(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2024년11월01일
(11) 등록번호 10-2725657
(24) 등록일자 2024년10월30일

(51) 국제특허분류(Int. Cl.)
G06F 16/332 (2019.01) *G06F 40/289* (2020.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06F 16/3329 (2019.01)
G06F 16/3325 (2019.01)
(21) 출원번호 10-2021-0135410
(22) 출원일자 2021년10월13일
심사청구일자 2021년10월13일
(65) 공개번호 10-2023-0052387
(43) 공개일자 2023년04월20일
(56) 선행기술조사문헌
KR1020200070198 A*
US20210294781 A1
한글문서 기반 QA 생성모델 만들기_part1(2018.12.08.) 1부.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
황선정
경기도 용인시 수지구 고기로 3, 101동 1001호(동천동, 원천마을 푸르지오아파트)
이근배
서울특별시 서초구 서운로 221, 103동 1203호(서초동, 래미안서초스위트아파트)
(74) 대리인
특허법인(유한)아이시스

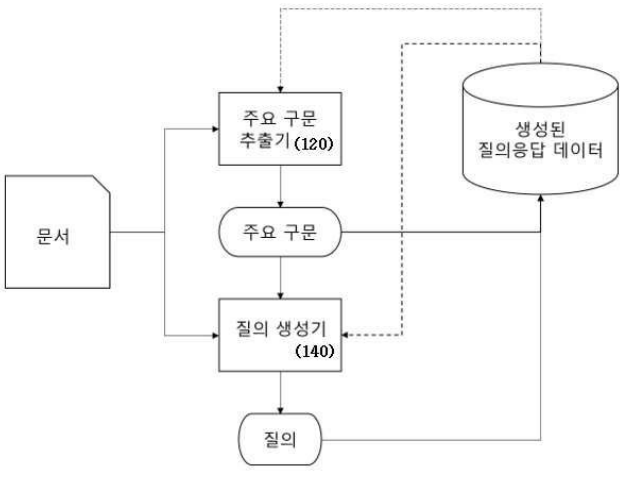
전체 청구항 수 : 총 10 항 심사관 : 안동진

(54) 발명의 명칭 **자동 질의응답 데이터 생성 방법 및 장치**

(57) 요 약

특정 분야를 위한 인공지능 질의응답 시스템 구축을 위해서는 해당 분야의 문서를 기반으로 한 질의-응답 형식의 대량의 데이터가 필요하다. 데이터 구축을 위해서는 해당 분야의 전문가가 필요하며 많은 시간과 비용이 소요된다. 또한 다양한 산업분야에서 자동 고객 응대 시스템이 상용화됨에 따라 대화형 질의응답 시스템의 필요성이 증대되고 있다. 본 특허는 입력된 문서를 토대로 단발성 질의응답 데이터와 대화형 질의응답 데이터를 생성하는 자동 질의응답 데이터 생성 기술에 대한 것이다.

대 표 도 - 도1



자동 질의응답 데이터 생성 시스템 (100)

(52) CPC특허분류
G06F 40/289 (2020.01)
G06N 20/00 (2021.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711126317
과제번호	2020-0-01789-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	High Performance Knowledge System 개발 및 인력양성
기 여 율	1/1
과제수행기관명	동국대학교산학협력단
연구기간	2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

자동 질의응답 데이터를 생성하는 방법에 있어서,

주요 구문 추출기에서, 응답 구문을 추출하는 단계;

질의 생성기에서, 상기 추출된 응답 구문에 대응하는 질의 구문을 생성하는 단계;

상기 응답 구문과 상기 질의 구문을 포함하는 질의응답 데이터를 생성하는 단계;

상기 생성된 질의응답 데이터를 대화 히스토리로 저장하는 단계;

상기 주요 구문 추출기에서, 상기 대화 히스토리를 기초로 새로운 응답 구문을 추출하는 단계;

상기 질의 생성기에서, 상기 대화 히스토리를 기초로 상기 새로운 응답 구문에 대응하는 새로운 질의 구문을 생성하는 단계; 및

상기 새로운 응답 구문과 상기 새로운 질의 구문을 포함하는 대화형 질의응답 데이터를 생성하는 단계를 포함하되,

상기 주요 구문 추출기 및 상기 질의 생성기는 자연어 처리를 하는 신경망 모델로서, 각각 이전 질의-응답 세트에 대한 정보를 더 입력받는 것을 특징으로 하는 방법.

청구항 2

제 1 항에 있어서, 상기 방법은,

상기 주요 구문 추출기에서, 첫 번째 응답 구문을 추출하는 단계;

상기 질의 생성기에서, 상기 첫 번째 응답 구문에 대응하는 첫 번째 질의 구문을 생성하는 단계; 및

상기 첫 번째 응답 구문과 상기 첫 번째 질의 구문을 포함하는 단발성 질의응답 데이터를 생성하는 단계를 더 포함하는 것을 특징으로 하는 방법.

청구항 3

제 2 항에 있어서, 상기 첫 번째 응답 구문을 추출하는 단계는,

상기 주요 구문 추출기에 텍스트 문서를 입력으로 넣는 단계를 더 포함하는 방법.

청구항 4

제 2 항에 있어서, 상기 첫 번째 질의 구문을 생성하는 단계는,

상기 질의 생성기에 텍스트 문서와 상기 첫 번째 응답 구문을 입력으로 넣는 단계를 더 포함하는 방법.

청구항 5

삭제

청구항 6

삭제

청구항 7

제 1 항에 있어서, 상기 새로운 응답 구문을 추출하는 단계는,

상기 주요 구문 추출기에 (i) 텍스트 문서 및 (ii) 상기 저장된 대화 히스토리를 입력으로 넣는 단계를 더 포함

하는 방법.

청구항 8

제 1 항에 있어서, 상기 새로운 질의 구문을 생성하는 단계는,

상기 질의 생성기에 (i) 텍스트 문서, (ii) 상기 저장된 대화 히스토리, 및 (iii) 상기 새로운 응답 구문을 입력으로 넣는 단계를 더 포함하는 방법.

청구항 9

제 1 항에 있어서, 상기 방법은,

상기 주요 구문 추출기에서, 응답 구문 추출을 중단하는 단계를 더 포함하는 방법.

청구항 10

제 9 항에 있어서, 상기 응답 구문 추출을 중단하는 단계는,

상기 주요 구문 추출기에서 추출된 하나 이상의 응답 구문이 상기 대화 히스토리에 포함된 응답 구문과 동일한지 여부를 결정하는 단계를 더 포함하는 방법.

청구항 11

제 1 항에 있어서, 상기 방법은,

상기 주요 구문 추출기가 텍스트 문서에 포함된 주요 구문을 응답 구문으로 추출하도록, 상기 주요 구문 추출기를 학습시키는 단계를 더 포함하는 방법.

청구항 12

제 1 항에 있어서, 상기 방법은,

상기 질의 생성기가 입력된 응답 구문을 답으로 하는 질의 구문을 생성하도록, 상기 질의 생성기를 학습시키는 단계를 더 포함하는 방법.

청구항 13

삭제

청구항 14

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 딥 러닝을 이용한 자동 질의응답 데이터 생성 기술에 대한 것으로, 주어진 문서로부터 자동으로 단발성 질의응답 데이터와 대화형 질의응답 데이터를 생성하는 기술이다.

배 경 기 술

[0002] 최근 딥 러닝 및 기계 학습 관련 기술을 이용한 질의응답 시스템이 개발되어 왔다. 질의응답 시스템은 자연어로 이루어진 질의에 대한 응답을 제공하거나, 자연어로 이루어진 자료(예를 들어, 문서, 음성 파일 들)로부터 질의 및/또는 응답을 제공할 수 있다. 최근 다양한 분야에서 자동 고객 응대 시스템을 도입하고 있으며, 이에 따른 질의응답 시스템 연구의 필요성이 증가하고 있다.

[0003] 대화형 질의응답(CQA: Conversational Question Answering)은 이전까지의 대화 내용을 고려하여 문서로부터 질의에 대한 답변을 찾아내는 자연어처리 기술이다. 챗봇, 대화 시스템, 지능형 가상 에이전트 등의 활용이 증가

함에 따라, CQA 기술의 중요성이 높아지고 있다.

[0004] 딥러닝 기반의 질의응답 시스템은 사전 학습을 통해 신뢰도를 높일 수 있다. 즉, CQA 시스템을 구축하기 위해서는 대량의 학습 데이터 (예를 들어, 대량의 대화형 질의응답 말뭉치 (Conversational QA Corpora))가 사용될 수 있다.

[0005] CQA를 위한 학습 데이터를 생성하는 방법으로서, 대화형 질의 생성(CQG: Conversational Question Generation) 방법이 제시되어 왔다. CQG 방법으로는, 주어진 응답에 대한 질의를 생성하는 answer-aware CQG과, 답변에 대한 단서 없이 문서로부터 유의미한 질의를 생성하는 answer-unaware CQG가 있다. answer-aware CQG는 답변이 존재하는 상태에서 질의를 생성할 수 있다. 따라서, answer-aware CQG를 단독으로 사용하여, CQA 말뭉치를 생성하는 것은 불가능하다.

[0006] 한편, answer-unaware CQG는 사전에 주어진 응답 없이 질의를 생성할 수 있기 때문에, 시스템과 챗봇 개발에 활용될 수 있다. 하지만 응답이 주어지지 않은 상태에서 생성된 질의들은 낮은 정확도를 보였고, 이러한 모델로부터 생성된 데이터는, 많은 오류를 포함하기 때문에, 새로운 CQA 시스템을 훈련시키기에 부적합할 수 있다.

[0007] CQA를 위한 학습 데이터를 생성하는 또 다른 방법으로서, 문서로부터 주요 구문을 병렬적으로 추출한 뒤 이를 기반으로 질의를 생성하는 방법이 연구되어 왔다. 하지만, CQA의 경우 대화 참여자들이 이전 대화 내용에 의존하여 질의응답을 이어나가기 때문에, 이전까지 이루어졌던 대화 내용과 중복되는 질의-응답 쌍을 생성하면 안되며, 이전 대화 내용과 다음에 발생할 질의의 맥락이 이어지도록 하는 응답 구문을 추출해야 한다. 따라서, 주요 구문 추출 방법을 통해 생성된 데이터는 새로운 CQA 시스템을 훈련시키기에 부적합할 수 있다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 한국공개특허 제10-2021-0083731호

발명의 내용

해결하려는 과제

[0009] 특정 분야를 위한 인공지능 질의응답 시스템 구축을 위해서는 해당 분야의 문서를 기반으로 한 질의-응답 형식의 대량의 훈련 데이터가 필요하다. 기존에는 질의응답 데이터 생성을 위해 해당 분야의 전문가들이 관련 문서로부터 질의와 응답을 생성했다. 이러한 방식은 많은 시간과 비용이 든다는 단점이 있다.

[0010] 또한 다양한 산업분야에서 자동 고객 응대 시스템이 상용화됨에 따라 대화형 질의응답 시스템의 필요성이 증대되고 있지만, 현재 활용 가능한 공개 대화형 질의응답 데이터는 한정적인 상태이다.

[0011] 본 특허에서는 딥러닝 기술을 활용하여 자동으로 단발성 질의응답 데이터와 대화형 질의응답 데이터를 생성하는 기술을 제안한다.

과제의 해결 수단

[0012] 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터를 생성하는 방법은, 주요 구문 추출기에서, 응답 구문을 추출하는 단계, 질의 생성기에서, 상기 추출된 응답 구문에 대응하는 질의 구문을 생성하는 단계, 및 상기 응답 구문과 상기 질의 구문을 포함하는 질의응답 데이터를 생성하는 단계를 포함할 수 있다.

[0013] 나아가, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터를 생성하는 방법은, 상기 주요 구문 추출기에서, 첫 번째 응답 구문을 추출하는 단계, 상기 질의 생성기에서, 상기 첫 번째 응답 구문에 대응하는 첫 번째 질의 구문을 생성하는 단계, 및 상기 첫 번째 응답 구문과 상기 첫 번째 질의 구문을 포함하는 단발성 질의응답 데이터를 생성하는 단계를 더 포함할 수 있다.

[0014] 나아가, 상기 첫 번째 응답 구문을 추출하는 단계는, 상기 주요 구문 추출기에 텍스트 문서를 입력으로 넣는 단계를 더 포함할 수 있다.

[0015] 나아가, 상기 첫 번째 질의 구문을 생성하는 단계는, 상기 질의 생성기에 텍스트 문서와 상기 첫 번째 응답 구

문을 입력으로 넣는 단계를 더 포함할 수 있다.

[0016] 나아가, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터를 생성하는 방법은, 상기 생성된 질의응답 데이터를 대화 히스토리로 저장하는 단계를 더 포함할 수 있다.

[0017] 나아가, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터를 생성하는 방법은, 상기 주요 구문 추출기에서, 상기 대화 히스토리를 기초로 새로운 응답 구문을 추출하는 단계, 상기 질의 생성기에서, 상기 대화 히스토리를 기초로 상기 새로운 응답 구문에 대응하는 새로운 질의 구문을 생성하는 단계, 및 상기 새로운 응답 구문과 상기 새로운 질의 구문을 포함하는 대화형 질의응답 데이터를 생성하는 단계를 더 포함할 수 있다.

[0018] 나아가, 상기 새로운 응답 구문을 추출하는 단계는, 상기 주요 구문 추출기에 (i) 텍스트 문서 및 (ii) 상기 저장된 대화 히스토리를 입력으로 넣는 단계를 더 포함할 수 있다.

[0019] 나아가, 상기 새로운 질의 구문을 생성하는 단계는, 상기 질의 생성기에 (i) 텍스트 문서, (ii) 상기 저장된 대화 히스토리, 및 (iii) 상기 새로운 응답 구문을 입력으로 넣는 단계를 더 포함할 수 있다.

[0020] 나아가, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터를 생성하는 방법은, 상기 주요 구문 추출기에서, 응답 구문 추출을 중단하는 단계를 더 포함할 수 있다..

[0021] 나아가, 상기 응답 구문 추출을 중단하는 단계는, 상기 주요 구문 추출기에서 추출된 하나 이상의 응답 구문이 상기 대화 히스토리에 포함된 응답 구문과 동일한지 여부를 결정하는 단계를 더 포함할 수 있다.

[0022] 나아가, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터를 생성하는 방법은, 상기 주요 구문 추출기가 텍스트 문서에 포함된 주요 구문을 응답 구문으로 추출하도록, 상기 주요 구문 추출기를 학습시키는 단계를 더 포함할 수 있다.

[0023] 나아가, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터를 생성하는 방법은, 상기 질의 생성기가 입력된 응답 구문을 답으로 하는 질의 구문을 생성하도록, 상기 질의 생성기를 학습시키는 단계를 더 포함할 수 있다.

[0024] 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터 생성 장치는 프로세서, 및 상기 프로세서와 결합되어 작동되는 메모리를 포함할 수 있다. 상기 프로세서는 응답 구문을 추출하는 단계, 상기 추출된 응답 구문에 대응하는 질의 구문을 생성하는 단계, 및 상기 응답 구문과 상기 질의 구문을 포함하는 질의응답 데이터를 생성하는 단계를 수행하도록 구성될 수 있다.

[0025] 나아가, 상기 프로세서는 상기 생성된 질의응답 데이터를 대화 히스토리로서 상기 메모리에 저장하는 단계를 더 수행하도록 구성될 수 있다.

발명의 효과

[0026] 특정 분야의 인공지능 질의응답 시스템 구축을 위해서는 해당 분야와 관련된 다량의 질의-응답 데이터가 필요하다. 자동 질의응답 데이터 생성 기술을 통해 훈련 데이터 생성을 자동화하여 데이터 생성에 필요한 시간 및 비용을 절감할 수 있다. 또한 다양한 분야에서의 질의응답 시스템 구축을 활성화하여 질의응답 시스템의 상용화에 기여한다.

도면의 간단한 설명

[0027] 도 1은 자동 질의응답 데이터 생성 시스템의 모식도이다.

도 2는 주요 구문 추출기의 구조도이다.

도 3은 본 발명의 몇몇 실시예에 따른, 질의응답 데이터를 생성하는 방법의 흐름도이다.

도 4는 본 발명의 몇몇 실시예에 따른, 단발성 질의응답 데이터를 생성하는 방법의 흐름도이다.

도 5는 본 발명의 몇몇 실시예에 따른, 대화형 질의응답 데이터를 생성하는 방법의 흐름도이다.

도 6은 본 발명의 몇몇 실시예에 따른, 자동 질의응답 생성 장치의 개념도이다.

발명을 실시하기 위한 구체적인 내용

[0028] 이하 첨부된 도면을 참조하여 본 발명의 바람직한 실시 예에 대해 상세히 설명한다.

[0029] 도 1은 자동 질의응답 데이터 생성 시스템의 구조도이다.

[0030] 도 1을 참조하면, 자동 질의응답 데이터 생성 시스템(100)은 문서로부터 주요 구문을 추출하는 주요 구문 추출기(120) 및 추출된 주요 구문에 대한 질의를 생성하는 질의 생성기(140)를 포함할 수 있다. 또한, 자동 질의응답 데이터 생성 시스템(100)은 주요 구문 및 질의를 이용하여 질의응답 데이터를 생성할 수 있다.

[0031] 문서(110)는 질의응답 시스템을 구축하고자 하는 분야의 텍스트 문서일 수 있다. 자동 질의응답 데이터 생성 시스템(100)은 해당 문서로부터 한 쌍 이상의 질의-응답 데이터를 생성 수 있다. 예를 들어, 문서는 도 1에 따른 자동 질의응답 데이터 생성 시스템(100)에 입력으로 입력되어, 한 쌍 이상의 질의-응답 데이터가 출력으로 생성될 수 있다.

[0032] 주요 구문 추출기(120)는 문서(예를 들어, 텍스트 문서)로부터 주요 구문을 추출하는 모듈이다. 예를 들어, 주요 구문 추출기(120)는 문서로부터 임의의 질의에 대한 응답이 될 수 있는 주요 구문을 추출할 수 있다.

[0033] 주요 구문 추출기(120)는 텍스트 문서와 이전까지 생성된 질의-응답 쌍들을 입력으로 받을 수 있다. 주요 구문 추출기(120)는, 이전까지의 질의-응답 쌍들을 통해 대화 맥락을 파악하여, 다음번에 사용자가 흥미를 가질 만한 주요 구문을 텍스트 문서로부터 추출할 수 있다. 이전에 생성된 질의-응답 쌍이 없는 경우, 주요 구문 추출기(120)는 텍스트 문서만을 활용하여 주요 구문을 추출할 수 있다.

[0034] 본 발명의 몇몇 실시예에 따르면, 주요 구문 추출기(120)는 맥락 관련 응답 추출(CAE: Contextual Answer Extraction) 모듈을 포함할 수 있다. 맥락 관련 응답 추출 모듈은 이전 대화 내용을 고려하여 사용자가 다음으로 흥미를 가질 것으로 예측되는 응답 후보를 문서로부터 추출할 수 있다.

[0035] 본 발명의 몇몇 실시예에 따르면, 주요 구문 추출기(120)는 BERT, XLNet, RoBERTa 등의 사전 훈련된 트랜스포머 인코더(Transformer encoder) 구조의 언어 모델을 사용할 수 있다. 또한, 주요 구문 추출기(120)는 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), 트랜스포머(Transformer) 등의 신경망 모델을 사용할 수 있다. 예를 들어, 주요 구문 추출기(120)는 BERT, XLNet, RoBERTa 등의 사전 훈련된 트랜스포머 인코더 구조의 언어 모델을 사용하며, CNN, RNN, Transformer 등의 신경망 모델을 사용하는 모듈일 수 있다.

[0036] 본 발명의 몇몇 실시예에 따르면, 주요 구문 추출기(120)는 이하의 도 2에서 자세히 설명되는 BERT-CAE 모델을 사용할 수 있다.

[0037] 주요 구문은 주요 구문 추출기(120)를 통해 문서로부터 추출된 구문이다. 예를 들어, 주요 구문은, 하나의 단어, 명사구, 형용사구, 부사구 등과 같이, 텍스트 문서 상에 존재하는 구문을 포함할 수 있다.

[0038] 질의 생성기(140)는 주요 구문을 답으로 하는 질의를 생성하는 모듈이다. 질의 생성기(140)는 텍스트 문서, 주요 구문, 이전까지 생성된 질의-응답 쌍들을 입력으로 받을 수 있다. 질의 생성기(140)는 입력된 텍스트 문서와 질의-응답 쌍들을 통해 대화 맥락을 파악하고, 주요 구문을 답으로 하는 대화형 질의를 생성할 수 있다. 이전에 생성된 질의-응답 쌍이 존재하지 않는 경우, 질의 생성기(140)는 텍스트 문서와 주요 구문만을 활용하여 질의를 생성할 수 있다.

[0039] 본 발명의 몇몇 실시예에 따르면, 질의 생성기(140)는 응답(즉, 주요 구문)을 알고 있는 상태에서 질의를 생성할 수 있다. 즉, 질의 생성기(140)는 answer-aware 대화형 질의 생성(CQG: Conversational Question Generation)모듈을 포함할 수 있다.

[0040] 본 발명의 몇몇 실시예에 따르면, 질의 생성기(140)는 T5, BART 등의 사전 훈련된 Transformer 구조의 언어 모델을 사용하는 모듈일 수 있다. 또한, 질의 생성기(140)는 CNN, RNN, Seq2seq, GPT 등의 언어 생성이 가능한 신경망 모델을 사용하는 모듈일 수 있다. 예를 들어, T5, BART 등의 사전 훈련된 Transformer 구조의 언어 모델을 사용하고, CNN, RNN, Seq2seq, GPT 등의 언어 생성이 가능한 신경망 모델을 사용할 수 있다.

[0041] 질의는 질의 생성기(140)로부터 생성된 의문문을 포함할 수 있다. 해당 질의를 생성할 때 이전까지 생성된 질의-응답 쌍이 활용된 경우에, 질의는 이전 대화 내용에 의존하는 간결한 대화 형태를 띌 수 있다.

[0042] 질의응답 데이터는 주요 구문 추출기(120)에서 추출된 주요 구문과 질의 생성기(140)에서 생성된 질의로부터 생성될 수 있다. 즉, 자동 질의응답 데이터 생성 시스템(100)에서, 하나의 텍스트 문서로부터 생성된 주요 구문과 질의는 다음 데이터 생성을 위해 축적될 수 있다. 질의-응답 데이터는 대화형 질의응답 시스템 훈련에 사용할 수 있다. 각 텍스트별로 생성된 첫번째 질의-응답 데이터는 단발성 질의응답 시스템 훈련에 사용될 수 있다.

[0043] 질의응답 데이터는 도 1에 따른 자동 질의응답 데이터 생성 시스템(100)에 저장될 수 있다. 예를 들어, 자동 질

의응답 데이터 생성 시스템(100)은 메모리를 포함할 수 있고, 생성된 질의응답 데이터는 메모리에 저장될 수 있다. 즉, 주요 구문 추출기(120)에서 추출된 '응답'과, 질의 생성기(140)에서 생성된 해당 응답에 대한 '질의'는 한 쌍의 '질의' 및 '응답'의 형태로 메모리에 저장될 수 있다.

[0044] 본 발명의 몇몇 실시예에 따르면, 자동 질의응답 데이터 생성 시스템(100)은 텍스트 문서로부터 단발성 질의응답 데이터를 생성할 수 있다. 구체적으로, 자동 질의응답 데이터 생성 시스템(100)은 주요 구문 추출기(120) 및 질의 생성기(140)를 사용하여, 텍스트 문서로부터 단발성 질의응답 데이터를 생성할 수 있다. 즉, 텍스트 문서에 대해 자동 질의응답 데이터 생성 시스템(100)을 최초로 사용되는 경어, 단발성 질의응답 데이터가 생성될 수 있다. 다시 말하면, 자동 질의응답 데이터 생성 시스템(100)이 저장된 질의응답 데이터를 사용하지 않고, 텍스트 문서로부터 질의응답 데이터를 생성하는 경우, 단발성 질의응답 데이터가 생성될 수 있다.

[0045] 본 발명의 몇몇 실시예에 따르면, 자동 질의응답 데이터 생성 시스템(100)은 텍스트 문서로부터 대화형 질의응답 데이터를 생성할 수 있다. 구체적으로, 자동 질의응답 데이터 생성 시스템(100)은 이전에 생성된 질의응답 데이터를 고려하여, 텍스트 문서로부터 대호형 질의응답 데이터를 생성할 수 있다. 다시 말하면, 자동 질의응답 데이터 생성 시스템(100)은 저장된 질의응답 데이터 및 텍스트 문서를 입력으로 사용하여, 대화형 질의응답 데이터를 생성할 수 있다.

[0046] 본 발명의 몇몇 실시예에 따르면, 자동 질의응답 데이터 생성 시스템(100)은 복수의 텍스트 문서로부터 단발성 질의응답 데이터 및 대화형 질의응답 데이터를 생성할 수 있다.

[0047] 본 발명의 몇몇 실시예에 따르면, 자동 질의응답 데이터 생성 시스템(100)에서 생성된 질의응답 데이터는 대화형 질의응답 말뭉치 (Conversational QA Corpora)로 사용될 수 있다. 즉, 자동 질의응답 데이터 생성 시스템(100)에서 생성된 질의응답 데이터는 질의응답 시스템을 학습시키기 위해 사용될 수 있다.

[0048] 본 발명에 따른 몇몇 실시예에서, 주요 구문 추출기(120)와 질의 생성기(140)의 훈련/실험을 위해, 기존에 공개된 대화형 질의응답 데이터를 각각의 입출력 형태에 맞게 변형하여 사용할 수 있다.

[0049] 도 2는 주요 구문 추출기의 구조도이다.

[0050] 도 2에 개시된 주요 구문 추출기는 도 1의 자동 질의응답 데이터 생성 시스템(100)에서 주요 구문 추출기(120)로 사용될 수 있다.

[0051] 도 2를 참조하면, 주요 구문 추출기는 BERT (Bidirectional Encoder Representations from Transformers)-CAE (Contextual Answer Extraction) 모듈을 포함할 수 있다. 다시 말하면, 주요 구문 추출기는 BERT-CAE 모델에 따라 동작할 수 있다. 예를 들어, BERT-CAE 모델은 질문-무조건 추출 답변 모델(question-unconditional extractive answer model) 구조를 기초로 생성될 수 있다.

[0052] 주요 구문 추출기에서 BERT-CAE 모듈은 대화 히스토리(conversational history)를 고려하여 주요구문 후보(즉, 응답 후보)를 추출할 수 있다. 즉, BERT-CAE 모듈은 사전 훈련된 언어 모델인 BERT를 활용하며, 문서와 대화 히스토리가 입력되었을 때 문서로부터 맥락에 맞는 응답 구문을 추출할 수 있다.

[0053] 따라서, 주요 구문 추출기는 이전까지 이루어졌던 대화 내용과 중복되는 질의-응답 쌍을 생성하지 않고, 이전 대화 내용과 다음에 발생할 질의의 맥락이 이어지도록 하는 응답 구문을 추출해낼 수 있다.

[0054] 주요 구문 추출기는 입력으로 대화 히스토리와 문서를 받아, 출력으로 주요 구문을 생성할 수 있다. 주요 구문 추출기에 입력된 대화 히스토리와 문서는 각각 대화 히스토리 세그먼트 및 문서 세그먼트로 가공(process)되어 BERT-CAE 모듈에 입력될 수 있다.

[0055] 다시 도 2를 참조하면, 주요 구문 추출기에 포함된 BERT-CAE 모듈은 n 개의 토큰으로 이루어진 대화 히스토리의 세그먼트(h1, ..., hn) 및 m 개의 토큰으로 이루어진 문서(document) 세그먼트(d1, ... , dm)을 입력으로 받을 수 있다.

[0056] BERT-CAE 모듈은 대화 히스토리 세그먼트들의 첫 번째 세그먼트의 앞에 <CLS>토큰을, 대화 히스토리 세그먼트들의 마지막 토큰 뒤에 <SEP>토큰을 추가하여, 대화 히스토리 세그먼트와 문서 세그먼트를 구분할 수 있다. 다시 말하면, BERT-CAE 모듈은 <CLS>토큰 및 <SEP>토큰을 사용하여 두 세그먼트가 별개의 세그먼트로 인식할 수 있다.

[0057] 또한, BERT-CAE 모듈은 대화 히스토리를 구성하는 질의-응답 쌍들을 구분하기 위해 스페셜 토큰인 <s>토큰을 질의 세그먼트 앞에 추가하고, 또 다른 스페셜 토큰인 </s>토큰을 응답 세그먼트 앞에 추가할 수 있다. 다시 말하

면, BERT-CAE 모듈은 <s>토큰 및 </s>토큰을 이용하여, 대화 히스토리 세그먼트에 포함된 질의 세그먼트와 응답 세그먼트를 구별할 수 있다.

[0058] 이하에서, 수학식 (1) 내지 (4)를 참조하여 BERT-CAE 모듈의 동작을 설명한다.

$$\text{start logits} = H * W_s + b_s \quad (1)$$

$$\text{start probability} = \text{Softmax}(\text{start logits}) \in R^l \quad (2)$$

$$\text{end logits} = H * W_e + b_e \quad (3)$$

$$\text{end probability} = \text{Softmax}(\text{end logits}) \in R^l \quad (4)$$

[0059]

[0060] 수학식 (1) 내지 (4)에서, H는 마지막 은닉층(hidden layer)의 출력 매트릭스(matrix)를 의미할 수 있다. 또한, W와 b는 각각 훈련 가능한 가중치(Weight)와 편차(bias)를 의미할 수 있다. H, W, 및 b는 행렬 크기는 다음과 같이 표현될 수 있다. 여기서, l은 BERT의 sequence length를, h는 hidden size를 의미한다.

$$H \in R^{l \times h}$$

$$W_s \, , W_e \in R^h$$

$$b_s, b_e \in R^l$$

[0061]

[0062] BERT-CAE 모듈에서, matrix H는 두 dense layer (또는, 완전 연결 계층 (fully connected layer))를 통과하여 각각 시작 로짓 (start logits)과 끝 로짓 (end logits)으로 변환될 수 있다.

[0063] 이후, BERT-CAE 모듈은 시작 로짓과 끝 로짓 각각에 소프트맥스(Softmax) 함수를 적용하여, 각 토큰이 응답 구문의 시작점(start position)과 끝점(end position)이 될 확률을 계산할 수 있다. 예를 들어, BERT-CAE 모듈은 문서 세그먼트(d1, ..., dm) 각각이 시작점(start position) 또는 끝점(end position)이 될 확률을 계산할 수 있다.

[0064] 본 발명의 몇몇 실시예에 따르면, 주요 구문 생성기에 포함된 BERT-CAE 모듈은 위의 수학식 (2) 및 (4)에서의 (start probability + end probability)를 기준으로 가장 확률이 높은 N 개의 응답 후보들을 추출할 수 있다.

[0065] BERT-CAE 모듈은 추출된 응답 후보군에서 이전까지 생성된 질의-응답 쌍에서의 응답 구문과 토큰의 위치가 겹치는 후보들은 제거할 수 있다. 예를 들어, BERT-CAE 모듈은, 대화 히스토리에 포함된 응답 구문을 기초로, 추출된 응답 후보군에서 이전까지 생성된 질의-응답 쌍에서 응답 구문과 토큰의 위치가 겹치는 후보들을 제거할 수 있다. 따라서, 본 발명의 자동 질의응답 데이터 생성 시스템은 같은 내용의 질의-응답 쌍이 생성되는 것을 방지할 수 있다.

[0066] BERT-CAE 모듈은 추출된 응답 후보군에서, 남은 응답 후보 중 확률이 가장 높은 응답 구문을 질의-응답 쌍의 응답 값으로 출력할 수 있다. 다시 말하면, BERT-CAE 모듈에서 출력된 응답 값은 주요 구문 추출기의 주요 구문으로 출력될 수 있다.

[0067] 본 발명의 몇몇 실시예에 따르면, BERT-CAE 모듈은 학습 단계에서 각각의 로짓에 크로스 엔트로피(Cross Entropy) 손실 함수를 적용하여 로스(loss) 값을 계산할 수 있다. 상기 계산된 로스 값을 활용한 역전파를 통해 BERT-CAE 모듈에서 사용되는 BERT-CAE 모델을 훈련시킬 수 있다.

[0068] 다시, 도 1을 참조하면, 주요 구문 추출기(120)는 질의 생성기(140)에 연결될 수 있다. 즉, 주요 구문 추출기(120)에 포함된 BERT-CAE 모듈은, 질의 생성기(140)에 포함된 answer-aware 대화형 질의 생성(CQG) 모듈에 연결될 수 있다.

[0069] 이 경우, 도 1의 자동 질의응답 데이터 생성 시스템(100)은 CAE 모듈과 answer-aware CQG 모듈이 결합된 형태일 수 있다. 즉, 자동 질의응답 데이터 생성 시스템(100)에서, BERT-CAE 모듈 (또는, CAE 모듈)에 문서와 대화 히스토리를 입력하면 해당 문서로부터 주요 구문 (즉, 응답(answer) 구문)이 출력될 수 있다. 이러한, 자동 질의응답 데이터 생성 시스템(100)은 대화형 질의-응답 생성(CQAG: Conversational Question-Answer Generation) 시스템으로 불릴 수 있다.

[0070] 추출된 응답(answer)은 문서, 대화 히스토리(conversational history)와 함께 answer-aware CQG 모듈 (또는, CQG 모듈)에 입력될 수 있다. CQG 모듈은 입력된 응답을 답으로 하는 대화형 질의(question)를 생성할 수 있다.

[0071] 두 모듈로부터 출력된 응답과 질의는 대화 히스토리에 저장되어, 새로운 질의-응답 쌍 생성을 위해 사용될 수 있다.

[0072] 본 발명의 몇몇 실시예에 따르면, 주요 구문 추출기에 포함된 CAE 모듈은 주어진 문서로부터 계속해서 응답 후보를 추출할 수 있다. 따라서, 효율적으로 응답 후보를 추출하기 위해, CAE 모듈은 추출 종료 시점을 지정할 수 있다.

[0073] 예를 들어, CAE 모듈은 가장 확률이 높은 N 개의 응답 후보들이 대화 히스토리에 포함된 이전 질의-응답 쌍의 응답 구문과 중복되는 경우, 주요 구문 추출 동작 또는 응답 추출 동작을 종료할 수 있다. 예를 들어, GAE 모듈에서 추출된 N개의 응답 후보들과, 대화 히스토리에 포함된 응답 구문의 토큰 위치가 모두 동일한 경우, 응답 추출 동작이 종료될 수 있다. 이 경우, 자동 질의응답 데이터 생성 시스템(100) (즉, CQAG 시스템)의 동작이 종료될 수 있다.

[0074] 또 다른 예로, CAE 모듈은 문서의 마지막 k 개의 토큰으로부터 응답 구문을 추출한 뒤, 새로운 질의-응답 쌍을 생성할 수 있다. 이 때, 새로운 질의-응답 쌍의 응답 구문이 상기 마지막 k 개의 토큰보다 앞선 토큰들로부터 추출되는 경우, CAE 모듈은 응답 구문 추출 동작을 종료할 수 있다. 이는 이전 응답보다 앞 쪽에서 응답 구문을 추출하여 질의-응답 쌍을 구성할 경우 대화 히스토리의 맥락으로부터 벗어날 가능성이 높기 때문이다. 이 경우, 자동 질의응답 데이터 생성 시스템(100) (즉, CQAG 시스템)의 동작이 종료될 수 있다.

[0075] 본 발명의 몇몇 실시예에 따른, 자동 질의응답 데이터 생성 시스템(100)은 대화 히스토리가 존재하지 않는 첫 번째 질의-응답 쌍을 생성할 수 있다. 이 경우, 주요 구문 추출기(120) 및 질의 생성기(140)에 저장된 질의응답 데이터가 입력되지 않을 수 있다. 즉, 주요 구문 추출기(120)는 대화 히스토리 없이 문서로부터 주요 구문을 추출하고, 질의 생성기(140)는 대화 히스토리 없이 문서로부터 추출된 주요 구문에 대한 질의를 생성할 수 있다.

[0076] 다시 말하면, 주요 구문 생성기에 포함된 CAE 모듈과 질의 생성기에 포함된CQG 모듈에 대화 히스토리 세그먼트가 입력되지 않을 수 있다. 즉, BERT-CAE 모델의 입력 값은 "<CLS><SEP>document 세그먼트<SEP>"가 될 수 있다.

[0077] 이하에서, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터 생성 시스템에 대해 설명한다. 보다 구체적으로, 본 발명에 따른 자동 질의응답 데이터 생성 시스템을 평가하기 위한 평가 방법이 제시된다.

[0078] CoQA는 문서와 이를 바탕으로 하는 대화형 질의-응답 쌍 들로 구성되어 있는 대화형 질의응답(CQA: Conversational Question Answering) 말뭉치이다. CoQA에서의 응답은 자유-형식 스팬(free-form span), yes, no, unknown의 4 가지 유형으로 분류되고, 각 응답에 대한 문서 내의 근거(rationale)가 함께 제공된다.

[0079] 본 발명에 따른 자동 질의응답 데이터 생성 시스템을 평가하기 위해서, 근거가 자유-형식(free-form)인 응답과의 F1 점수가 가장 높게 측정 되는 구문을 추출하여 응답 구문으로 사용할 수 있고, 수정된 말 뭉치를 CoQA-span으로 지칭할 수 있다.

[0080] CAE 모듈은 문서 내에 존재하는 응답 구문의 시작점(start position)과 끝점(end position)을 예측하는 것을 목적으로 한다. 따라서, yes, no, unknown 유형의 질의-응답 쌍은 CAE의 응답 추출 훈련에 직접적으로 사용되지 않았지만, 정확한 대화 맥락을 활용하기 위해서 대화 히스토리에 포함될 수 있다.

[0081] 위의 표 1 및 표 2에서 BERT-CAE 모델의 평가를 위해 CoQA-span 학습 말뭉치의 10%를 평가 말뭉치로 사용될 수 있다.

[0082] 예를 들어, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터 생성 시스템(즉, CQAG 시스템)은 새로운 CQA 말뭉치 생성을 위해 CoQA와 QuAC에 존재하는 문서들을 활용할 수 있다. 이때, CoQA는 Children's Story, Literature, Mid/High School textbook, News, Wikipedia 등 다양한 도메인의 문서들을 포함한다. QuAC는 Wikipedia 문서만으로 구성되어 있다.

[0083] 다시 표 1을 참조하면, Sequential F1은 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터 생성 시스템을 평가하기 위한 평가 지표이다.

[0084] 먼저, 일반적인 F1 점수(Score)는 아래와 같이 계산될 수 있다.

[0085] $$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

[0086] 여기서, TP는 참 긍정 (true positive)의 개수, FP는 거짓 긍정(false positive)의 개수, 그리고, FN은 거짓 부정(false negative)의 개수일 수 있다. 이때, 참 긍정(True Positive)은 모델이 긍정 클래스(positive class)를 올바르게 예측한 결과이고, 참 부정(True negative)는 모델이 부정 클래스(negative class)를 올바르게 예측한 결과이다. 또한, 거짓 긍정(Fales Positive)는 모델이 긍정 클래스를 잘못 예측한 결과이며, 거짓 부정(False Negative)은 모델이 부정 클래스를 잘못 예측한 결과이다.

[0087] F1 점수는 F Score 또는 F Measure라고 불릴 수 있다. 다시 말해, F1 점수는 정밀도(precision)와 재현율(recall) 사이의 균형(balance)을 나타낼 수 있다.

[0088] 일반적인 응답 추출 (AE: Answer Extraction) 모듈은 문서 단위 또는 문장 단위로 정답의 개수만큼의 응답 구문을 추출하고, 추출된 응답 구문들과 정답 응답 구문들 간의 Soft F1 점수를 측정하여 모델 성능을 평가하였다. 그러나, 본 발명에 따른 자동 질의응답 데이터 생성 시스템(즉, CQA 시스템)의 경우, 질의-응답 쌍들 간의 선후관계가 존재하기 때문에 Soft F1 점수를 CAE 모델 평가에 사용하는 것은 부적절하다.

[0089] 따라서, 본 명세서에서는 새로운 순차적 F1 (Sequential F1) 점수 측정 방법을 제시한다. 순차적 F1 점수 측정 방법에 따르면, 질의-응답 쌍들 간의 선후관계를 고려하여 CAE 모듈의 성능을 평가할 수 있다.

[0090] 예를 들어, 입력 문서에 N 개의 정답 응답 구문들이 존재하는 경우, i 번째 응답(pred_i)을 추출하고자 하는 CAE 모듈의 입출력은 다음과 같다.

[0091] $$\text{pred}_i = \text{CAE}(\text{conversational history}_i,\ \text{document})$$
$$\text{conversational history}_i = Q_{i-k}; A_{i-k}; \ldots; Q_{i-1}; A_{i-1}$$

[0092] i 번째 응답을 추출하기 위해 입력되는 대화 히스토리(conversational history)는 순차적으로 나열된 이전 k 개의 질의응답 쌍을 포함할 수 있다. 자연스러운 대화 맥락을 위해서는, i 번째 질의-응답 쌍을 구성할 응답 구문은 이전까지 추출되었던 응답 구문과 중복되지 않아야 한다. 다만, 현재 대화 히스토리를 고려했을 때 i 번째 정답 응답과는 다른 다양한 질의-응답 쌍이 발생할 수 있다. 예를 들어, 아래의 표 4의 G-CoQA 예시를 참조하면, Q1-A1 다음으로 A2가 아닌 A3 또는 A4가 추출되더라도 대화 맥락에 맞는 질의를 생성할 수 있다.

[0093] 즉, 순차적 F1 점수는 BERT-CAE 모듈에서 i 번째로 추출된 응답 구문 pred_i와 i 번째 이후의 정답 응답 구문들 간의 F1 점수를 측정하기 위한 것으로서, 다음과 같이 정의될 수 있다.

[0094] $$\text{Gold}_i = \{A_i, A_{i+1}, \ldots, A_N\}$$
$$\text{Sequential F1 score of pred}_i = \max_{\text{gold} \in \text{Gold}_i} \text{F1}(\text{pred}_i, \text{gold})$$

[0095] 여기서 Gold_i는 i 번째 이후의 정답 응답 구문들의 집합이다. 즉, pred_i 와 Gold_i의 각 원소 간의 F1 점수를 측정하여 가장 높은 값이 pred_i의 순차적 F1 점수가 된다. 말뭉치 단위의 Sequential F1 점수는 전체 예측 응답 구문들의 순차적 F1 점수의 평균 값이다.

[0096] 본 발명의 몇몇 실시예에 따르면, 주요 구문 추출기에 포함되는 BERT-CAE 모듈을 학습하기 위해, Hugging Face에서 제공하는 사전 훈련 모델인 bert-large-uncased을 활용할 수 있다. 예를 들어, BERT-CAE 모델의 max sequence length는 384, max history length 는 64, conversational history turn의 개수는 2로 설정될 수 있다. 훈련 단계에서의 매개변수인 learning rate, batch size, epoch은 각각 3e-5, 24, 2로 설정될 수 있다. 또한, 질의 생성기에 포함되는 CQG 모듈을 위한 answer-aware CQG 모델로는 T5-large(T5- CQG)가 사용될 수 있다. 이 경우, 본 발명의 자동 질의응답 데이터 생성 시스템(CQAG 시스템)은 훈련된 BERT-CAE 모델과 T5-CQG 모델을 포함할 수 있다.

[0097] 예를 들어, 본 발명의 CQAG 시스템이 생성한 G-CoQA 말뭉치 및 G-QuAC 말뭉치의 타당성을 검증하기 위해서, CoQA 챌린지 리더보드2에 공개된 xlnet-augmentation 모델이 사용될 수 있다.

[0098] 이하에서는, 설명의 편의상 CoQA-span, G-CoQA, G-QuAC의 학습 말뭉치로 훈련된 xlnet-augmentation 모델을 각각 CoQA-span CQA, G-CoQA CQA, G-QuAC CQA로 명명하였다

[0099] 이하에서, 표 1 내지 표 3을 참조하여, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터 생성 시스템에 대해 설명한다.

표 1

[0100]

모델	Sequential F1
AE	27.4
BERT-CAE	58.5

[0101] 표 1은 CoQA-span 말뭉치에 대한 모델 성능 비교 결과를 나타낸다. 구체적으로, 표 1은 질문 무조건 추출 답변 모델(question-unconditional extractive answer model)을 사용하는 일반적인 응답 추출 (AE: Answer Extraction) 모듈과 본 발명에 따른 BERT-CAE 모듈의 CoQA-span 말뭉치에 대한 Sequential F1 점수를 나타낸 것이다.

[0102] 즉, conversational history를 고려하지 않고 응답을 추출한 일반적인 응답 추출 모듈에 비하여, 본 발명의 BERT-CAE 모듈이 31.1 높은 성능을 보였다.

[0103] 이와 같은, 두 모듈의 극명한 성능 차이는 CQA 말뭉치를 위한 응답 추출에 conversational history를 고려하는 것이, 필수 조건이라는 것을 나타낸다. 또한, 본 발명에 따른 BERT-CAE 모듈이 높은 질의 CQA 말뭉치를 생성할 수 있다는 것을 보여준다.

표 2

[0104]

	CoQA	G-CoQA	G-QuAC
질의-응답 개수	15.1	17.8	18.8

[0105] 표 2는 CQA 말뭉치별 문서 단위의 평균 질의-응답 쌍의 개수를 나타낸다. CoQA 말뭉치보다, CQAG 시스템을 통해 생성한 말뭉치들 (G-CoQA, G-QuAC)이 평균적으로 더 많은 질의-응답 쌍을 포함하고 있음을 확인할 수 있다.

[0106] 즉, 본 발명에 따른 CQAG 시스템에 의해 생성된 말뭉치는, CoQA 보다, 문서로부터 더욱 다양한 정보를 추출할 수 있다. 따라서, 본 발명에 따른 CQAG 시스템에 의해 생성된 말뭉치를 사용하는 경우, 대화형 질의응답 시스템 (즉, CQA 시스템)의 질의 대응 능력 향상을 더욱 증가시킬 수 있다.

표 3

[0107]

	평가 말뭉치		
모델	CoQA	G-CoQA	G-QuAC
CoQA-span CQA	68.4	80.3	81.7
G-CoQA CQA	56.2	88.9	88.8
G-QuAC CQA	55.1	87.2	90.1

[0108] 표 3은 CQA 모델의 F1 점수를 나타낸다. 보다 구체적으로, 표 3은 서로 다른 말뭉치로 훈련된 CQA 모델의 각 평

가 말뭉치에 대한 F1 점수를 보여준다. CoQA-span CQA 모델의 각 평가 말뭉치에 대한 F1 점수는 각각 68.4, 80.3, 81.7로서, CoQA-span 평가 말뭉치에 대한 성능이 가장 낮게 측정되었다. 이는, 사람이 직접 구축 및 검증한 말뭉치가 CQAG 시스템을 통해 생성된 말뭉치보다 풀기 어려운 질의를 포함하고 있다고 해석될 수 있다.

[0109] 표 3을 참조하면, 검증된 말뭉치로 훈련된 모델인 CoQA-span CQA의 G-CoQA, G-QuAC 말뭉치에 대한 응답 예측 점수 (즉, 80.3 및 81.7)가 CoQA-span 말뭉치에 대한 점수(즉, 68.4)보다 각각 11.9, 13.3 높다. 이는, 본 발명에 따른 자동 질의응답 데이터 생성 시스템(즉, CQAG 시스템)이 생성한 말뭉치에서, 대화 히스토리, 질의, 응답 사이에 일정한 관계가 있다는 것을 나타낸다. 다시 말하면, 본 발명에 따른 CQAG 시스템이 생성한 말뭉치가 높은 질(quality)의 대화형 질의응답 데이터라는 것을 확인할 수 있다.

[0110] 다시 표 3을 참조하면, G-CoQA CQA와 G-QuAC CQA 모델의 G-CoQA, G-QuAC 평가 말뭉치에 대한 성능이 모델별로 유사한 수치를 보인다. 이는, 하나의 CQA 말뭉치로 훈련된 자동 질의응답 데이터 생성 시스템(즉, CQAG 시스템)은 서로 다른 문서들로부터 유사한 스타일의 CQA 말뭉치를 생성할 수 있다는 것을 나타낸다. 따라서, 본 발명에 몇몇 실시예에 따른, 자동 질의응답 데이터 생성 시스템은 효율적으로 높은 질의 대화형 질의응답 데이터를 생성할 수 있다.

[0111] 이하에서 표 4를 참조하여, 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터 생성 시스템 및 생성된 말뭉치에 대해 설명한다.

표 4

[0112]

Document
Once there was a beautiful fish named Asta. Asta lived in the ocean. There were lots of other fish in the ocean where Asta lived. They played all day long. One day, a bottle floated by over the heads of Asta and his friends. They looked up and saw the bottle. "What is it?" said Asta's friend Sharkie. "It looks like a bird's belly," said Asta. But when they swam closer, it was not a bird's belly. It was hard and clear, and there was something inside it.The bottle floated above them. They wanted to open it. They wanted to see what was inside. So they caught the bottle and carried it down to the bottom of the ocean. They cracked it open on a rock. When they got it open, they found what was inside. It was a note. ...

CoQA	G-CoQA
Q1 : what was the name of the fish	Q1 : What was the name of the fish?
A1 : Asta.	A1 : Asta.
Q2 : What looked like a birds belly	Q2 : Where did Asta live?
A2 : a bottle	A2 : in the ocean
Q3 : who said that	Q3 : What else did he play with?who said that
A3 : Asta.	A3 : lots of other fish
Q4 : Was Sharkie a friend?	Q4 : How long did they play?
A4 : Yes	A4 : all day long
Q5 : did they get the bottle?	Q5 : What did they see?
A5 : Yes	A5 : a bottle
Q6 : What was in it	Q6 : What did it look like?
A6 : a note	A6: a bird's belly
	Q7 : What was it like?
	A7 : hard and clear
	Q8 : What was inside?
	A8 : note

[0113] 표 4은 본 발명의 몇몇 실시예에 따른 자동 질의응답 데이터 생성 시스템(즉, CQAG 시스템)으로 생성한 말뭉치의 일 예시이다.

[0114] 표 4는 문서(document), CoQA 말뭉치, 및 G-CoQA 말뭉치를 포함한다. 본 발명에 몇몇 실시예에 따른 자동 질의응답 데이터 생성 시스템은 표 4의 문서를 입력으로 받아, 표 4의 G-CoQA 말뭉치를 생성할 수 있다.

[0115] 한편, CoQA 말뭉치는 표 4의 문서와 함께 제공된 CQA 말뭉치에서, 자유-형식 스팬(free-form span) 응답을 추출한 CoQA-span 말뭉치이다. CoQA 말뭉치는 본 발명의 자동 질의응답 데이터 생성 시스템에 의해 생성된 G-CoQA와

비교하기 위해 제시되었다.

[0116] 표 4에서, 본 발명의 자동 질의응답 데이터 생성 시스템에 의해 생성된 G-CoQA 말뭉치는 Q1-A1부터 Q8-A8까지의 8개의 질의응답 쌍을 포함한다. 반면에, CoQA 말뭉치는 Q1-A1부터 Q6-A6까지의 6개의 질의응답 쌍을 포함한다. 즉, 본 발명에 따른 자동 질의응답 데이터 생성 시스템은 동일한 문서로부터, 문서와 함께 제공되는 CoQA 말뭉치보다, 더 많은 질의응답 데이터를 생성할 수 있다. 따라서, 본 발명에 따른 자동 질의응답 데이터 생성 시스템은 양질의 질의응답 데이터를 생성할 수 있다.

[0117] 이하, 도 3 내지 도 6를 참조하여, 본 발명의 몇몇 실시예에 따른, 자동 질의응답 생성 장치 및 방법에 대해 설명한다.

[0118] 도 3은 본 발명의 몇몇 실시예에 따른, 질의응답 데이터를 생성하는 방법의 흐름도이다.

[0119] 예를 들어, 도 3의 질의응답 데이터를 생성하는 방법은, 아래의 도 6의 자동 질의응답 데이터 생성 장치에 의해 수행될 수 있다.

[0120] 도 3을 참조하면, 단계 S310에서, 주요 구문 추출기는 응답 구문을 추출할 수 있다. 예를 들어, 주요 구문 추출기는 텍스트 문서로부터 주요 구문을 추출하고, 추출된 주요 구문을 응답 구문으로 결정할 수 있다.

[0121] 예를 들어, 주요 구문 추출기는 텍스트 문서로부터 주요 구문을 추출하고, 추출된 주요 구문이 이전에 추출된 응답 구문과 동일한지 여부를 판단할 수 있다. 추출된 주요 구문이 이전에 추출된 응답 구문과 다른 경우, 주요 구문 추출기는 추출된 주요 구문을 응답 구문으로 결정할 수 있다.

[0122] 예를 들어, 주요 구문 추출기는 추출된 주요 구문이 이전에 추출된 응답 구문과 상이한지 여부를 판단하기 위해, 이전에 생성된 질의응답 데이터(또는, 저장된 대화 히스토리)를 사용할 수 있다.

[0123] 예를 들어, 주요 구문 추출기는 텍스트 문서를 입력으로 받을 수 있다. 즉, 주요 구문 추출기는 텍스트 문서로부터 주요 구문 또는 응답 구문을 추출할 수 있다.

[0124] 예를 들어, 주요 구문 추출기는 (i) 텍스트 문서 및 (ii) 이전에 생성된 질의응답 데이터(즉, 대화 히스토리)를 입력으로 받을 수 있다. 예를 들어, 주요 구문 추출기는 입력된 텍스트 문서로부터 주요 구문을 추출하고, 추출된 주요 구문이 대화 히스토리에 저장된 응답 구문과 동일한지 여부를 판단할 수 있다.

[0125] 또 다른 예로, 주요 구문 추출기는, 대화 히스토리를 고려하여, 텍스트 문서로부터 주요 구문을 추출할 수 있다.

[0126] 또한, 주요 구문 추출기는 도 1 내지 도 2에서 설명된 BERT-CAE 모듈을 포함할 수 있다. 즉, 주요 구문 추출기는 BERT-CAE 모듈로서 동작할 수 있다.

[0127] 단계 S320에서, 질의 생성기는 상기 주요 구문 추출기에서 추출된 응답 구문에 대응하는 질의 구문을 생성할 수 있다. 즉, 질의 생성기는 상기 응답 구문을 답으로 하는 질의 구문을 생성할 수 있다.

[0128] 예를 들어, 질의 생성기는 (i) 텍스트 문서 및 (ii) 단계 S310에서 생성된 응답 구문를 입력으로 받을 수 있다. 즉, 질의 생성기는 텍스트 문서를 고려하여, 입력된 응답 구문에 대한 질의를 생성할 수 있다.

[0129] 예를 들어, 주요 구문 추출기는 (i) 텍스트 문서, (ii) 단계 S310에서 생성된 응답 구문, 및 (iii) 이전에 생성된 질의응답 데이터(즉, 대화 히스토리)를 입력으로 받을 수 있다. 즉, 질의 생성기는 텍스트 문서 및 대화 히스토리를 고려하여, 입력된 응답 구문에 대한 질의를 생성할 수 있다.

[0130] 단계 S330에서, 상기 응답 구문과 상기 질의 구문을 포함하는 질의응답 데이터를 생성할 수 있다. 즉, 생성된 질의응답 데이터는 응답 구문과 질의 구문을 한 쌍으로 포함할 수 있다.

[0131] 본 발명의 몇몇 실시예에 따르면, 단계 S330에서 생성된 질의응답 데이터는 순차적으로 저장될 수 있다. 예를 들어, 상기 생성된 질의응답 데이터는 대화 히스토리로서 저장될 수 있다.

[0132] 다시 말하면, 대화 히스토리에는 복수의 질의 구문 및 응답 구문의 쌍이 포함될 수 있다. 대화 히스토리에 포함된 질의-응답 쌍은 생성된 순서에 따라 순차적으로 저장될 수 있다. 예를 들어, 대화 히스토리에 포함된 각각의 질의-응답 쌍은 생성된 순서를 나타내는 정보를 더 포함할 수 있다. 예를 들어, 대화 히스토리는 제1 질의 구문(Q1)-제1 응답 구문(A1), 제2 질의 구문(Q2)-제2 응답 구문(A2), ... 제n 질의 구문(Qn)-제n 응답 구문(An)의 형식으로 저장될 수 있다.

[0133] 본 발명의 몇몇 실시예에 따르면, 도 3에 따른 자동 질의응답 데이터 생성 방법은 주요 구문 추출기를 학습/훈련시키는 단계를 더 포함할 수 있다. 예를 들어, 상기 주요 구문 추출기가 텍스트 문서에 포함된 주요 구문을 응답 구문으로 추출하도록, 상기 주요 구문 추출기를 학습시킬 수 있다.

[0134] 본 발명의 몇몇 실시예에 따르면, 도 3에 따른 자동 질의응답 데이터 생성 방법은 질의 생성기를 학습/훈련시키는 단계를 더 포함할 수 있다. 예를 들어, 상기 질의 생성기가 입력된 응답 구문을 답으로 하는 질의 구문을 생성하도록, 상기 질의 생성기를 학습시킬 수 있다.

[0135] 본 발명의 몇몇 실시예에 따르면, 도 3에 따른 자동 질의응답 데이터 생성 방법은 상기 주요 구문 추출기의 응답 구문 추출을 중단하는 단계를 더 포함할 수 있다.

[0136] 예를 들어, 상기 주요 구문 추출기에서 추출된 하나 이상의 응답 구문이 상기 대화 히스토리에 포함된 응답 구문과 모두 동일한 경우에, 상기 주요 구문 추출기의 응답 구문 추출하는 단계를 중단시킬 수 있다.

[0137] 이 경우, 동일한 응답 구문이 지속적으로 추출되는 것이 방지되어, 자동 질의응답 데이터 생성 효율을 증가시킬 수 있다.

[0138] 도 4는 본 발명의 몇몇 실시예에 따른, 단발성 질의응답 데이터를 생성하는 방법의 흐름도이다.

[0139] 예를 들어, 도 4의 단발성 질의응답 데이터를 생성하는 방법은, 아래의 도 6의 자동 질의응답 데이터 생성 장치에 의해 수행될 수 있다.

[0140] 도 4을 참조하면, 단계 S410에서, 주요 구문 추출기는 첫 번째 응답 구문을 추출할 수 있다. 예를 들어, 주요 구문 추출기는 입력된 문서로부터 추출된 최초의 주요 구문을 첫 번째 응답 구문으로 결정할 수 있다.

[0141] 예를 들어, 첫 번째 응답 구문을 추출하는 단계는, 상기 주요 구문 추출기에 텍스트 문서를 입력으로 넣는 단계를 더 포함할 수 있다.

[0142] 예를 들어, 단발성 질의응답 데이터를 생성하는 경우, 주요 구문 추출기에 텍스트 문서만을 입력으로 넣고, 대화 히스토리를 입력으로 넣지 않을 수 있다.

[0143] 단계 S420에서, 질의 생성기는 상기 첫 번째 응답 구문에 대응하는 첫 번째 질의 구문을 생성할 수 있다.

[0144] 예를 들어, 상기 첫 번째 질의 구문을 생성하는 단계는, 상기 질의 생성기에 텍스트 문서와 상기 첫 번째 응답 구문을 입력으로 넣는 단계를 더 포함할 수 있다.

[0145] 예를 들어, 단발성 질의응답 데이터를 생성하는 경우, 질의 생성기에 텍스트 문서 및 첫 번째 응답 구문만을 입력으로 넣고, 대화 히스토리를 입력으로 넣지 않을 수 있다.

[0146] 단계 S430에서는, 단계 S410 및 S420에서 추출 및 생성된, 상기 첫 번째 응답 구문과 상기 첫 번째 질의 구문을 포함하는 단발성 질의응답 데이터를 생성할 수 있다.

[0147] 예를 들어, 상기 생성된 단발성 질의응답 데이터는 대화 히스토리로 저장되어, 이하의 도 5에서 설명될 대화형 질의응답 데이터를 생성하기 위해 사용될 수 있다.

[0148] 다시 말하면, 특정 텍스트 문서로부터 질의응답 데이터 생성을 시작하는 경우, 이전에 생성된 대화 히스토리가 없을 수 있다. 즉, 이전에 생성된 대화 히스토리가 없는 경우에 생성된 질의응답 데이터는 단발성 질의응답 데이터로 불릴 수 있다. 이와 같이, 본 발명에 따르면, 저장된 대화 히스토리가 없는 경우에도 질의응답 데이터(즉, 단발성 질의응답 데이터)를 생성할 수 있다.

[0149] 도 5는 본 발명의 몇몇 실시예에 따른, 대화형 질의응답 데이터를 생성하는 방법의 흐름도이다.

[0150] 예를 들어, 도 5의 대화형 질의응답 데이터를 생성하는 방법은, 아래의 도 6의 자동 질의응답 데이터 생성 장치에 의해 수행될 수 있다.

[0151] 도 5을 참조하면, 단계 S510에서, 이전에 생성된 질의응답 데이터를 대화 히스토리로 저장할 수 있다.

[0152] 예를 들어, 대화 히스토리는 도 4에서 생성된 단발성 질의응답 데이터를 포함할 수 있다. 예를 들어, 대화 히스토리는, 이하의 도 5에서 생성된 대화형 질의응답 데이터를 더 포함할 수 있다. 예를 들어, 대화 히스토리는 단발성 질의응답 데이터와 대화형 질의응답 데이터를 더 포함할 수 있다.

[0153] 다시 말하면, 최초 생성된 단발성 질의응답 데이터에, 순차적으로 질의응답 데이터가 추가되는 경우, 대화 히스

토리에 저장된 질의응답 데이터 전체를 대화형 질의응답 데이터로 볼 수 있다.

[0154] 단계 S520에서, 주요 구문 추출기는 저장된 대화 히스토리를 기초로 새로운 응답 구문을 추출할 수 있다.

[0155] 예를 들어, 상기 새로운 응답 구문을 추출하는 단계는, 주요 구문 추출기에 (i) 텍스트 문서 및 (ii) 상기 저장된 대화 히스토리를 입력으로 넣는 단계를 더 포함할 수 있다.

[0156] 단계 S530에서, 질의 생성기는, 상기 대화 히스토리를 기초로 상기 새로운 응답 구문에 대응하는 새로운 질의 구문을 생성할 수 있다.

[0157] 예를 들어, 상기 새로운 질의 구문을 생성하는 단계는, 상기 질의 생성기에 (i) 텍스트 문서, (ii) 상기 저장된 대화 히스토리, 및 (iii) 상기 새로운 응답 구문을 입력으로 넣는 단계를 더 포함할 수 있다.

[0158] 단계 S540에서, 상기 새로운 응답 구문과 상기 새로운 질의 구문을 포함하는 대화형 질의응답 데이터를 생성할 수 있다.

[0159] 대화형 질의응답 데이터는 다시 대화 히스토리로 저장되어, 새로운 대화형 질의응답 데이터를 생성하는데 사용될 수 있다.

[0160] 도 6은 본 발명의 몇몇 실시예에 따른, 자동 질의응답 생성 장치의 개념도이다.

[0161] 도 6을 참조하면, 자동 질의응답 데이터 생성 장치(1000)는 메모리(1200) 및 프로세서(1400)를 포함할 수 있다.

[0162] 메모리(1200) 및 프로세서(1400)는 각각 별도의 칩으로 구현되거나, 하나의 칩을 통해 구현될 수 있다. 메모리(1200) 및 프로세서(1400)는 서로 유기적으로 결합되어 작동될 수 있다. 예를 들어, 프로세서(1400)는 메모리(1200)에 저장된 데이터를 사용할 수 있고, 프로세서(1400)에서 출력된 데이터는 다시 메모리(1200)에 저장될 수 있다. 또한, 메모리(1200)는, 휘발성 및/또는 비휘발성 메모리를 포함할 수 있다. 메모리(1200)는, 프로세서(1400)에 의해 실행되는 명령들(instructions) 또는 프로그램을 저장할 수 있다. 또한, 프로세서(1400)는, 소프트웨어를 구동하여 프로세서(1400)에 연결된 자동 질의응답 데이터 생성 장치(1000)를 제어할 수 있다. 또한, 프로세서(1400)는 본 발명과 관련된 다양한 연산, 처리, 데이터 생성, 가공 등의 동작을 수행할 수 있다.

[0163] 프로세서(1400)는 주요 구문 추출기(120)와 질의 생성기(140)를 포함할 수 있다. 주요 구문 추출기(120)와 질의 생성기(140)는 각각 별개의 모듈로 구현되거나, 하나의 모듈로 통합되어 구현될 수 있다.

[0164] 도 6에 따른 자동 질의응답 데이터 생성 장치(1000)는 도 1 내지 도 5에서 설명된 자동 질의응답 생성 방법을 수행할 수 있다. 예를 들어, 프로세서(1400)에 포함된 주요 구문 추출기(120) 및 질의 생성기(140)는 각각 도 1 내지 도 5에서 상술한 동작을 수행할 수 있다.

[0165] 예를 들어, 본 발명에 따른 자동 질의응답 데이터 생성 장치(1000)는 프로세서(1400) 및 상기 프로세서(1400)와 결합되어 작동되는 메모리(1200)를 포함하는 할 수 있다.

[0166] 상기 프로세서(1400)는 응답 구문을 추출하는 단계, 상기 추출된 응답 구문에 대응하는 질의 구문을 생성하는 단계, 상기 응답 구문과 상기 질의 구문을 포함하는 질의응답 데이터를 생성하는 단계를 수행하도록 구성될 수 있다.

[0167] 또한, 상기 프로세서는, 상기 생성된 질의응답 데이터를 대화 히스토리로서 상기 메모리(1200)에 저장하는 단계를 더 수행하도록 구성될 수 있다.

[0168] 본 명세서의 기술적 특징은 CRM(computer readable medium)을 기초로 구현될 수 있다. 예를 들어, 도 1 내지 도 5을 참조하여 설명된 자동 질의응답 데이터 생성 시스템 및 방법은 CRM(computer readable medium)을 기초로 구현될 수 있다.

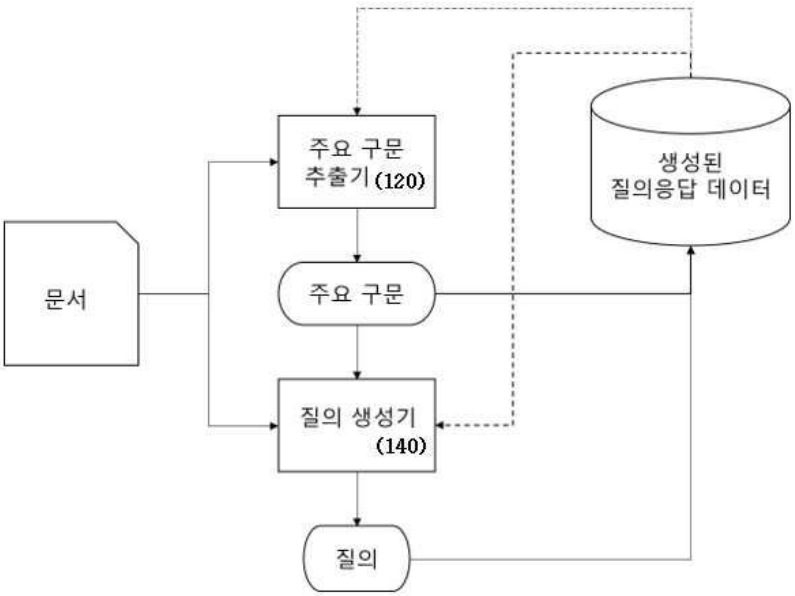
[0169] 예를 들어, 본 명세서에 의해 제안되는 CRM은 프로세서에 의해 실행될 수 있는 명령들(instructions)을 포함할 수 있다. CRM에 저장된 명령들이 프로세서에 의해 실행(execute)되는 경우, 프로세서 또는 프로세서를 포함한 장치들은 특정 동작을 수행할 수 있다. 예를 들어, CRM에 저장된 명령들은 도 6의 프로세서(1400)에 의해 실행되어 자동 질의응답 데이터 생성 장치(1000)가 특정 동작을 수행하게 할 수 있다.

[0170] 예를 들어, 본 발명에 따른 CRM에 저장된 명령들이 실행되는 경우, 프로세서는 응답 구문을 추출하고, 상기 추출된 응답 구문에 대응하는 질의 구문을 생성하고, 상기 응답 구문과 상기 질의 구문을 포함하는 질의응답 데이터를 생성하고, 상기 생성된 질의응답 데이터를 대화 히스토리로서 저장하는 동작을 수행할 수 있다.

[0171] 이상, 첨부된 도면을 참조로 하여 본 발명의 몇몇 실시예를 설명하였다. 그러나, 본 발명이 속하는 기술분야의 통상의 기술자는, 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고, 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이지 않은 것으로 이해되어야 한다.

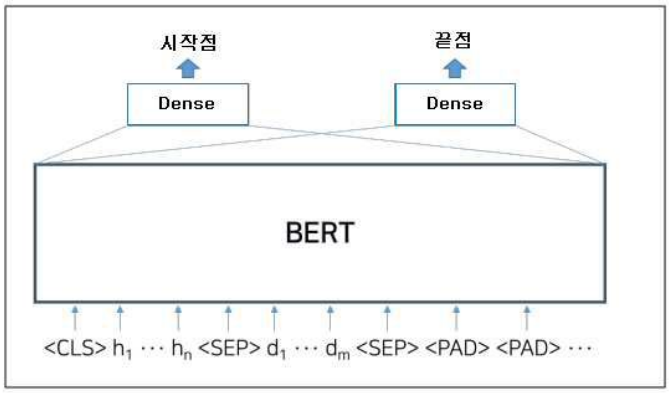
도면

도면1



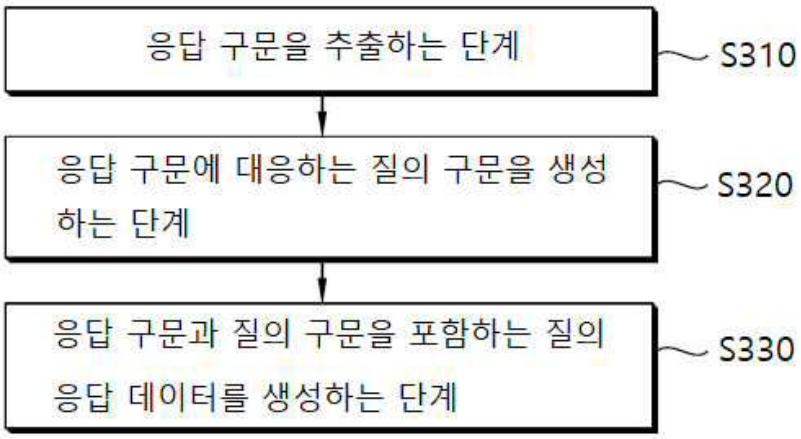
자동 질의응답 데이터 생성 시스템 (100)

도면2

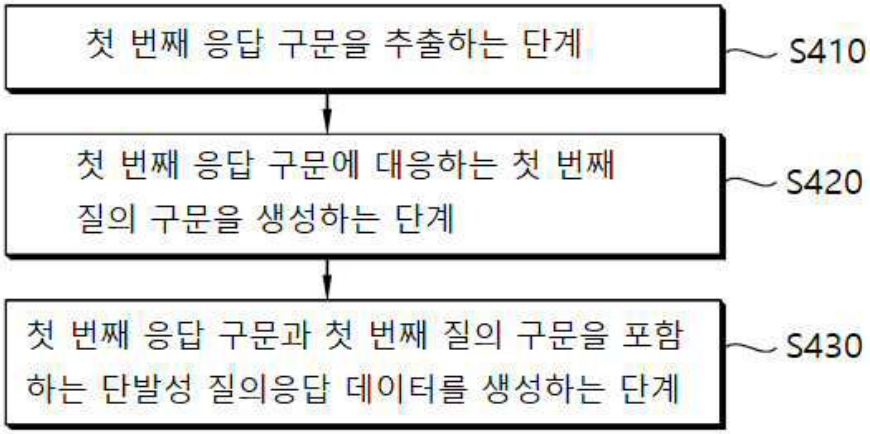


주요 구문 추출기

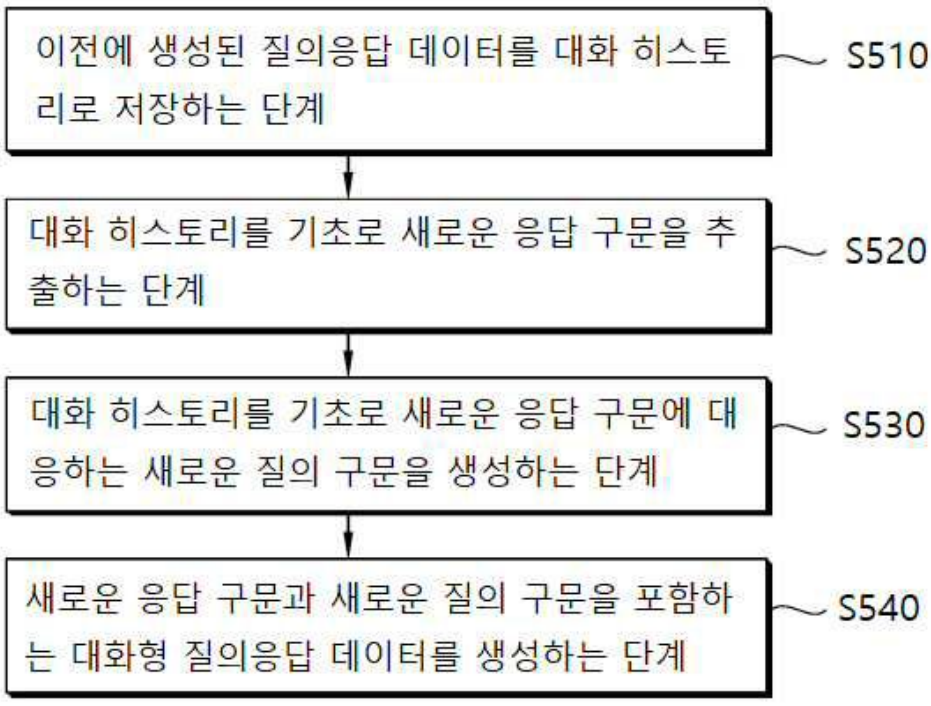
도면3


응답 구문을 추출하는 단계
S310
응답 구문에 대응하는 질의 구문을 생성하는 단계
S320
응답 구문과 질의 구문을 포함하는 질의 응답 데이터를 생성하는 단계
S330

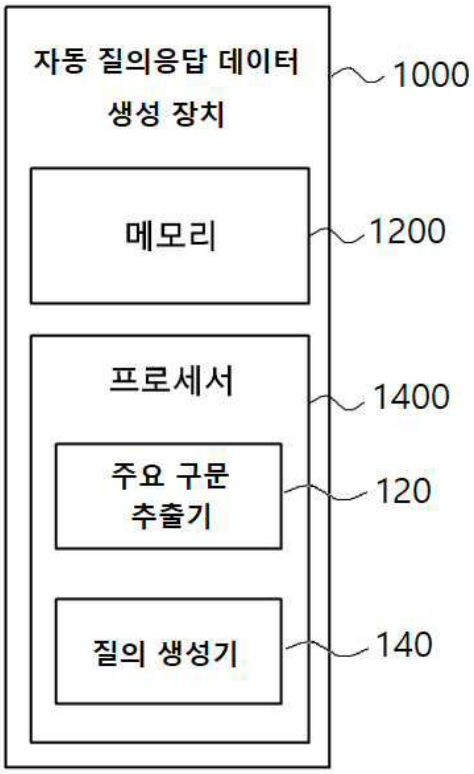
도면4


첫 번째 응답 구문을 추출하는 단계
S410
첫 번째 응답 구문에 대응하는 첫 번째 질의 구문을 생성하는 단계
S420
첫 번째 응답 구문과 첫 번째 질의 구문을 포함하는 단발성 질의응답 데이터를 생성하는 단계
S430

도면5


이전에 생성된 질의응답 데이터를 대화 히스토리로 저장하는 단계
S510
대화 히스토리를 기초로 새로운 응답 구문을 추출하는 단계
S520
대화 히스토리를 기초로 새로운 응답 구문에 대응하는 새로운 질의 구문을 생성하는 단계
S530
새로운 응답 구문과 새로운 질의 구문을 포함하는 대화형 질의응답 데이터를 생성하는 단계
S540

도면6


자동 질의응답 데이터 생성 장치
1000
메모리
1200
프로세서
1400
주요 구문 추출기
120
질의 생성기
140