인공지능 튜터링 시스템을 위한 대화 기반 교육 데이터 구축 및 품질 평가

전예림 [1], 황금하 [2], 최승권 [2], 조민수 [2]
[1] 순천향대학교 AI 빅데이터학과 학부생
[2] 한국전자통신연구원 언어지능연구실

9261190@sch.ac.kr, hgh@etri.re.kr, choisk@etri.re.kr, mscho@etri.re.kr

Building and quality assessing conversation-based training data for artificial intelligence tutoring systems

Ye-Lim Jeon[1], Jinxia Huang[2], Sung-Kwon Choi[2], Minsoo Cho[2]
[1]Dept. of Computer Science, Han-Kook University
[2]Language Intelligent Research Section, ETRI

요 약

교육 분야에서는 각 학생의 특성과 요구에 부응하는 개인화 교육의 중요성이 증가하고 있다. 이에 따라 인공지능 기반의 튜터링 시스템, 특히 대화 기반의 튜터링이 주목받고 있다. 본 연구는 GPT-3.5-turbo 를 사용하여 데이터를 생성하는 과정에서 프롬프트 설계의 중요성과 인간의 감수 과정의 필요성을 확인했다. 또한, 자동 평가 방법을 제안하여 데이터의 품질과 유용성을 평가하였다.

1. 서론

최근 교육 분야에서 각 학생의 특성과 요구에 맞는 개인화된 교육의 필요성이 높아지고 있다. 그 중, 대화 기반의 인공지능 튜터링 시스템은 대화를 통해 실제 교육 환경과 유사한 학습 경험을 제공하는 것으로 평가받고 있다. [1, 2]

OpenAI 의 GPT-3.5-turbo 는 대규모 데이터셋을 기반으로 학습되었기 때문에, 다양한 대화 상황과 맥락을 파악하고 적절한 응답을 생성할 수 있다. 따라서, 본 연구는 GPT-3.5-turbo 를 활용하여 인공지능 튜터링 시스템을 위한 대화 기반 교육 데이터의 구축 방법과 구축된 데이터의 품질 평가하는 방법론을 제시하고자 한다.

2. 관련 연구

기존에 공개된 대화 기반 튜터링 데이터 세트 RACE[3]에서 학생은 교사가 묻는 질문에만 대답하는 수동적인 태도를 지녔다. 본 논문에서는 [3]을 확장한 DIRECT[4]를 기반으로 데이터를 생성하고 평가한다.

본 논문은 시나리오를 기반으로 맞춤형 교육을 제공할 수 있는 시스템의 베이스 라인을 따른다. [5]

3. 데이터 자동 구축을 위한 최적의 프롬프트 전략

데이터 자동 구축 과정에서 교육 데이터의 완전성과 일관성을 최대화하기 위한 최적의 프롬프트 구성을 도출하고자 했다.

우선, DIRECT[4] 데이터 중 5 개의 데이터만 뽑아 chatGPT 로 대화 구축을 진행하여 최적의 프롬프트 구성을 찾는다. 독해지문, 연습 문제, 정답에 대한 정보를 바탕으로 대화를 시작하며 학생에게 질문을 하고 피드백을 제공하며 토론을 진행한다. 학생의 응답에 대한 피드백은 학생이 스스로 답을 찾아갈 수 있도록 도와준다. 또한 학생은 능동적으로 대화에 참여하며 지문과 관련된 주제에 관해 토론을 주도할 수 있다. 이런 상황을 가정하여 chatGPT 프롬프트로 입력하고 수정하며 대화 생성의 성능을 확인한다. 이렇게 찾은 최적의 프롬프트를 GPT-3.5-turbo 모델에 입력하여 대용량의 교육용 대화 데이터를 구축한다.

이렇게 생성된 대화에서 부족한 측면은 크게 다음 3 가지 유형으로 확인됐다. 연습문제 그대로 교사가 질문하기, 토론을 진행하지 않은 상태로 대화 끝나기, 질문 유형 태깅 안 하기로 나뉘었다. 이런 유형의 대화 생성하는 것을 줄이기 위해 한 번 언급한 부분을 제대로 수행하지 않을 때 프롬프트의 마지막에 다시 언급을 추가하면 성능이 좋아졌다. 또한, One-shot learning 방식을 도입하였을 때 일관성 있게 대화를 생성하는 것을 확인할 수 있었다.

GPT-3.5-turbo 모델에 입력된 데이터 순서에 따라 성능이 저조해지는 것이 아닌 랜덤하게 일부 데이터에서 안 좋은 성능이 나타나는 특성이 나타났다. 20 개

의 데이터를 순차적으로 입력하였을 때 3 번째, 14 번째, 17 번째 데이터와 관련하여 프롬프트대로 대화 생성을 미흡하게 이루어지는 점을 발견했다. 이러한 결과는 입력 데이터의 순서와 성능 간의 직접적인 상관관계가 존재하지 않음을 시사한다.

4. 자동 생성된 교육 데이터의 인간 감수의 필요성

자동으로 생성된 교육 데이터의 정확성과 효율성을 보장하기 위해 인간의 감수 과정이 필요한지 연구하였다.

결과적으로 자동으로 생성하더라도, 얼마나 정교하게 프롬프트 구성하든 간에 인간의 감수 과정은 불가피하게 필요한 것으로 보였다. 대화 생성이 잘 이루어지지 않은 지문의 빈도는 낮출 수 있어도 모든 지문이 Instruction 대로 생성할 수는 없었다. 또한 학생의 응답이 정답인데 틀렸다고 피드백을 주는 대화를 생성하는 등의 문제도 빈번하게 발생하였다. 태깅을 진행하는 부분을 자동으로 맡겼을 때 다음과 같이 “Hi there! Let's discuss the passage and answer some questions based on it. Are you ready?”를 (Common sense)로 태깅하거나 “Is there any way to use laptops in a healthier way?”를 (Inference)로 부적절하게 태깅하는 패턴이 보였다. 따라서 대화 생성 후에 사람의 검수 과정이 필요하다고 판단하였다.

5. 데이터 자동 평가

완성된 데이터를 자동 평가하기 위한 평가항목은 다음과 같다. 첫째, 한 대화 당 평균 10 턴의 대화 턴으로 구성해야 한다. 둘째, 학생은 한 대화 당 최소한 1 개 이상의 오답을 포함해야 하며, 평균적으로 1~2 개의 오답을 답해야 한다. 셋째, 학생 질문에 각각 공감, 상식, 추론, 비판적 사고, 개인적 연결 등 5 가지 질문 유형 태깅이 포함되어 있어야 하며 각각 20%씩 분포되어 있는다. 넷째, 교사의 발화는 20 단어 미만이어야 한다. 이러한 평가항목을 기반으로 생성된 대화의 특성을 객관적으로 평가할 수 있다. 대화 데이터의 전체적인 분포 특성을 분석하여 구축된 데이터의 품질과 신뢰성을 평가하고자 하였다. 다음 표는 20 개 데이터에 대한 GPT 로 생성된 데이터와 사람의 검수 과정을 거친 데이터의 자동 평가 결과를 비교했다.

<표 1> 데이터 자동 평가 결과.

	GPT-3.5-turbo	Human
Number of dialogues	20	20
Number of turns	220	200
Number of turns with missing taggings	201	4
Missing tagging in keys	{1029, 1049, 1051…}	{1049, 1198, 1378, 149}
Skill type distribution		
Empathy	33.33% (20%)	14.42% (20%)
Common sense	33.33% (20%)	36.54% (20%)
Inference	00.00% (20%)	15.38% (20%)
Critical Thinking	00.00% (20%)	16.35% (20%)
Experience	33.33% (20%)	17.31% (20%)
Average of tutor tokens	29.2	24.93

표 1 을 분석해보면, 사람이 검수한 데이터가 전체적으로 평가항목을 만족한다. 태깅이 안 되어 있는 발화가 포함된 턴 수가 201 개인 반면 사람이 검수한 데이터는 4 개로 현저하게 줄었다. 또한 5 개의 질문 유형이 균일하게 20%에 가깝게 분포되어 있으며 교사의 평균 단어 수가 줄어든 것을 확인할 수 있다.

6. 결론

이러한 연구를 통해, 대화 기반 교육 데이터의 구축과 평가에 대한 중요한 고려 사항을 찾고자 하였다. 자동으로 교육용 대화 데이터를 생성할 때, One-shot learning 방식의 도입이 일관성 있는 대화 생성에 도움이 되는 것으로 나타났다. 또한, 자동 생성된 데이터의 정확성을 보장하기 위해서는 사람의 검수를 통해 부적절한 내용이나 오류를 검증하는 작업이 필수적인 것으로 나타났다. 마지막으로, 생성된 대화가 평가항목을 충족시키는지를 객관적으로 확인하는 방법을 제안하였다. 이를 통해 데이터의 품질을 평가하고 유용성을 입증할 수 있다.

이로써 본 연구는 대화 기반 교육 데이터의 구축과 평가에 관한 유의한 인사이트를 제공하였으며, 개인화된 교육 환경을 위한 인공지능 튜터링 시스템의 발전에 기여할 수 있을 것으로 기대한다.

Acknowledgement

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

참고문헌

[1] Roger S. Pressman "Software Engineering A Practltiners' Approach" 3rd Ed. McGraw Hill

[2] M. Ventura, M. Chang, P. Foltz, N. Mukhi, J. Yarbro, A. P. Salverda, et al., "Preliminary evaluations of a dialogue-based digital tutor", Proc. Int. Conf. Artif. Intell. Educ., pp. 480-483, Jun. 2018.

[3] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, ‘‘RACE: Large-scale ReAding comprehension dataset from examinations,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 785–794, doi: 10.18653/v1/D17-1082.

[4] J. -X. Huang, Y. Lee and O. -W. Kwon, "DIRECT: Toward Dialogue-Based Reading Comprehension Tutoring," in IEEE Access, vol. 11, pp. 8978-8987, 2023, doi: 10.1109/ACCESS.2022.3233224.

[5] 이정민, 조민수, 김현, 권오욱, 황금하. (2023). 한국어 교육을 위한 간편 챗봇 : 학습자 발화 분류와 대화 생성 모델을 이용하여. 한국정보과학회 학술발표논문집, (), 917-919.