강의를 제공하는 온라인 교육 서비스의 이용은 개인의 선택으로 이루어졌다. 하지만 현재 많은 교육기관이 대면 강의 대신 비대면 강의를 필수적으로 제공해야 하는 상황을 맞이했다. 비대면 온라인 강의의 만족도 조사를 따르면 과반수의 학생이 비대면 온라인 강의에 만족하는 것으로 나타난다. 그러나 실시간 소통 부족의 이유로 수업 수행에 대한 두려움과 스트레스 비율은 ‘다소 그렇다’가 상대적으로 높게 나타났다[1].

현재 교육 및 여러 산업에서 많이 활용되는 챗봇은 특정한 주제만 다루는 closed domain 챗봇으로, 항상 정확한 정보를 일관되게 전달해야 하므로 대화 시나리오와 사용된 단어를 기반으로 정해진 규칙에 따라서 미리 준비한 답변을 제공하는 방법을 많이 채택한다. 이러한 규칙 기반 챗봇은 구현이 간단하고 정해진 답변을 하므로 안정적이지만 정의된 규칙에 따라 답변하기 때문에 비슷한 질문을 이해하지 못하거나 예상치 못한 질문은 답변하지 못하는 경우가 빈번하다. 이러한 문제점을 해결하기 위해 closed domain에 NLP와 기계학습을 적용한 AI 챗봇이 연구되고 있다[2]. 그러나, closed domain에 활용하기 위해 방대한 학습 데이터가 필요하고 답변의 정확성을 보장하진 않기 때문에 보완할 점이 많다.

본 연구는 강의자가 제공하는 교안을 효과적으로 검색 및 분석하여 수강자에게 제공하는 기술을 확보하는 데 그 목적이 있다. 예를 들어, 사용자가 Node.js 강의 내용 중, ‘에러처리 미들웨어임을 표시하는 매개변수는 무엇인가?’ 라고 질문을 했을 때, 제공된 강의자의 교안

에서 ‘(에러처리 미들웨어에 관련된 내용)’을 찾은 다음, 정답에 해당하는 부분인 ‘next’를 반환한다. 또는, ‘REST API가 무엇인가?’라고 질문을 했을 때 이에 해당하는 답변을 반환한다. 이를 위해 다음 두 가지 시스템이 구축되어야 한다. 첫 번째로, 강의자가 제공한 교안을 읽어야 한다. 우리는 대부분 강의자료가 PDF인 것을 고려해 PDF 문서를 텍스트로 변환했다. 두 번째로, 텍스트의 내용을 검색하고 질문에 해당하는 대답이 있는 위치를 반환하는 기계독해 시스템이 필요하다.

우리는 기계독해 시스템 구축을 위해 BERT 언어모델 [3]을 KorQuAD 1.0[4] 데이터에 맞게 fine-tuning 했다. 하지만, KorQuAD 1.0 학습 데이터는 강의 내용 질의응답 용도로 만들어진 데이터가 아니기 때문에 주로 짧은 답변을 선택하는 문제가 있었다. 해당 문제를 해결하기 위해 우리는 298쌍의 강의 내용과 관련된 질의응답 데이터를 제작했으며 해당 데이터는 답변의 평균 길이가 KorQuAD 1.0 보다 2.7배 길기 때문에 강의 질의응답에 조금 더 적절했다. 우리는 강의에 관련된 답변의 정확도를 높이기 위해서 자체 제작한 298쌍의 학습 데이터를 이용해 추가로 모델을 fine-tuning 했다. 우리는 실험 결과를 통해 우리가 제안한 학습 데이터를 활용했을 경우 9%정도의 exact matching 성능 향상을 보이는 걸 확인했다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고, 3장에서는 데이터의 전처리 및 질의응답 시스템의 전체적인 구조에 대해 상세히 설명한다. 4장에서는 제안모형의 성능 평가 결과를 분석하고, 마지막으로 5장에서 향후 연구 방향에 대해 논한다.

2. 관련 연구

2.1 비대면 온라인 강의 인식 조사

[1]에서, 2020학년도 1학기에 비대면 온라인 강의를 수행한 대학생들의 수업 참여도와 만족도, 선호도, 수업의 효과, 수행 환경 등을 조사한 결과를 기반으로 비대면 교육의 나아갈 방향성을 모색하고자 하였다. 온라인 강의는 온라인으로 수업 자료와 영상이 제공되기 때문에 반복 학습이 수월해지며, 복습이 기존 대면 수업의 필기와는 달리 언제든지 원하는 수업 내용을 재생하여 볼 수 있어 효과가 탁월하다. 비대면 온라인 강의에 대한 부정적 반응 요인 간의 관련성에서는 시험과 과제 수행, 충분하지 않은 설명, 실시간 소통 부족의 이유가 두려움과 스트레스의 주를 이루고 있다. 따라서 위 내용을 바탕으로 본 연구를 고안했다.

2.2 기계독해 기반 질의응답 시스템

기계독해는 기계가 주어진 지문과 질문을 이해하여 지문 내에서 답변 영역을 찾아야 하는 자연어 처리 과제로 자동 질의응답 기술의 핵심 토대가 되는 기술이다. 한국어 기계독해를 위한 표준 데이터셋으로는 KorQuAD 1.0이 있다. 이는 위키피디아 문서와 그에 대한 질문, 문서 내 답변이 되는 영역으로 구성된 10만 건 이상의 데이터인 SQuAD[5]를 번역한 질문-답변 70,000여 건과 직접 생성한 질문 4,000여 건으로 이루어져 있는 K-QuAD[6]과는 달리 자연스러운 한국어 문장을 위해 번역을 사용하지 않고 모든 질문을 국문으로 새로 생성하면서, 대표성을 가질 수 있도록 SQuAD의 데이터 구축 방식과 같은 기준을 사용했다. 따라서 모델학습에 이용할 수 있을 뿐만 아니라 여러 모델 간의 성능 평가를 위한 객관적 기준이 된다. 데이터를 처음부터 학습하는 것은 큰 비용이 따르므로, 본 연구의 KorQuAD 1.0을 BERT 언어모델에 fine-tuning 한 기계독해 모형은 적은 비용으로 고품질의 모형을 얻을 수 있다.

2.3 BERT

BERT(Bidirectional Encoder Representations from Transformer)[3]는 구글에서 개발한 자연어 처리 언어모델이다. BERT는 사전 학습된 대용량의 레이블이 지정되지 않은 텍스트에서 심층 양방향 표현을 사전 학습하도록 설계되었다. 사전 훈련된 BERT는 단 하나의 추가 출력 계층으로 조정되어 광범위한 모델을 생성할 수 있다. 다른 사전학습 모델인 OpenAI GPT, ELMo[7]와 비교하여 BERT는 데이터를 양방향으로 학습하여 문맥을 더 잘 파악할 수 있다.

3. 시스템 제안

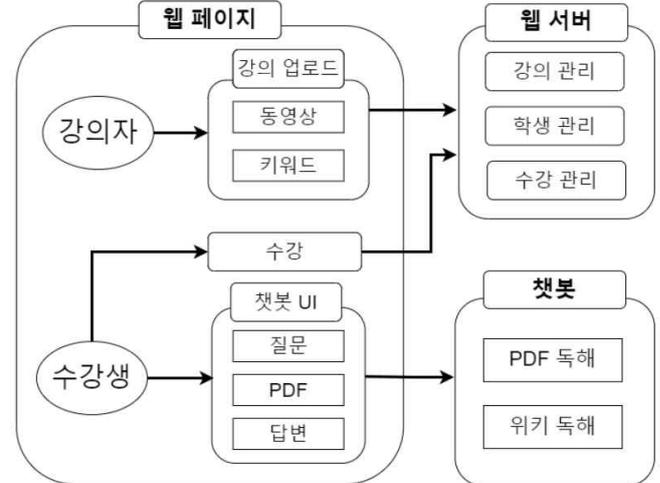


그림 1. 챗봇 시스템 구조

[그림1]은 전체 시스템의 구성을 보여준다. 본 시스템에서는 질의응답을 위해서 BERT 기반의 기계독해를 활용하며, BERT에 입력되는 문서는 강의자가 서버에 올린 강의자료와 위키피디아 문서로 구분된다. (1) 강의자료 기계독해는 사용자의 질문을 강의자료 안에서 찾는다. 사용자는 강의 내용과 공지사항에 관련된 질문을 할 수 있다. (2) 위키피디아 기계독해는 업로드 되어 있는 강의자료가 없을 때 질문과 관련된 위키피디아 문서를 가져와서 진행된다.

3.1 강의자료 기계독해



그림 2. PDF 기계독해 구조

[그림2]에서 강의자료 기계독해는 사용자의 질문에 대한 답을 서버에 업로드 되어 있는 강의자료 안에서 찾는다. 강의자는 각 강의 동영상마다 강의 내용과 관련된 자료를 지정하여 기계독해가 수행되는 문서의 범위를 제한할 수 있다. 지정된 자료에서 텍스트를 추출하여 사용자 질문과 함께 BERT에 입력하여 예측한 답변을 사용자에게 전달한다.

3.2 위키피디아 기계독해

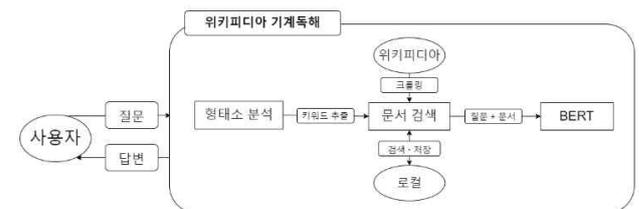


그림 3. 위키피디아 기계독해 구조

위키피디아 기계독해는 강의자료가 없을 때 수행된다. 사용자 질문이 입력되면 형태소 분석을 통해서 질문 내의 명사, 고유명사, 외래어 등 키워드가 될 수 있는 단어들을 추출한다. 추출된 단어에 대한 위키피디아 문서들을 우선으로 로컬 저장 공간에서 검색하여 가져오고 없으면 위키피디아에서 해당 단어들을 검색해서 문서들을 가져온 뒤 재사용을 위해서 로컬 저장소에 저장한다. 가져온 문서와 사용자 질문을 BERT에 입력하여 답을 예측한 뒤 사용자에게 전달한다.

3.3 강의자료 데이터 셋

이 시스템에서 BERT를 fine-tuning 할 때 사용한 KorQuAD 1.0의 답변 유형은 "무엇인가?"와 같이 대상을 묻는 유형이 55.4%, 인물, 시간, 장소를 묻는 유형이 대다수를 차지하고 방법, 원인을 묻는 유형은 적은 비중을 차지한다.[4] KorQuAD 1.0은 대부분의 답변 유형이 단답형이기 때문에 KorQuAD 1.0만 사용하여 fine-tuning 한 모델은 앞서 제시된 유형에 대해서 잘 예측하지만, 어떤 단어의 정의를 묻는 말, 원인을 묻는 말 등 문장형 답변이 필요한 질문에 대해서 잘 예측하지 못하는 경향이 있다. 이를 보완하여 문장형 답변을 포함하는 데이터 셋을 구축하여 추가로 fine-tuning을 진행했다.

```
{
  "answers": [
    {
      "text": "훌륭한 오픈 소스가 있음에도 기존 시스템을 사용하는 기업",
      "answer_start": 604
    }
  ],
  "id": "71",
  "question": "금방치 패턴의 예는?"
}

"context": "설계원리 Single Responsibility Principle(단일책임원리) - 한 클래스는 하나의 일만 가져야 하며, 클래스"
```

그림 4. 강의자료 데이터 셋 구조

이 데이터 셋은 298쌍의 질의응답으로 구성되어 있다, 298개의 답변에서 문장형 답변이 230개, 단답형이 68개를 차지하고 있다.

4. 실험 및 결과

4.1 실험 환경

문장형 답변을 포함한 데이터 셋이 KorQuAD 1.0에 fine-tuning 된 BERT 모델에 어떤 영향을 미치는지 확인하기 위해서 298쌍의 질의응답을 228/70쌍으로 분할하고 학습과 테스트에 교차 적용하여 실험을 진행했다. 228쌍의 질의응답은 문장형 답변 179개, 단답형 답변 49개로 구성되어 있고, 70쌍의 질의응답은 문장형 답변 51개, 단답형 19개로 구성되어 있다.

4.2 성능 및 테스트

강의자료 데이터를 학습하기 전 KorQuAD 1.0에 fine-tuning된 모델은 개발 데이터에 대해서 EX 69%, F1 89%의 성능을 보였다. 이 모델을 이용하여 본 연구에서 제작한 데이터로 테스트한 결과는 [표1] [표2]와 같다.

표1 Train data 228개, Validation data 70개

	exact match	F1 score
학습 전	22	47
학습 후	31	63

표2 Train data 70개, Validation data 228개

	exact match	F1 score
학습 전	17	43
학습 후	26	53

[표1] [표2]의 “학습 전” 모델은 우리가 제안한 학습 데이터를 활용해 2차 fine-tuning을 진행하기 전 모델을 의미한다. 반면 “학습 후” 모델은 우리가 제안한 학습 데이터를 정해진 개수만큼 활용해 2차 fine-tuning을 진행한 모델로 대략 9~10점 정도의 성능향상을 보이는 걸 확인했다. 이로써 본 연구에서 제안한 강의용 질의응답 데이터가 성능향상에 유효한 영향을 미친다는 것을 확인할 수 있었다.

표3. 강의자료 데이터 셋에 대한 질의응답 표

질문	학습 전 답변	학습 후 답변
인공지능의 예시는?	동영상 추천	유튜브에서 인공지능이 동영상을 추천하는 예시가 있다.
블루스의 음악적인 특징?	즉흥 연주	블루스 노트라는 반응을 적극적으로 활용하고 12마디의 블루스를 반복 연주하면서 즉흥 연주
다형성은 무엇인가?	함수명이지만	동일한 함수명이지만 상황에 따라 다른 기능을 수행한다.

[표3]은 같은 질문에 대해서 강의자료 데이터 셋 학습 전, 학습 후 질의응답을 수행한 결과표다. 본 연구에서 제안한 문장형 답변을 학습한 모델은 상대적으로 답변을 문장형으로 예측하는 결과를 보인다.

5. 결론

본 논문에서는 e-러닝 서비스산업 분야에서 강의자가 제공하는 교안을 효과적으로 검색 및 분석하여 수강자에게 제공하는 기술을 확보하기 위해 기계독해 기반 질의응답 시스템을 제안하였다. 온라인 강의 환경은 과거에서부터 점점 증가해왔지만, 이번 코로나19 사태를 겪으면서 그 시장이 큰 규모로 커질 발판이 됐다. 불가피한 온라인 강의 환경에서 강의자와 사용자 사이에 의사소통이 제한됨으로써 강의자가 전달하고자 하는 바를 사용자가 잘 이해하지 못하는 경우가 다분하다. 이를 해결하기 위해 BERT 모델을 이용하여 수강자에게 답변을 제공할 수 있는 시스템을 설계했다.

KorQuAD 1.0의 답변 유형은 대상, 인물, 시간, 장소 등 단답형 답변으로 대부분 구성되어있다. 본 연구에서는 정의나 원인, 등 문장형 답변이 필요한 경우를 보완하기 위해 문장형 답변을 포함하는 데이터 셋을 구축하여 추가로 fine-tuning을 진행했고, 적은 양의 데이터로도 모델은 약간의 성능향상의 결과를 보였다.

현재 수강자의 모든 질의의 정확한 답변을 하기는 무리가 있다. 하지만 위키피디아와 강의 자료를 통한 기계 독해로 수강자들은 그전에는 어려웠던 즉각적인 소통을 기대할 수 있어 효율적이고 유연한 학습 환경을 제공할 수 있다.

참고문헌

- [1] 황요한, 김창수, “코로나 19로 인한 비대면 온라인 강의 인식 조사: 대학생의 만족도와 불안도를 중심으로”, 언어학 29.1 pp.71-91,2021.
- [2] Hulth, Anette, “Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction” Kista: Institutionen för data-och systemvetenskap (tills m KTH),p.111, 2004.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint at arXiv:1810.04805v1, 2018.
- [4] S. Lim, M. Kim, and J. Lee, “KorQuAD: Korean QA Dataset for Machine Comprehension,” Proceeding of the Conference of the Korea Information Science Society, pp.539-541, 2018.
- [5] P. Rajpurkar, et al., SQuAD: 100,000+ Questions for Machine Comprehension of Text, 2016.
- [6] Lee, Kyungjae, et al., Semi-supervised Training Data Generation for Multilingual Question Answeringm LREC, 2018.
- [7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Deep contextualized word representations, arXiv:1802.05365v1 [cs.CL] 15 Feb 2018.