(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0034679
(43) 공개일자 2021년03월30일

(51) 국제특허분류(Int. Cl.)
G06F 16/36 (2019.01) *G06F 40/205* (2020.01)
G06F 40/295 (2020.01)
(52) CPC특허분류
G06F 16/36 (2019.01)
G06F 40/205 (2020.01)
(21) 출원번호 10-2021-7008154
(22) 출원일자(국제) 2020년07월06일
심사청구일자 2021년03월18일
(85) 번역문제출일자 2021년03월18일
(86) 국제출원번호 PCT/US2020/040890
(87) 국제공개번호 WO 2021/007159
국제공개일자 2021년01월14일
(30) 우선권주장
16/504,068 2019년07월05일 미국(US)

(71) 출원인
구글 엘엘씨
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
(72) 발명자
잇터 댄
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
위 샤오
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
리 팡타오
미국 캘리포니아 마운틴 뷰 엠피시어터 파크웨이 1600 (우:94043)
(74) 대리인
박장원

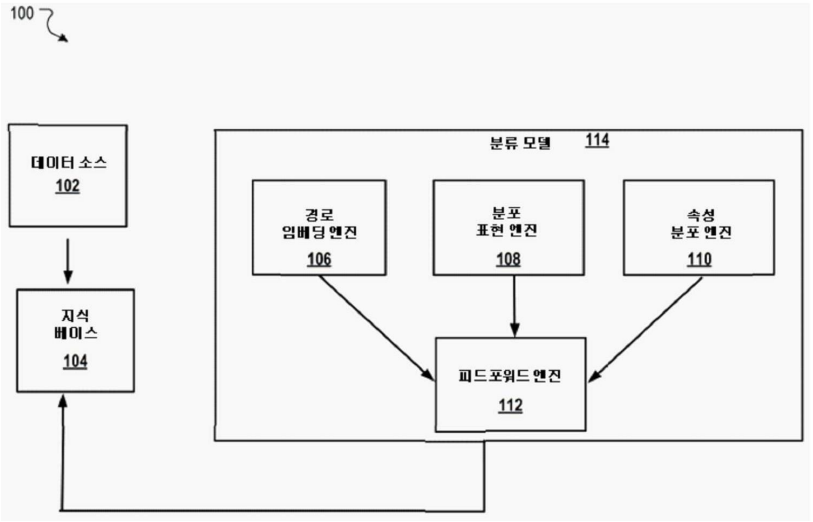
전체 청구항 수 : 총 24 항

(54) 발명의 명칭 **엔티티-속성 관계 식별**

(57) 요 약

방법, 시스템 및 장치는 텍스트 코퍼스에서 엔티티-속성 관계를 쉽게 식별할 수 있는 컴퓨터 저장 매체에 인코딩된 컴퓨터 프로그램을 포함한다. 방법은 후보 엔티티-속성 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함한다. 이 방법은 엔티티 및 속성을 포함하는 문장들에서 단어들에 대한 임베딩을 생성하는 단계를 포함할 수 있다. 이 단계는 또한 엔티티와 연관된 다른 속성들에 기초하여 엔티티에 대한 속성 분포 임베딩을 생성하는 단계와 속성의 공지 엔티티들과 연관된 공지 속성들에 기초하여 속성에 대한 속성 분산 임베딩을 생성하는 단계를 포함할 수 있다. 이러한 임베딩에 기초하여, 피드 포워드 네트워크는 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정할 수 있다.

대 표 도



(52) CPC특허분류
G06F 40/295 (2020.01)

명 세 서

청구범위

청구항 1

컴퓨터 구현 방법으로서,

엔티티 및 속성을 정의하는 엔티티-속성 후보 쌍을 획득하는 단계와, 상기 속성은 엔티티의 후보 속성이고;

엔티티 및 속성을 포함하는 문장 세트에 기초하여, 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하고, 상기 결정하는 단계는:

엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계;

공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계, 상기 엔티티에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍의 엔티티와 연관된 다른 속성에 기초하여 엔티티에 대한 임베딩을 지정하고;

공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계, 상기 속성에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍내의 속성의 공지 엔티티들과 연관된 공지 속성에 기초하는 속성들에 대한 임베딩을 지정하고; 및

문장 세트 내의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 컴퓨터 구현 방법.

청구항 2

제1항에 있어서,

상기 엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계는,

문장 세트에서 엔티티와 속성 사이에 단어들에 대한 제1 임베딩을 지정하는 제1 벡터 표현을 생성하는 단계;

문장 세트에 기초하여 엔티티에 대한 제2 임베딩을 지정하는 제2 벡터 표현을 생성하는 단계; 및

문장 세트에 기초하여 속성에 대한 제3 임베딩을 지정하는 제3 벡터 표현을 생성하는 단계를 포함하는 컴퓨터 구현 방법.

청구항 3

제2항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계를 포함하고; 그리고

상기 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계를 포함하는 컴퓨터 구현 방법.

청구항 4

제3항에 있어서,

상기 문장 세트의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는,

제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보

쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 컴퓨터 구현 방법.

청구항 5

제4항에 있어서,

상기 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는 피드 포워드 네트워크를 사용하여 수행되는 컴퓨터 구현 방법.

청구항 6

제5항에 있어서,

상기 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍 내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는,

제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현을 연결함으로써 단일 벡터 표현을 생성하는 단계;

단일 벡터 표현을 피드 포워드 네트워크에 입력하는 단계; 및

피드 포워드 네트워크에 의해 단일 벡터 표현을 사용하여, 엔티티-속성 후보 쌍의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 컴퓨터 구현 방법.

청구항 7

제3항 내지 제6항 중 어느 한 항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계는,

공지 엔티티-속성 쌍의 엔티티와 연관된 속성 세트를 식별하는 단계, 상기 속성 세트는 속성을 포함하지 않으며; 및

속성 세트에 있는 속성들의 가중치 합을 계산함으로써 엔티티에 대한 속성 분포 임베딩을 생성하는 단계를 포함하는 컴퓨터 구현 방법.

청구항 8

제3항 내지 제7항 중 어느 한 항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계는,

속성을 사용하여, 공지 엔티티-속성 쌍 중에서 엔티티 세트를 식별하는 단계;

엔티티 세트의 각 엔티티에 대해, 엔티티와 연관된 속성 세트를 식별하는 단계, 상기 속성 세트는 속성을 포함하지 않으며; 및

속성 세트에 있는 속성들의 가중치 합을 계산함으로써 속성에 대한 속성 분포 임베딩을 생성하는 단계를 포함하는 컴퓨터 구현 방법.

청구항 9

시스템으로서,

명령들을 저장하는 하나 이상의 메모리 디바이스와; 그리고

하나 이상의 메모리 디바이스와 상호 작용하고 명령들의 실행시 동작들을 수행하도록 구성된 하나 이상의 데이터 처리 장치를 포함하고, 상기 동작들은:

엔티티 및 속성을 정의하는 엔티티-속성 후보 쌍을 획득하는 단계와, 상기 속성은 엔티티의 후보 속성이고;

엔티티 및 속성을 포함하는 문장 세트에 기초하여, 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하고, 상기 결정하는 단계는:

엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계;

공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계, 상기 엔티티에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍의 엔티티와 연관된 다른 속성에 기초하여 엔티티에 대한 임베딩을 지정하고;

공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계, 상기 속성에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍내의 속성의 공지 엔티티들과 연관된 공지 속성에 기초하는 속성들에 대한 임베딩을 지정하고; 그리고

문장 세트의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 시스템.

청구항 10

제9항에 있어서,

상기 엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계는,

문장 세트에서 엔티티와 속성 사이에 단어들에 대한 제1 임베딩을 지정하는 제1 벡터 표현을 생성하는 단계;

문장 세트에 기초하여 엔티티에 대한 제2 임베딩을 지정하는 제2 벡터 표현을 생성하는 단계; 및

문장 세트에 기초하여 속성에 대한 제3 임베딩을 지정하는 제3 벡터 표현을 생성하는 단계를 포함하는 시스템.

청구항 11

제10항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계를 포함하고; 그리고

상기 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계를 포함하는 시스템.

청구항 12

제11항에 있어서,

상기 문장 세트의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는,

제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 시스템.

청구항 13

제12항에 있어서,

상기 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는 피드 포워드 네트워크를 사용하여 수행되는 시스템.

청구항 14

제13항에 있어서,

상기 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍 내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는,

제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현을 연결함으로써 단일 벡터 표현을 생성하는 단계;

단일 벡터 표현을 피드 포워드 네트워크에 입력하는 단계; 및

피드 포워드 네트워크에 의해 단일 벡터 표현을 사용하여, 엔티티-속성 후보 쌍의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 시스템.

청구항 15

제11항 내지 제14항 중 어느 한 항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계는,

공지 엔티티-속성 쌍의 엔티티와 연관된 속성 세트를 식별하는 단계, 상기 속성 세트는 속성을 포함하지 않으며; 및

속성 세트에 있는 속성들의 가중치 합을 계산함으로써 엔티티에 대한 속성 분포 임베딩을 생성하는 단계를 포함하는 시스템.

청구항 16

제11항 내지 제15항 중 어느 한 항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계는,

속성을 사용하여, 공지 엔티티-속성 쌍 중에서 엔티티 세트를 식별하는 단계;

엔티티 세트의 각 엔티티에 대해, 엔티티와 연관된 속성 세트를 식별하는 단계, 상기 속성 세트는 속성을 포함하지 않으며; 및

속성 세트에 있는 속성들의 가중치 합을 계산함으로써 속성에 대한 속성 분포 임베딩을 생성하는 단계를 포함하는 시스템.

청구항 17

하나 이상의 데이터 처리 장치에 의해 실행될 때 하나 이상의 데이터 처리 장치로 하여금 동작들을 수행하게 하는 명령들을 저장하는 비-일시적 컴퓨터 판독 가능 매체로서, 상기 동작들은:

엔티티 및 속성을 정의하는 엔티티-속성 후보 쌍을 획득하는 단계와, 상기 속성은 엔티티의 후보 속성이고;

엔티티 및 속성을 포함하는 문장 세트에 기초하여, 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하고, 상기 결정하는 단계는:

엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계;

공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계, 상기 엔티티에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍의 엔티티와 연관된 다른 속성에 기초하여 엔티티에 대한 임베딩을 지정하고;

공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계, 상기 속성에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍내의 속성의 공지 엔티티들과 연관된 공지 속성에 기초하는 속성들에 대한 임베딩을 지정하고; 및

문장 세트 내의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 비-일시적 컴퓨터 판독 가능 매체.

청구항 18

제17항에 있어서,

상기 엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계는,

문장 세트에서 엔티티와 속성 사이에 단어들에 대한 제1 임베딩을 지정하는 제1 벡터 표현을 생성하는 단계;

문장 세트에 기초하여 엔티티에 대한 제2 임베딩을 지정하는 제2 벡터 표현을 생성하는 단계; 및

문장 세트에 기초하여 속성에 대한 제3 임베딩을 지정하는 제3 벡터 표현을 생성하는 단계를 포함하는 비-일시적 컴퓨터 판독 가능 매체.

청구항 19

제18항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계를 포함하고; 그리고

상기 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계를 포함하는 비-일시적 컴퓨터 판독 가능 매체.

청구항 20

제19항에 있어서,

상기 문장 세트의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는,

제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 비-일시적 컴퓨터 판독 가능 매체.

청구항 21

제20항에 있어서,

상기 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는 피드 포워드 네트워크를 사용하여 수행되는 비-일시적 컴퓨터 판독 가능 매체.

청구항 22

제21항에 있어서,

상기 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍 내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는,

제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현을 연결함으로써 단일 벡터 표현을 생성하는 단계;

단일 벡터 표현을 피드 포워드 네트워크에 입력하는 단계; 및

피드 포워드 네트워크에 의해 단일 벡터 표현을 사용하여, 엔티티-속성 후보 쌍의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하는 비-일시적 컴퓨터 판독 가능 매체.

청구항 23

제19항 내지 제22항 중 어느 한 항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계는,

공지 엔티티-속성 쌍의 엔티티와 연관된 속성 세트를 식별하는 단계, 상기 속성 세트는 속성을 포함하지 않으며; 및

속성 세트에 있는 속성들의 가중치 합을 계산함으로써 엔티티에 대한 속성 분포 임베딩을 생성하는 단계를 포함하는 비-일시적 컴퓨터 판독 가능 매체.

청구항 24

제19항 내지 제23항 중 어느 한 항에 있어서,

상기 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계는,

속성을 사용하여, 공지 엔티티-속성 쌍 중에서 엔티티 세트를 식별하는 단계;

엔티티 세트의 각 엔티티에 대해, 엔티티와 연관된 속성 세트를 식별하는 단계, 상기 속성 세트는 속성을 포함하지 않으며; 및

속성 세트에 있는 속성들의 가중치 합을 계산함으로써 속성에 대한 속성 분포 임베딩을 생성하는 단계를 포함하는 비-일시적 컴퓨터 판독 가능 매체.

발명의 설명

기 술 분 야

[0001] 본 출원은 국제 출원이며 2019년 7월 5에 출원된 미국 출원 번호 16/504,068의 이점을 청구한다. 전술한 출원의 개시는 그 전체가 참고로 본 명세서에 포함된다.

[0002] 본 명세서는 텍스트 코퍼스(말뭉치)에서 엔티티-귀속 관계를 식별하는 것에 관한 것이다.

배 경 기 술

[0003] 검색 기반 애플리케이션(예를 들어, 검색 엔진, 지식 베이스)의 목표는 사용자의 정보 요구와 관련된 리소스(예를 들어, 웹 페이지, 이미지, 텍스트 문서 및 멀티미디어 컨텐츠)를 식별하여 사용자에게 가장 유용한 방식으로 리소스에 관한 정보를 제시하는 것이다. 검색 기반 애플리케이션이 식별된 리소스에 관한 정보를 제시할 수 있는 하나의 방법은 구조화된 검색 결과의 형태이다. 구조화된 검색 결과는 일반적으로 사용자 요청(예를 들어, 쿼리)에 지정된 엔터티에 대한 답변과 함께 속성 목록을 제시한다. 예를 들어, "케빈 듀란트"에 대한 쿼리에 응답하여, 구조화된 검색 결과에는 급여, 팀, 출생 연도, 가족 등과 같은 "케빈 듀란트"에 대한 속성과 이들 속성에 관한 정보를 제공하는 답변이 함께 포함될 수 있다.

[0004] 이러한 구조화된 검색 결과를 구성하는 것은 일반적으로 엔티티-속성 관계를 식별하는 것을 필요로 한다. 엔티티-속성 관계는 한 쌍의 용어 사이의 텍스트 관계의 특별한 케이스이다. 한 쌍의 용어내의 제1 용어는 사람, 장소, 조직, 개념 등일 수 엔티티이다. 한 쌍의 용어내의 제2 용어는 그 엔티티의 일부 양태 또는 특성을 설명하는 문자열인 속성이다. 속성의 예는 개인의 "생년월일", 국가의 "인구", 운동 선수의 "급여" 또는 조직의 "CEO"를 포함할 수 있다.

[0005] 그러나, 적절한 검색 결과를 식별하기 위해서는 대량의 데이터가 종종 처리되어야 한다. 이것은 많은 양의 처리 능력을 사용할 수 있다. 또한 이헐게 많은 양의 데이터를 처리해야 하는 경우에는 결과를 얻는 속도가 느려지거나 프로세스가 손상되어 결과의 품질 저하를 야기할 수 있다.

발명의 내용

[0006] 일반적으로, 본 명세서에 기술된 주제의 하나의 혁신적인 측면은 엔티티 및 속성을 정의하는 엔티티-속성 후보 쌍을 획득하는 단계와, 상기 속성은 엔티티의 후보 속성이고; 엔티티 및 속성을 포함하는 문장 세트에 기초하여, 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하고, 상기

결정하는 단계는: 엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계와; 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계와, 상기 엔티티에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍의 엔티티와 연관된 다른 속성에 기초하여 엔티티에 대한 임베딩을 지정하고; 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계와, 상기 속성에 대한 속성 분포 임베딩은 공지 엔티티-속성 쌍내의 속성의 공지 엔티티들과 연관된 공지 속성에 기초하는 속성들에 대한 임베딩을 지정하고; 그리고 문장 세트 내의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계의 동작들을 포함하는 방법으로 구현될 수 있다. 이 양태의 다른 실시예는 방법들의 동작들을 수행하도록 구성된 대응하는 시스템, 디바이스, 장치 및 컴퓨터 프로그램을 포함한다. 컴퓨터 프로그램(예를 들어, 명령들)은 컴퓨터 저장 디바이스에 인코딩될 수 있다. 이들 및 다른 실시예는 각각 선택적으로 다음 특징들 중 하나 이상을 포함할 수 있다.

[0007] 일부 구현에서, 엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계는 문장 세트에서 엔티티와 속성 사이에 단어들에 대한 제1 임베딩을 지정하는 제1 벡터 표현을 생성하는 단계와; 문장 세트에 기초하여 엔티티에 대한 제2 임베딩을 지정하는 제2 벡터 표현을 생성하는 단계와; 그리고 문장 세트에 기초하여 속성에 대한 제3 임베딩을 지정하는 제3 벡터 표현을 생성하는 단계를 포함한다.

[0008] 일부 구현에서, 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계를 포함한다.

[0009] 일부 구현에서, 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계는 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계를 포함한다.

[0010] 일부 실시예에서, 문장 세트의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함한다.

[0011] 일부 구현에서, 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는 피드 포워드 네트워크를 사용하여 수행된다.

[0012] 일부 구현에서, 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현에 기초하여, 엔티티-속성 후보 쌍 내의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계는 제1 벡터 표현, 제2 벡터 표현, 제3 벡터 표현, 제4 벡터 표현 및 제5 벡터 표현을 연결함으로써 단일 벡터 표현을 생성하는 단계와; 단일 벡터 표현을 피드 포워드 네트워크에 입력하는 단계와; 그리고 피드 포워드 네트워크에 의해 단일 벡터 표현을 사용하여, 엔티티-속성 후보 쌍의 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함한다.

[0013] 일부 구현에서, 공지 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성하는 단계는 공지 엔티티-속성 쌍의 엔티티와 연관된 속성 세트를 식별하는 단계와, 상기 속성 세트는 속성을 포함하지 않으며; 그리고 속성 세트에 있는 속성들의 가중치 합을 계산함으로써 엔티티에 대한 속성 분포 임베딩을 생성하는 단계를 포함한다.

[0014] 일부 구현에서, 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성하는 단계는 속성을 사용하여, 공지 엔티티-속성 쌍 중에서 엔티티 세트를 식별하는 단계와; 엔티티 세트의 각 엔티티에 대해, 엔티티와 연관된 속성 세트를 식별하는 단계와, 상기 속성 세트는 속성을 포함하지 않으며; 그리고 속성 세트에 있는 속성들의 가중치 합을 계산함으로써 속성에 대한 속성 분포 임베딩을 생성하는 단계를 포함한다..

[0015] 본 명세서에 설명된 주제의 특정 실시 예는 종래 기술의 모델 기반 엔티티-속성 식별 기술에 비해 더 정확한 엔티티-속성 관계를 식별하는 이점을 실현하기 위해 구현될 수 있다. 선행 기술 엔티티 속성 식별 기술은 이러한 용어가 나타나는 데이터(예를 들어, 문장)를 기반으로 엔티티와 속성을 표현하여 엔티티-속성 관계를 식별하는 다양한 모델 기반 접근 방식(예를 들어, 자연어 처리(NLP) 기능, 원격 감독 및 전통적인 기계 학습 모델)을 사

용한다. 대조적으로, 이 사양에 설명된 혁신은 이러한 용어가 나타나는 데이터에서 엔티티 및 속성이 설명되는 방식에 대한 정보를 사용하는 것뿐만 아니라 공지 다른 속성을 사용하여 엔티티 및 속성을 표현함으로써 데이터 세트에서 엔티티-속성 관계를 식별한다. 이 용어와 연관된다. 이를 통해 유사한 엔터티가 공유하는 속성으로 엔터티 및 속성을 표현할 수 있으며, 이러한 용어가 나타나는 문장을 고려하는 것만으로는 식별할 수없는 엔터티 속성 관계를 식별하는 정확도가 향상된다. 따라서 사용자가 검색을 수행하면 더 유용한 정보를 얻을 수 있다. 특히, 검색어의 범위가 넓을수록 더 관련성 높은 정보가 식별된다. 따라서 반복 검색이 반드시 필요한 것은 아니다. 따라서 원하는 검색 결과를 얻기 위해 더 적은 처리가 필요할 수 있다.

[0016] 예를 들어, 데이터 세트에 "레코드"속성을 사용하여 설명되는 "호날두" 및 "메시"라는 두 개의 엔티티가있는 문장과 "목표" 속성을 사용하여 엔티티 "메시"가 설명되는 문장이 포함된 시나리오를 고려해 보자. 그러한 시나리오에서, 종래 기술은 다음과 같은 엔티티 속성 쌍(호날두, 레코드),(메시, 레코드) 및(메시, 목표)을 식별할 수 있다. 이 명세서에 기술된 혁신은 이러한 용어가 데이터 세트에서 어떻게 사용되는지에 의해 쉽게 식별되지 않을 수 있는 엔티티-속성 관계를 식별함으로써 이러한 선행 기술 접근 방식을 뛰어 넘는다. 위의 예를 사용하여 이 사양에 설명된 혁신은 "호날두"와 "메시"가 "기록" 속성을 공유하기 때문에 유사한 엔티티라고 판단한 다음 "골" 속성을 사용하여 "기록"속성을 나타낸다.

[0017] 이러한 방식으로, 예를 들어,이 명세서에 설명된 혁신은 그러한 관계가 데이터 세트에서 쉽게 식별되지 않을 수 있더라도 엔티티-속성 관계(예를 들어, 크리스티아노, 골)를 식별할 수 있게 한다.

도면의 간단한 설명

[0018] 도 1은 엔티티-속성 관계를 추출하기 위한 예시적인 환경의 블록도이다.

도 2는 엔티티-속성 관계를 식별하기 위한 예시적인 프로세스의 흐름도이다.

도 3은 예시적인 컴퓨터 시스템의 블록도이다.

다양한 도면에서 유사한 참조 번호 및 지정은 유사한 요소를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0019] 본 명세서는 텍스트 코퍼스(말뭉치)에서 엔티티-속성 관계를 식별하는 것에 관한 것이다.

[0020] 본 명세서에서 더 설명하는 바와 같이, 후보 엔티티-속성 쌍(속성이 엔티티의 후보 속성인 경우)이 분류 모델에 입력된다. 경로 임베딩 엔진, 분포 (distributional) 표현 엔진, 속성 분포 엔진 및 피드 포워드 네트워크를 포함하는 분류 모델은 후보 엔티티-속성 쌍에 있는 속성이 후보 엔티티-속성 쌍에 있는 엔티티의 실제 속성인지 여부를 결정한다.

[0021] 경로 임베딩 엔진은 데이터 세트의 문장 세트(예를 들어, 30개 이상의 문장)에서 엔티티와 속성의 공동(joint) 발생을 연결하는 경로들 또는 단어들의 임베딩을 나타내는 벡터를 생성한다. 분포 표현 엔진은 이들 용어가 문장 세트에 나타나는 컨텍스트에 기초하여 엔티티 및 속성 용어들에 대한 임베딩을 나타내는 벡터를 생성한다. 속성 분포 엔진은 엔티티에 대한 임베딩을 나타내는 벡터와 속성에 대한 임베딩을 나타내는 다른 벡터를 생성한다. 엔티티에 대한 속성 배포 엔진의 임베딩은 데이터 세트에서 엔티티와 연관된 것으로 알려진 다른 속성(예를 들어, 후보 속성 이외의 속성들)에 기초한다. 속성에 대한 속성 분포 엔진의 임베딩은 후보 속성의 공지(된) 엔티티들과 연관된 다른 속성에 기초한다.

[0022] 분류 모델은 경로 임베딩 엔진, 분포 표현 엔진 및 속성 분포 엔진으로부터의 벡터 표현을 단일 벡터 표현으로 연결(concatenates)한다. 따라서 이러한 연결된 벡터들에 단일 벡터 표현이 사용되기 때문에 더 적은 데이터가 저장되어야 한다. 그런 다음 분류 모델은 단일 벡터 표현을 사용하여 후보 엔티티-속성 쌍내의 속성이 후보 엔티티-속성 쌍내의 엔티티의 실제 속성인지 여부를 결정하는 피드 포워드 네트워크로 단일 벡터 표현을 입력한다. 만약 피드 포워드 네트워크가 후보 엔티티-속성 쌍내의 속성이 후보 엔티티-속성 쌍내의 엔티티의 실제 속성이라고 결정하면, 그 후보 엔티티-속성 쌍은 다른 공지/실제 엔티티-속성 쌍과 함께 지식 베이스에 저장된다.

[0023] 이러한 특징들과 추가 특징들은 도 1 내지 도 3을 참조하여 아래에서 더 상세히 설명된다.

[0024] 도 1은 엔티티 속성 관계를 추출하기 위한 예시적인 환경의 블록도이다. 환경(100)은 지식 베이스(104)에 있는 후보 엔티티-속성 쌍들에 대해, 후보 엔티티-속성 쌍의 속성이 그 후보 쌍의 엔티티의 실제 속성인지 여부를 결

정하는 분류 모델(114)을 포함한다. 일부 구현에서, 분류 모델(114)은 신경망 모델이고, 그 구성 요소들/엔진들(및 그들 각각의 동작들)은 아래에서 설명된다. 분류 모델(114)은 또한 다른 유형의 지도 및/또는 비지도 기계 학습 모델을 사용하여 구현될 수 있음을 이해할 것이다.

[0025] 하나 이상의 비-일시적 데이터 저장 매체(예를 들어, 하드 드라이브(들), 플래시 메모리 등)에 저장된 하나 이상의 데이터베이스(또는 다른 적절한 데이터 저장 구조)를 포함할 수 있는 지식 베이스(104)는 후보 엔티티-속성 쌍 세트를 저장한다. 후보 엔티티 속성 쌍은 뉴스 웹 사이트, 데이터 수집 플랫폼, 소셜 미디어 플랫폼 등과 같은 임의의 컨텐츠 소스를 포함할 수 있는 데이터 소스(102)로부터 획득된 텍스트 문서(예를 들어, 웹 페이지, 뉴스 기사 등) 형태의 컨텐츠 세트를 사용하여 획득될 수 있다. 일부 구현에서, 데이터 소스(102)는 데이터 수집 플랫폼으로부터 뉴스 기사를 획득한다. 일부 구현에서, 데이터 소스(102)는 모델(예를 들어, 지도 또는 기계 학습 모델 , 자연어 처리 모델)을 사용하여, 기사에서 문장들을 추출하고, 품사 및 종속성 구문 분석 트리 태그를 사용하여 상기 추출된 문장을 예를 들어 엔티티 및 속성으로 토큰화 및 라벨링함으로써 후보 엔티티-속성 쌍 세트를 생성할 수 있다. 일부 구현에서, 데이터 소스(102)는 추출된 문장을 기계 학습 모델에 입력할 수 있으며, 이는 예를 들어 일련의 트레이닝 문장 및 그와 관련된 엔티티-속성 쌍을 사용하여 트레이닝될 수 있다. 그런 다음 이러한 기계 학습 모델은 입력 추출된 문장에 대한 후보 엔티티-속성 쌍을 출력할 수 있다.

[0026] 데이터 소스(102)는 지식 베이스(104)에, 후보 엔티티-속성 쌍의 단어들을 포함하는 데이터 소스(102)에 의해 추출된 문장들과 함께 후보 엔티티-속성 쌍을 저장한다. 일부 구현에서, 후보 엔티티-속성 쌍들은 엔티티 및 속성이 존재하는 문장의 수가 임계 문장 수(예를 들어, 30개의 문장)를 만족(예를 들어, 충족 또는 초과)하는 경우에만 지식 베이스(104)에 저장된다. 따라서 데이터 저장 여부를 결정하기 위한 이 임계값을 설정함으로써 저장되는 데이터의 양이 감소된다.

[0027] 분류 모델(114)은 (지식 베이스(104)에 저장된) 후보 엔티티-속성 쌍내의 속성이 그 후보 엔티티-속성 쌍내의 엔티티의 실제 속성인지 여부를 결정한다. 분류 모델(114)은 경로 임베딩 엔진(106), 분포 표현 엔진(108), 속성 분포 엔진(110) 및 피드 포워드 네트워크(112)를 포함한다. 본 명세서에서 사용되는 바와같이 엔진이라는 용어는 일련의 작업(task)을 수행하는 데이터 처리 장치를 지칭한다. 후보 엔티티-속성 쌍의 속성이 엔티티의 실제 속성인지 여부를 결정하는 분류 모델(114)의 이들 엔진 각각의 동작은 도 2를 참조하여 설명된다.

[0028] 도 2는 엔티티 속성 관계를 식별하기 위한 예시적인 프로세스(200)의 흐름도이다. 프로세스(200)의 동작들은 도 1에 설명되고 도시된 시스템의 구성 요소에 의해 수행되는 것으로 아래에서 설명된다. 프로세스(200)의 동작들은 예시 목적으로만 아래에서 설명된다. 프로세스(200)의 동작들은 임의의 적절한 디바이스 또는 시스템, 예를 들어 임의의 적절한 데이터 처리 장치에 의해 수행될 수 있다. 프로세스(200)의 동작들은 또한 비-일시적 컴퓨터 판독 가능 매체에 저장된 명령들(명령어)로 구현될 수 있다. 명령들의 실행은 하나 이상의 데이터 처리 장치로 하여금 프로세스(200)의 동작들을 수행하게 한다.

[0029] 지식 베이스(104)는 도 1을 참조하여 설명된 바와 같이(202에서) 데이터 소스(104)로부터 엔티티-속성 후보 쌍을 획득한다.

[0030] 지식 베이스(104)는 도 1을 참조하여 전술한 바와 같이(204에서) 후보 엔티티-속성 쌍의 속성 및 엔티티의 단어들을 포함하는 문장 세트를 데이터 소스(102)로부터 획득한다.

[0031] 분류 모델(114)은 문장 세트 및 후보 엔티티-속성 쌍에 기초하여, 후보 속성이 후보 엔티티의 실제 속성인지 여부를 결정한다. 일부 구현에서, 문장 세트는 다수의 문장, 예를 들어 30개 이상의 문장일 수 있다. 분류 모델은 다음 동작들을 수행하여 이러한 결정을 내린다. 동작들은 (1) 동작(206, 208 및 210)을 참조하여 아래에서 더 자세히 설명되는, 엔티티 및 속성을 포함하는 문장 세트의 단어들에 대한 임베딩을 생성하는 단계와; (2) 동작(212)을 참조하여 아래에서 더 자세히 설명되는, 엔티티-속성 쌍을 사용하여, 엔티티에 대한 속성 분포 임베딩을 생성하는 단계와; (3) 동작(214)을 참조하여 아래에서 더 자세히 설명되는, 공지 엔티티-속성 쌍을 사용하여, 속성에 대한 속성 분포 임베딩을 생성하는 단계; 및 (4) 문장 세트의 단어들에 대한 임베딩, 엔티티에 대한 속성 분포 임베딩 및 속성에 대한 속성 분포 임베딩에 기초하여, 엔티티-속성 후보 쌍의 속성이 엔티티-속성 후보 쌍의 엔티티의 실제 속성인지 여부를 결정하는 단계를 포함하며, 이는 동작(216)을 참조하여 아래에서 더 자세히 설명된다. 동작(206-216)은 아래에 설명되어 있다.

[0032] 경로 임베딩 엔진(106)은 문장 세트에 있는 엔티티와 속성 사이의 제1 단어 임베딩을 지정하는 제1 벡터 표현을 생성한다(206). 경로 임베딩 엔진(106)은 문장 세트에 있는 이들 용어의 공동 발생을 연결하는 경로들 또는 단어들을 임베딩함으로써 후보 엔티티-속성 용어들 간의 관계를 검출한다. 예를 들어, "뱀은 파충류이다(snake is

a reptile)"라는 문구의 경우, 경로 임베딩 엔진(106)은 경로 "is a"에 대한 임베딩을 생성하는데, 이는 예를 들어 종-속(genus-species) 관계를 검출하는데 사용되고 이어서 다른 엔티티 속성 쌍들을 식별할 수 있다. 이러한 경로를 생성함으로써 이러한 용어의 분석 및 검출을 수행하기 위해 처리량이 감소된다.

[0033] 경로 임베딩 엔진(106)은 문장 세트에서 엔티티와 속성 사이의 단어 임베딩을 생성하기 위해 다음 동작들을 수행한다. 문장 세트에 있는 각 문장에 대해, 경로 임베딩 엔진(106)은 먼저 엔티티와 속성 사이의 종속성 경로(단어 세트를 지정함)를 추출한다. 경로 임베딩 엔진(106)은 문장을 문자열에서 목록으로 변환하는데, 여기서 첫 번째 용어는 엔티티이고 마지막 용어는 속성이다(또는 대안적으로, 첫 번째 용어는 속성이고 마지막 용어는 엔티티이다). 종속성 경로에서 각 용어(에지로도 지칭됨)은 용어의 기본형, 품사 태그, 종속성 라벨 및 종속성 경로 방향(왼쪽, 오른쪽 또는 루트)와 같은 피처들을 사용하여 표현된다. 이들 피처 각각은 임베딩되고 연결되어 아래 방정식과 같이 벡터 시퀀스(V_l, V_{pos}, V_{dep}, V_{dir})를 포함하는 용어 또는 에지(V_e)에 대한 벡터 표현을 생성한다.

[0034] $$\vec{v}_e = [\vec{v}_l, \vec{v}_{pos}, \vec{v}_{dep}, \vec{v}_{dir}]$$

[0035] 이어서 경로 임베딩 엔진(106)은 각 경로의 항들 또는 에지들에 대한 벡터 시퀀스를 LSTM(장단기 기억) 네트워크에 입력하여 아래 방정식으로 도시된 바와같이 문장(V_s)에 대한 단일 벡터 표현을 생성한다.

[0036] $$\vec{v}_s = LSTM(\vec{v}_e^{(1)} \ldots \vec{v}_e^{(k)})$$

[0037] 이러한 단일 벡터 표현은 저장, 처리 또는 전송해야 하는 데이터의 양을 감소시킨다. 이는 단일 벡터가 단일 벡터가 나타내는 모든 벡터에 필요한 데이터보다 훨씬 적은 데이터를 사용하기 때문이다.

[0038] 마지막으로, 경로 임베딩 엔진(106)은 문장 세트에 있는 모든 문장에 대한 단일 벡터 표현을 어텐션 메커니즘에 입력하여 아래 방정식으로 도시된 바와 같이 문장 표현($V_{sents(e,a)}$)의 가중치 평균을 결정한다.

[0039] $$\vec{v}_{sents(e,a)} = ATTN(\vec{v}_s^{(1)} \ldots \vec{v}_s^{(n)})$$

[0040] 분포 표현 엔진(108)은 문장 세트에 기초하여 엔티티에 대한 제2 벡터 표현 및 속성에 대한 제3 벡터 표현을 생성한다(208 및 210). 분포 표현 엔진(108)은 문장 세트에서 후보 엔티티-속성 쌍의 엔티티 및 속성이 발생하는 컨텍스트에 기초하여 후보 엔티티-속성 용어 간의 관계를 검출한다. 이 처리에 사용되는 벡터 기반 접근 방식은 처리량을 줄여 결과를 얻는 속도를 높인다. 예를 들어, 분포 표현 엔진(108)은 엔티티 "뉴욕"이 이 엔티티가 미국의 도시 또는 주를 지칭함을 제안하는 방식으로 문장 세트에서 사용된다고 결정할 수 있다. 다른 예로, 분포 표현 엔진(108)은 속성 "수도"가 이 속성이 주 또는 국가 내의 중요한 도시를 지칭함을 제안하는 방식으로 문장 세트에서 사용된다고 결정할 수 있다. 따라서, 분포 표현 엔진(108)은 엔티티가 나타나는 컨텍스트(즉, 문장 세트)를 사용하여 엔티티(V_e)에 대한 임베딩을 지정하는 벡터 표현을 생성한다. 유사하게, 분포 표현 엔진(108)은 속성이 나타나는 문장 세트를 사용하여 속성에 대한 임베딩을 지정하는 벡터 표현 (V_a)을 생성한다.

[0041] 속성 분포 엔진(110)은 공지 엔티티-속성 쌍을 사용하여 엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현을 생성한다(212). 지식 베이스(104)에 저장되는 공지 엔티티-속성 쌍은 (예를 들어, 분류 모델(114)에 의한 사전 처리를 사용하거나 인간 평가에 기초하여) 엔티티-속성 쌍에 있는 각 속성이 엔티티-속성 쌍에 있는 엔티티의 실제 속성임을 확인한 엔티티-속성 쌍이다.

[0042] 일부 구현에서, 속성 분포 엔진(110)은 다음 동작들을 수행하여, 엔티티가 연관되어 있는 공지 엔티티-속성 쌍 중 일부(예를 들어, 가장 일반적인) 또는 다른 공지 속성의 전부를 사용하여 엔티티에 대한 임베딩을 지정하는 속성 분포 임베딩을 결정한다. 엔티티-속성 후보 쌍의 엔티티에 대해, 속성 분포 엔진(110)은 공지 엔티티-속성 쌍의 엔티티와 연관된 다른 속성(즉, 엔티티-속성 후보 쌍에 포함된 것과 다른 속성)을 식별한다. 예를 들어, 후보 엔티티-속성 쌍(마이클 조던, 유명한)의 엔티티 "마이클 조던"에 대해, 속성 분포 엔진(110)은 (마이클 조던, 부자) 및 (마이클 조던, 기록)과 같은 마이크로 조던에 대한 공지 엔티티-속성 쌍을 사용하여 부자 및 기록과 같은 속성을 식별할 수 있다.

[0043] 그런 다음 속성 분포 엔진(110)은 (이전 단락에 설명된 바와같이) 식별된 공지 속성들의 가중치 합을 계산하여 엔티티에 대한 임베딩을 생성하는데, 여기서 가중치는 아래 방정식으로 도시된 바와같이 어텐션(주의) 메커니즘

을 사용하여 학습된다.

[0044] $$\vec{v}_{\mathbf{e}} = ATTN(\mathcal{E}(\alpha_1)...\mathcal{E}(\alpha_m))$$

[0045] 속성 분포 엔진(110)은 공지 엔티티-속성 쌍을 사용하여 속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현을 생성한다(214). 일부 구현에서, 속성 분포 엔진(110)은 다음 동작들을 수행하여, 후보 속성의 공지 엔티티들과 연관된 공지 속성의 일부(예를 들어, 가장 일반적인) 또는 모든 공지 속성에 기초하여 속성에 대한 표현을 결정한다. 엔티티-속성 후보 쌍의 속성에 대해, 속성 분포 엔진(110)은 속성을 갖는 공지 엔티티-속성 쌍 중에서 공지 엔티티들을 식별한다. 식별된 공지 엔티티 각각에 대해, 속성 분포 엔진(110)은 공지 엔티티-속성 쌍에서 엔티티와 연관된 다른 속성(즉, 엔티티-속성 후보 쌍에 포함된 것과 다른 속성)을 식별한다. 일부 구현에서, 속성 분포 엔진(110)은 (1) 각 속성과 연관된 공지 엔티티의 수에 기초하여 속성들의 순위를 매기고(예를 들어, 더 적은 수의 엔티티와 연관된 속성보다 더 많은 수의 엔티티와 연관된 속성에 더 높은 순위 할당) 및 (2) 이어서 순위에 기초하여 임계 속성 수를 선택(예를 들어, 상위 5개의 가장 높은 순위의 속성 선택)함으로써 식별된 속성들 중에서 속성 서브세트를 식별할 수 있다..

[0046] 이어서 속성 분포 엔진(110)은 (이전 단락에 설명된 바와같이) 식별된 공지 속성들의 (전체 또는 서브세트의) 가중치 합을 계산하여 속성에 대한 임베딩을 생성하는데, 여기서 가중치는 아래 방정식에 도시된 바와같이 어텐션 메커니즘을 사용하여/통해 학습된다.

[0047] $$\vec{v}_{\boldsymbol{a}} = ATTN(\mathcal{E}(\alpha_1)...\mathcal{E}(\alpha_m))$$

[0048] 동작(206 내지 214)에 의해 생성된 임베딩은 일반적으로 개별 엔진(106, 108, 110)에 의해 병렬로 생성된다는 것을 알 수 있을 것이다.

[0049] 피드 포워드 네트워크(112)는 벡터 표현에 기초하여, 엔티티-속성 후보 쌍의 속성이 엔티티-속성 후보 쌍의 엔티티의 실제 속성인지 여부를 결정한다(216). 일부 구현에서, 피드 포워드 네트워크(112)는 경로 임베딩 엔진(106), 분포 표현 엔진(108) 및 속성 분포 엔진(110)에 의해 출력된 각각의 벡터 표현을 아래 방정식으로 도시된 바와 같이 단일 벡터 표현($V_{(e,a)}$)으로 연결한다 :

[0050] $$\vec{v}_{(e,a)} = [\vec{v}_{\mathbf{e}}, \vec{v}_{e}, \vec{v}_{sents(e,a)}, \vec{v}_{a}, \vec{v}_{\mathbf{a}}]$$

[0051] 이 입력 단일 벡터 표현을 사용하여, 피드 포워드 네트워크(112)는 후보 엔티티-속성 쌍의 속성이 후보 엔티티-속성 쌍의 엔티티의 실제 속성인지 여부를 출력한다. 일부 구현에서, 피드 포워드 네트워크(112)의 출력은 바이너리(2진)일 수 있다. 예를 들어, 피드 포워드 네트워크(112)는 후보 엔티티-속성 쌍의 속성이 후보 엔티티-속성 쌍의 엔티티의 실제 속성인 경우 "예"를 출력하고 후보 엔티티-속성 쌍의 속성이 후보 엔티티-속성 쌍의 엔티티의 실제 속성이 아닌 경우 "아니오"를 출력할 수 있다. 이러한 바이너리 출력을 생성하기 위해, 이러한 결과를 얻기 위해 필요한 처리량은 전술한 프로세스에 의해 수행된 다양한 단순화로 인해 감소된다. 일부 구현에서, 피드 포워드 네트워크(112)의 출력은 신뢰도 값, 예를 들어 0 내지 1 범위의 값일 수 있는데, 여기서 0은 후보 엔티티-속성 쌍의 속성이 후보 엔티티-속성 쌍에 있는 엔티티의 실제 속성이 아님을 지정하고, 1은 후보 엔티티-속성 쌍의 속성이 후보 엔티티-속성 쌍에 있는 엔티티의 실제 속성임을 지정한다.

[0052] 일부 구현에서, 피드 포워드 네트워크(112)는 원격 지도를 사용하여 트레이닝된다. 트레이닝은 분류 모델(114)을 사용하여, 참(true) 쌍으로 식별된(즉, 후보 엔티티-속성 쌍의 속성이 예를 들어, 인간 평가 또는 피드 포워드 네트워크 (112)에 의한 사전 처리에 기초하여, 후보 엔티티-속성 쌍의 엔티티에 있는 실제 속성으로 식별된) 후보 엔티티-속성 쌍에 대해 전술한 처리를 수행한다.

[0053] 피드 포워드 네트워크(112)의 출력이 엔티티-속성 후보 쌍의 속성이 실제 속성이거나(예를 들어, 피드 포워드 네트워크(112)가 전술한 바와 같이 "예" 표시자를 출력하는 경우) 또는 그럴 가능성이 높은 것으로(예를 들어, 피드 포워드 네트워크(112)가 전술한 바와 같이 0.8과 같은 특정 임계값을 충족하거나 초과하는 신뢰값을 출력하는 경우) 지정하는 경우, 피드 포워드 네트워크(112)는 엔티티-속성 후보 쌍을 지식 베이스(104)에 실제 엔티티-속성 쌍으로서 저장한다. 이것은 저장된 데이터에 대한 확실성을 제공한다. 게다가, 관련 데이터만 저장되도록 하여 저장된 데이터의 품질이 향상시키고 저장되는 전체 데이터 양을 감소시킨다.

[0054] 도 3은 전술한 동작들을 수행하는데 사용될 수 있는 예시적인 컴퓨터 시스템 (300)의 블록도이다. 시스템(300)

은 프로세서(310), 메모리(320), 저장(storage) 디바이스(330) 및 입/출력 디바이스(340)를 포함한다. 구성 요소(310, 320, 330 및 340) 각각은 예를 들어 시스템 버스(350)를 사용하여 상호 연결될 수 있다. 프로세서(310)는 시스템(300) 내에서 실행하기 위한 명령들을 처리할 수 있다. 일부 구현에서, 프로세서(310)는 단일 스레드 프로세서이다. 다른 구현에서, 프로세서(310)는 다중 스레드 프로세서이다. 프로세서(310)는 메모리(320) 또는 저장 디바이스(330)에 저장된 명령들을 처리할 수 있다.

[0055] 메모리(320)는 시스템(300) 내에 정보를 저장한다. 일 구현에서, 메모리 (320)는 컴퓨터 판독 가능 매체이다. 일부 구현에서, 메모리(320)는 휘발성 메모리 유닛이다. 다른 구현에서, 메모리(320)는 비-휘발성 메모리 유닛이다.

[0056] 저장 디바이스(330)는 시스템(300)에 대용량 스토리지를 제공할 수 있다. 일부 구현에서, 저장 디바이스(330)는 컴퓨터 판독 가능 매체이다. 다양한 다른 구현에서, 저장 디바이스(330)는 예를 들어, 하드 디스크 디바이스, 광 디스크 디바이스, 다수의 컴퓨팅 디바이스(예를 들어, 클라우드 저장 디바이스)에 의해 네트워크를 통해 공유되는 저장 디바이스, 또는 일부 다른 대용량 저장 디바이스를 포함할 수 있다.

[0057] 입/출력 디바이스(340)는 시스템(300)에 대한 입/출력 동작을 제공한다. 일부 구현에서, 입/출력 디바이스(340)는 네트워크 인터페이스 디바이스, 예를 들어 이더넷 카드, 직렬 통신 디바이스(예를 들어 RS-232 포트) 및/또는 무선 인터페이스 디바이스(예를 들어, 802.11 카드) 중 하나 이상을 포함할 수 있다. 다른 구현에서, 입/출력 디바이스는 입력 데이터를 수신하고 출력 데이터를 다른 입/출력 디바이스, 예를 들어 키보드, 프린터 및 디스플레이 디바이스(360)로 전송하도록 구성된 드라이버 디바이스를 포함할 수 있다. 그러나 모바일 컴퓨팅 디바이스, 모바일 통신 디바이스, 셋탑 박스 텔레비전 클라이언트 디바이스 등과 같은 다른 구현도 사용될 수 있다.

[0058] 예시적인 처리 시스템이 도 3에서 설명되었지만, 본 명세서에서 설명된 주제및 기능적 동작들의 구현은 다른 유형의 디지털 전자 회로, 본 명세서에 개시된 구조 및 그의 구조적 등가물을 포함하는 컴퓨터 소프트웨어, 펌웨어 또는 하드웨어 또는 이들 중 하나 이상의 조합으로 구현될 수 있다.

[0059] 본 명세서에 기술된 주제 및 동작들의 실시예는 디지털 전자 회로, 본 명세서에소 설명된 구조 및 그의 구조적 등가물을 포함하는 컴퓨터 소프트웨어, 펌웨어 또는 하드웨어, 또는 이들 중 하나 이상의 조합으로 구현될 수 있다. 본 명세서에 설명된 주제의 실시예는 하나 이상의 컴퓨터 프로그램, 즉 데이터 처리 장치에 의해 실행되거나 데이터 처리 장치의 동작을 제어하기 위해 컴퓨터 저장 미디어(또는 매체)에 인코딩된 컴퓨터 프로그램 명령의 하나 이상의 모듈로 구현될 수 있다. 대안적으로 또는 추가적으로, 프로그램 명령은 인위적으로 생성된 전파 신호, 예를 들어, 데이터 처리 장치에 의한 실행을 위해 적절한 수신기 장치로의 전송을 위해 정보를 인코딩하도록 생성된 기계 생성의 전기, 광학 또는 전자기 신호에 인코딩될 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독 가능 저장 디바이스, 컴퓨터 판독 가능 저장 기판, 랜덤 또는 직렬 액세스 메모리 어레이 또는 디바이스, 또는 이들 중 하나 이상의 조합이거나 이에 포함될 수 있다. 더욱이, 컴퓨터 저장 매체는 전파 신호가 아니지만, 컴퓨터 저장 매체는 인위적으로 생성된 전파 신호로 인코딩된 컴퓨터 프로그램 명령의 소스 또는 목적지일 수 있다. 컴퓨터 저장 매체는 또한 하나 이상의 개별 물리적 구성 요소 또는 매체(예를 들어, 다수의 CD, 디스크 또는 기타 저장 디바이스)일 수 있거나 그에 포함될 수 있다.

[0060] 본 명세서에 설명된 동작들은 하나 이상의 컴퓨터 판독 가능 저장 디바이스에 저장되거나 다른 소스로부터 수신된 데이터에 대해 데이터 처리 장치에 의해 수행되는 동작들로 구현될 수 있다.

[0061] "데이터 처리 장치"라는 용어는 예를 들어 프로그램 가능한 프로세서, 컴퓨터, 시스템 온 칩, 또는 전술한 것의 다수 또는 조합을 포함하여 데이터를 처리하기 위한 모든 종류의 장치, 디바이스 및 기계를 포함한다. 장치는 예를 들어 FPGA(필드 프로그래밍 가능 게이트 어레이) 또는 ASIC(애플리케이션 특정 집적 회로)와 같은 특수 목적 논리 회로를 포함할 수 있다. 장치는 또한 하드웨어에 추가하여, 해당 컴퓨터 프로그램에 대한 실행 환경을 생성하는 코드, 예를 들어, 프로세서 펌웨어, 프로토콜 스택, 데이터베이스 관리 시스템, 운영 체제, 크로스 플랫폼 런타임, 가상 머신 또는 이들 중 하나 이상의 조합을 구성하는 코드를 포함할 수 있다. 장치 및 실행 환경은 웹 서비스, 분산 컴퓨팅 및 그리드 컴퓨팅 인프라와 같은 다양한 컴퓨팅 모델 인프라를 구현할 수 있다.

[0062] 컴퓨터 프로그램(프로그램, 소프트웨어, 소프트웨어 애플리케이션, 스크립트 또는 코드라고도 함)은 컴파일 또는 해석 언어, 선언적 또는 절차적 언어를 포함한 모든 형태의 프로그래밍 언어로 작성될 수 있으며, 독립 실행형 프로그램 또는 모듈, 구성 요소, 서브 루틴, 객체 또는 컴퓨팅 환경에서 사용하기에 적합한 다른 유닛을 포함하여 임의의 형태로 배포될 수 있다. 컴퓨터 프로그램은 파일 시스템의 파일에 해당할 수 있지만 반드시 그럴

필요는 없다. 프로그램은 다른 프로그램 또는 데이터(예를 들어, 마크 업 언어 문서에 저장된 하나 이상의 스크립트)를 보유하는 파일의 일부, 해당 프로그램 전용 단일 파일 또는 다수의 조정 파일(예를 들어, 하나 이상의 모듈, 서브 프로그램 또는 코드 일부를 저장하는 파일)에 저장될 수 있다. 컴퓨터 프로그램은 하나의 컴퓨터 또는 하나의 사이트에 위치하거나 여러 사이트에 분산되고 통신 네트워크로 상호 연결된 다수의 컴퓨터에서 실행되도록 배포될 수 있다.

[0063] 본 명세서에 설명된 프로세스 및 로직 흐름은 입력 데이터에 대해 동작하고 출력을 생성함으로써 동작들을 수행하기 위해 하나 이상의 컴퓨터 프로그램을 실행하는 하나 이상의 프로그래밍 가능한 프로세서에 의해 수행될 수 있다. 프로세스 및 로직 흐름은 또한 FPGA 또는 ASIC과 같은 특수 목적 로직 회로로 구현될 수 있다.

[0064] 컴퓨터 프로그램의 실행에 적합한 프로세서는 예로서 범용 및 특수 목적의 마이크로 프로세서를 모두 포함한다. 일반적으로 프로세서는 판독 전용 메모리 나 랜덤 액세스 메모리 또는 둘 다에서 명령들과 데이터를 수신한다. 컴퓨터의 필수 요소는 명령들에 따라 동작을 수행하기 위한 프로세서 및 명령 및 데이터를 저장하기 위한 하나 이상의 메모리 디바이스이다. 일반적으로, 컴퓨터는 또한 데이터를 저장하기 위한 하나 이상의 대용량 저장 디바이스, 예를 들어 자기, 광 자기 디스크 또는 광 디스크로부터 데이터를 수신하거나 데이터를 전송하거나 둘 모두를 포함하거나 작동 가능하게 결합된다. 그러나 컴퓨터에는 이러한 장치가 필요하지 않다. 게다가, 컴퓨터는 다른 디바이스, 예를 들어 모바일 전화, PDA, 모바일 오디오 또는 비디오 플레이어, 게임 콘솔, GPS 수신기 또는 휴대용 저장 디바이스(예를 들어, 범용 직렬 버스(USB) 플래시 드라이브)에 내장될 수 있다. 컴퓨터 프로그램 명령 및 데이터를 저장하는데 적합한 디바이스는 반도체 메모리 디바이스(예를 들어 EPROM, EEPROM 및 플래시 메모리 디바이스), 자기 디스크(예를 들어 내부 하드 디스크 또는 이동식 디스크), 광 자기 디스크, 및 CD ROM 및 DVD-ROM 디스크를 포함하여 모든 형태의 비-휘발성 메모리, 매체 및 메모리 디바이스를 포함한다. 프로세서와 메모리는 특수 목적 논리 회로에 의해 보완되거나 이에 통합될 수 있다.

[0065] 사용자와의 상호 작용을 제공하기 위해, 본 명세서에 설명된 주제의 실시 예는 사용자에게 정보를 디스플레이하기 위한 디스플레이 디바이스, 예를 들어 CRT(음극선 관) 또는 LCD(액정 디스플레이) 모니터, 사용자가 컴퓨터에 입력을 제공할 수 있는 키보드 및 포인팅 디바이스, 예를 들어 마우스 또는 트랙볼을 갖는 컴퓨터에서 구현될 수 있다. 사용자와의 상호 작용을 제공하기 위해 다른 종류의 디바이스를 사용할 수도 있는데, 예를 들어, 사용자에게 제공되는 피드백은 예를 들어 시각적 피드백, 청각적 피드백 또는 촉각적 피드백과 같은 모든 형태의 감각 피드백일 수 있으며, 사용자로부터의 입력은 음향, 음성 또는 촉각 입력을 포함한 모든 형태로 수신될 수 있다. 게다가, 컴퓨터는 사용자가 사용하는 디바이스로 문서를 송수신함으로써. 예를 들어, 웹 브라우저에서 수신된 요청에 응답하여 사용자 클라이언트 디바이스의 웹 브라우저에 웹 페이지를 전송함으로써 사용자와 상호 작용할 수 있다.

[0066] 본 명세서에 기술된 주제의 실시예는 백엔드 구성 요소(예를 들어, 데이터 서버) 또는 미들웨어 구성 요소(예를 들어, 애플리케이션 서버), 또는 프런트 엔드 구성 요소(예를 들어, 사용자가 본 명세서에 설명된 주제의 구현과 상호 작용할 수있는 웹 브라우저 또는 그래픽 사용자 인터페이스를 가진 클라이언트 컴퓨터), 또는 이러한 백엔드, 미들웨어 또는 프런트 엔드 구성 요소 중 하나 이상의 조합을 포함하는 컴퓨팅 시스템에서 구현될 수 있다. 시스템의 구성 요소는 디지털 데이터 통신의 모든 형태 또는 매체, 예를 들어 통신 네트워크에 의해 상호 연결될 수 있다. 통신 네트워크의 예는 근거리 통신망("LAN") 및 광역 통신망("WAN"), 네트워크 간(예를 들어, 인터넷) 및 피어-투-피어 네트워크(예를 들어, 애드혹 피어- 투 피어 네트워크)를 포함한다.

[0067] 컴퓨팅 시스템은 클라이언트 및 서버를 포함할 수 있다. 클라이언트와 서버는 일반적으로 서로 떨어져 있으며 일반적으로 통신 네트워크를 통해 상호 작용한다. 클라이언트와 서버의 관계는 각 컴퓨터에서 실행되고 서로 클라이언트-서버 관계를 갖는 컴퓨터 프로그램으로 인해 발생한다. 일부 실시예에서, 서버는 데이터(예를 들어, HTML 페이지)를 (예를 들어, 클라이언트 디바이스와 상호 작용하는 사용자에게 데이터를 디스플레이하고 사용자로부터 사용자 입력을 수신하기 위해) 클라이언트 디바이스로 전송한다. 클라이언트 디바이스에서 생성된 데이터(예를 들어, 사용자 상호 작용의 결과)는 서버의 클라이언트 디바이스로부터 수신될 수 있다.

[0068] 본 명세서는 많은 특정 구현 세부 사항을 포함하지만, 이들은 임의의 발명의 범위 또는 청구될 수 있는 것에 대한 제한으로 해석되어서는 안되며, 오히려 특정 발명의 특정 실시 예에 특정한 특징의 설명으로 해석되어야 한다. 별개의 실시예의 맥락에서 본 명세서에 설명된 특정 특징은 또한 단일 실시예에서 조합하여 구현될 수 있다. 반대로, 단일 실시예의 맥락에서 설명된 다양한 특징은 또한 다수의 실시 예에서 개별적으로 또는 임의의 적절한 하위 조합으로 구현될 수 있다. 더욱이, 피처들이 특정 조합으로 작용하는 것으로 위에서 설명되고 심지어 처음에 그렇게 주장될 수도 있지만, 청구된 조합으로부터의 하나 이상의 피처는 경우에 따라 조합으로부터

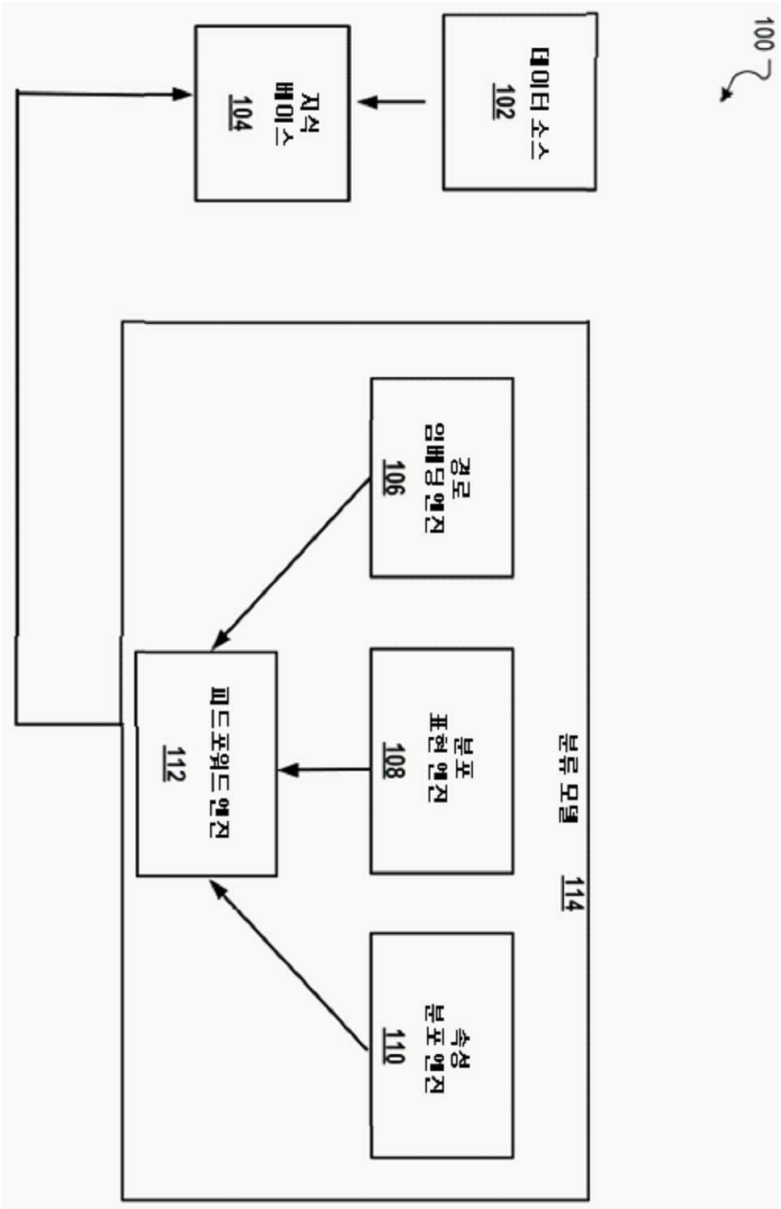
배제될 수 있고, 청구된 조합은 하위 조합 또는 하위 조합의 변형으로 지정될 수 있다.

[0069] 유사하게, 동작들이 특정 순서로 도면에 도시되어 있지만, 이는 바람직한 결과를 달성하기 위해 그러한 동작들이 도시된 특정 순서 또는 순차적인 순서로 수행되거나 모든 예시된 동작이 수행될 것을 요구하는 것으로 이해되어서는 안된다. 특정 상황에서는 멀티 태스킹 및 병렬 처리가 유리할 수 있다. 더욱이, 위에서 설명된 실시예들에서 다양한 시스템 구성 요소들의 분리는 모든 실시예에서 그러한 분리를 요구하는 것으로 이해되어서는 안되며, 설명된 프로그램 구성 요소 및 시스템은 일반적으로 단일 소프트웨어 제품으로 함께 통합되거나 다수의 소프트웨어 제품으로 패키지될 수 있음을 이해해야 한다.

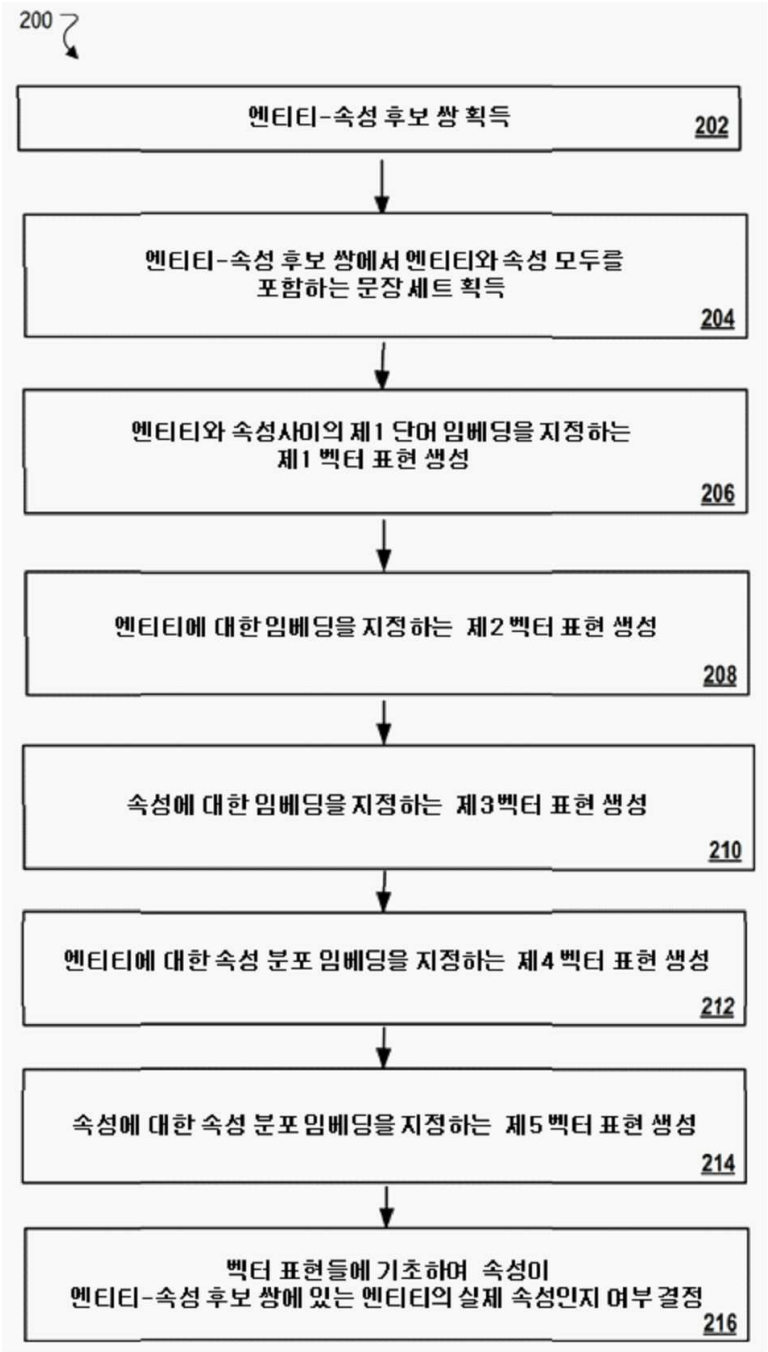
[0070] 따라서, 주제의 특정 실시예들이 설명되었다. 다른 실시예는 다음의 청구항의 범위 내에 있다. 일부 경우, 청구범위에 언급된 동작들은 다른 순서로 수행될 수 있으며 여전히 바람직한 결과를 얻을 수 있다. 또한, 첨부된 도면에 도시된 프로세스들은 바람직한 결과를 얻기 위해 도시된 특정 순서 또는 순차적인 순서를 반드시 필요로하지 않는다. 특정 구현에서는, 멀티 태스킹 및 병렬 처리가 유리할 수 있다.

도면

도면1



도면2


200
엔티티-속성 후보 쌍 획득
202
엔티티-속성 후보 쌍에서 엔티티와 속성 모두를 포함하는 문장 세트 획득
204
엔티티와 속성사이의 제1 단어 임베딩을 지정하는 제1 벡터 표현 생성
206
엔티티에 대한 임베딩을 지정하는 제2 벡터 표현 생성
208
속성에 대한 임베딩을 지정하는 제3벡터 표현 생성
210
엔티티에 대한 속성 분포 임베딩을 지정하는 제4 벡터 표현 생성
212
속성에 대한 속성 분포 임베딩을 지정하는 제5 벡터 표현 생성
214
벡터 표현들에 기초하여 속성이 엔티티-속성 후보 쌍에 있는 엔티티의 실제 속성인지 여부 결정
216

도면3