

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0075862
(43) 공개일자 2025년05월29일

(51) 국제특허분류(Int. Cl.)

G06F 18/21 (2023.01) G06F 16/332 (2025.01)
G06F 16/338 (2019.01) G06N 3/0475 (2023.01)
G06N 3/096 (2023.01)

(52) CPC특허분류

G06F 18/217 (2023.01)
G06F 16/3329 (2025.01)

(21) 출원번호 10-2023-0162875

(22) 출원일자 2023년11월22일

심사청구일자 2023년11월22일

(71) 출원인

고려대학교 산학협력단

서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)

(72) 발명자

임희석

서울특별시 성북구 안암로 145 자연계캠퍼스 애기
능생활관 311호 NLP&AI연구실

문현석

서울특별시 강남구 자곡로 260, 422동 903호

(74) 대리인

김동용

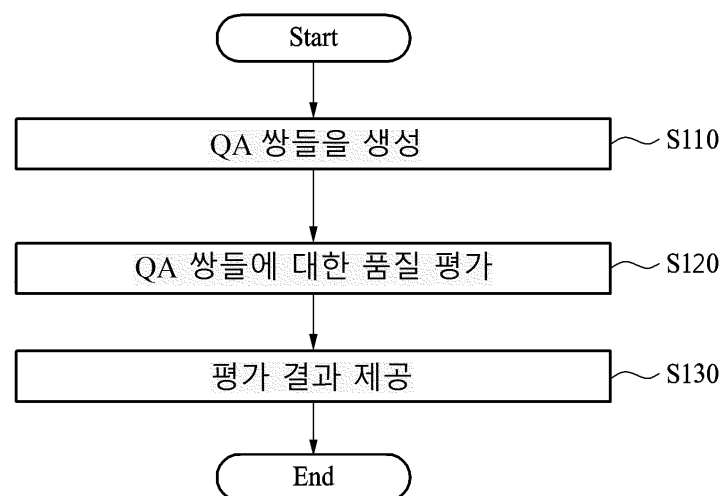
전체 청구항 수 : 총 3 항

(54) 발명의 명칭 교육용 질의응답 데이터 평가 장치 및 방법

(57) 요약

질의응답(Question-Answer, QA) 데이터 평가 장치 및 방법이 개시된다. 상기 질의응답(QA) 데이터 평가 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 주어진 지문(passage)에 대응하는 질의응답(QA) 쌍이 생성되는 단계, 및 상기 질의응답 쌍에 대한 품질이 평가되는 단계를 포함하고, 상기 질의응답(QA) 쌍이 생성되는 단계는, 사전학습된 질의응답 생성(Question-Answer Generation, QAG) 모델에 상기 주어진 지문과 질문 시작 토큰을 입력함으로써 생성된다.

대표도 - 도3



(52) CPC특허분류

G06F 16/338 (2019.01)

G06N 3/0475 (2023.01)

G06N 3/096 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193711
과제번호	2018-0-01405-006
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1345365402
과제번호	2021R1A6A1A03045425
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	이공학학술연구기반구축
연구과제명	Human-inspired AI 연구소
기 여 율	1/2
과제수행기관명	고려대학교
연구기간	2023.03.01 ~ 2024.02.29

명세서

청구범위

청구항 1

적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 질의응답 데이터 평가 방법에 있어서,

주어진 지문(passage)에 대응하는 질의응답(Question-Answer, QA) 쌍이 생성되는 단계; 및

상기 질의응답 쌍에 대한 품질이 평가되는 단계를 포함하고,

상기 질의응답(QA) 쌍이 생성되는 단계는, 사전학습된 질의응답 생성(Question-Answer Generation, QAG) 모델에 상기 주어진 지문과 질문 시작 토큰을 입력함으로써 생성되는,

질의응답 데이터 평가 방법.

청구항 2

제1항에 있어서,

상기 품질이 평가되는 단계는 수학적식을 이용하여 품질(S_{GI})을 평가하고,

상기 수학적식은

$$S_{GI} = \frac{1}{N_c} \sum_{j=1}^{N_c} \left[\frac{1}{N_{c_j}} \sum_{l=2}^{N_{c_j}-1} prob_{j,l}^c \right]$$

이고,

상기 $prob_{j,l}^c$ 는 생성된 질의응답(QA) 쌍의 집합(C) 내의 j번째 질의응답(QA) 쌍에서 1번째 토큰이 생성될 확률이고,

상기 N_c 는 생성된 질의응답(QA) 쌍의 개수이고,

상기 N_{c_j} 는 j번째 질의응답(QA) 쌍에 포함된 토큰의 개수인,

질의응답 데이터 평가 방법.

청구항 3

제2항에 있어서,

상기 품질을 사용자에게 제공하는 단계를 더 포함하는,

질의응답 데이터 평가 방법.

발명의 설명

기술 분야

[0001] 본 발명은 인공 신경망 기반 질의응답 평가 시스템에 관한 것이며, 질의에 대한 응답이 적절한지에 대한 평가 및 질의응답 그 자체의 교육용으로써의 적합성 평가를 위한 장치 및 방법에 관한 것이다.

배경 기술

[0002] 주어진 지문(passage)에 대한 적절한 질의를 던지고, 이에 대한 응답을 유도하는 것은, 아이들의 문해력 증진에 매우 효과적인 방법이다. 이를 위하여는, 지문을 읽고 질문할만한 내용을 선정해야 하며, 응답이 적절한지에 대해서도 사람이 직접 평가하여야 한다. 그러나, 모든 과정을 사람이 직접 수행하는 것은 큰 노동력이 요구된다. 이를 위하여, ChatGPT 등과 같은 거대 언어 모델(Large Language Model, LLM)을 활용하는 것도 효과적인 방법이 될 수 있으나, 지속적인 API 비용이 요구된다는 점에서 바람직하지 않다.

[0003] 현재까지, 지문에 대한 질의응답 데이터를 생성하는 방법론은 다수 등장하였으나, 생성된 질의응답 데이터의 품

질이 좋은지, 실제로 교육 목적으로 사용하기에 적합한지 등에 대한 평가 방법에 대한 적절한 논의는 진행된 바가 거의 없다.

[0004] ChatGPT 등의 거대 언어 모델(LLM)이 뛰어난 성능을 보이면서, 이를 활용한 데이터 선별 및 평가 기술을 활용할 수 있으나, 아직 환각(hallucination) 등의 문제가 남아 있어 ChatGPT 등의 결과를 온전히 신뢰할 수 없을뿐더러, 사용에 따라 즉각적인 API 사용료가 부과되기 때문에 지속적인 활용에 한계가 있다.

[0005] 이에, ChatGPT보다 적은 크기의 모델을 활용하면서도 ChatGPT의 성능과 유사한 수준의 평가 능력을 보이는 기술 개발이 필요한 실정이다.

선행기술문헌

비특허문헌

[0006] (비특허문헌 0001) Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74-81.

(비특허문헌 0002) Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.)

(비특허문헌 0003) Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5073-5085.

(비특허문헌 0004) Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is ai's turn to ask humans a question: Question-answer pair generation for children's storybooks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 731-744.

(비특허문헌 0005) Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, Songeun Lee, Changwoo Chun, Sungsoo Park, and Heuseok Lim. 2023. Towards diverse and effective question-answer pair generation from children storybooks.

발명의 내용

해결하려는 과제

[0007] 본 발명이 이루고자 하는 기술적인 과제는 교육용 질의응답 데이터 자동평가 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0008] 본 발명의 일 실시예에 따른 질의응답(Question-Answer, QA) 데이터 평가 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 주어진 지문(passage)에 대응하는 질의응답(QA) 쌍이 생성되는 단계, 및 상기 질의응답 쌍에 대한 품질이 평가되는 단계를 포함하고, 상기 질의응답(QA) 쌍이 생성되는 단계는, 사전 학습된 질의응답 생성(Question-Answer Generation, QAG) 모델에 상기 주어진 지문과 질문 시작 토큰을 입력함으로써 생성된다.

발명의 효과

[0009] 본 발명의 실시예에 따른 교육용 질의응답 데이터 자동평가 장치 및 방법에 의할 경우, 지문에 대한 질의응답 쌍의 적절성 및 지문과 질의에 대한 응답의 적절성을 평가할 수 있는 효과가 있다.

도면의 간단한 설명

[0010] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 사람 평가와의 상관관계를 고려한 실험 결과인 표 1을 도시한다.

도 2는 본 발명의 일 실시예에 따른, 질의응답 (데이터) 평가 방법에 관한 사례 연구 결과를 도시한다.

도 3은 본 발명의 일 실시예에 따른, 질의응답 (데이터) 평가 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0011] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.
- [0012] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.
- [0013] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.
- [0014] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.
- [0015] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0016] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0017] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.
- [0018] 질의응답 생성(Question-Answer Generation, QAG)에 대한 평가, 즉 생성된 질의응답 데이터에 대한 평가는 후보 질의응답(Question-Answer, QA) 집합 $C = \{(q_j^c, a_j^c)\}_{j=1}^{N_c}$ 의 품질(quality)를 측정함으로써 수행될 수 있다. 이때, 후보 질의응답(QA) 집합 C는 주어진 지문(passage) P에 대해 생성된 것일 수 있다. 기존의 기법은 C와 GT(Ground-Truth) 질의응답(QA) 집합 $R = \{(q_i^r, a_i^r)\}_{i=1}^{N_r}$ 사이의 텍스트 유사도(textual similarity)를 측정한다. 본 명세서에서는, 텍스트 유사도를 Metric으로 표기한다. 또한, 기존 기법은 ROUGE(Chin-Yew Lin, 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74-81.)와 BERTscore(Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.)의 두 측정치를 사용한다. 본 발명에서는, 다중-참조(multi-reference)와 다중 후보 설정

(multi-candidate setting)을 고려하여, C를 평가하는 두 가지 전략을 제안한다.

[0019] **Concat-Metric.** 종합적인 평가를 위해, Zhao 등은 모든 질의응답(QA) 쌍들을, 각 질의응답(QA) 쌍 집합에 대한 단일의 시퀀스로 결합(concatenates)하고 그들 사이에서의 Metric을 평가한다(Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5073-5085.). 이 경우, C의 추정된 품질(s_{Concat} 로 표기됨)은 수학적 식 1과 같이 계산될 수 있다. 수학적 식 1에서, $[\dots]$ 은 모든 요소(elements)에 대한 시퀀스화된 병합(sequentialized concatenation)을 의미할 수 있다.

[0020] [수학적 식 1]

$$\begin{aligned} r_i &= [q_i^r \ a_i^r], \ r = [r_1, \dots, r_{n_r}] \\ c_j &= [q_j^c \ a_j^c], \ c = [c_1, \dots, c_{n_c}] \\ s_{\text{Concat}} &= \mathbf{Metric}(r, c) \end{aligned}$$

[0021]

[0022] **MAP@N-Metric.** Yao 등(Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is ai's turn to ask humans a question: Question-answer pair generation for children's story books. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 731-744.), Eo 등(Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, Songeun Lee, Changwoo Chun, Sungsoo Park, and Heuseok Lim. 2023. Towards diverse and effective question-answer pair generation from children storybooks.), 그리고 Xu 등(Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5073-5085.)은, R 내의 각 질의응답(QA) 쌍에 대하여, C와 가장 유사한 질의응답(QA) 쌍을 찾는다. 다시 말하면, R 내의 각 질의응답(QA) 쌍에 대한 가장 높은 Metric을 계산할 수 있고, 이는 C 내의 질의응답(QA) 쌍과의 비교를 통해 도출될 수 있다. C의 추정된 품질(s_{MAP} 로 표기함)은 수학적 식 2와 같다.

[0023] [수학적 식 2]

$$\begin{aligned} \text{metric}_{i,j} &= \mathbf{Metric}([q_i^r \ a_i^r], [q_j^c \ a_j^c]) \\ s_{\text{MAP}} &= \frac{1}{N_r} \sum_{i=1}^{N_r} \max_j \{\text{metric}_{i,j}\}_{j=1}^{N_c} \end{aligned}$$

[0024]

[0025] 사람에게 의한 평가를 적용할 때, 질의응답 생성(QAG) 시스템은 일반적으로 다음 측면을 기준으로 평가한다. (i) 질의응답(QA) 쌍이 지문과 관련이 있는지를 결정하는, 지문과의 관련성(Relevancy to the passage), (ii) 답변이 질문에 대한 적절한 응답으로 간주될 수 있는지 여부를 보여주는, 답변의 답변 가능성(Answerability of the answer), 그리고 (iii) 생성된 질의응답(QA) 쌍의 문법적 타당성(Grammatical plausibility). 그러나, GT와의 유사성을 측정하는 기존의 자동 평가 방법이 위 요구 사항을 충족시키는 데 제한이 있으며, 질의응답(QA)의 객관적인 품질을 평가하지 않은채 질의응답(QA)이 GT와 유사한지만을 평가한다는 문제가 있다.

[0026] 이하에서는, 본 발명의 일 실시예에 따른 질의응답 평가 방법(또는 질의응답 데이터 평가 방법)에 대해 구체적으로 설명한다. 제안하는 질의응답 평가 방법은 GI(Generative Interpretation)로 명명될 수도 있다.

[0027] GI는 생성된 질의응답(QA) 쌍의 적절성(adequacy)을 평가한다. 적절성은 지문과의 연관성(또는 관련성) 및 QA 간의 연결성(connectivity)을 포괄하는 개념으로 이해될 수 있다. BARTscore(Yuan 등(Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, 34:27263-27277.))와 유사하게, 본 발명에서는 질의응답(QA) 생성 모델을 채택하고 질의응답(QA) 생성의 교사 강제 로짓(teacher-forced logits)을 사용한다. 특히, 본 발명에서는, 지문과 질문 시작 토큰(question start tokens)을 제공함으로써, QA 생성 모델 θ 가 연결된 시퀀스의 QA 쌍을 반환하도록 학습한다. 이때, L_{CE} 는 수학적 식 3과 같다.

[0028] [수학식 3]

$$\mathcal{L}_{CE} = -\frac{1}{N_r} \sum_{i=1}^{N_r} \prod_{l=1}^{N_{r_i}} \mathbf{P}_{\theta}(r_{i,l} | r_{i,<l}, q_{i,<n_s}^r, P)$$

[0030] 질문 시작 토큰의 개수는 하이퍼파라미터(hyper-parameter) n_s 로 미리 설정될 수 있다. 본 발명에서는, 질문 타입 편향(question type bias)을 완화하기 위하여, 각 질문의 시작 토큰을 입력 시퀀스의 일부로 제공한다.

[0031] GI는 θ 에 의해 계산된 후보 질의응답(QA) 쌍의 교차 강제 로짓에 의해 추정된다. 보다 명확하게, $c_j = [c_{j,1}, \dots, c_{j,n_c}]$ 내의 l^{th} 토큰의 확률(θ 에 의해 생성될 확률을 의미할 수 있음)을 $prob_{j,l}^c$ 로 표기한다(수학식 4). 그러면, C에 대한 스코어 s_{GI} 는 수학식 5와 같다.

[0032] [수학식 4]

$$prob_{j,l}^c = \mathbf{P}_{\theta}(c_{j,l} | c_{j,<l}, q_{j,<n_s}^c, P)$$

[0034] [수학식 5]

$$s_{GI} = \frac{1}{N_c} \sum_{j=1}^{N_c} \left[\frac{1}{N_{c_j}} \sum_{l=2}^{N_{c_j}-1} prob_{j,l}^c \right]$$

[0036] 여기서 c_j 는 j번째로 생성된 질의응답(QA) 쌍을, q_j 는 질문의 전체 길이를 의미하고, N_{c_j} 는 j번째 질의응답(QA) 쌍의 토큰의 개수를, N_c 는 생성된 질의응답(QA) 쌍의 개수를 의미할 수 있다.

[0037] GI는, 어떠한 질의응답(QA) 쌍들도 GT와 독립적으로 평가할 수 있는, 참조 없는 평가 방법(reference-free evaluation method)으로 동작한다. 특히, c_j 내의 질문 위치(question position)의 로짓은 질의응답(QA) 쌍의 지문과의 연관성을 결정하고, c_j 내의 답변 위치(answer position)는 c_j 내 답변에 대한 응답가능성을 반영한다. 또한, 로짓은 생성 확률(generation possibility)을 반영하기 때문에, GI를 통해 질의응답(QA) 쌍의 가독성(readability)을 판단할 수도 있다.

[0038] 학습 단계(training phase)와 달리, GI는, 단일 이상치(single outlier)에 의한 확률 저하를 방지하기 위해, 확률의 평균으로 계산된다. 또한, [BOS] 및 [EOS] 위치에서의 확률은 의도하지 않은 편향을 완화하기 위해 계산에서 배제된다.

[0039] GI의 높은 성능은 θ 의 광범위한 언어 능력(linguistic capability)에 기인할 수 있다. GI의 타당성을 명확히 하기 위해, θ 의 생성 출력과 C 간의 텍스트 유사성을 추정하는, 다른 베이스라인(baseline) 평가 방법 GI_{ss} 을 구축한다. GI와 GI_{ss} 를 비교함으로써, 질의응답 생성(QAG)을 평가함에 있어, 질의응답 생성(QAG) 모델 능력에 대한 의존성이 완화된 상태에서, GI의 유효성을 검증할 수 있다.

[0040] 실험 결과

[0041] Fairytale QA dataset(Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqaan authentic dataset for narrative comprehension. arXiv preprint arXiv:2203.13947.)을 사용한다. 4 개의 질의응답 생성(QAG) 시스템에 의해 생성된, 20 개의 지문들에 대한 질의응답(QA) 쌍들에 대한 사람 평가(human evaluation)를 진행한다. Eo 등(Sugyeong Eo, Hyeonseok Moon, Jinsung Kim, Yuna Hur, Jeongwook Kim, Songeun Lee, Changwoo Chun, Sungsoo Park, and Heuseok Lim. 2023. Towards diverse and effective question-answer pair generation from children storybooks.)에서와 같이, 4가지 측면: QA 쌍과 지문의 연관성을 추정하는 Relevancy P-QA (Rel P-QA), 질문과 이에 대응하는 대답이 정확하게 매칭되었는지 여부를 평가하는 Relevancy Q-A (Rel Q-A), 교육 분야(educational field)에서 QA 쌍의 실용적인 이용 가능성(practical usability)을 추정하는 Usability (Usb), 문법적 정확성을 가르키는 Readability (Rdb)를 평가한다.

[0042] GI를 구축하기 위해, 두 개의 사전학습된 언어 모델, T5와 BART를 채용하고, ROUGE 추정에서 ROUGE-L F1-score

를 측정하고, BERTscore를 위한 F1-score를 측정한다.

- [0043] GT 간의 유사성을 추정하는 것은 의심스러운 결과를 낳을 수 있다. 표 1(도 1 참조)의 클래스 (a)에 도시된 바와 같이, GT 유사성을 기반으로 한 메트릭(metric)이 특히 관련성 측면에서 낮은 인간 평가를 보여줄 수 있다. 이는 GT 간의 유사성이 주어진 질문과의 관련성을 결정하는 데 심각한 어려움을 겪는다는 것을 시사한다. 대신, 이는 이용가능성 측면과 매우 높은 상관 관계를 보인다. 이러한 결과는, 기존의 평가 방법은 타당하지 않음을 나타낸다.
- [0044] 참고로, 표 1에서, BERTscore는 BS로, content-wise evaluation은 (P)로, QA-wise evaluation은 (QA)로, Rel P-QA와 Rel Q-A의 평균은 Rel Avg로 나타내었다. 또한, 실험에서 설정된 n_s 의 값은 4이다.
- [0045] ChatGPT는 준수한 평가 결과를 제공한다. 표 1의 클래스 (b)에서의 결과는, ChatGPT가 질의응답 생성(QAG)에 대한 준수한 평가자임을 보여준다. 또한, 질의응답(QA) 기반 평가(QA-wise evaluation)(i.e. GPT#(QA))가 ChatGPT의 평가 성능을 높이는 데 크게 기여한다는 것을 발견하였다. 특히, GPT4는 뛰어난 성능을 보이는 반면, GPT3.5는 상대적으로 중간 정도의 성능을 보이며, 이는 질의응답 생성(QAG)에 대한 평가의 어려움을 시사한다.
- [0046] GI는 신뢰성 있는(trust-worthy) 평가 결과를 제공한다. 표 1의 모든 결과들을 고려할 때, 특히 클래스 (c)와 관련하여, GI는 사람 평가와의 높은 일치성을 보인다. GI-BART는 기존의 평가 방법을 능가하며, 심지어 GPT3.5의 성능도 뛰어넘는다.
- [0047] 특히, GI는 GPT3.5와 비교 가능한 성능을 보이는 반면, GI_{SS}는 ROUGE를 능가하지 못한다. 이러한 결과는, GI에서 적용된 방법론이 각 질의응답(QA) 쌍에 대한 보다 객관적이고 사람 평가와 유사한 평가를 가능케 한다는 것을 나타낸다.
- [0048] GI를 추정할 때, 의도하지 않은 편향을 유발할 수 있는 비논리적인 요인으로 간주하여, 입력으로 제공함으로써 "질문 시작 토큰"을 배제하고, [BOS] 및 [EOS] 위치에서의 로짓을 배제한다. 도 2는 이와 관련된 사례 연구를 도시한다. 질문 시작 토큰의 수(n_s)를 조절하여, QA 쌍을 평가할 때 관련 없는 로짓을 배제함으로써, 더 높은 성능을 얻을 수 있다는 것을 발견하였다. 유사하게, [BOS] 및 [EOS] 위치에서의 로짓도 의도하지 않은 편향을 유발하고 인간 일치를 감소시킨다는 것을 발견하였다.
- [0049] 도 3은 본 발명의 일 실시예에 따른, 질의응답 (데이터) 평가 방법을 설명하기 위한 흐름도이다.
- [0050] 도 3은 참조하면, 질의응답 평가 방법은 적어도 프로세서(Processor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 다시 말하면, 질의응답 평가 방법을 이루는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 질의응답 평가 장치로 명명될 수도 있는 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 랩톱 컴퓨터, 태블릿 PC 등을 포함할 수 있다. 또한, 컴퓨팅 장치는 물리적으로 구분된 복수의 장치로 구현될 수도 있다.
- [0051] 우선, 적어도 하나의 질의응답(QA) 쌍이 생성된다(S110). 질의응답(QA) 쌍은 소정의 질의응답 생성(QAG) 모델에 의해 생성될 수 있다. 다만, 본 발명이 질의응답(QA) 쌍을 생성하는 모델에 제한되는 것은 아니다. 본 발명이 생성된 질의응답(QA) 쌍의 품질을 평가하는 방법에 관한 발명이기 때문이다. 따라서, 컴퓨팅 장치에는 다양한 질의응답 생성(QAG) 모델들이 저장되어 있을 수 있고, 질의응답 쌍을 생성하기 위한 복수의 질문들이 저장되어 있을 수 있다.
- [0052] 이때, 질의응답(QA) 쌍을 생성하기 위하여, 질의응답 생성(QAG) 모델에는 질문 이외에 질문 시작 토큰이 더 입력될 수 있다. 질문 시작 토큰은 What (is), Can you tell me 등과 같이 질의응답(QA) 쌍에 포함되는 질문을 시작하는 문구를 의미할 수 있다. 또한, 질문 시작 토큰의 개수는 미리 설정될 수 있으며, 이에 따라 질문 시작 토큰의 개수에 대응하는 개수의 질의응답 쌍이 생성될 수 있다.
- [0053] 생성된 질의응답(QA) 쌍에 대한 품질 평가가 수행된다(S120). 이를 위해, 사전에 학습이 완료된 평가 모델이 컴퓨팅 장치에 미리 저장되어 있을 수 있다. 또는, (미세조정된) 거대 언어 모델(LLM)이 평가 모델로 이용될 수도 있다. 실시예에 따라, 품질 평가는 앞서 기재된 수학적 식 5에 따라 수행될 수 있다.
- [0054] 질의응답(QA) 쌍에 대한 품질 평가 결과는, 소정의 출력 장치(예컨대, 디스플레이 장치 등)를 통해 사용자에게 제공될 수 있다(S130).
- [0055] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어,

프로세서, 컨트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0056] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0057] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

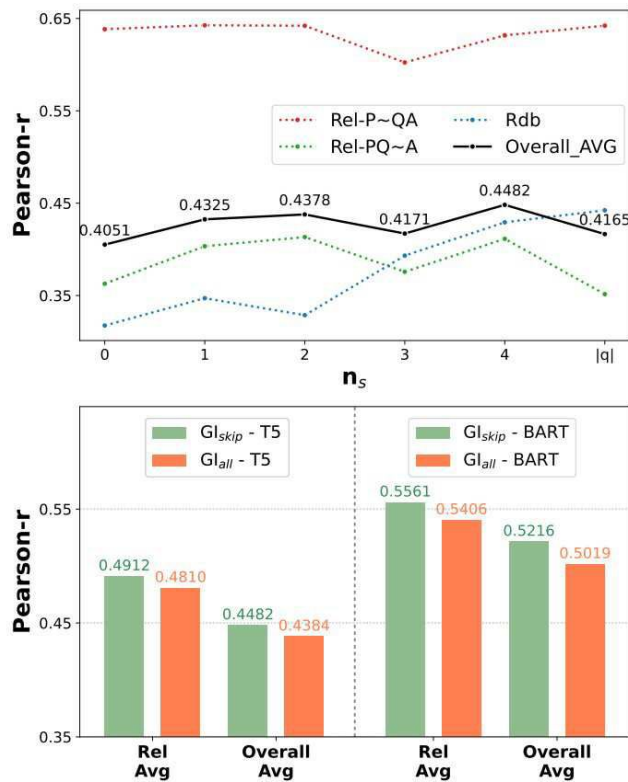
[0058] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

도면1

CLASS	Evaluation Method	Rel P-QA	Rel Q-A	Rel Avg	Usb	Rdb	Overall Avg
(a) GT Similarity	MAP@N-ROUGE	0.25708	0.28495	0.27101	0.44938	0.27702	0.31711
	Concat-ROUGE	0.23987	0.29014	0.26501	0.44401	0.27455	0.31214
	MAP@N-BS	0.31421	0.34464	0.32943	0.45315	0.33133	0.36083
	Concat-BS	0.30326	0.31240	0.30783	0.44174	0.33771	0.34878
(b) ChatGPT	GPT3.5 (P)	0.71699	0.36462	0.54080	0.13104	0.43409	0.41168
	GPT3.5 (QA)	0.73321	0.44916	0.59118	0.36482	0.35204	0.47481
	GPT4 (P)	0.70115	0.50391	0.60253	0.41532	0.14611	0.44162
	GPT4 (QA)	0.78532	0.64633	0.71583	0.47354	0.39561	0.57520
(c) GI	GI - T5	0.63169	0.41141	0.52155	0.32029	0.42928	0.44817
	GI - BART	0.64525	0.46689	0.55607	0.40833	0.44438	0.49121
	GI _{SS} - T5	0.28245	0.23667	0.25956	0.25465	0.29370	0.26687
	GI _{SS} - BART	0.17870	0.19743	0.18806	0.15051	0.33684	0.21587

도면2



도면3

