(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년02월09일
(11) 등록번호 10-2498403
(24) 등록일자 2023년02월07일

(51) 국제특허분류(Int. Cl.)
G06N 5/02 (2023.01) *G06F 16/22* (2019.01)
G06F 16/242 (2019.01) *G06F 16/2452* (2019.01)
G06F 16/248 (2019.01)
(52) CPC특허분류
G06N 5/02 (2023.01)
G06F 16/2228 (2019.01)
(21) 출원번호 10-2021-0013642
(22) 출원일자 2021년01월29일
심사청구일자 2021년01월29일
(65) 공개번호 10-2022-0109978
(43) 공개일자 2022년08월05일
(56) 선행기술조사문헌
US20180336198 A1*
(뒷면에 계속)
전체 청구항 수 : 총 14 항

(73) 특허권자
포항공과대학교 산학협력단
경상북도 포항시 남구 청암로 77 (지곡동)
(72) 발명자
한욱신
경상북도 포항시 남구 청암로 77 창의IT융합공학과 (지곡동, 포항공과대학교)
강혁규
경상북도 포항시 남구 효자동길5번길 23, 201호(효자동)
(뒷면에 계속)
(74) 대리인
특허법인이룸리온

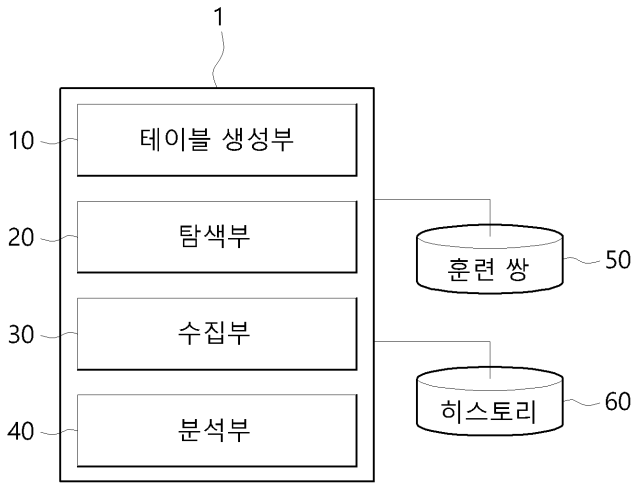
심사관 : 서광훈

(54) 발명의 명칭 **자연어를 SQL로 변환하는 시스템을 위한 훈련 세트 수집 장치 및 그 방법**

(57) 요 약

본 발명은 자연어를 SQL로 변환하는 NL2SQL 시스템을 위한 훈련 세트 수집 및 분석에 관한 것으로 본 발명에 따르면 대상 데이터베이스로부터 샘플링을 통해 발췌 테이블을 생성하고, 생성한 발췌 테이블에 대한 사용자의 자연어 질문을 SQL 쿼리로 변환하여 수행하고, 수행한 결과를 시각화 함으로써 SQL 전문가가 아니더라도 고품질의 NL2SQL 시스템 훈련 세트를 수집할 수 있도록 함으로써 비용을 절약하면서도 개발자가 오류 원인을 쉽게 분석하도록 해주는 효과가 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/243 (2019.01)
G06F 16/24522 (2019.01)
G06F 16/248 (2019.01)

(72) 발명자
김현지
울산광역시 동구 월봉12길 50, C동 308호 (화정동, 송정타워맨션3차)
나인혁
서울특별시 서대문구 연희로32길 48, 104동 105호 (연희동, 연희동성원아파트)

(56) 선행기술조사문헌
US20200004831 A1*
US20200334233 A1
US20200301925 A1
US20200175304 A1
*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호	1711103171
과제번호	2018-0-01398-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발사업(SW스타랩)
연구과제명	대화 가능하고 자동으로 튜닝하는 DBMS의 개발
기 여 율	1/1
과제수행기관명	포항공과대학교 산학협력단
연구기간	2020.01.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

주어진 데이터베이스에 샘플링을 수행하여 상기 데이터베이스의 일부로 구성되는 발췌 테이블(Table Excerpts)을 생성하는 테이블 생성부;

상기 발췌 테이블에 대해 생성된 사용자의 자연어 질문을 NL2SQL(Natural Language to SQL) 시스템을 이용하여 SQL 쿼리로 번역하고, 상기 SQL 쿼리를 수행한 결과를 표시하는 탐색부; 및

상기 SQL 쿼리를 수행한 결과 사용자에 의해 부정확한 것으로 표시된 훈련 쌍인 자연어 질문과 SQL 쿼리의 쌍을 분석하여 모델 해석(Model Interpretation) 수행, 오류 하이라이팅(Error Highlighting) 수행 및 상기 NL2SQL 시스템을 제외한 다른 NL2SQL 시스템의 제안을 수행하는 분석부;를 포함하는, NL2SQL 시스템을 위한 훈련 세트 수집 장치.

청구항 2

제1항에 있어서,

상기 테이블 생성부는 상기 발췌 테이블을 생성하기 위해 상기 주어진 데이터베이스에 관계 샘플링, 속성 샘플링 및 튜플 샘플링을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 장치.

청구항 3

제1항에 있어서,

상기 탐색부에서 상기 SQL 쿼리를 수행한 결과를 저장하기 위한 히스토리 데이터베이스; 및 상기 탐색부에서 상기 SQL 쿼리를 수행한 결과 중 정확한 것으로 표시된 자연어 질문과 SQL 쿼리의 쌍을 저장하는 훈련 쌍 데이터베이스;를 더 포함하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 장치.

청구항 4

제1항에 있어서,

상기 테이블 생성부는 상기 발췌 테이블을 교란(Perturbing)하여 새로운 발췌 테이블을 생성하고,

상기 탐색부는 상기 새로운 발췌 테이블에 의해 상기 SQL 쿼리를 검증하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 장치.

청구항 5

제1항에 있어서,

상기 분석부는 통합 그라디언트 방법과 컨덕턴스 방법에 의해 상기 부정확한 훈련 쌍을 분석하여 상기 모델 해석을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 장치.

청구항 6

제1항에 있어서,

상기 분석부는 상기 부정확한 훈련 쌍의 SQL 쿼리와 주어진 정확한 쿼리를 질의 변환(Query Rewriting)을 사용하여 정규화하고 정규화된 두 쿼리들의 문법 구조를 비교하여 상기 부정확한 훈련 쌍의 SQL 쿼리의 잘못된 부분을 시각화 함으로써 상기 오류 하이라이팅을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 장치.

청구항 7

제1항에 있어서,

상기 분석부는 상기 사용자의 자연어 질문을 상기 NL2SQL 시스템을 제외한 다른 복수의 NL2SQL 시스템을 이용하여 번역하고, 번역된 SQL 쿼리가 주어진 정확한 SQL 쿼리와 일치하는 상기 다른 NL2SQL 시스템의 제안을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 장치.

청구항 8

하나 이상의 프로세서를 포함하는 훈련 세트 수집 장치에 의해 수행되는 훈련 세트 수집 방법에 있어서:

(a) 주어진 데이터베이스에 샘플링을 수행하여 상기 데이터베이스의 일부로 구성되는 발췌 테이블(Table Excerpts)을 생성하는 단계;

(b) 상기 발췌 테이블에 대해 생성된 사용자의 자연어 질문을 NL2SQL(Natural Language to SQL) 시스템을 이용하여 SQL 쿼리로 번역하고, 상기 SQL 쿼리를 수행한 결과를 표시하는 단계; 및

(c) 상기 SQL 쿼리를 수행한 결과 사용자에 의해 부정확한 것으로 표시된 훈련 쌍인 자연어 질문과 SQL 쿼리의 쌍을 분석하여 모델 해석(Model Interpretation) 수행, 오류 하이라이팅(Error Highlighting) 수행 및 상기 NL2SQL 시스템을 제외한 다른 NL2SQL 시스템의 제안을 수행하는 단계;를 포함하는 NL2SQL 시스템을 위한 훈련 세트 수집 방법.

청구항 9

제8항에 있어서,

상기 (a) 단계는 상기 발췌 테이블을 생성하기 위해 상기 주어진 데이터베이스에 관계 샘플링, 속성 샘플링 및 튜플 샘플링을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 방법.

청구항 10

제8항에 있어서,

상기 (b)단계 이후에, 상기 SQL 쿼리를 수행한 결과를 히스토리 데이터베이스에 저장하는 단계; 및 상기 SQL 쿼리를 수행한 결과 중 정확한 것으로 표시된 자연어 질문과 SQL 쿼리의 쌍을 훈련 쌍 데이터베이스에 저장하는 단계;를 더 포함하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 방법.

청구항 11

제8항에 있어서,

상기 (a)단계는 상기 발췌 테이블을 교란(Perturbing)하여 새로운 발췌 테이블을 생성하고,

상기 (b)단계는 상기 새로운 발췌 테이블에 의해 상기 SQL 쿼리를 검증하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 방법.

청구항 12

제8항에 있어서,

상기 (c)단계는 통합 그라디언트 방법과 컨덕턴스 방법에 의해 상기 부정확한 훈련 쌍을 분석하여 상기 모델 해석을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 방법.

청구항 13

제8항에 있어서,

상기 (c)단계는 상기 부정확한 훈련 쌍의 SQL 쿼리와 주어진 정확한 쿼리를 질의 변환(Query Rewriting)을 사용하여 정규화하고 정규화된 두 쿼리들의 문법 구조를 비교하여 상기 부정확한 훈련 쌍의 SQL 쿼리의 잘못된 부분을 시각화 함으로써 상기 오류 하이라이팅을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 방법.

청구항 14

제8항에 있어서,

상기 (c)단계는 상기 사용자의 자연어 질문을 상기 NL2SQL 시스템을 제외한 다른 복수의 NL2SQL 시스템을 이용하여 번역하고, 번역된 SQL 쿼리가 주어진 정확한 SQL 쿼리와 일치하는 상기 다른 NL2SQL 시스템의 제안을 수행하는 것을 특징으로 하는, NL2SQL 시스템을 위한 훈련 세트 수집 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 데이터베이스 관리 시스템에 관한 것으로, 특히 자연어를 SQL로 변환하기 위한 시스템에 관한다.

배 경 기 술

[0003] 관계형 데이터베이스와 자연어(Natural Language) 질문에 대해, 자연어를 SQL로 번역(NL2SQL)하는 것은 자연어 질문에 대응하는 SQL 서술(Statement)을 찾는 것이다. 최근에는 이러한 NL2SQL을 위해 딥러닝(Deep Learning) 기술을 이용하는 많은 방법들이 개발되어 왔다.

[0004] NL2SQL 시스템을 개발하는데 있어 주요 작업은 좋은 인코더와 디코더를 설계하는 것이지만, NL2SQL 시스템의 품질을 향상하기 위해서는 크게 두 가지가 필요하다. 첫째로는 많은 수의 고품질, 라벨이 붙여진 훈련 데이터와, 둘째로는 잘못된 번역을 찾기 위한 툴이다. 시스템의 정확성은 훈련 세트에 크게 의존하기 때문이다.

[0005] 많은 훈련 데이터를 얻기 위해 크라우드 소싱(Crowdsourcing)이 사용되지만 다른 머신러닝 문제들에 비해 NL2SQL 시스템은 작업자들이 SQL 쿼리(Queries)와 주어진 데이터베이스 스키마(Schema)에 대해 잘 알아야 하기 때문에 비용이 많이 드는 문제가 있다.

[0006] 또한 대표적인 NL2SQL 모델들이 인코더 디코더 모델을 사용하지만 각 입력 특징(Feature)들이 출력에 어떤 역할을 하는지 적절히 파악하기 힘든 문제도 있다.

[0007] 본 발명의 발명자들은 이러한 종래 기술의 NL2SQL 시스템의 자료수집 및 분석의 문제점들을 해결하기 위해 연구 노력해 왔다. 작업자들이 쉽게 데이터를 탐색하고 수집할 수 있도록 데이터들을 시각화하고 잘못 번역된 쿼리들을 분석할 수 있도록 함으로써 비용을 절약할 수 있는 데이터베이스 자연어 인터페이스 시스템을 완성하기 위해 많은 노력 끝에 본 발명을 완성하기에 이르렀다.

발명의 내용

해결하려는 과제

[0009] 본 발명의 목적은 저비용으로 고품질의 훈련 세트를 획득하고 잘못 번역된 쿼리들을 분석하기 위한 장치 및 그

방법을 제공하는 것이다.

[0010] 한편, 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론 할 수 있는 범위 내에서 추가적으로 고려될 것이다.

과제의 해결 수단

[0012] 본 발명에 따른 NL2SQL 시스템을 위한 훈련 세트 수집 장치는,

[0013] 주어진 데이터베이스에서 상기 데이터베이스의 일부로 구성되는 발췌 테이블(Table Excerpts)을 생성하는 테이블 생성부; 상기 발췌 테이블에 대해 생성된 사용자의 자연어 질문을 NL2SQL(Natural Language to SQL) 시스템을 이용하여 SQL 쿼리로 번역하고, 상기 SQL 쿼리를 수행한 결과를 표시하는 탐색부; 및 상기 SQL 쿼리를 수행한 결과 사용자에 의해 부정확한 것으로 표시된 훈련 쌍인 자연어 질문과 SQL 쿼리의 쌍을 분석하여 모델 해석(Model Interpretation), 오류 하이라이팅(Error Highlighting) 또는 시스템을 제안하는 분석부;를 포함한다.

[0014] 상기 테이블 생성부는 상기 발췌 테이블을 생성하기 위해 상기 주어진 데이터베이스에서 관계 샘플링, 속성 샘플링 및 튜플 샘플링을 수행하는 것을 특징으로 한다.

[0015] 상기 탐색부에서 상기 SQL 쿼리를 수행한 결과를 저장하기 위한 히스토리 데이터베이스; 및 상기 탐색부에서 상기 SQL 쿼리를 수행한 결과 중 정확한 것으로 표시된 자연어 질문과 SQL 쿼리의 쌍을 저장하는 훈련 쌍 데이터베이스;를 더 포함하는 것을 특징으로 한다.

[0016] 상기 테이블 생성부는 상기 발췌 테이블을 교란(Perturbing)하여 새로운 발췌 테이블을 생성하고, 상기 탐색부는 상기 새로운 발췌 테이블에 의해 상기 SQL 쿼리를 검증하는 것을 특징으로 한다.

[0017] 상기 분석부는 통합 그라디언트 방법과 컨덕턴스 방법에 의해 상기 부정확한 훈련 쌍을 분석하여 상기 모델 해석을 수행하는 것을 특징으로 한다.

[0018] 상기 분석부는 상기 부정확한 훈련 쌍의 SQL 쿼리와 주어진 정확한 쿼리를 질의 변환(Query Rewriting)을 사용하여 정규화하고 정규화된 두 쿼리들의 문법 구조를 비교하여 상기 부정확한 훈련 쌍의 SQL 쿼리의 잘못된 부분을 시각화 함으로써 상기 오류 하이라이팅을 수행하는 것을 특징으로 한다.

[0019] 상기 분석부는 상기 사용자의 자연어 질문을 상기 NL2SQL 시스템 외의 다른 복수의 NL2SQL 시스템을 이용하여 번역하고, 번역된 SQL 쿼리가 주어진 정확한 SQL 쿼리와 일치하는 NL2SQL 시스템을 추천하는 것을 특징으로 한다.

[0020] 본 발명의 다른 실시예에 따른 훈련 세트 수집 방법은,

[0021] (a) 주어진 데이터베이스에서 상기 데이터베이스의 일부로 구성되는 발췌 테이블(Table Excerpts)을 생성하는 단계; (b) 상기 발췌 테이블에 대해 생성된 사용자의 자연어 질문을 NL2SQL(Natural Language to SQL) 시스템을 이용하여 SQL 쿼리로 번역하고, 상기 SQL 쿼리를 수행한 결과를 표시하는 단계; 및 (c) 상기 SQL 쿼리를 수행한 결과 사용자에 의해 부정확한 것으로 표시된 훈련 쌍인 자연어 질문과 SQL 쿼리의 쌍을 분석하여 모델 해석(Model Interpretation), 오류 하이라이팅(Error Highlighting) 또는 시스템을 제안하는 단계;를 포함한다.

[0022] 상기 (a) 단계는 상기 발췌 테이블을 생성하기 위해 상기 주어진 데이터베이스에서 관계 샘플링, 속성 샘플링 및 튜플 샘플링을 수행하는 것을 특징으로 한다.

[0023] 상기 (b)단계 이후에, 상기 SQL 쿼리를 수행한 결과를 히스토리 데이터베이스에 저장하는 단계; 및 상기 SQL 쿼리를 수행한 결과 중 정확한 것으로 표시된 자연어 질문과 SQL 쿼리의 쌍을 훈련 쌍 데이터베이스에 저장하는 단계;를 더 포함하는 것을 특징으로 한다.

[0024] 상기 (a)단계는 상기 발췌 테이블을 교란(Perturbing)하여 새로운 발췌 테이블을 생성하고, 상기 (b)단계는 상기 새로운 발췌 테이블에 의해 상기 SQL 쿼리를 검증하는 것을 특징으로 한다.

[0025] 상기 (c)단계는 통합 그라디언트 방법과 컨덕턴스 방법에 의해 상기 부정확한 훈련 쌍을 분석하여 상기 모델 해석을 수행하는 것을 특징으로 한다.

[0026] 상기 (c)단계는 상기 부정확한 훈련 쌍의 SQL 쿼리와 주어진 정확한 쿼리를 질의 변환(Query Rewriting)을 사용하여 정규화하고 정규화된 두 쿼리들의 문법 구조를 비교하여 상기 부정확한 훈련 쌍의 SQL 쿼리의 잘못된 부분을 시각화 함으로써 상기 오류 하이라이팅을 수행하는 것을 특징으로 한다.

[0027] 상기 (c)단계는 상기 사용자의 자연어 질문을 상기 NL2SQL 시스템 외의 다른 복수의 NL2SQL 시스템을 이용하여 번역하고, 번역된 SQL 쿼리가 주어진 정확한 SQL 쿼리와 일치하는 NL2SQL 시스템을 추천하는 것을 특징으로 한다.

발명의 효과

[0029] 본 발명에 따르면 자연어를 SQL로 변환하는 시스템을 훈련하기 위한 고품질의 훈련 세트를 수집할 수 있는 효과가 있다.

[0030] 또한 훈련을 위한 데이터들을 분석하기 위한 시각화 자료들을 제공함으로써 작업자들이 보다 쉽게 훈련 데이터를 수집하고 잘못된 데이터들을 수정할 수 있는 장점도 있다.

[0031] 한편, 여기에서 명시적으로 언급되지 않은 효과라 하더라도, 본 발명의 기술적 특징에 의해 기대되는 이하의 명세서에서 기재된 효과 및 그 잠정적인 효과는 본 발명의 명세서에 기재된 것과 같이 취급됨을 첨언한다.

도면의 간단한 설명

[0033] 도 1은 본 발명의 바람직한 어느 실시예에 따른 훈련 세트 수집 장치의 개략적인 구조도이다.

도 2 및 도 3은 본 발명의 바람직한 어느 실시예에 따른 훈련 세트 수집 장치의 실행 예를 나타낸다.

도 4는 본 발명의 바람직한 다른 실시예에 따른 훈련 세트 수집 방법의 개략적인 흐름도이다.

※ 첨부된 도면은 본 발명의 기술사상에 대한 이해를 위하여 참조로서 예시된 것임을 밝히며, 그것에 의해 본 발명의 권리범위가 제한되지는 아니한다

발명을 실시하기 위한 구체적인 내용

[0034] 이하, 도면을 참조하여 본 발명의 다양한 실시예가 안내하는 본 발명의 구성과 그 구성으로부터 비롯되는 효과에 대해 살펴본다. 본 발명을 설명함에 있어서 관련된 공지기능에 대하여 이 분야의 기술자에게 자명한 사항으로서 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략한다.

[0035] '제1', '제2' 등의 용어는 다양한 구성요소를 설명하는데 사용될 수 있지만, 상기 구성요소는 위 용어에 의해 한정되어서는 안 된다. 위 용어는 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용될 수 있다. 예를 들어, 본 발명의 권리범위를 벗어나지 않으면서 '제1구성요소'는 '제2구성요소'로 명명될 수 있고, 유사하게 '제2구성요소'도 '제1구성요소'로 명명될 수 있다. 또한, 단수의 표현은 문맥상 명백하게 다르게 표현하지 않는 한, 복수의 표현을 포함한다. 본 발명의 실시예에서 사용되는 용어는 다르게 정의되지 않는 한, 해당 기술분야에서 통상의 지식을 가진 자에게 통상적으로 알려진 의미로 해석될 수 있다.

[0036] 이하, 도면을 참조하여 본 발명의 다양한 실시예가 안내하는 본 발명의 구성과 그 구성으로부터 비롯되는 효과에 대해 살펴본다.

[0038] 도 1은 본 발명의 바람직한 어느 실시예에 따른 훈련 세트 수집 장치의 개략적인 구조도이다.

[0039] 본 발명에 따른 NL2SQL 시스템을 위한 훈련 세트 수집 장치(1)는 테이블 생성부(10), 탐색부(20), 수집부(30) 및 분석부(40)를 포함한다. 또한 훈련쌍 데이터베이스(50) 및 히스토리(History) 데이터베이스(60)를 더 포함할 수 있다.

[0040] 테이블 생성부(10)는 주어진 데이터베이스 D에서 샘플링을 통해 발췌 테이블(Table Excerpts)을 생성한다.

[0041] 작고 직관적인 발췌 테이블 생성을 위해서는 데이터베이스 D에서 1) 관계(Relation) 샘플링, 2) 속성(Attribute) 샘플링, 그리고 3) 튜플(Tuple) 샘플링을 수행한다.

[0042] 관계/속성 샘플링을 위해서는 누적된 SQL 쿼리를 이용하는데 서로 다른 관계와 속성의 관련성(Relevance)을 계산하기 위함이다. 두 스키마 엔트리(즉, 관계 혹은 속성)가 같이 나타나는 빈도가 높을수록 더 연관되어 있다고 판단할 수 있다.

[0043] 이를 위해 각 노드(Node)가 관계 또는 속성이고 각 에지(Edge)는 두 관계들 사이의 결합가능한 관계 또는 속성의 관계에 대한 연계(affiliation)를 암시하는 스키마 그래프로부터 연결된 서브그래프를 추출한다. 특히, SQL 쿼리 로그에 나타나는 빈도에 비례하는 확률로 관계를 무작위로 선택한다. 그 다음 이전에 선택된 관계와 결합 가능한 새 관계를 무작위로 반복적으로 선택한다. 여기서 선택 확률은 선택된 관계들의 공동 출현(co-

occurrence) 빈도와 비례하게 된다.

[0044] 관계들을 선택한 후에 이와 비슷하게 선택된 관계들의 공동 출현 빈도를 고려하여 속성들을 선택한다.

[0045] 샘플링 단계가 끝나면 이를 이용하여 발췌 테이블을 생성하게 된다. 이는 1) 모든 선택된 관계들 사이의 모든 바깥(outer) 결합 수행, 2) 테이블 t_D를 얻기 위해 모든 선택된 속성들에 대한 프로젝션(Projection), 3) t_D로부터 일부 튜플들을 샘플링함으로써 생성된다.

[0046] 다음 표 1은 IMDB(International Movie DataBase)로부터 생성된 발췌 테이블의 예를 나타낸다.

[0047] [표 1]

Cast					
Movie.mid	Movie.title	Movie. relased_year	Actor. aid	Actor. name	Actor. birth_year
2627	101 Sexiest Cele...	2005	1718	Brad Pitt	1963
2627	101 Sexiest Cele...	2005	7131	Sunshine	-
2632	2013 MTV..	2013	1718	Brad Pitt	1963
13	1001 Inventions...	2015	3267	Khalid Abdalla	1980
28	120	2008	7172	Yasar Abravaya	1990

[0048]

[0049] 우선 관계 샘플링에 의해 세 속성(영화(Movie), 캐스팅(Cast) 및 배우(Actor))을 선택하는데, 캐스팅은 영화와 배우 사이의 다수-대-다수(many-to-many) 관계를 나타낸다. 다음으로 속성 샘플링에 의해 여섯 개의 속성들을 선택한다. 그 다음 세 관계들의 모든 외부 결합과 여섯 속성들에 대한 프로젝션을 수행함으로써 테이블 t_D를 얻을 수 있다. 마지막으로 t_D로부터 다섯 개의 튜플들을 샘플링하면 발췌 테이블인 표 1을 얻게 된다.

[0050] 보다 직관적인 발췌 테이블을 얻기 위해서는 발췌 테이블은 동일한 부분 키 값을 가지는 튜플들을 포함해야 한다. 표 1에서 키 {Movie.mid, Actor.aid}를 고려해보면, 첫 두 튜플은 부분 키 값이 2627로 동일하므로 동일한 영화를 나타낸다. 이는 작업자들이 발췌 테이블에 대해 의미있는 그룹과 쿼리 집합을 제시할 수 있게 해준다.

[0051] 탐색부(20)는 사용자의 질문을 SQL 쿼리로 번역하여 데이터베이스를 탐색한다. 이를 위해 사용자는 NL2SQL 시스템과 데이터베이스를 선택할 수 있다.

[0052] 도 2는 탐색부(20)에서 사용자에게 표시하는 화면의 한 예이다.

[0053] 탐색부(20)는 우선 선택된 데이터베이스에서 무작위로 발췌 테이블(210)을 선택하여 보여준다.

[0054] 탐색부(20)가 발췌 테이블(210)에 대해 자연어 질문 q_{nl}(220) 수신하면 이를 선택된 NL2SQL 시스템을 이용하여 SQL 쿼리 q_{sql}로 번역하고, 실행 결과(230)를 시각화한다.

[0055] 예를 들어 사용자가 “브래드 피트가 출연한 영화의 제목을 찾아라(Find the titles of the movies which “Brad Pitt” starred in)”라는 자연어 질문(220)을 입력한다면 이는 “SELECT T1.title FROM movie AS T1 JOIN cast AS T3 ON T1.mid = T3.msid JOIN actor AS T2 ON T3.aid = T2.aid WHERE T2.name = “Brad Pitt”” 라는 SQL 쿼리로 변환되고 탐색부(20)는 발췌 테이블(220)에서 해당 영화들의 제목을 찾아 출력(230)한다. 결과를 보면 사용자는 정확한지 여부를 체크(240)하게 된다.

[0056] 이 과정에서 발췌 테이블에 따라 잘못된 쿼리가 정확한 쿼리와 동일한 응답을 생성할 수도 있다. 예를 들면 표 1의 발췌 테이블 t에서 사용자가 “연기자가 몇 명인가?(How many performers are there?)”라는 질문을 했을 때 이를 잘못 번역하여 “SELECT count(*) FROM movie AS T1”과 같은 SQL 쿼리를 생성할 수 있다. 이 SQL 쿼리에 의해 영화의 수를 응답으로 생성하더라도 사용자는 이를 정확한 응답으로 판단하여 번역이 정확하다고 마크할 것이다. 표 1에 나타난 배우의 수와 영화의 수는 모두 3이기 때문이다.

[0057] 이러한 오류를 줄이기 위해 테이블 생성부(10)는 다른 발췌 테이블 t'를 생성하여 SQL 쿼리를 검증한다. 발췌

테이블 t'는 발췌 테이블 t를 교란하여 생성하고 이를 다른 사용자에 의해 검증하도록 할 수 있다.

[0058] 이렇게 해서 얻어지는 SQL 쿼리들은 결국 발췌 테이블에 관한 쿼리가 아니라 원래의 데이테베이스에 대한 쿼리가 된다. 하나의 발췌 테이블에 대한 간단한 쿼리뿐 아니라 결합을 포함한 복잡한 쿼리들을 수집할 수 있기 때문이다.

[0059] 탐색부(20)에서는 1) 선택된 테이블들의 결합, 2) 선택된 열(columns)에 대한 프로젝션, 3) 샘플링을 포함하는 SQL 쿼리들을 수행하여 발췌 테이블을 생성하므로, 발췌 테이블에 대한 SQL 쿼리 q1을 원본 데이터베이스에 대한 SQL 쿼리 q2로 변환할 수 있다.

[0060] 변환은 1) 테이블이 아닌 q1의 FROM 절에 결합 술어를 대입하고, 2) q1의 열을 원본 데이터베이스의 대응하는 열로 대체함으로써 이루어진다.

[0061] 예를 들면, 표 1에서 “영화 120에서 누가 연기했는가?”라는 질문은 NL2SQL 시스템에 의해 “SELECT actor_name FROM table_excerpt WHERE movie_title='120'”이라는 SQL 쿼리로 변환되고, 이는 “SELECT actor.name FROM actor JOIN cast ON actor.aid = cast.aid JOIN movie ON cast.mid = movie.mid WHERE movie.title = '120'.”으로 다시 변환될 수 있다.

[0062] 즉 앞의 쿼리가 사용자에게는 보여지지만 뒤의 변환된 쿼리가 탐색부(20)에 의해 수집될 것이다. 물론 발췌 테이블에 대한 변환 결과와 원본 데이터베이스에 대한 변환 결과 모두 분석될 수 있다.

[0063] 수집부(30)는 훈련 쌍(Training Pairs)을 수집한다. 훈련 쌍 수집은 사용자가 다른 작업을 하는 동안 백그라운드 작업으로 수행되므로 사용자에게는 보이지 않는다.

[0064] 탐색부(20)에서 사용자가 정확하다고 표시한 훈련 쌍은 훈련 쌍 데이터베이스(50)에 저장된다. 만일 사용자가 부정확하다고 표시한 훈련 쌍은 히스토리 데이터베이스(60)에 저장되고 분석부(40)로도 전달된다.

[0065] 수집부(30)에서 자연어 질문, SQL 쿼리, 피드백으로 이루어진 묶음(q_{nl}, q_{sql}, feedback)은 모두 히스토리 데이터베이스(60)에 저장한다.

[0066] 분석부(40)는 사용자가 분석하고자 하는 질문을 선택받아 선택된 질문에 대한 정확한 SQL 쿼리를 수신한다. 그리고 모델 해석(Model Interpretation), 오류 하이라이팅(Error Highlighting), 시스템 제안을 수행한다.

[0067] 도 3은 분석부(40)에서 사용자에게 표시하는 화면의 한 예이다.

[0068] 분석부(40)는 모델 해석을 위해 통합 그라디언트(Integrated Gradients) 방법을 사용한다. 통합 그라디언트 방법은 자연어 또는 스키마의 각 단어(w)들이 모델 출력에 기여하는 정도를 평가한다.

[0069] 분석부(40)는 단어 w와 디코더에서 생성된 출력 시퀀스의 각 요소들과의 기여도를 계산하는데, w의 기여도는 w에 의한 모든 기여도를 더하여 구해진다. 각 단어는 그 기여도에 따라 색을 달리한 글자로 시각화된다. 예를 들면 녹색은 긍정적인 기여를, 적색은 부정적인 기여를 의미할 수 있다.

[0070] 분석부(40)는 또한 모델의 내부 히든 유닛들(Internal Hidden Units)의 중요성을 시각화 하기 위해 컨덕턴스 방법(Conductance Method) 방법을 사용할 수 있다. 분석부(40)는 각 히든 상태(State)의 컨덕턴스를 평가한다. 히든 상태들은 시퀀스(Sequences)와 레이어(Layer)들의 요소(Element)들에 교차로 존재하기 때문에(예를 들어, 시퀀스 길이가 5이고 레이어 수가 3이면 5x3=15 개의 히든 상태가 존재한다), 컨덕턴스는 x축이 시퀀스의 요소들이고 y축은 레이어인 히트 맵(Heat Map)으로 시각화된다.

[0071] 이러한 방법들에 의해 모델의 행동들이 이해될 수 있으므로 잘못된 번역들을 분석함으로써 모델을 향상시키고 문제 해결을 보다 쉽게 할 수 있다.

[0072] 도 3에서 사용자는 스크롤바 상에 표시된 붉은 선(330)을 통해 오류 발생 위치를 쉽게 찾을 수 있다. 모든 오류들은 히스토리 데이터베이스(60)에 저장되어 있으므로 이를 로드하여 오류 발생 위치를 찾을 수 있다.

[0073] 오류 검증을 위해 사용자는 정답 쿼리(310)를 입력한다. “연기자가 몇 명인가?(How many performers are there?)”에 대한 정답 쿼리(310)는 “SELECT count(*) FROM actor AS T1”이 될 것이다.

[0074] 사용자가 “모든 연기자의 수”를 질문에 대해 자연어 질문에 포함된 “연기자”는 “배우”와 관계를 가진다. 만일 분석부(40)에 의해 분석한 “연기자”라는 단어의 기여도(320)가 너무 낮다면 모델이 그 단어를 충분히 이해하지 못하고 있음을 알 수 있다. 도 3의 예에서 단어의 기여도(320)는 색의 농도로 표현된다. 따라서 모델의

인코더 구조를 미리 훈련된 언어 모델로 바꿈으로써 모델을 수정할 수 있을 것이다.

[0075] 분석부(40)는 다음으로 오류 하이라이팅을 수행한다.

[0076] 분석부(40)는 선택된 SQL 쿼리를 주어진 정확한 쿼리와 비교하여 선택된 SQL 쿼리의 어느 부분이 잘못되었는지 시각화한다. 분석부(40)는 두 쿼리들을 질의 변환(Query Rewriting)을 사용하여 정규화하고 정규화된 두 쿼리들의 문법 구조를 비교하게 된다.

[0077] 마지막으로 분석부(40)는 자연어 질문을 사용자가 선택하지 않은 다른 NL2SQL 시스템을 사용하여 번역한 다음 정확한 번역을 생성한 NL2SQL 시스템을 추천한다. 만일 정확한 번역을 제공하는 시스템이 없다면 분석부(40)는 각 시스템의 번역된 SQL 쿼리와 정답(Ground Truth) 사이의 유사도 점수를 계산하여 가장 정확한 NL2SQL 시스템을 추천하게 된다. 두 SQL 쿼리 사이의 유사도 점수를 계산하기 위해 분석부(40)는 두 SQL 쿼리를 정규화한다.

[0078] 번역의 정확성을 결정하기 위해 분석부(40)는 멀티-스테이지(multi-staged) 확인(Validation) 툴을 사용하는데, 이는 번역된 쿼리와 주어진 정확한 SQL 쿼리 사이의 의미 동등성을 결정한다.

[0079] 분석부(40)는 우선 주어진 데이터베이스 인스턴스와 데이터베이스 테스트 기법으로부터 얻은 데이터베이스 인스턴스들에서 두 쿼리를 수행하고, 수행된 결과를 비교한다. 만일 수행 결과가 다르다면 두 쿼리들은 동등하지 않은 것으로 결정된다. 만일 수행 결과가 같다면 두 SQL 쿼리들의 의미 동등성을 위해 증명 툴(prover)에 의해 둘을 비교한다. 만일 증명 툴이 동등성을 판단하는데 실패한다면 분석부(40)는 두 쿼리들을 질의 변환에 의해 변환하여 그들의 문법 구조를 비교할 수 있다. 질의 변환에는 RDBMS(Relational Database Management System) 등이 사용될 수 있다.

[0080] 도 4는 본 발명의 바람직한 다른 실시예에 따른 훈련 세트 수집 방법의 개략적인 흐름도이다.

[0081] 본 발명의 훈련 세트 수집 방법은 하나 이상의 프로세서 및 제어부를 포함하는 훈련 세트 수집 장치에 의해 수행될 수 있다.

[0082] 훈련 세트 수집을 위해 우선 주어진 데이터베이스에서 샘플링을 통해 발췌 테이블을 생성한다(S10).

[0083] 발췌 테이블 생성을 위해서는 데이터베이스에서 1) 관계(Relation) 샘플링, 2) 속성(Attribute) 샘플링, 그리고 3) 튜플(Tuple) 샘플링을 수행할 수 있다. 테이블 생성을 위한 자세한 방법은 앞서 설명한 바와 같다.

[0084] 발췌 테이블이 생성되면 사용자는 테이블 탐색 단계를 수행한다(S20).

[0085] 테이블 탐색 단계에서는 사용자의 자연어 질문을 SQL 쿼리로 번역하여 데이터베이스를 탐색한다. 이를 위해 사용자는 NL2SQL 시스템과 데이터베이스를 선택할 수 있다.

[0086] SQL 쿼리의 번역 오류를 줄이기 위해 테이블 생성 단계(10)와 탐색(20) 단계를 반복할 수 있다. 이 때 새로운 발췌 테이블은 이전 테이블을 교란하여 생성하고 사용자가 다시 검증하도록 할 수 있다.

[0087] 탐색이 끝난 후 자연어 질문과 SQL 쿼리 쌍인 훈련 쌍들은 데이터베이스에 저장된다(S30).

[0088] 사용자가 정확한 것으로 체크한 훈련 쌍은 훈련 쌍 데이터베이스에 저장되고, 부정확하다고 체크한 훈련 쌍은 히스토리 데이터베이스에 저장되고 오류 원인이 분석된다.

[0089] 분석 단계에서는 모델 해석이 수행된다(S40).

[0090] 사용자가 부정확하다고 체크한 훈련 쌍을 선택하면 그에 대한 정확한 SQL 쿼리를 수신하고, 모델 해석을 위해서는 통합 그라디언트 방법과 컨덕턴스 방법을 사용하여 자연어 질문의 단어들의 출력 기여도와 모델 내부의 히든 상태들을 분석하게 된다.

[0091] 이러한 방법들에 의해 모델의 행동들이 이해될 수 있으므로 잘못된 번역들을 분석함으로써 모델을 향상시키고 문제 해결을 보다 쉽게 할 수 있다.

[0092] 분석 단계에서 오류 하이라이팅 또한 수행될 수 있다(S50).

[0093] 오류 하이라이팅 단계에서는 선택된 SQL 쿼리를 주어진 정확한 쿼리와 비교하여 선택된 SQL 쿼리의 어느 부분이 잘못되었는지 시각화함으로써 두 쿼리들의 문법 구조를 비교할 수 있다.

[0094] 마지막으로 분석 결과에 의해 보다 나은 시스템을 제안할 수 있다(S60).

[0095] 자연어 질문을 사용자가 선택하지 않은 다른 NL2SQL 시스템을 사용하여 번역한 다음 정확한 번역을 생성한

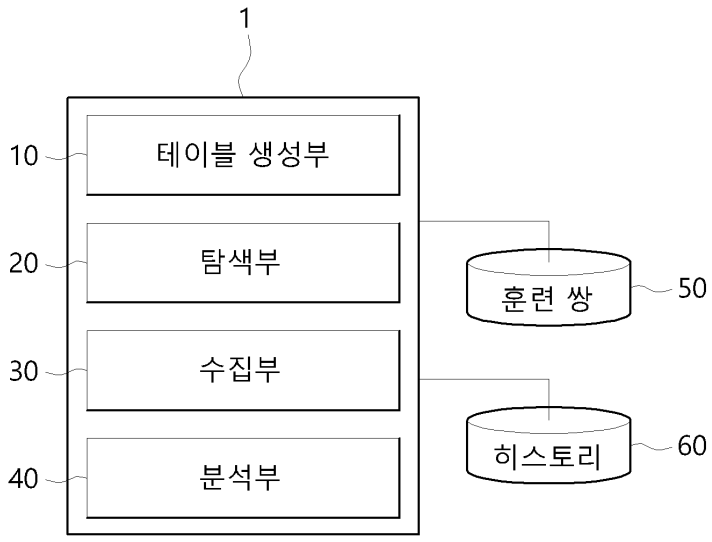
NL2SQL 시스템을 추천한다. 만일 정확한 번역을 제공하는 시스템이 없다면 각 시스템의 번역된 SQL 쿼리와 정답(Ground Truth) 사이의 유사도 점수를 계산하여 가장 정확한 NL2SQL 시스템을 추천하게 된다.

[0096] 이상과 같은 본 발명에 따른 NL2SQL 시스템을 위한 훈련 세트 수집 장치 및 방법에 따르면 NL2SQL을 위한 고품질의 훈련 데이터를 SQL 전문가가 아닌 사용자도 수집할 수 있으며 개발자들이 잘못된 SQL 번역을 쉽게 분석할 수 있도록 해주는 장점이 있다.

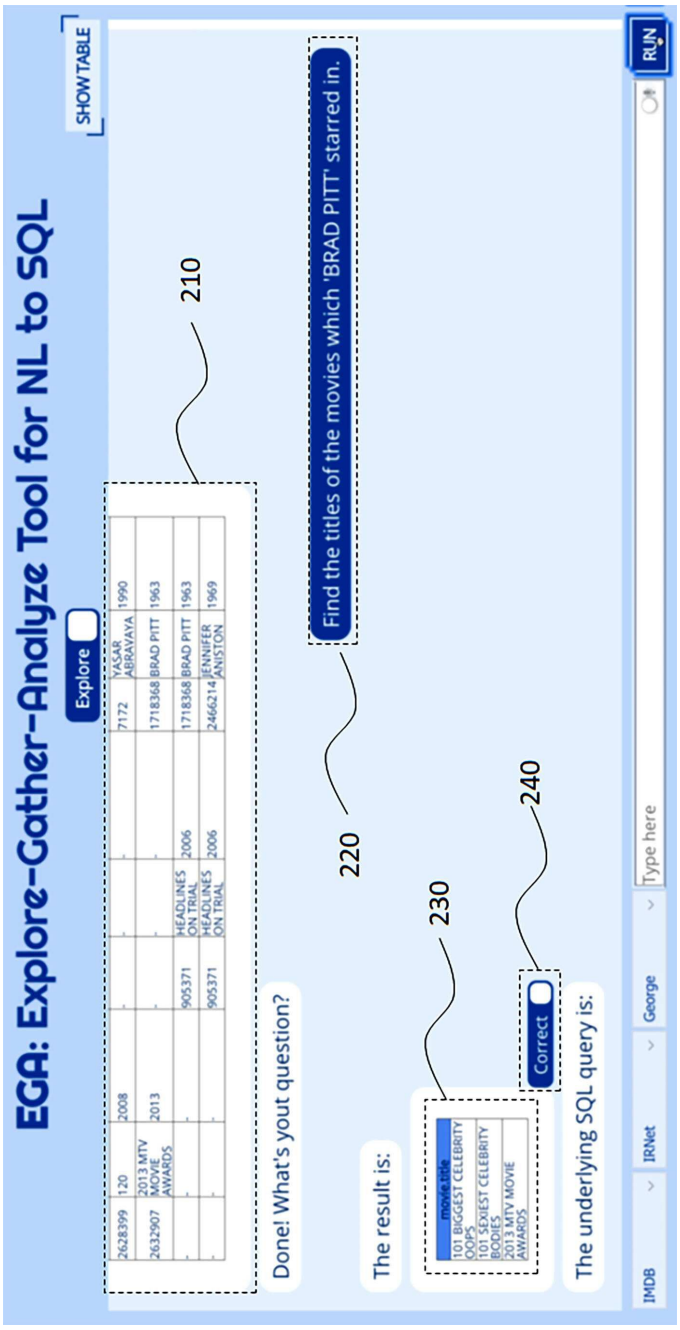
[0098] 본 발명의 보호범위가 이상에서 명시적으로 설명한 실시예의 기재와 표현에 제한되는 것은 아니다. 또한, 본 발명이 속하는 기술분야에서 자명한 변경이나 치환으로 말미암아 본 발명이 보호범위가 제한될 수도 없음을 다시 한 번 첨언한다.

도면

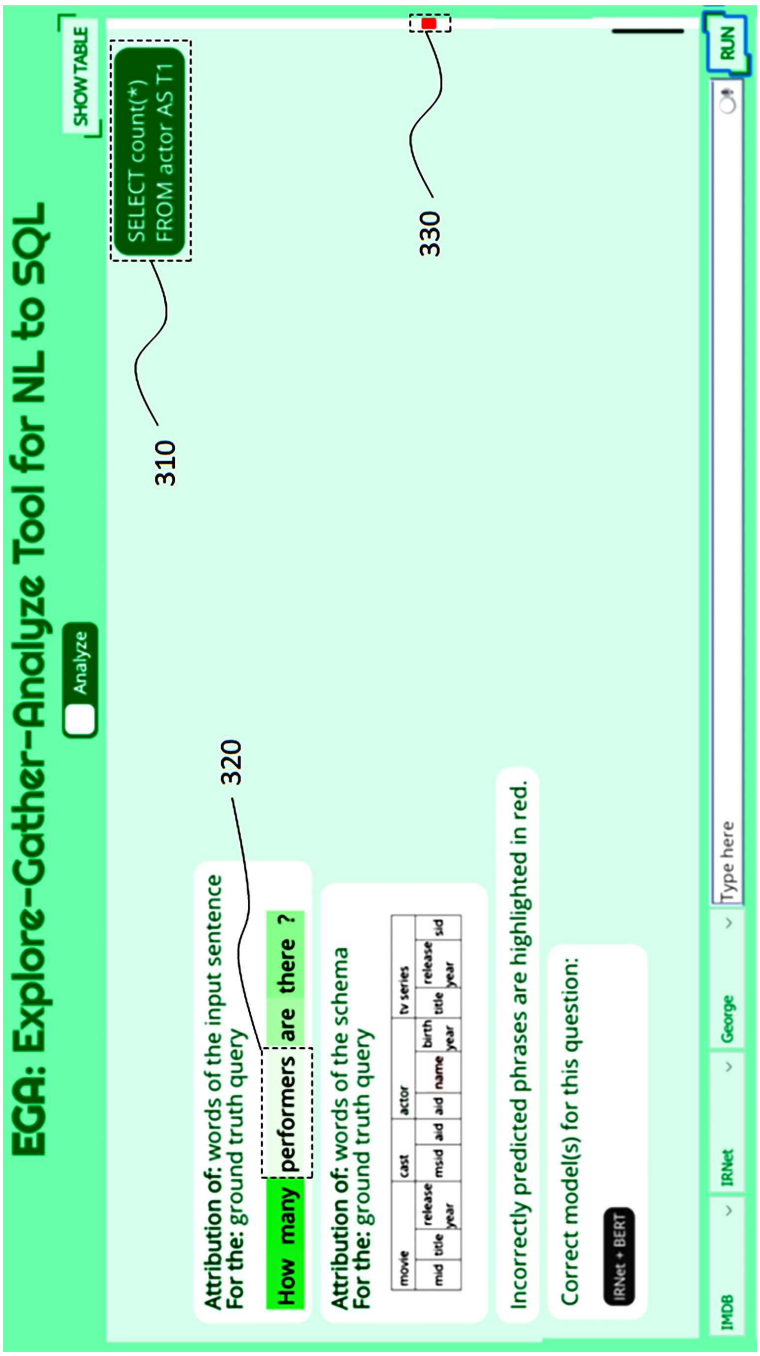
도면1



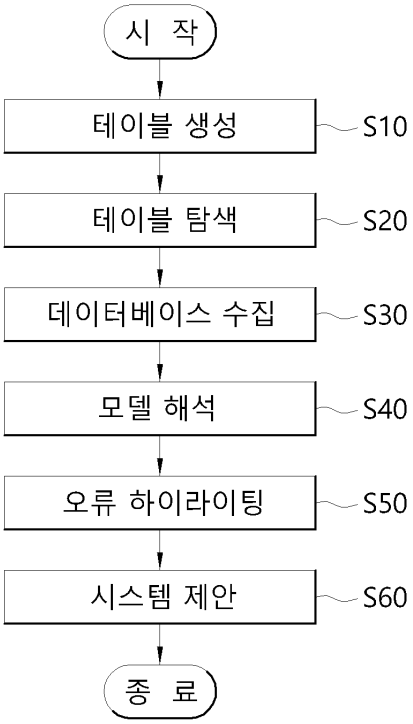
도면2


EGA: Explore-Gather-Analyze Tool for NL to SQL
SHOW TABLE
Explore
RUN
Type here
IMDB
IRNet
George
Find the titles of the movies which 'BRAD PITT' starred in.
Done! What's yout question?
The result is:
movie title
101 BIGGEST CELEBRITY OOPS
101 SEXIEST CELEBRITY BODIES
2013 MTV MOVIE AWARDS
Correct
The underlying SQL query is:
210
220
230
240

도면3


EGA: Explore-Gather-Analyze Tool for NL to SQL
Analyze
SHOW TABLE
SELECT count(*)
FROM actor AS T1
310
330
320
Attribution of: words of the input sentence
For the: ground truth query
How many performers are there ?
Attribution of: words of the schema
For the: ground truth query
movie
mid title release year
cast
msid aid
actor
aid name birth year
tv series
sid title release year
Incorrectly predicted phrases are highlighted in red.
Correct model(s) for this question:
IRNet + BERT
IMDB
IRNet
George
Type here
RUN

도면4


시 작
테이블 생성
S10
테이블 탐색
S20
데이터베이스 수집
S30
모델 해석
S40
오류 하이라이팅
S50
시스템 제안
S60
종 료