(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0101216
(43) 공개일자 2024년07월02일

(51) 국제특허분류(Int. Cl.)
G06F 16/332 (2019.01) *G06F 16/33* (2019.01)
G06F 16/36 (2019.01) *G06F 18/2325* (2023.01)
G06N 3/0455 (2023.01)
(52) CPC특허분류
G06F 16/3329 (2019.01)
G06F 16/3347 (2019.01)
(21) 출원번호 10-2022-0183663
(22) 출원일자 2022년12월23일
심사청구일자 2022년12월23일

(71) 출원인
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
이근배
경상북도 포항시 남구 청암로 77
강덕형
경상북도 포항시 남구 청암로 77
(74) 대리인
인비전 특허법인

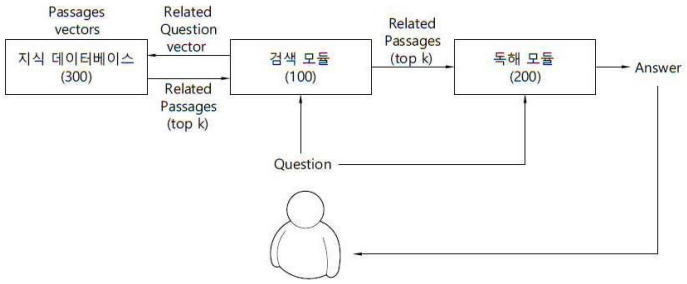
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **오픈 도메인 질의응답 시스템 및 방법**

(57) 요 약

본 발명에 따른 오픈 도메인 질의응답 시스템은 지식 데이터베이스로부터 텍스트 및 테이블 정보를 포함하는 데이터를 획득하며 상기 테이블 정보의 구조적 정보가 상실되지 않도록 상기 테이블의 맥락에 대한 맥락 벡터를 생성하는 맥락 엔코더, 및 외부로부터 제공되는 질의에 대한 데이터를 획득하고 상기 질의에 대한 질의 벡터를 생성하는 질의 엔코더를 포함하는 검색 모듈을 구비하며, 상기 검색 모듈은 상기 맥락 벡터 및 상기 질의 벡터를 동일한 벡터 공간상에 표현하고 상기 질의 백터와 코사인 유사도가 높은 맥락 벡터에 대응되는 맥락을 검색한다.

대 표 도



(52) CPC특허분류
G06F 16/36 (2019.01)
G06F 18/2325 (2023.01)
G06N 3/0455 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711160307
과제번호 2020-0-01789-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성(R&D)
연구과제명 High Pefomace Kowledge System 개발 및 인력양성
기 여 율 1/2
과제수행기관명 동국대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711152953
과제번호 2021-0-00354-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 차세대인공지능핵심원천기술개발(R&D)
연구과제명 비정형 텍스트를 학습하여 쟁점별 사실과 논리적 근거 추론이 가능한 인공지능 원천기술
기 여 율 1/2
과제수행기관명 창원대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

오픈 도메인 질의응답 시스템에 있어서,

지식 데이터베이스로부터 텍스트 및 테이블 정보를 포함하는 데이터를 획득하며 상기 테이블 정보의 구조적 정보가 상실되지 않도록 상기 테이블의 맥락에 대한 맥락 벡터를 생성하는 맥락 엔코더, 및

외부로부터 제공되는 질의에 대한 데이터를 획득하고 상기 질의에 대한 질의 벡터를 생성하는 질의 엔코더를 포함하는 검색 모듈을 구비하며,

상기 검색 모듈은

상기 맥락 벡터 및 상기 질의 벡터를 동일한 벡터 공간상에 표현하고 상기 질의 백터와 코사인 유사도가 높은 맥락 벡터에 대응되는 맥락을 검색하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 2

제1 항에 있어서,

상기 맥락 엔코더 및 상기 질의 엔코더는

BERT(Bidirectional Encoder Representations from Transformers) 엔코더와,

입력 토큰을 수치화하는 임베딩 레이어를 각각 포함하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 3

제2 항에 있어서,

상기 맥락 엔코더의 임베딩 레이어는

상기 테이블의 구조적 정보를 나타낼 수 있는 선형화를 수행하는 열 임베딩(Column embeddings) 레이어를 포함하여,

상기 테이블이 선형화를 거친 이후에는

트랜스포머 기반의 BERT(Bidirectional Encoder Representations from Transformers) 엔코더로 입력되도록 하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 4

제3 항에 있어서,

상기 테이블의 선형화에서, 상기 열 임베딩 레이어는

상기 테이블의 값 부분에 대하여 각 행별로 행에 속하는 값을 ','를 통해 구분하고 각 행은 '▣'을 통해 구분하여 텍스트 형태의 테이블 맥락을 생성하며,

상기 테이블 맥락의 각 토큰에 대하여 해당 토큰이 속하는 열의 인덱스에 대응되는 열 임베딩 값을 반환하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 5

제4 항에 있어서,

상기 테이블의 선형화 이전에 상기 열 임베딩 레이어는

상기 텍스트 및 테이블 정보를 포함하는 데이터를 유사한 길이로 분할하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 6

제3 항에 있어서,

상기 맥락 엔코더의 임베딩 레이어는

행 임베딩(Row embeddings) 레이어를 적용하지 않는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 7

제1 항에 있어서,

상기 검색 모듈에 의해 검색된 질의 벡터와 코사인 유사도가 높은 k개의 맥락과, 상기 질의에 대한 데이터를 제공받아 상기 질의에 대한 응답을 생성하는 독해 모듈을 더 포함하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 8

제7 항에 있어서,

상기 독해 모듈은

상기 k개의 맥락 벡터에 각각 대응하는 k개의 맥락을 입력받아 맥락 표현을 독립적으로 생성하고,

독립적으로 생성된 맥락 표현들을 서로 연결하여 k개의 맥락에 대한 정보를 하나의 백터로 생성하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 9

제7 항에 있어서,

상기 독해 모듈은

FiD(Fusion-in-Decoder)모델을 포함하는 것을 특징으로 하는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 10

제1 항에 있어서,

상기 검색 모듈은

DPR(Dense Passage Retrieval) 모델에 테이블의 구조적 정보를 나타낼 수 있는 임베딩 레이어를 추가하여 구축되는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

청구항 11

오픈 도메인의 질의응답 방법에 있어서,

지식 데이터베이스로부터 텍스트 및 테이블 정보를 포함하는 데이터를 획득하고 맥락 엔코더를 기반으로 상기 테이블 정보의 구조적 정보가 상실되지 않도록 상기 테이블의 맥락에 대한 맥락 벡터를 생성하는 단계;

외부로부터 제공되는 질의에 대한 데이터를 획득하고 질의 엔코더를 기반으로 상기 질의에 대한 질의 벡터를 생성하는 단계; 및

상기 맥락 벡터 및 상기 질의 벡터를 동일한 벡터 공간상에 표현하고 상기 질의 백터와 코사인 유사도가 높은 맥락 벡터에 대응되는 맥락을 검색하는 단계를 포함하는 것을 특징으로 하는 오픈 도메인의 질의응답 방법.

청구항 12

제11 항에 있어서

상기 맥락 엔코더 및 상기 질의 엔코더는

BERT(Bidirectional Encoder Representations from Transformers) 엔코더와,

입력 토큰을 수치화하는 임베딩 레이어를 각각 포함하는 것을 특징으로 하는 오픈 도메인의 질의응답 방법.

청구항 13

제12 항에 있어서,

상기 맥락 엔코더의 임베딩 레이어는

상기 테이블의 구조적 정보를 나타낼 수 있는 선형화를 수행하는 열 임베딩(Column embeddings) 레이어를 포함하여,

상기 테이블이 선형화를 거친 이후에는

트랜스포머 기반의 BERT(Bidirectional Encoder Representations from Transformers) 엔코더로 입력되도록 하는 것을 특징으로 하는 오픈 도메인 질의응답 방법.

청구항 14

제13 항에 있어서,

상기 테이블을 선형화하는 단계는

상기 테이블의 값 부분에 대하여 각 행별로 행에 속하는 값을 ','를 통해 구분하고 각 행은 '◙'을 통해 구분하여 텍스트 형태의 테이블 맥락을 생성하는 단계와,

상기 테이블 맥락의 각 토큰에 대하여 해당 토큰이 속하는 열의 인덱스에 대응되는 열 임베딩 값을 반환하는 단계를 포함하는 것을 특징으로 하는 오픈 도메인 질의응답 방법.

청구항 15

제14 항에 있어서,

상기 테이블의 선형화하는 단계 이전에

상기 텍스트 및 테이블 정보를 포함하는 데이터를 유사한 길이로 분할하는 단계를 더 포함하는 것을 특징으로 하는 오픈 도메인 질의응답 방법.

청구항 16

제13 항에 있어서,

상기 맥락 엔코더의 임베딩 레이어는

행 임베딩(Row embeddings) 레이어를 적용하지 않는 것을 특징으로 하는 오픈 도메인 질의응답 방법.

청구항 17

제11 항에 있어서,

상기 검색하는 단계하는 단계 이후에,

상기 검색 모듈에 의해 검색된 질의 벡터와 코사인 유사도가 높은 k개의 상기 맥락 벡터에 대응하는 k개의 맥락과, 상기 질의에 대한 데이터가 제공받아 상기 질의에 대한 응답을 생성하는 단계를 더 포함하는 것을 특징으로 하는 오픈 도메인 질의응답 방법.

청구항 18

제17 항에 있어서,

상기 질의에 대한 응답을 생성하는 단계는

상기 k개의 맥락 벡터를 입력받아 맥락 표현을 독립적으로 생성하고,

독립적으로 생성된 맥락 표현들을 서로 연결하여 k개의 맥락에 대한 정보를 하나의 백터로 생성하는 것을 특징으로 하는 오픈 도메인 질의응답 방법.

청구항 19

제17 항에 있어서,

상기 질의에 대한 응답을 생성하는 단계에서는

FiD(Fusion-in-Decoder)모델을 포함하는 독해 모듈이 적용되는 것을 특징으로 하는 것을 특징으로 하는 오픈 도메인 질의응답 방법.

청구항 20

제11 항에 있어서,

상기 맥락 벡터 및 상기 질의 벡터를 생성하는 단계에서는

DPR(Dense Passage Retrieval) 모델에 테이블의 구조적 정보를 나타낼 수 있는 임베딩 레이어를 추가하여 구축되는 검색 모듈이 적용되는 것을 특징으로 하는 오픈 도메인 질의응답 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 질의응답 시스템 및 방법에 관한 것으로, 보다 상세하게는 주어진 질의에 대한 대답을 제공하는 오픈 도메인 질의응답 시스템 및 방법에 관한 것이다.

[0002]

배 경 기 술

[0003] 일반적으로 오픈 도메인 질의응답 시스템은 주어진 질문에 대응하는 문서를 검색하고, 검색된 문서로부터 적합한 응답을 찾아 제공하는 태스크를 의미한다. 이에, 다양한 산업 분야에서는 오픈 도메인 질의응답 시스템을 이용하여, 시스템에 접근하는 사용자를 보조 및 지원하기 위한 기술에 대한 연구개발을 활발하게 수행하고 있다.

[0004] 종래의 오픈 도메인 질의응답 시스템에 기술은 "대한민국 공개특허공보 제10-2022-0022701호(지식 그래프 추론 기반의 오픈 도메인 질문 응답 시스템, 2022.02.28.)"에 의해 공개되어 있다. 상기 공개발명은 자연어 질문과 연관된 지식 베이스를 통해 질문에 대한 답변을 추론하는 것을 특징으로 한다.

[0005] 종래의 질의응답 기술은 대부분 문서의 텍스트 부분만을 활용하여 질의에 대한 답변을 도출하는 기술을 주로 연구하고 있다. 그러나 최근에는 문서 내의 텍스트 정보뿐만 아니라 테이블 부분을 활용하여 답변을 도출하기 위한 테이블-텍스트 오픈 도메인 질의응답 기술이 연구 및 개발되고 있다. 이러한 테이블-텍스트 오픈 도메인 질의응답 기술은 시스템이 테이블을 이해할 때에, 테이블의 값이 어떠한 행과 열에 속하는지 알려주는 구조적 정보가 중요하다. 그러나 종래의 테이블-텍스트 오픈 도메인 질의응답 기술은 테이블을 텍스트 형태로 변환하는 과정에서 테이블의 구조적 정보가 손실되는 문제점이 있었다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) "대한민국 공개특허공보 제10-2022-0022701호(지식 그래프 추론 기반의 오픈 도메인 질문 응답 시스템, 2022.02.28.)"

발명의 내용

해결하려는 과제

[0007] 본 발명의 목적은 테이블의 구조적 정보를 명시적으로 활용하지 않는 문제점을 해결하여 테이블과 텍스트 정보를 모두 오픈 도메인의 질의응답의 맥락으로 활용할 수 있는 오픈 도메인 질의응답 시스템 및 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0008] 본 발명에 따른 오픈 도메인 질의응답 시스템은 지식 데이터베이스로부터 텍스트 및 테이블 정보를 포함하는 데이터를 획득하며 상기 테이블 정보의 구조적 정보가 상실되지 않도록 상기 테이블의 맥락에 대한 맥락 벡터를 생성하는 맥락 엔코더, 및 외부로부터 제공되는 질의에 대한 데이터를 획득하고 상기 질의에 대한 질의 벡터를 생성하는 질의 엔코더를 포함하는 검색 모듈을 구비하며, 상기 검색 모듈은 상기 맥락 벡터 및 상기 질의 벡터를 동일한 벡터 공간상에 표현하고 상기 질의 백터와 코사인 유사도가 높은 맥락 벡터에 대응되는 맥락을 검색한다.

[0009] 상기 맥락 엔코더 및 상기 질의 엔코더는 BERT(Bidirectional Encoder Representations from Transformers) 엔코더와, 입력 토큰을 수치화하는 임베딩 레이어를 각각 포함할 수 있다.

[0010] 상기 맥락 엔코더의 임베딩 레이어는 상기 테이블의 구조적 정보를 나타낼 수 있는 선형화를 수행하는 열 임베딩(Column embeddings) 레이어를 포함하여, 상기 테이블이 선형화를 거친 이후에는 트랜스포머 기반의 BERT(Bidirectional Encoder Representations from Transformers) 엔코더로 입력되도록 할 수 있다.

[0011] 상기 테이블의 선형화에서, 상기 열 임베딩 레이어는 상기 테이블의 값 부분에 대하여 각 행별로 행에 속하는 값을 ','를 통해 구분하고 각 행은 '▣'을 통해 구분하여 텍스트 형태의 테이블 맥락을 생성하며, 상기 테이블 맥락의 각 토큰에 대하여 해당 토큰이 속하는 열의 인덱스에 대응되는 열 임베딩 값을 반환할 수 있다.

[0012] 상기 테이블의 선형화 이전에 상기 열 임베딩 레이어는 상기 텍스트 및 테이블 정보를 포함하는 데이터를 유사한 길이로 분할할 수 있다.

[0013] 상기 맥락 엔코더의 임베딩 레이어는 행 임베딩(Row embeddings) 레이어를 적용하지 않을 수 있다.

[0014] 상기 오픈 도메인 질의응답 시스템은 상기 검색 모듈에 의해 검색된 질의 벡터와 코사인 유사도가 높은 k개의 맥락과, 상기 질의에 대한 데이터를 제공받아 상기 질의에 대한 응답을 생성하는 독해 모듈을 더 포함할 수 있

다.

[0015] 상기 독해 모듈은 상기 k개의 맥락 벡터에 각각 대응하는 k개의 맥락을 입력받아 맥락 표현을 독립적으로 생성하고, 독립적으로 생성된 맥락 표현들을 서로 연결하여 k개의 맥락에 대한 정보를 하나의 백터로 생성할 수 있다.

[0016] 상기 독해 모듈은 FiD(Fusion-in-Decoder)모델을 포함할 수 있다.

[0017] 상기 검색 모듈은 DPR(Dense Passage Retrieval) 모델에 테이블의 구조적 정보를 나타낼 수 있는 임베딩 레이어를 추가하여 구축될 수 있다.

[0018] 한편, 본 발명에 따른 오픈 도메인의 질의응답 방법은 지식 데이터베이스로부터 텍스트 및 테이블 정보를 포함하는 데이터를 획득하고 맥락 엔코더를 기반으로 상기 테이블 정보의 구조적 정보가 상실되지 않도록 상기 테이블의 맥락에 대한 맥락 벡터를 생성하는 단계 및 외부로부터 제공되는 질의에 대한 데이터를 획득하고 질의 엔코더를 기반으로 상기 질의에 대한 질의 벡터를 생성하는 단계 및 상기 맥락 벡터 및 상기 질의 벡터를 동일한 벡터 공간상에 표현하고 상기 질의 백터와 코사인 유사도가 높은 맥락 벡터에 대응되는 맥락을 검색하는 단계를 포함한다.

[0019] 상기 맥락 엔코더 및 상기 질의 엔코더는 BERT(Bidirectional Encoder Representations from Transformers) 엔코더와, 입력 토큰을 수치화하는 임베딩 레이어를 각각 포함할 수 있다.

[0020] 상기 맥락 엔코더의 임베딩 레이어는 상기 테이블의 구조적 정보를 나타낼 수 있는 선형화를 수행하는 열 임베딩(Column embeddings) 레이어를 포함하여, 상기 테이블이 선형화를 거친 이후에는 트랜스포머 기반의 BERT(Bidirectional Encoder Representations from Transformers) 엔코더로 입력되도록 할 수 있다.

[0021] 상기 테이블을 선형화하는 단계는 상기 테이블의 값 부분에 대하여 각 행별로 행에 속하는 값을 ','를 통해 구분하고 각 행은 '[]'을 통해 구분하여 텍스트 형태의 테이블 맥락을 생성하는 단계와, 상기 테이블 맥락의 각 토큰에 대하여 해당 토큰이 속하는 열의 인덱스에 대응되는 열 임베딩 값을 반환하는 단계를 포함할 수 있다.

[0022] 상기 오픈 도메인 질의응답 방법은 상기 테이블의 선형화하는 단계 이전에, 상기 텍스트 및 테이블 정보를 포함하는 데이터를 유사한 길이로 분할하는 단계를 더 포함할 수 있다.

[0023] 상기 맥락 엔코더의 임베딩 레이어는 행 임베딩(Row embeddings) 레이어를 적용하지 않을 수 있다.

[0024] 상기 오픈 도메인 질의응답 방법은 상기 검색하는 단계하는 단계 이후에, 상기 검색 모듈에 의해 검색된 질의 벡터와 코사인 유사도가 높은 k개의 상기 맥락 벡터에 대응하는 k개의 맥락과, 상기 질의에 대한 데이터가 제공받아 상기 질의에 대한 응답을 생성하는 단계를 더 포함할 수 있다.

[0025] 상기 질의에 대한 응답을 생성하는 단계는 상기 k개의 맥락 벡터를 입력받아 맥락 표현을 독립적으로 생성하고, 독립적으로 생성된 맥락 표현들을 서로 연결하여 k개의 맥락에 대한 정보를 하나의 백터로 생성할 수 있다.

[0026] 상기 질의에 대한 응답을 생성하는 단계에서는 FiD(Fusion-in-Decoder)모델을 포함하는 독해 모듈이 적용될 수 있다.

[0027] 상기 맥락 벡터 및 상기 질의 벡터를 생성하는 단계에서는 DPR(Dense Passage Retrieval) 모델에 테이블의 구조적 정보를 나타낼 수 있는 임베딩 레이어를 추가하여 구축되는 검색 모듈이 적용될 수 있다.

발명의 효과

[0028] 본 발명에 따른 오픈 도메인 질의응답 시스템 및 방법은 종래의 텍스트 정보의 활용뿐만 아니라 테이블의 구조적 정보를 활용한 테이블 정보를 활용하여, 질의에 대한 응답을 보다 효과적으로 도출하는 효과가 있다.

[0029] 이상과 같은 본 발명의 기술적 효과는 이상에서 언급한 효과로 제한되지 않으며, 언급되지 않은 또 다른 기술적 효과들은 아래의 기재로부터 당업자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0030] 도 1은 본 실시예에 따른 오픈 도메인 질의응답 시스템을 개략적으로 나타낸 개념도이고,

도 2는 본 실시예에 따른 오픈 도메인 질의응답 시스템의 검색 모듈의 구조를 개략적으로 나타낸 개념도이고,

도 3은 본 실시예에 따른 오픈 도메인 질의응답 시스템의 맥락 엔코더 및 질의 엔코더의 구조를 개략적을 나타

낸 개념도이고,

도 4는 본 실시예에 따른 오픈 도메인 질의응답 시스템의 맥락 엔코더에서 테이블이 임베딩 레이어를 거쳐 수치화되는 프로세스를 나타낸 개념도이다.

발명을 실시하기 위한 구체적인 내용

[0031] 이하 첨부된 도면을 참조하여 본 발명의 실시예를 상세히 설명한다. 그러나 본 실시예는 이하에서 개시되는 실시예에 한정되는 것이 아니라 서로 다양한 형태로 구현될 수 있으며, 단지 본 실시예는 본 발명의 개시가 완전하도록 하며, 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이다. 도면에서의 요소의 형상 등은 보다 명확한 설명을 위하여 과장되게 표현된 부분이 있을 수 있으며, 도면 상에서 동일 부호로 표시된 요소는 동일 요소를 의미한다.

[0032] 도 1은 본 실시예에 따른 오픈 도메인 질의응답 시스템을 개략적을 나타낸 개념도이고, 도 2는 본 실시예에 따른 질의응답 시스템의 검색 모듈의 구조를 개략적으로 나타낸 개념도이다. 그리고 도 3은 본 실시예에 따른 오픈 도메인 질의응답 시스템의 맥락 엔코더 및 질의 엔코더의 구조를 개략적을 나타낸 개념도이다.

[0033] 도 1 내지 도 3에 도시된 바와 같이, 본 실시예에 따른 오픈 도메인 질의응답 시스템(1000, 이하, 질의응답 시스템이라 칭한다.)은 특정 도메인에 한정되지 않고, 질의응답 시스템(1000)으로 접근하는 사용자로부터 제공되는 질의에 대한 응답을 도출하여 사용자에게 전달할 수 있다.

[0034] 이러한 질의응답 시스템(1000)은 검색 모듈(100) 및 독해 모듈(200)을 포함하여, 지식 데이터베이스(300)로부터 질의에 대한 응답을 도출할 수 있다.

[0035] 여기서, 질의응답 시스템(1000)은 지식 데이터베이스(300)로부터 문서를 획득하거나 지식 데이터베이스에 액세스(Access) 가능하게 마련될 수 있다. 여기서, 지식 데이터베이스(300)는 질의에 대한 응답을 도출하기 위한 문서 데이터베이스로, 텍스트 정보 및 테이블 정보를 포함할 수 있다. 일례로, 지식 데이터베이스는 Wikipedia, Book corpus, News, CommonCrawl 및 Reddit 등을 포함할 수 있다. 그러나 이는 본 실시예를 설명하기 위한 것으로 지식 데이터베이스(300)의 종류는 한정하지 않는다.

[0036] 또한, 지식 데이터베이스(300)는 상술된 데이터베이스 자체를 의미하거나, 지식 데이터베이스(300)로부터 획득된 데이터를 재가공한 별도의 데이터베이스일 수 있다.

[0037] 한편, 검색 모듈(100)은 질의응답 시스템(1000)에 접근하는 사용자로부터 질의에 대한 정보가 제공될 수 있다. 검색 모듈(100)은 사전에 훈련된 모델로, 사용자의 질의가 입력되면 질의와 관련성이 높은 텍스트 맥락 및 테이블 맥락을 지식 데이터베이스(300)로부터 검색할 수 있다.

[0038] 그리고 지식 데이터베이스(300)로부터 검색된 텍스트 맥락 및 테이블 맥락은 독해 모듈(200)로 제공된다. 이에, 독해 모듈(200)은 사용자로부터 제공되는 질의에 대한 정보를 기반으로 검색된 텍스트 맥락 및 테이블 맥락으로부터 정답을 추출하게 된다.

[0039] 이에, 질의응답 시스템(1000)은 추출된 정답을 재구성하여 질의를 제공한 사용자에게 질의에 대한 응답을 제공할 수 있다.

[0040] 한편, 검색 모듈(100)은 DPR(Dense Passage Retrieval) 모델에 테이블의 구조적 정보를 나타낼 수 있는 임베딩 레이어를 추가하여 구축될 수 있다.

[0041] 일반적으로, DPR 모델은 두 개의 BERT(Bidirectional Encoder Representations from Transformers) 엔코더로 구성될 수 있다. 본 실시예에서는 일례로, 두 개 BERT 엔코더가 맥락 엔코더(110, Passage encoder) 및 질의 엔코더(120, Question encoder)를 포함할 수 있다.

[0042] 맥락 엔코더(110)는 학습 과정에서 지식 데이터베이스(300)로부터 텍스트 맥락 및 테이블 맥락을 제공받을 수 있다. 이에, 맥락 엔코더(110)는 텍스트 맥락 및 테이블 맥락을 고차원의 벡터로 변환할 수 있다. 여기서, 고차원의 벡터는 맥락 엔코더(110)로 입력되는 입력 토큰에 대응하는 출력 벡터이다.

[0043] 이러한 맥락 엔코더(110)는 입력 토큰을 임베딩 레이어를 통해 수치화한다. 그리고 수치화된 입력 토큰은 트랜스포머 기반의 BERT 엔코더를 통과하게 된다. 이에, BERT 엔코더는 맥락 엔코더(110)로 입력되는 입력 토큰에 대응하는 고차원의 벡터를 생성할 수 있다.

[0044] 여기서, 맥락 엔코더(110)의 임베딩 레이어는 토큰 임베딩(Token embeddings), 분할 임베딩(Segment

embeddings), 포지션 임베딩(Position embeddind) 및 열 임베딩(Column embeddings) 레이어들을 포함할 수 있다.

[0045] 토큰 임베딩 레이어는 입력 토큰에 대응하는 토큰 ID를 찾고, 토큰 ID에 대응하는 고차원의 임베딩 벡터를 반환한다. 그리고 모델에 입력되는 맥락은 겹치지 않는 단위인 세그멘트(Segment)로 분할될 수 있는데, 분할 임베딩 레이어는 입력 토큰에 대응하는 세스멘트 ID를 찾아 이에 대응되는 고차원의 임베딩 벡터를 반환한다. 또한, 포지션 임베딩 레이어는 해당 토큰의 맥락 상에서의 위치 정보에 대응되는 고차원 임베딩 벡터를 반환한다, 그리고 열 임베딩 레이어는 해당 토큰이 테이블 맥락의 일부인 경우에, 해당 토큰이 테이블 상에서 대응하는 열의 위치에 대응하는 고차원 임베딩 벡터를 반환한다.

[0046] 맥락 엔코더(110)는 임베딩 레이어를 거치며 얻어진 벡터들을 더함으로서 토큰이 임베딩 벡터들의 합으로 수치화되도록 한다. 그리고 수치화된 벡터는 맥락 엔코더(110)의 이후 레이어를 통과하며 고차원의 벡터로 표현될 수 있다.

[0047] 그리고 맥락 엔코더(110)는 학습 과정에서 지식 데이터베이스(300)로부터 취득된 텍스트 맥락 및 테이블 맥락을 고차원의 벡터로 변환하고, 변환된 고차원의 벡터를 벡터 공간 상에 표현할 수 있다.

[0048] 그리고, 질의 엔코더(120)는 질의응답 시스템(1000)의 운용에서 사용자로부터 질의가 제공될 수 있다. 이에, 질의 엔코더(120)는 사용자로부터 제공되는 질의를 고차원의 벡터로 변환할 수 있다. 여기서, 고차원의 벡터는 질의 엔코더(120)로 입력되는 입력 토큰에 대응하는 출력 벡터이다.

[0049] 이러한 질의 엔코더(120)는 입력 토큰을 임베딩 레이어를 통해 수치화한다. 그리고 수치화된 입력 토큰은 트랜스포머 기반의 BERT 엔코더를 통과하게 된다. 이에, BERT 엔코더는 질의 엔코더로 입력되는 입력 토큰에 대응하는 고차원의 벡터를 생성할 수 있다.

[0050] 여기서, 질의 엔코더(120)의 임베딩 레이어는 토큰 임베딩(Token embeddings), 분할 임베딩(Segment embeddings) 및 포지션 임베딩(Position embeddind) 레이어들을 포함할 수 있다.

[0051] 이에, 질의 엔코더(120)는 질의응답 시스템(1000)의 운용에서 학습자로부터 취득된 질의를 고차원의 벡터로 변환하고, 변환된 고차원의 벡터를 벡터 공간 상에 표현할 수 있다.

[0052] 이때, 학습 과정에서 생성된 맥락 벡터 및 운용 과정에서 생성된 질의 벡터는 모두 같은 공간 상에 표현될 수 있다.

[0053] 이에, 검색 모듈(100)은 질의 벡터와 코사인 유사도가 높은 k개의 맥락 벡터에 각각 대응하는 k개의 맥락을 검색하고, 검색된 맥락을 결과로서 독해 모듈(200)로 제공할 수 있다.

[0054] 한편, 독해 모듈(200)은 운용 과정에서 사용자로부터 제공되는 질의에 대한 정보와, 검색 모듈(100)로부터 제공되는 질의 벡터와 코사인 유사도가 높은 k개의 맥락 벡터 각각 대응하는 k개의 맥락을 수신한다.

[0055] 이에, 독해 모듈(200)은 사용자로부터 제공되는 질의와, 검색 모듈에서 제공된 k개의 맥락을 기반으로 질의에 대한 정답을 생성한다.

[0056] 이러한 FiD 방식의 독해 모듈(200)은 엔코더와 디코더로 구성될 수 있다. 여기서, 엔코더는 검색 모듈(100)로부터 제공되는 k개의 맥락를 입력받아 맥락 표현을 독립적으로 생성할 수 있다.

[0057] 여기서, 독립적으로 생성된 맥락 표현들은 서로 연결되어 k개의 맥락에 대한 정보를 포함하는 하나의 벡터로 생성될 수 있다. 그리고 디코더는 엔코더에서 생성된 하나의 벡터를 입력으로 받아, 질의에 대한 정답을 생성할 수 있다.

[0058] 한편, 이하에서는 맥락 엔코더의 임베딩 레이어에 포함되는 열 임베딩 레이어에 대하여 상세히 설명하도록 한다. 다만, 상술된 구성요소에 대해서는 상세한 설명을 생략하고 동일한 참조부호를 부여하여 설명하도록 한다.

[0059] 도 4는 본 실시예에 따른 오픈 도메인 질의응답 시스템의 맥락 엔코더에서 테이블이 임베딩 레이어를 거쳐 수치화되는 프로세스를 나타낸 개념도이다.

[0060] 도 4에 도시된 바와 같이, 본 실시예에 따른 맥락 엔코더(110)의 임베딩 레이어에는 테이블의 정보가 손실되는 것을 방지하기 위한 열 임베딩 레이어가 존재한다. 이때, 행 임베딩(Row embeddings) 레이어를 추가하지 않은 이유는 열과 행 임베딩 레이어를 동시에 사용하는 경우에 임베딩 레이어의 학습이 원활하지 않기 때문이다.

[0061] 한편, 맥락에 대한 고차원의 벡터를 생성하기 위해, 지식 데이터베이스(300)로부터 제공되는 테이블이 텍스트 형태로 선형화되는 과정을 살펴보면, 선형화된 테이블의 헤더를 제외한 테이블의 값 부분을 우선 각 행별로 행에 속하는 값을 ','를 통해 구분하고 각 행은 '▣'을 통해 구분하여, 텍스트함으로서 이루어진다.

[0062] 그리고 선형화된 테이블 맥락의 각 토큰에 대하여 열 임베딩 레이어는 해당 토큰이 속하는 열의 인덱스(예를 들어, 1부터 시작)에 대응되는 열 임베딩 값을 반환한다. 그리고 해당 토큰이 테이블 맥락에 속하지 않는 경우에, 기본값을 0번째 인덱스에 대응되는 열 임베딩 값으로 변환할 수 있다. 이에, 텍스트 형태로 선형화된 테이블은 맥락 엔코더(110)의 BERT 엔코더를 통과하며 고차원의 벡터로 변환될 수 있다. 그럼에 불구하고, 테이블 정보는 열 임베딩 레이어에 의해 맥락 엔코더(110)를 거치면서도 테이블의 구조적 정보가 상실되지 않게 된다.

[0063] 그리고 질의응답 시스템(1000)의 운용에서는 고차원 벡터가 FAISS(Facebook AI Similarity Search)을 통해 인덱싱이 수행된다. 그리고 주어진 질의를 질의 엔코더(120)를 이용하여 고차원의 벡터로 변환하고, FAISS를 이용하여 해당 벡터와 코사인 유사도가 높은 k개의 맥락 벡터에 대응하는 k개의 맥락을 검색할 수 있다.

[0064] 한편, 맥락 엔코더(110)에서 취급하는 테이블 및 텍스트의 정보가 대량일 경우에, 데이터의 선형화 이전에 해당 데이터를 유사한 길이로 분할해야 할 필요성이 있다.

[0065] 특히, 본 실시예에서는 설명하는 열 임베딩 레이어은 입력 토큰의 위치를 표현하는 포지션 임베딩 레이어와 다르게 입력 토큰이 테이블의 어떠한 열)과 연관되는지 관계 정보를 표현하기 위한 것이다.

[0066] 이에, 열 임베딩 레이어의 임베딩 레이어 값을 계산하기 위해서는 입력 토큰뿐만 아니라 외부정보가 필요하다. 일례로, 포지션 임베딩 레이어의 경우에는 임베딩 레이어 값을 계산하기 위해 입력되는 시퀀스(Sequence)만으로 충분하다.

[0067] 그러나 열 임베딩 레이어의 경우에는 임베딩 레이어 값을 계산하기 위해 입력되는 시퀀스뿐만 아니라 시퀀스가 나타내는 테이블의 구조적 정보가 요구된다. 즉, 시퀀스만으로 대응되는 열 임베딩 값을 도출하기에는 모호성이 존재하기 때문이다.

[0068] 일반적으로 시퀀스는 테이블이 직렬화하는 과정을 통해 얻어진다. 테이블의 선형화에서는 테이블의 각 행에 속하는 값을 ','로 구분하고 열에 속하는 값은 '▣'으로 구분하는 과정을 거친다. 이에, 시퀀스의 길이가 긴 경우에 BERT 모델에서 처리할 수 있는 토큰 수 제한(Token number limit)인 512를 넘을 수 있다. 즉 테이블의 크기가 큰 경우에 시퀀스를 BERT 모델이 처리할 수 있는 길이로 나누어 BERT 모델로 제공해야 할 필요가 있다.

[0069] 일례로, 아래 표 1과 같이 테이블이 C1부터 C100까지 100개의 열을 가지고 열에 대응되는 값이 A/B/C/D로 동일하다고 가정할 수 있다.

표 1

[0070] Title

C1	· · ·	C100
A/B/C/D	· · ·	A/B/C/D

[0071]

[0072] 이에, 표 1의 테이블을 선형화하면 "[CLS] Title [SEP] C1, C2, · · · C100 [SEP] A/B/C/D, A/B/C/D, · · · A/B/C/D"와 같은 시퀀스X를 생성할 수 있다. 이때, 시퀀스X의 길이는 1003(3+100+1+(100*8)+99)이게 된다. 즉, BERT 모델에서 처리할 수 있는 최대 길이인 512를 넘게 된다.

[0073] 이에, 본 실시예에 따른 질의응답 시스템(1000)이 해당 직렬화된 시퀀스X를 처리하기 위해서는 BERT 모델의 토큰 수 제한을 만족하도록 시퀀스X를 나누게 된다. 이때, 질의응답 시스템(1000)은 테이블의 값을 제외한 부분은 그대로 두고 테이블을 나눌 수 있다.

[0074] 일례로, 시퀀스X는 시퀀스X_1, 시퀀스X_2, 시퀀스X_3으로 나눠질 수 있고, 각각의 시퀀스는 아래의 표 2와 같다.

표 2

[0075]

Sequence X-1	[CLS] Title [SEP] C1, C2, · · · C100 [SEP] A/B/C/D, · · · , A/B
Sequence X-2	[CLS] Title [SEP] C1, C2, · · · C100 [SEP] /C/D/A/B, · · ·
Sequence X-3	[CLS] Title [SEP] C1, C2, · · · C100 [SEP] · · · A/B/C/D

[0076]

[0077] 여기서, 시퀀스 X_2의 B/C 토큰의 경우에 C46에 대응되는 것을 알 수 있다. 즉, 이러한 테이블의 구조적 정보는 시퀀스만으로는 알 수 없다.

[0078] 이에, 질의응답 시스템(1000)의 작동에서는 테이블의 선형화에서 시퀀스에 속하는 토큰의 값이 어떠한 열에 속하는지 미리 계산될 필요성이 있다. 이에, 열 임베딩 레이어의 계산에서는 외부정보가 필요할 수 있으며, 입력 시퀀스만으로 계산하는 포지션 임베딩 레이어와는 구분될 수 있다.

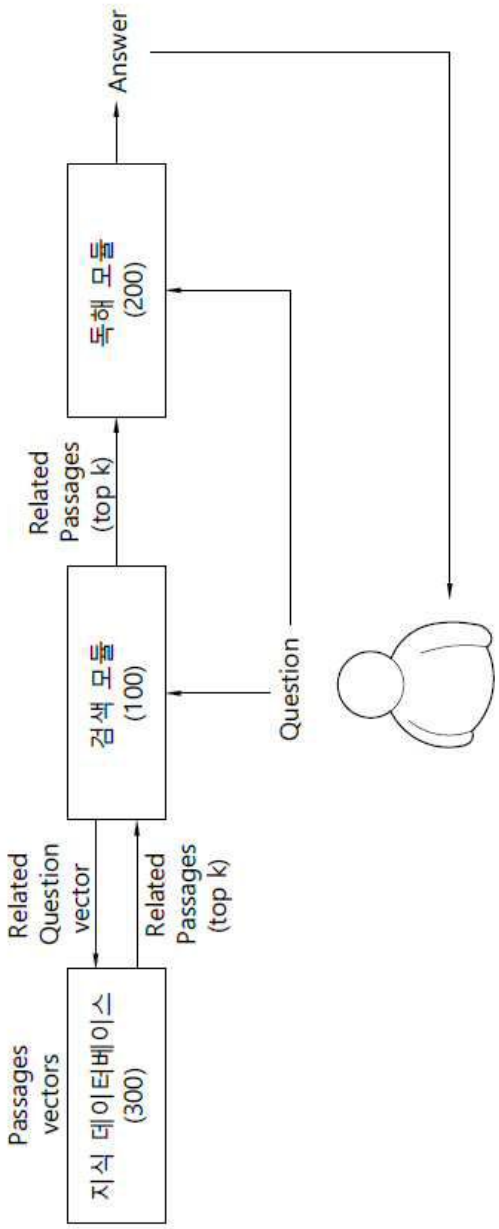
[0079] 이러한 계산은 선형화된 시퀀스를 나누기 전에 시퀀스 내에 등장하는 토큰의 값이 어떠한 열에 속하는지 미리 저장해두고, 추후 질의응답 시스템(1000)이 작동할 때 토큰에 대응되는 열 정보를 미리 저장된 정보로부터 불러옴으로서 수행된다.

[0080] 이에, 본 발명에 따른 오픈 도메인 질의응답 시스템 및 방법은 종래의 텍스트 정보의 활용뿐만 아니라 테이블의 구조적 정보를 활용한 테이블 정보를 활용하여, 질의에 대한 응답을 보다 효과적으로 도출하는 효과가 있다.

[0081] 앞에서 설명되고, 도면에 도시된 본 발명의 일 실시예는 본 발명의 기술적 사상을 한정하는 것으로 해석되어서는 안 된다. 본 발명의 보호범위는 청구범위에 기재된 사항에 의하여만 제한되고, 본 발명의 기술분야에서 통상의 지식을 가진 자는 본 발명의 기술적 사상을 다양한 형태로 개량 변경하는 것이 가능하다. 따라서 이러한 개량 및 변경은 통상의 지식을 가진 자에게 자명한 것인 한 본 발명의 보호범위에 속하게 될 것이다.

도면

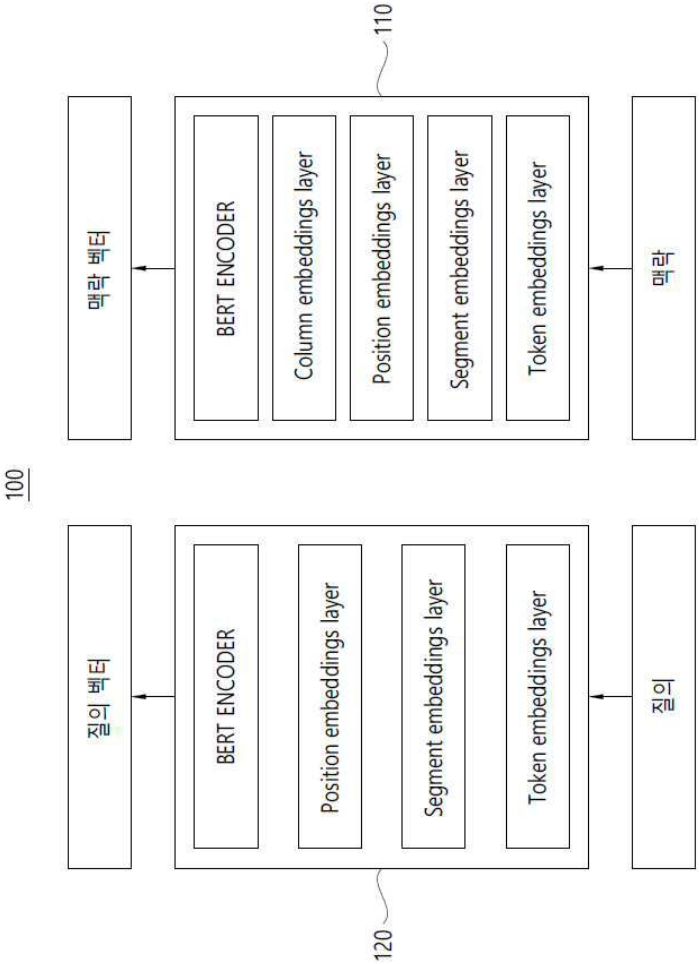
도면1


Answer
독해 모듈
(200)
Related
Passages
(top k)
검색 모듈
(100)
Question
Related
Question
vector
Related
Passages
(top k)
Passages
vectors
지식 데이터베이스
(300)

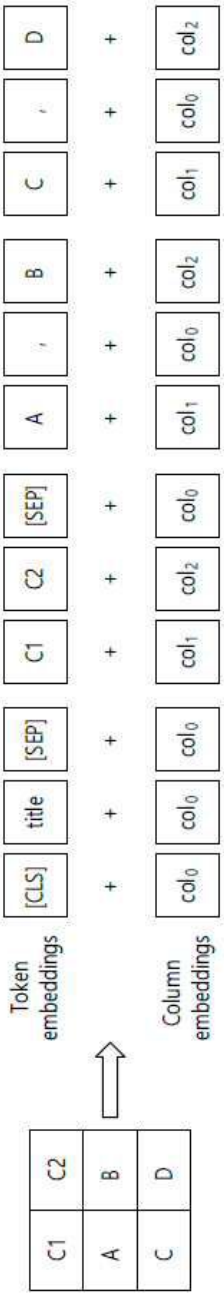
도면2


100
110
120
맥락 벡터
질의 벡터
BERT ENCODER
EMBEDDING LAYER
맥락
질의
C
CLS
E [CLS]

도면3


100
110
120
맥락 벡터
BERT ENCODER
Column embeddings layer
Position embeddings layer
Segment embeddings layer
Token embeddings layer
맥락
질의 벡터
BERT ENCODER
Position embeddings layer
Segment embeddings layer
Token embeddings layer
질의

도면4


C1
C2
A
B
C
D
Token embeddings
Column embeddings
[CLS]
title
[SEP]
C1
C2
[SEP]
A
,
B
C
,
D
col0
col1
col2