ChatGPT 를 이용한 독해 튜터링 대화 데이터 확장

권현유 [1], 최승권 [2], 황금하 [2], 권오욱 [2]
[1] 부산대학교 공공정책학부 학부생
[2] 한국전자통신연구원 언어지능연구실

klairekwon@gmail.com, choisk@etri.re.kr, hgh@etri.re.kr, ohwoog@etri.re.kr

Data Augmentation of English Reading Comprehension Tutoring Dialogs using ChatGPT

Hyunyou Kwon[1], Sung-Kwon Choi[2], Jinxia Huang[2], Oh-Woog Kwon[2]
[1]Dept. of Public Policy and Management, Pusan National University
[2]Language Intelligence Research Section, ETRI

요 약

대화형 독해 튜터링 시스템을 위한 학생주도 대화 데이터셋 생성 및 확장에 ChatGPT 의 활용 가능성을 평가하였다. 단순히 수동으로만 구축한 기존의 데이터셋과 ChatGPT 에 의해 반자동으로 확장된 데이터셋을 비교한 결과, 구축량, 소요 시간, 비용 및 반복 작업 측면에서 ChatGPT 가 가진 유용성을 알 수 있었다. 그러나, 유형별 배분의 편중과, 부적절한 데이터 생성 등의 한계도 나타났다. Chat GPT 의 빠른 발전이 예상됨에 따라 대화형 튜터링 분야에 ChatGPT 에 의한 반자동 데이터 확장 방법이 널리 활용될 것으로 기대된다.

1. 서론

기존의 대화형 독해 튜터링에서는 컴퓨터 튜터가 질문하면 학생이 응답하는 수동적인 독해 튜터링 대화 방식[1]이 주를 이루고 있다. 그러나, 학생이 능동적으로 질문을 주도하는 대화형 독해 학습방식에 대한 연구는 매우 부족한 실정이다.

본 논문에서는 학생 입장에서 수동적인 기존 데이터셋을 활용하여 학생 주도의 능동적인 학습용 대화 데이터의 생성 및 확장에 GPT-3.5 기반 ChatGPT[2]의 활용 가능성을 검증하고자 한다.

2. 학생 주도 대화 방식의 학생 질문 유형 분류

학생주도 영어 독해 튜터링 대화시스템의 데이터 생성 및 확장을 위하여 RACE[3]를 토대로 구축된 DIRECT 데이터[1]를 기초자료로 활용하였다. DIRECT 데이터는 영어 지문과 각 지문당 튜터 주도형 질문 약 3~5 개의 형태로 구성되어 있다. 본 연구에서는 DIRECT 데이터의 지문 중 9 개의 지문을 선별하여 활용하였다.

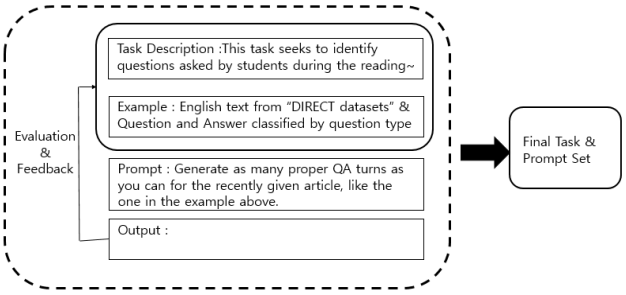
본 논문에서는 학생 주도 질문을 두가지 유형으로 분류하였다. 지문의 내용에 관한 질문 유형인 general 유형과 학생의 개인적인 특성이 가미된 질문 유형인 specific 유형이다. General 유형은 단어의 뜻을 묻는 word, 문장이나 단락의 내용을 묻는 reasoning, 전체 지문의 의미나 주제를 묻는 subject 로 세분화하였다. Specific 유형은 감정 표현인 attitude, 개인적 의견이 포함된 opinion, 관련성이 낮은 unrelated 로 세분화하였다. 상기 두 가지 유형을 바탕으로 질문을 생성, 분류 및 확장하였으며, 이러한 분류체계는 표 1 과 같다.

<표 1> 독해 튜터링 대화 시스템의 학생 질문 유형 및 분류

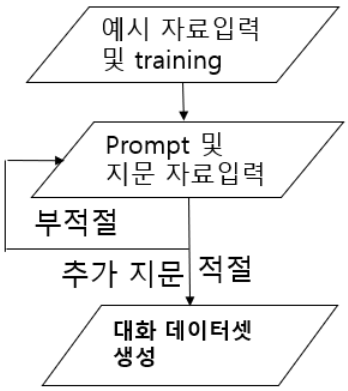
Types	Category
General Question	Word, Reasoning, Subject
Specific Question	Attitude, Opinion, Unrelated

3. ChatGPT 를 이용한 질문 생성 및 데이터 확장 파이프라인

먼저 ChatGPT 에게 task description 과 함께 예시 지문, 유형별 정의, 유형별 예시 QA 를 제공한다. 다음으로 prompt 와 함께 새로운 지문을 제공하고 도출된 output 은 유형별로 valid/invalid 여부를 판별하고 원인을 찾아 이를 보완하여 적절한 대화를 생성할 것을 요청하였다. 이처럼 task 와 prompt 세트를 구축하는 과정을 그림 1 에 도식적으로 나타내었다. ChatGPT 는 동일한 prompt 의 반복적인 수행에서 매우 높은 효율을 나타내었다. ChatGPT 를 활용한 초등학생 수준에 적합한 학생 주도 전체 독해 데이터에 대한 대화의 생성 및 확장 과정을 그림 2 에 도식적으로 나타내었다.



(그림 1) Task & prompt set 구축 과정



(그림 2) ChatGPT 를 이용한 대화 데이터 생성 및 확장 과정의 도식적 체계도

4. 결과 고찰

ChatGPT 를 사용하여 얻은 데이터셋과 사람이 생성한 데이터셋을 유형별로 분류하고, 초등학생 수준에서의 답변 적합성 여부를 판단하여 valid/invalid 로 구분하였다(표 2).

사람은 9 개의 지문에서 general 질문 35 개와 specific 질문 19 개, 지문당 평균 6 개의 질문을 생성하였다. 모든 질문이 초등학생 수준에 적합하였고, 유형별로 비슷한 수의 질문을 생성하였다. 반면 ChatGPT 는 general 62 개와 specific 73 개, 지문당 평균 15 개로 사람 대비 2.5 배 많은 질문을 생성하였으나 일부 부적합한 질문을 생성하였다. 특히 specific 의 경우 ChatGPT 는 약 3 배의 질문을 생성하였으나, 오류 비율이 약 40%로 높았다. 또한 reasoning 및 subject 질문은 상대적으로 적었다.

Invalid 를 유형별로 살펴보면, reasoning, attitude, opinion, unrelated 의 경우 지문과 무관한 내용을 제시하거나 지문에 없는 내용을 지어내는 오류가 주로 나타났다. 예외적으로 subject 의 경우에는 reasoning 유형의 질문이 생성되는 오류가 한 차례 발생하였다. 그러나 유사 자료에 대한 동일 작업 내용의 수행에서는 ChatGPT 가 매우 높은 효율을 보여주었으므로 향후 대량의 데이터 처리에서 활용도가 높을 것으로 기대된다.

<표 2> DIRECT 데이터를 기반으로 사람이 추출한 데이터와 ChatGPT 가 생성한 데이터 자료의 유형별 통계적 비교

Dataset			Human	ChatGPT
General Question	word	valid	9	23
		invalid	0	0
	reasoning	valid	13	22
		invalid	0	4
	subject	valid	13	12
		invalid	0	1
	Sub_sum	valid	35	57
		invalid	0	5
Specific Question	attitude	valid	6	31
		invalid	0	5
	opinion	valid	13	10
		invalid	0	4
	unrelated	valid	0	11
		invalid	0	12
	Sub_sum	valid	19	52
		invalid	0	21
Total_sum		valid	54	109
		invalid	0	26

5. 결론

ChatGPT 를 이용하여 초등학생 대상 학생주도 대화형 튜터링 시스템 구축을 위한 데이터셋 생성 및 확장에의 활용 가능성을 평가하였다. 사람이 생성한 데이터셋과 ChatGPT 의 자료를 비교한 결과, 데이터셋 구축량, 소요 시간, 비용 및 반복 작업 측면에서 ChatGPT 가 가지는 유용성을 알 수 있었다. 그러나 reasoning 및 subject 유형 데이터 생성 부족으로 유형별 배분이 편중되었고, 일부 부적합 데이터가 생성되었다. 따라서, 현재까지는 ChatGPT 를 활용한 데이터셋에 사람의 개입이 필요하다는 한계도 보여주었다. ChatGPT 의 빠른 발전이 예상됨에 따라 대화형 튜터링 분야에 널리 활용될 것으로 기대된다.

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]] Jin-Xia Huang, Yohan Lee, Oh-Woog Kwon, "DIRECT: Toward Dialogue-Based Reading Comprehension Tutoring", IEEE Acess, Vol.11, pp.8978-8987,2023

[2] ChatGPT: Optimizing Language Models for Dialogue, https://openai.com/blog/chatgpt/

[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