(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0114256
(43) 공개일자 2021년09월23일

(51) 국제특허분류(Int. Cl.)
G06F 40/205 (2020.01) *G06F 16/36* (2019.01)
G06N 5/02 (2006.01)
(52) CPC특허분류
G06F 40/205 (2020.01)
G06F 16/367 (2019.01)
(21) 출원번호 10-2020-0029742
(22) 출원일자 2020년03월10일
심사청구일자 2020년03월10일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
이경호
서울특별시 서대문구 연세로 50, 제4공학관 D917호(신촌동, 연세대학교)
서승민
서울특별시 서대문구 연세로 50, 제4공학관 D816호(신촌동, 연세대학교)
조은주
서울특별시 관악구 남부순환로 1935-3, 609호(봉천동, 두성오피스텔)
(74) 대리인
특허법인우인

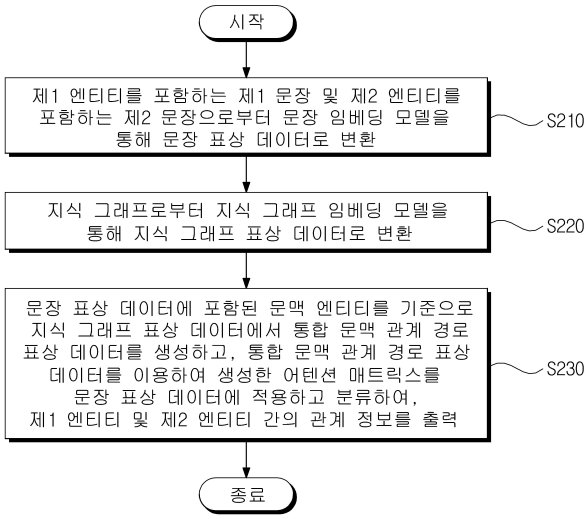
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **지식 그래프를 이용한 상황 인지형 다중 문장 관계 추출 방법 및 장치**

(57) 요 약

본 실시예들은 문장 임베딩과 지식 그래프 임베딩을 결합한 모델을 통해 어텐션 매트릭스를 적용하여 여러 문장에 걸쳐 나타나는 엔티티 쌍 사이의 관계를 정확하게 추출할 수 있는 다중 문장 관계 추출 방법 및 장치를 제공한다.

대 표 도 - 도5



(52) CPC특허분류
G06N 5/02 (2019.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 2019R1A2B5B01070555
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 중견연구자지원사업
연구과제명 (후속)동적으로 진화하는 지식 그래프의 정보 품질 개선을 위한 심층 표현 학습(2/3)
기 여 율 1/1
과제수행기관명 연세대학교 산학협력단
연구기간 2020.03.01 ~ 2021.02.28

명 세 서

청구범위

청구항 1

컴퓨팅 디바이스에 의한 다중 문장 관계 추출 방법에 있어서,

제1 엔티티를 포함하는 제1 문장 및 제2 엔티티를 포함하는 제2 문장으로부터 문장 임베딩 모델을 통해 문장 표상 데이터로 변환하는 단계;

지식 그래프로부터 지식 그래프 임베딩 모델을 통해 지식 그래프 표상 데이터로 변환하는 단계; 및

상기 문장 표상 데이터에 포함된 문맥 엔티티를 기준으로 상기 지식 그래프 표상 데이터에서 통합 문맥 관계 경로 표상 데이터를 생성하고, 상기 통합 문맥 관계 경로 표상 데이터를 이용하여 생성한 어텐션 매트릭스를 상기 문장 표상 데이터에 적용하고 분류하여, 상기 제1 엔티티 및 상기 제2 엔티티 간의 관계 정보를 출력하는 단계

를 포함하는 다중 문장 관계 추출 방법.

청구항 2

제1항에 있어서,

상기 문장 임베딩 모델은 문장에 포함된 단어를 변환한 단어 표상 데이터에 상기 단어의 위치 정보를 결합하는 것을 특징으로 하는 다중 문장 관계 추출 방법.

청구항 3

제1항에 있어서,

상기 지식 그래프 임베딩 모델은 (i) 헤드 엔티티 및 테일 엔티티 간의 직접적인 관계에 대한 제1 점수, 및 (ii) 상기 헤드 엔티티 및 상기 테일 엔티티 간에 존재 가능한 다양한 길이를 갖는 복수의 경로에 대한 제2 점수를 기준으로 상기 지식 그래프 표상 데이터로 변환하는 것을 특징으로 하는 다중 문장 관계 추출 방법.

청구항 4

제1항에 있어서,

상기 관계 정보를 출력하는 단계는,

상기 문장 표상 데이터 및 상기 지식 그래프 표상 데이터로부터 상기 제1 엔티티 및 상기 제2 엔티티 간에 상기 문맥 엔티티를 통과하는 단일 및 다중 홉의 문맥 관계 경로를 추출하고,

상기 단일 및 다중 홉의 문맥 관계 경로 중에서 상기 문맥 엔티티를 통과할 확률에 따라 복수의 주요 문맥 관계 경로를 선별하는 것을 특징으로 하는 다중 문장 관계 추출 방법.

청구항 5

제4항에 있어서,

상기 관계 정보를 출력하는 단계는,

상기 문맥 엔티티를 통과할 확률을 고려하여 상기 복수의 주요 문맥 관계 경로를 통합하여 상기 통합 문맥 관계 경로를 생성하는 것을 특징으로 하는 다중 문장 관계 추출 방법.

청구항 6

제5항에 있어서,

상기 관계 정보를 출력하는 단계는,

장단기 메모리(Long Short Term Memory, LSTM) 모델을 통해 상기 통합 문맥 관계 경로를 생성하는 것을 특징으

로 하는 다중 문장 관계 추출 방법.

청구항 7

제1항에 있어서,

상기 어텐션 매트릭스는,

(i) 상기 제1 엔티티에 대한 제1 어텐션 가중치, (ii) 상기 제2 엔티티에 대한 제2 어텐션 가중치, 및 (iii) 상기 통합 문맥 관계 경로 표상 데이터에 대한 제3 어텐션 가중치를 적용한 것을 특징으로 하는 다중 문장 관계 추출 방법.

청구항 8

제1항에 있어서,

상기 관계 정보를 출력하는 단계는,

상기 어텐션 매트릭스를 상기 문장 표상 데이터에 적용한 최종 표상 데이터에 대해서 (i) 복수의 문장에 각각 포함된 엔티티 쌍에 대한 관계 확률을 정의한 제1 손실 함수 및 (ii) 문장에 포함된 엔티티와 상기 지식 그래프 임베딩 모델에 포함된 유사한 엔티티를 매핑하는 제2 손실 함수를 최적화하도록 학습된 분류 모델을 적용하는 것을 특징으로 하는 다중 문장 관계 추출 방법.

청구항 9

하나 이상의 프로세서 및 상기 하나 이상의 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 다중 문장 관계 추출 장치에 있어서,

상기 프로세서는 제1 엔티티를 포함하는 제1 문장 및 제2 엔티티를 포함하는 제2 문장으로부터 문장 임베딩 모델을 통해 문장 표상 데이터로 변환하고,

상기 지식 그래프로부터 지식 그래프 임베딩 모델을 통해 지식 그래프 표상 데이터로 변환하고,

상기 프로세서는 상기 문장 표상 데이터에 포함된 문맥 엔티티를 기준으로 상기 지식 그래프 표상 데이터에서 통합 문맥 관계 경로 표상 데이터를 생성하고, 상기 통합 문맥 관계 경로 표상 데이터를 이용하여 생성한 어텐션 매트릭스를 상기 문장 표상 데이터에 적용하고 분류하여, 상기 제1 엔티티 및 상기 제2 엔티티 간의 관계 정보를 출력하는 것을 특징으로 하는 다중 문장 관계 추출 장치.

청구항 10

제9항에 있어서,

상기 문장 임베딩 모델은 문장에 포함된 단어를 변환한 단어 표상 데이터에 상기 단어의 위치 정보를 결합하는 것을 특징으로 하는 다중 문장 관계 추출 장치.

청구항 11

제9항에 있어서,

상기 지식 그래프 임베딩 모델은 (i) 헤드 엔티티 및 테일 엔티티 간의 직접적인 관계에 대한 제1 점수, 및 (ii) 상기 헤드 엔티티 및 상기 테일 엔티티 간에 존재 가능한 다양한 길이를 갖는 복수의 경로에 대한 제2 점수를 기준으로 상기 지식 그래프 표상 데이터로 변환하는 것을 특징으로 하는 다중 문장 관계 추출 장치.

청구항 12

제9항에 있어서,

상기 프로세서는,

상기 문장 표상 데이터 및 상기 지식 그래프 표상 데이터로부터 상기 제1 엔티티 및 상기 제2 엔티티 간에 상기 문맥 엔티티를 통과하는 단일 및 다중 홉의 문맥 관계 경로를 추출하고,

상기 단일 및 다중 홉의 문맥 관계 경로 중에서 상기 문맥 엔티티를 통과할 확률에 따라 복수의 주요 문맥 관계

경로를 선별하고,

상기 문맥 엔티티를 통과할 확률을 고려하여 상기 복수의 주요 문맥 관계 경로를 통합하여 상기 통합 문맥 관계 경로를 생성하는 것을 특징으로 하는 다중 문장 관계 추출 장치.

청구항 13

제9항에 있어서,

상기 어텐션 매트릭스는,

(i) 상기 제1 엔티티에 대한 제1 어텐션 가중치, (ii) 상기 제2 엔티티에 대한 제2 어텐션 가중치, 및 (iii) 상기 통합 문맥 관계 경로 표상 데이터에 대한 제3 어텐션 가중치를 적용한 것을 특징으로 하는 다중 문장 관계 추출 장치.

청구항 14

제9항에 있어서,

상기 프로세서는,

상기 어텐션 매트릭스를 상기 문장 표상 데이터에 적용한 최종 표상 데이터에 대해서 (i) 복수의 문장에 각각 포함된 엔티티 쌍에 대한 관계 확률을 정의한 제1 손실 함수 및 (ii) 문장에 포함된 엔티티와 상기 지식 그래프 임베딩 모델에 포함된 유사한 엔티티를 매핑하는 제2 손실 함수를 최적화하도록 학습된 분류 모델을 적용하는 것을 특징으로 하는 다중 문장 관계 추출 장치.

발명의 설명

기 술 분 야

[0001] 본 발명이 속하는 기술 분야는 다중 문장에 포함된 엔티티 간의 관계를 추출하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0003] 정보 추출 기술은 비구조적 또는 반구조적인 텍스트에서 의미있는 정보를 추출하여 구조화시킴으로써 질의응답, 문서 요약, 기계 독해 등의 여러 자연어 처리에 활용된다. 관계 추출(Relation Extraction)은 정보 추출 기술 중 하나이며, 정보 추출과 같은 목적으로 다양한 텍스트 데이터에서 엔티티(Entity) 사이의 상관 관계를 도출한다.

[0004] 전통적인 관계 추출 기술은 주석이 달리지 않은 하나의 문장에 참여하는 엔티티 쌍 사이의 이진 상관 관계(Binary Relation)를 찾는 것을 대상으로 하며 대체적으로 기계학습 및 딥러닝에 기반한다. 감독 학습 기반의 기존 기술들은 성공적인 관계 추출을 이뤄냈지만, 이들은 특정 도메인에 국한되며 실제 문장에 대한 정답 데이터의 부족으로 과대적합 문제를 겪기 쉽기 때문에 웹-스케일 관계 추출을 지원하지 못한다는 문제점이 있다.

[0005] 관계 추출 작업에서는 대상 엔티티(Target Entity) 간 의미적 관계가 잘 담기도록 문장을 모델링할 필요가 있다. 이를 위해 대표적으로 컨볼루션 신경망(Convolution Neural Network, CNN)과 순환 신경망(Recurrent Neural Network, RNN)이 사용된다. 다중 문장에 걸친 관계 추출은 더욱 길이가 긴 입력 시퀀스를 다루고 여러 문장에서 제시된 두 엔티티와 관련된 많은 단어들을 고려해야 하므로 기존에 제시된 네트워크 모델로는 한계가 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국공개특허공보 제10-2018-0077691호 (2018.07.09)

(특허문헌 0002) 한국공개특허공보 제10-2017-0030297호 (2017.03.17)

발명의 내용

해결하려는 과제

[0007] 본 발명의 실시예들은 문장 임베딩과 지식 그래프 임베딩을 결합한 모델을 통해 어텐션 매트릭스를 적용하여 여러 문장에 걸쳐 나타나는 엔티티 쌍 사이의 관계를 추출하는 데 주된 목적이 있다.

[0008] 본 발명의 명시되지 않은 또 다른 목적들은 하기의 상세한 설명 및 그 효과로부터 용이하게 추론할 수 있는 범위 내에서 추가적으로 고려될 수 있다.

과제의 해결 수단

[0009] 본 실시예의 일 측면에 의하면, 컴퓨팅 디바이스에 의한 다중 문장 관계 추출 방법에 있어서, 제1 엔티티를 포함하는 제1 문장 및 제2 엔티티를 포함하는 제2 문장으로부터 문장 임베딩 모델을 통해 문장 표상 데이터로 변환하는 단계, 지식 그래프로부터 지식 그래프 임베딩 모델을 통해 지식 그래프 표상 데이터로 변환하는 단계, 및 상기 문장 표상 데이터에 포함된 문맥 엔티티를 기준으로 상기 지식 그래프 표상 데이터에서 통합 문맥 관계 경로 표상 데이터를 생성하고, 상기 통합 문맥 관계 경로 표상 데이터를 이용하여 생성한 어텐션 매트릭스를 상기 문장 표상 데이터에 적용하고 분류하여, 상기 제1 엔티티 및 상기 제2 엔티티 간의 관계 정보를 출력하는 단계를 포함하는 다중 문장 관계 추출 방법을 제공한다.

[0010] 본 실시예의 다른 측면에 의하면, 하나 이상의 프로세서 및 상기 하나 이상의 프로세서에 의해 실행되는 하나 이상의 프로그램을 저장하는 메모리를 포함하는 다중 문장 관계 추출 장치에 있어서, 상기 프로세서는 제1 엔티티를 포함하는 제1 문장 및 제2 엔티티를 포함하는 제2 문장으로부터 문장 임베딩 모델을 통해 문장 표상 데이터로 변환하고, 상기 지식 그래프로부터 지식 그래프 임베딩 모델을 통해 지식 그래프 표상 데이터로 변환하고, 상기 프로세서는 상기 문장 표상 데이터에 포함된 문맥 엔티티를 기준으로 상기 지식 그래프 표상 데이터에서 통합 문맥 관계 경로 표상 데이터를 생성하고, 상기 통합 문맥 관계 경로 표상 데이터를 이용하여 생성한 어텐션 매트릭스를 상기 문장 표상 데이터에 적용하고 분류하여, 상기 제1 엔티티 및 상기 제2 엔티티 간의 관계 정보를 출력하는 것을 특징으로 하는 다중 문장 관계 추출 장치를 제공한다.

발명의 효과

[0011] 이상에서 설명한 바와 같이 본 발명의 실시예들에 의하면, 문장 임베딩과 지식 그래프 임베딩을 결합한 모델을 통해 어텐션 매트릭스를 적용하여 여러 문장에 걸쳐 나타나는 엔티티 쌍 사이의 관계를 정확하게 추출할 수 있는 효과가 있다.

[0012] 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 취급된다.

도면의 간단한 설명

[0013] 도 1은 다중 문장에 각각 포함된 엔티티의 관계를 예시한 도면이다.

도 2는 본 발명의 일 실시예에 따른 다중 문장 관계 추출 장치를 예시한 블록도이다.

도 3 및 도 4는 본 발명의 일 실시예에 따른 다중 문장 관계 추출 장치가 관계를 추출하는 메커니즘을 예시한 개념도이다.

도 5는 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법을 예시한 흐름도이다.

도 6은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법의 구체적인 동작을 예시한 도면이다.

도 7은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 문장을 임베딩하는 동작을 예시한 도면이다.

도 8은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 지식 그래프를 임베딩하는 동작을 예시한 도면이다.

도 9 및 도 10은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 문맥 관계 경로를 인코딩하는 동작을 예시한 도면이다.

도 11은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 문맥 관계 경로를 이용하여 어텐션을 적용

하는 동작을 예시한 도면이다.

도 12는 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 입력된 자연어에 대해서 지식 그래프 임베딩 모델에서 유사한 엔티티를 매핑하는 것을 예시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0014] 이하, 본 발명을 설명함에 있어서 관련된 공지기능에 대하여 이 분야의 기술자에게 자명한 사항으로서 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하고, 본 발명의 일부 실시예들을 예시적인 도면을 통해 상세하게 설명한다.

[0015] 도 1은 다중 문장에 각각 포함된 엔티티의 관계를 예시한 도면이다.

[0016] 도 1의 (a)를 참조하면, "우잠은 다가오는 2016년 말라얄람어의 드라마 영화이다."라는 첫번째 문장이 있고, "그 우잠이라는 필름은 프리스비라지 수쿠마란을 주연으로. 지투조셉에 의해 각본되고 감독되었다."라는 두번째 문장이 있을 때, 말라얄람어와 프리스비라지 수쿠마란 간의 관계를 추출해야 하는 예제에서, 다중 문장의 엔티티 간에 관계로 추출되어야 할 정답은 <Malaylam, language_spoken, Prithvirj Sukumaran>이다.

[0017] 하지만 기존의 방식들은 <Malayalam, no relation, Prithviraj Sukumaran>으로 파악할 뿐 정답을 추출하지 못한다. 다중 문장에 내재된 심층 문맥을 파악하기 위해서는 다중 문장의 엔티티 간의 구조적 관계 및 정보를 활용할 필요가 있다.

[0018] 도 1의 (b)를 참조하면, "빅토리아 슈미트는 뉴질랜드의 연극, 영화, 텔레비전 배우이다. 그리고 그녀는 사모네 웨딩의 알리샤 역으로 유명하다."라는 문장이 있을 때 뉴질랜드와 사모네 웨딩 사이에 film_film_country라는 정답 관계가 추출되어야 하는 예제에서, 다중 문장의 엔티티 간에 관계로 추출되어야 할 정답은 <New Zealand, film_film_country, Siones Wedding>이다.

[0019] 문장에서 바로 알 수 있는 정보로는 (사모네 웨딩이 연극인지, 영화인지, tv프로그램인지에 해당하는) 관계 추출을 위한 중요한 컨텍스트를 알아낼 수 없다. 기존의 방식들은 < New Zealand, tv_tv_program_country, Sines Wedding> 오답을 추출한다. 주어진 컨텍스트로는 중요한 컨텍스트를 파악하기 곤란하므로, 관계 추출시 단어 중요도에 있어서 도움을 받기 위해서 배경 지식(외부 지식)을 활용할 필요가 있다.

[0020] 여러 문장에 걸쳐 나타나는 엔티티 쌍 사이의 관계를 정확하게 추출하기 위해서, 본 실시예에 따른 다중 문장 관계 추출 장치는 문장 임베딩과 지식 그래프 임베딩을 결합한 모델을 통해 어텐션 매트릭스를 적용하는 방식으로 문제를 해결한다.

[0021] 도 2는 본 발명의 일 실시예에 따른 다중 문장 관계 추출 장치를 예시한 블록도이다.

[0022] 다중 문장 관계 추출 장치(110)는 적어도 하나의 프로세서(120), 컴퓨터 판독 가능한 저장매체(130) 및 통신 버스(170)를 포함한다.

[0023] 프로세서(120)는 다중 문장 관계 추출 장치(110)로 동작하도록 제어할 수 있다. 예컨대, 프로세서(120)는 컴퓨터 판독 가능한 저장 매체(130)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(120)에 의해 실행되는 경우 다중 문장 관계 추출 장치(110)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0024] 컴퓨터 판독 가능한 저장 매체(130)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장 매체(130)에 저장된 프로그램(140)은 프로세서(120)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독한 가능 저장 매체(130)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 지식 그래프 완성 장치(110)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0025] 통신 버스(170)는 프로세서(120), 컴퓨터 판독 가능한 저장 매체(140)를 포함하여 다중 문장 관계 추출 장치(110)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0026] 다중 문장 관계 추출 장치(110)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입

출력 인터페이스(150) 및 하나 이상의 통신 인터페이스(160)를 포함할 수 있다. 입출력 인터페이스(150) 및 통신 인터페이스(160)는 통신 버스(170)에 연결된다. 입출력 장치는 입출력 인터페이스(150)를 통해 다중 문장 관계 추출 장치(110)의 다른 컴포넌트들에 연결될 수 있다.

[0027] 도 3 및 도 4는 본 발명의 일 실시예에 따른 다중 문장 관계 추출 장치가 관계를 추출하는 메커니즘을 예시한 개념도이다.

[0028] 중요 문맥을 모델링하기 위해 구조적 정보인 지식 그래프를 이용한다. 여러 단어 간의 의존 관계인 다중 홉 경로를 파악하고, 연관된 정보 유형 등을 추가적인 배경지식으로 사용한다.

[0029] 지식 그래프는 엔티티와 관련된 실세계의 사실을 구조화하고 기계 이해를 위해 사실을 유기적으로 연결하는 데이터 모델이다. 지식 그래프는 개방형 연계 데이터(Linked Open Data)와 같은 거대한 이종의 소스 지식들을 통합하고 상호 운용을 가능하게 한다. 지식 그래프는 팩트(Fact)이라 불리는 트리플 <h,r,t>의 집합으로 구성되며, 헤드 엔티티 h와 테일 엔티티 t는 의미적 관계 r로 연결되어 있다. 지식 그래프 내 각각의 사실은 일반적으로 방향성을 지닌 엣지(Edge)인 관계(Relationship) <h,r,t>로 표현되며, 노드(Node)인 헤드 엔티티 h와 테일 엔티티 t는 엣지 타입인 의미적 관계(Semantic Relation)를 통해 서로 연결된다.

[0030] 도 4를 참고하면, 활용 가능한 지식 그래프의 구조적 정보를 예시적으로 나타낸다. 지식 그래프는 문장에서 등장한 문맥 단어가 포함되는 중요 문맥 관계 경로 및 배경 지식 정보를 포함한다. 구조적 정보를 사용하면 관계 추출의 성능을 높일 수 있다. 게다가 관계 추출을 위한 문장 모델링을 수행할 때 어텐션을 사용한다.

[0031] 다중 문장 관계 추출 장치는 제1 엔티티를 포함하는 제1 문장 및 제2 엔티티를 포함하는 제2 문장으로부터 문장 임베딩 모델을 통해 문장 표상 데이터로 변환한다. 다중 문장 관계 추출 장치는 지식 그래프로부터 지식 그래프 임베딩 모델을 통해 지식 그래프 표상 데이터로 변환한다. 다중 문장 관계 추출 장치는 문장 표상 데이터에 포함된 문맥 엔티티를 기준으로 지식 그래프 표상 데이터에서 통합 문맥 관계 경로 표상 데이터를 생성하고, 통합 문맥 관계 경로 표상 데이터를 이용하여 생성한 어텐션 매트릭스를 문장 표상 데이터에 적용하고 분류하여, 제1 엔티티 및 상기 제2 엔티티 간의 관계 정보를 출력한다.

[0032] 도 5는 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법을 예시한 흐름도이다. 다중 문장 관계 추출 방법은 다중 문장 관계 추출 장치에 의해 수행될 수 있다.

[0033] 다중 문장 관계 추출 방법은 연속된 문장들(문단) $S= \{s_1, s_2, \ldots, s_n\}$의 대상 엔티티 (e_1, e_2)에서 의미적 관계 r을 예측한다.

[0034] 단계 S210에서 프로세서는 제1 엔티티를 포함하는 제1 문장 및 제2 엔티티를 포함하는 제2 문장으로부터 문장 임베딩 모델을 통해 문장 표상 데이터로 변환한다. 단어 토큰 $T = \{t_1, \ldots, t_p, t_q, \ldots, t_r\}$이고, 첫 문장 s_1은 $t_1 \sim t_p$를 포함하고, 마지막 문장 s_n은 $t_q \sim t_r$을 포함한다.

[0035] 단계 S220에서 프로세서는 지식 그래프로부터 지식 그래프 임베딩 모델을 통해 지식 그래프 표상 데이터로 변환한다.

[0036] 단계 S230에서 프로세서는 문장 표상 데이터에 포함된 문맥 엔티티를 기준으로 지식 그래프 표상 데이터에서 통합 문맥 관계 경로 표상 데이터를 생성하고, 통합 문맥 관계 경로 표상 데이터를 이용하여 생성한 어텐션 매트릭스를 문장 표상 데이터에 적용하고 분류하여, 제1 엔티티 및 제2 엔티티 간의 관계 정보를 출력한다.

[0037] 도 6은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법의 구체적인 동작을 예시한 도면이다.

[0038] 다중 문장 관계 추출 방법은 (1) 문장 임베딩, (2) 경로 기반의 지식 그래프 임베딩, (3) 문맥 관계 경로 인코딩, (4) 문장 표상 데이터와 경로 임베딩 값을 이용한 지식 어텐션 매트릭스 생성, 및 (5)최종 표현을 통해 선형 분류를 거쳐 관계를 추출하는 동작을 수행한다.

[0039] 다중 문장 관계 추출 방법에 적용되는 임베딩 모델은 다양한 일반 데이터를 벡터 등의 표상 데이터로 변경하는 모델이다. 임베딩 모델은 데이터의 차원을 변경하여 처리 가능한 형태의 데이터로 변환하는 임베딩을 수행한다.

[0040] 다중 문장 관계 추출 방법에 적용되는 어텐션 모델은 예측 과정에서 특정 영역을 집중하여 관련된 영역에 어텐션 가중치를 부여하는 모델이다. 어텐션 메커니즘은 키-값 자료를 통해 매핑된 값을 추출할 수 있다. 주어진 쿼리에 대한 키의 유사도를 산출하고 키에 매핑된 값을 더해 반환한다. 어텐션 가중치는 스케일된 내적에 의해 쿼리-키 쌍과 함께 산출된다.

[0041] 도 7은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 문장을 임베딩하는 동작을 예시한 도면이다.

[0042] 문장 임베딩 모델에서 다중 문장이 입력되고, 트랜스포머 모델(Transformer)을 통한 문장 표상 데이터(Sentence Representation)이 출력된다. 문장 임베딩 모델은 문장에 포함된 단어를 변환한 단어 표상 데이터 s에 단어의 위치 정보를 결합한다. 각각의 절대적 위치 p를 결합한다.

수학식 1

[0043] $$x_i = s_i \oplus p_i$$

[0044] 들어온 토큰 값 즉, 단어 임베딩과 위치 임베딩 값이 합쳐진 벡터값들은 N 개의 트랜스포머 블록을 거친다. 셀프 어텐션을 적용하므로, 값, 키, 쿼리로 X값을 각각 프로젝션을 수행한다. 벡터 곱을 수행하고 스케일링한 후 선형 함수를 적용한다. Z는 문장들에 대해서 분산된 표상 데이터이다.

수학식 2

[0045] $$z_i^{(k)} = z_i^{(k-1)} + Transformer_{\,k}\left(z_i^{(k-1)}\right)\left(z_i^{(0)} = x_i\right)$$

수학식 3

[0046] $$z_i = \sum_{j=1}^{l} \frac{\exp e_{ij}}{\sum_{k=1}^{N} \exp e_{ij}} \left(x_j W^V\right)$$

수학식 4

[0047] $$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d}}$$

[0048] W는 모델의 파라이터이고, e_{ij}는 i와 j번째 정보에 간의 셀프 어텐션 점수이며, 각 위치별 유사도를 구하고 가중치로 반영한다.

[0049] 도 8은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 지식 그래프를 임베딩하는 동작을 예시한 도면이다.

[0050] 지식 그래프 임베딩 모델에서 트리플 세트가 입력되고, 입력받은 구조적 데이터를 통해 학습된 벡터값이 출력된다.

[0051] 지식 그래프 임베딩 모델은 (i) 헤드 엔티티 및 테일 엔티티 간의 직접적인 관계에 대한 제1 점수, 및 (ii) 헤드 엔티티 및 테일 엔티티 간에 존재 가능한 다양한 길이를 갖는 복수의 경로에 대한 제2 점수를 기준으로 지식 그래프 표상 데이터로 변환한다. 제1 점수는 수학식 6과 같이 표현되고, 제2 점수는 수학식 7과 같이 표현된다.

수학식 5

[0052] $$G(h,r,t)=E(h,r,t)+E(h,\Pi,t)$$

수학식 6

[0053] $$E(h,r,t)=||h+r-t||$$

수학식 7

[0054] $$E(h,\Pi,t)=\frac{1}{Z}\sum_{\pi_i\in\Pi(h,t)}R(\pi_i|h,t)E(h,\pi_i,t)$$

[0055] π는 헤드 엔티티와 테일 엔티티 간의 존재할 수 있는 경로이고, π 개수만큼의 에너지 점수를 산출한다. R은 신뢰도 가중치이고, Z는 모든 경로의 신뢰도 가중치를 합한 값이다.

[0056] 도 9 및 도 10은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 문맥 관계 경로를 인코딩하는 동작을 예시한 도면이다.

[0057] 문장이 주어져 있을 때, 그 문장의 대상 엔티티 사이에 존재할 수 있는 많은 관계 경로 중에서 문맥 엔티티를 거치는 주요 관계 경로를 추출하며, 주요한 문맥 관계 경로들이 추출되었다면 이를 하나의 표상 데이터로 LSTM 모델을 거쳐 통합한다.

[0058] 다중 문장 관계 추출 방법은 문장 표상 데이터 및 지식 그래프 표상 데이터로부터 제1 엔티티 및 제2 엔티티 간에 문맥 엔티티를 통과하는 단일 및 다중 홉의 문맥 관계 경로를 추출하고, 단일 및 다중 홉의 문맥 관계 경로 중에서 문맥 엔티티를 통과할 확률에 따라 복수의 주요 문맥 관계 경로를 선별한다.

수학식 8

[0059] $$P(t|h,\pi_i)=\begin{cases}1, & if\ h=t\\ 0, & otherwise\end{cases}$$

$$P(t|h,\pi_i)=\sum P(e_{n-1}|h,\pi_i')\cdot P(t|e_{n-1},r_n)$$

$$P(t|e_{n-1},r_n)=r_n(e_{n-1},t)/|r_n(e_{n-1},t')|$$

[0060] 문맥 관계 경로를 추출하는 것은 먼저 대상 엔티티들을 연결하는 관계 경로 중 문맥 엔티티가 포함된 경로 π를 통해 연결될 확률을 구한다. 이는 e_{n-1} 엔티티가 있고, r_n 관계가 연결되어 있을 때 t로 갈 확률을 구한다. 그리고 연속적인 경로에 대해서 하나의 표상 데이터로 통합시킨다.

[0061] 다중 문장 관계 추출 방법은 문맥 엔티티를 통과할 확률을 고려하여 복수의 주요 문맥 관계 경로를 통합하여 통합 문맥 관계 경로를 생성한다. 다중 문장 관계 추출 방법은 장단기 메모리(Long Short Term Memory, LSTM) 모델을 통해 통합 문맥 관계 경로를 생성한다.

[0062] LSTM 모델은 시간적 순서에 따른 데이터를 인코딩을 위해 사용된다. LSTM은 데이터를 분석하고 벡터를

생성한다.

[0063] LSTM 모델은 은닉 레어어에 여러 개의 게이트가 연결된 셀을 추가한 구조이다. 은닉 레이어는 입력 게이트(Input Gate), 출력 게이트(Output Gate), 망각 게이트(Forget Gate)를 포함하는 셀, 즉 메모리 블록(Memory Block)을 갖는다. 망각 게이트는 과거 정보를 잊기를 위한 게이트이고, 입력 게이트는 현재 정보를 기억하기 위한 게이트이다. 게이트는 각각 세기 및 방향을 가진다. 셀은 컨베이어 벨트 역할을 하고, 상태가 오래 경과하더라도 그래디언트가 비교적 전파를 유지할 수 있다. 망각 게이트는 이전 상태의 정보를 얼마나 기억할 것인지를 결정하는 단계로 0에서 1 사이 값이 출력되고 0이면 이전 상태의 정보를 완전히 잊는 것이고 1이면 이전의 정보를 온전히 기억하는 것이 될 수 있다.

[0064] 도 11은 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 문맥 관계 경로를 이용하여 어텐션을 적용하는 동작을 예시한 도면이다.

[0065] 어텐션 매트릭스는 (i) 제1 문장의 제1 엔티티에 대한 제1 어텐션 가중치, (ii) 제2 문장의 제2 엔티티에 대한 제2 어텐션 가중치, 및 (iii) 통합 문맥 관계 경로 표상 데이터에 대한 제3 어텐션 가중치를 적용한다.

수학식 9

[0066]
$$c = z_1 * (a_1 + b_1 + c_1) + \cdots + z_n * (a_n + b_n + c_n) = \frac{\sum z_i * (a_i + b_i + c_i)}{d}$$

[0067] 가능한 문맥 관계 경로에 대한 벡터값들과 대상 엔티티의 지식 그래프 임베딩 값과의 벡터곱을 수행하여 지식 에텐션 매트릭스를 만들고, 최종 표상 데이터를 만들 땐 더해서 각각의 단어 표상 데이터와 더한 후에 평균합 풀링을 수행한다. n개의 값이 있을 때 이들의 평균을 대표값으로 풀링하여 상위 레이어에서 활용한다.

[0068] 도 12는 본 발명의 다른 실시예에 따른 다중 문장 관계 추출 방법이 입력된 자연어에 대해서 지식 그래프 임베딩 모델에서 유사한 엔티티를 매핑하는 것을 예시한 도면이다.

[0069] 다중 문장 관계 추출 방법은 어텐션 매트릭스를 문장 표상 데이터에 적용한 최종 표상 데이터에 대해서 (i) 복수의 문장에 각각 포함된 엔티티 쌍에 대한 관계 확률을 정의한 제1 손실 함수 및 (ii) 문장에 포함된 엔티티와 상기 지식 그래프 임베딩 모델에 포함된 유사한 엔티티를 매핑하는 제2 손실 함수를 최적화하도록 학습된 분류 모델을 적용한다. 제1 손실 함수는 수학식 11과 같이 표현되고, 제2 손실 함수는 수학식 13과 같이 표현된다.

수학식 10

[0070]
$$L = L_{RE} + L_{EL}$$

수학식 11

[0071]
$$L_{RE} = \sum_{i=1}^{|T|} logP(r_i|X, e_1, e_2)$$

수학식 12

[0072] $$P(r|X, e_1, e_2) = \text{softmax}(cW^L + b)$$

[0073] 다중 문장 관계 추출 방법은 최종 표상 데이터를 가지고 소프트맥스 함수를 통해 분류를 수행하며, 추가적인 제2 손실 함수를 적용한다.

수학식 13

[0074] $$L_{EL} = -\sum \|e - \hat{e}\|^2$$

수학식 14

[0075] $$\hat{e} = tanh\left(W\begin{pmatrix} m \\ c \end{pmatrix}\right)$$

[0076] $\hat{e}$ 는 텍스트 기반의 엔티티 표상 데이터이고, W는 파라미터 매트릭스이고, m은 언급된 자연어 표상 데이터이고, c는 문맥 표상 데이터이다.

[0077] 실시예들에 의하면 문장 임베딩과 지식 그래프 임베딩을 결합한 모델을 통해 어텐션 매트릭스를 적용하여 여러 문장에 걸쳐 나타나는 엔티티 쌍 사이의 관계를 정확하게 추출할 수 있다.

[0078] 다중 문장 관계 추출 장치는 하드웨어, 펌웨어, 소프트웨어 또는 이들의 조합에 의해 로직회로 내에서 구현될 수 있고, 범용 또는 특정 목적 컴퓨터를 이용하여 구현될 수도 있다. 장치는 고정배선형(Hardwired) 기기, 필드 프로그램 가능한 게이트 어레이(Field Programmable Gate Array, FPGA), 주문형 반도체(Application Specific Integrated Circuit, ASIC) 등을 이용하여 구현될 수 있다. 또한, 장치는 하나 이상의 프로세서 및 컨트롤러를 포함한 시스템온칩(System on Chip, SoC)으로 구현될 수 있다.

[0079] 다중 문장 관계 추출 장치는 하드웨어적 요소가 마련된 컴퓨팅 디바이스 또는 서버에 소프트웨어, 하드웨어, 또는 이들의 조합하는 형태로 탑재될 수 있다. 컴퓨팅 디바이스 또는 서버는 각종 기기 또는 유무선 통신망과 통신을 수행하기 위한 통신 모뎀 등의 통신장치, 프로그램을 실행하기 위한 데이터를 저장하는 메모리, 프로그램을 실행하여 연산 및 명령하기 위한 마이크로프로세서 등을 전부 또는 일부 포함한 다양한 장치를 의미할 수 있다.

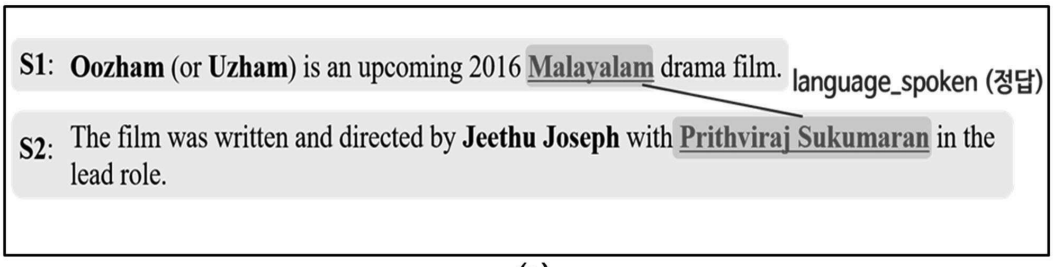
[0080] 도 5에서는 각각의 과정을 순차적으로 실행하는 것으로 기재하고 있으나 이는 예시적으로 설명한 것에 불과하고, 이 분야의 기술자라면 본 발명의 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 도 5에 기재된 순서를 변경하여 실행하거나 또는 하나 이상의 과정을 병렬적으로 실행하거나 다른 과정을 추가하는 것으로 다양하게 수정 및 변형하여 적용 가능할 것이다.

[0081] 본 실시예들에 따른 동작은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능한 매체에 기록될 수 있다. 컴퓨터 판독 가능한 매체는 실행을 위해 프로세서에 명령어를 제공하는 데 참여한 임의의 매체를 나타낸다. 컴퓨터 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 또는 이들의 조합을 포함할 수 있다. 예를 들면, 자기 매체, 광기록 매체, 메모리 등이 있을 수 있다. 컴퓨터 프로그램은 네트워크로 연결된 컴퓨터 시스템 상에 분산되어 분산 방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수도 있다. 본 실시예를 구현하기 위한 기능적인(Functional) 프로그램, 코드, 및 코드 세그먼트들은 본 실시예가 속하는 기술분야의 프로그래머들에 의해 용이하게 추론될 수 있을 것이다.

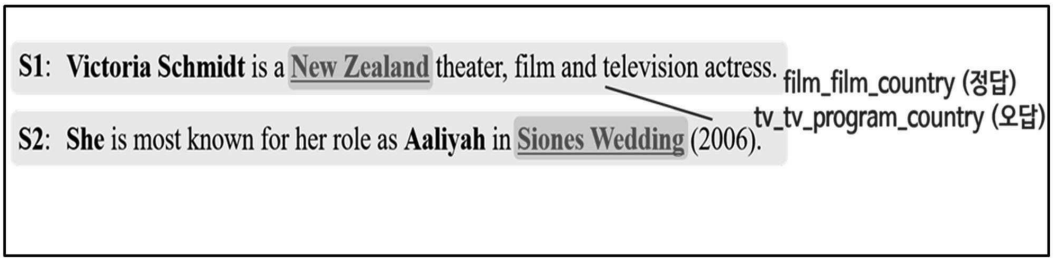
[0082] 본 실시예들은 본 실시예의 기술 사상을 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

도면

도면1

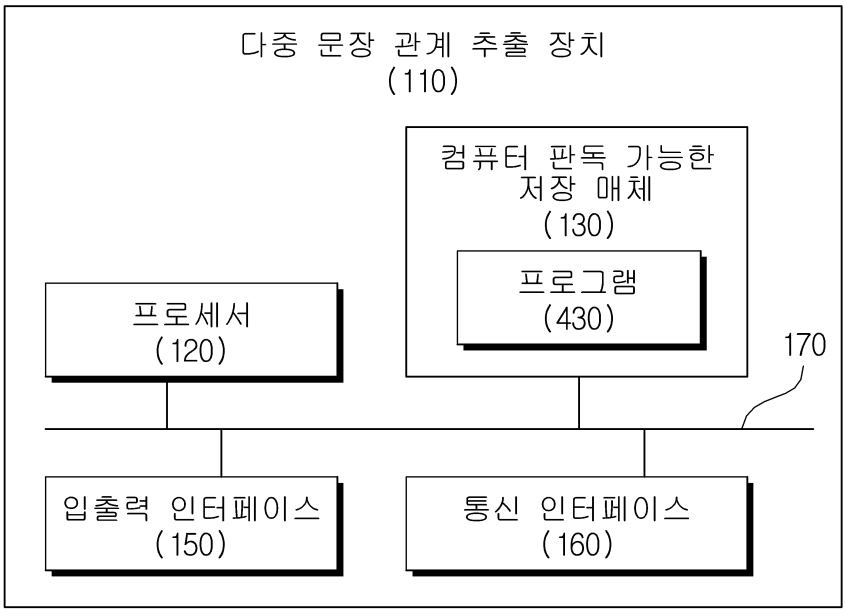


(a)

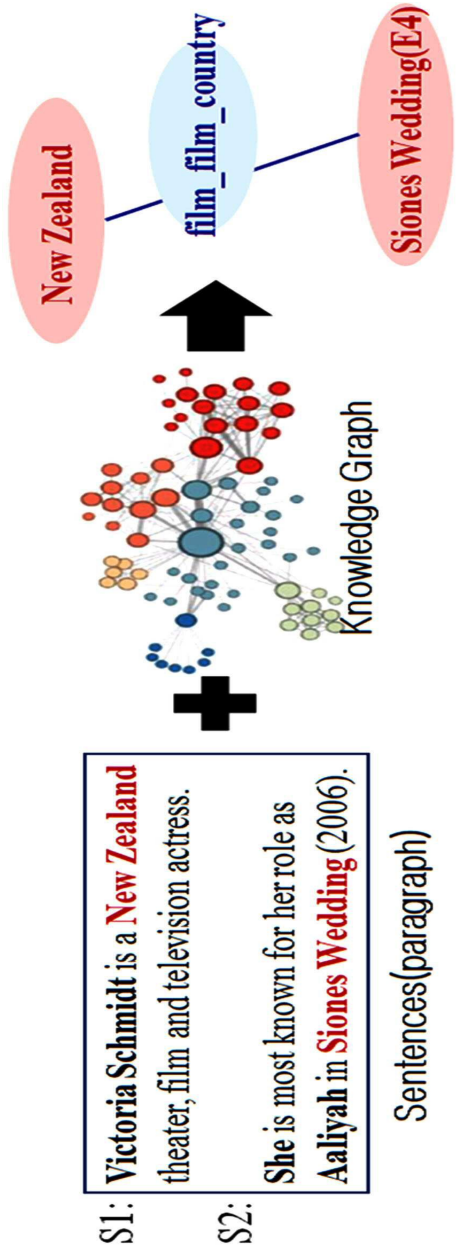


(b)

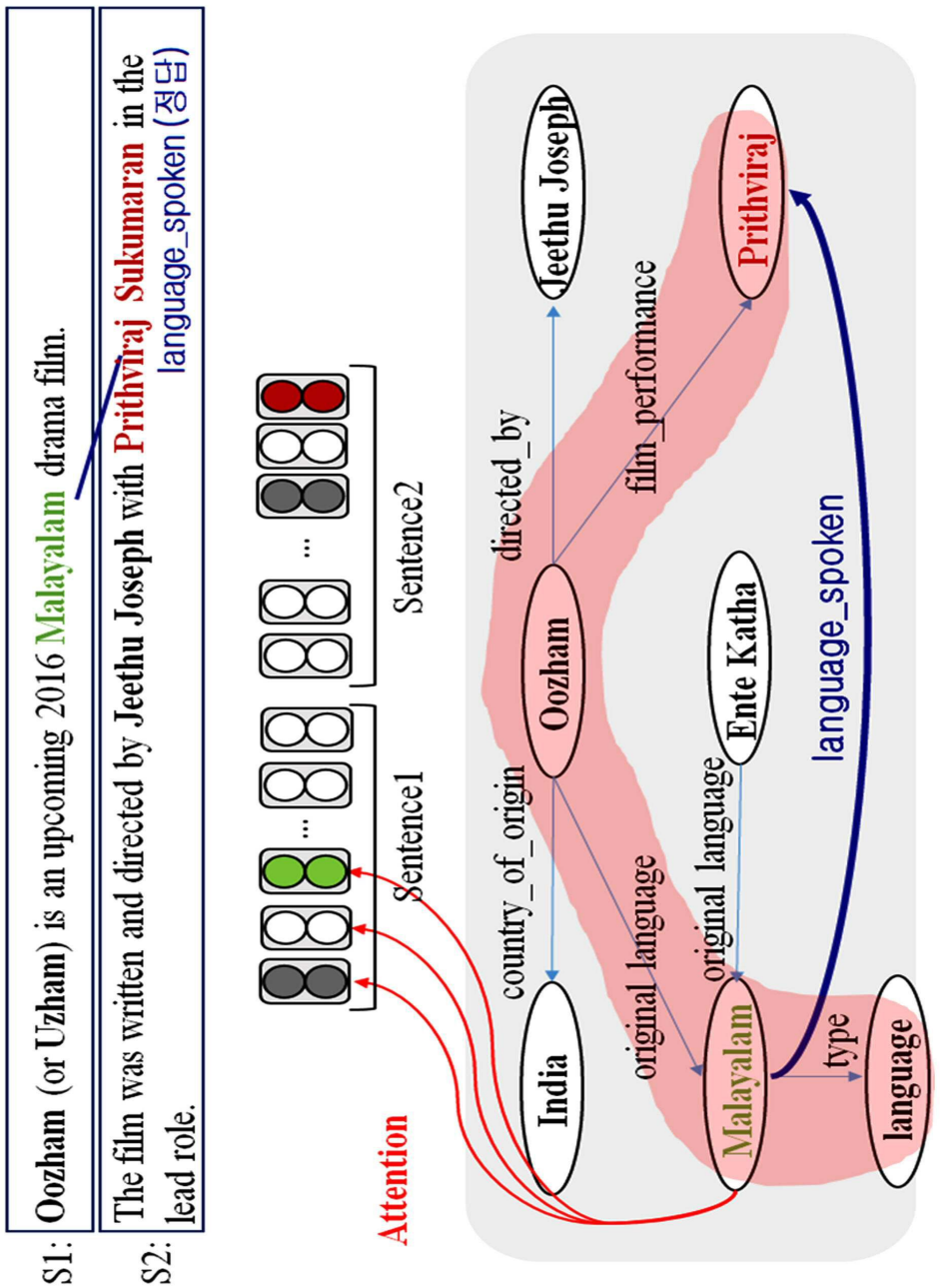
도면2



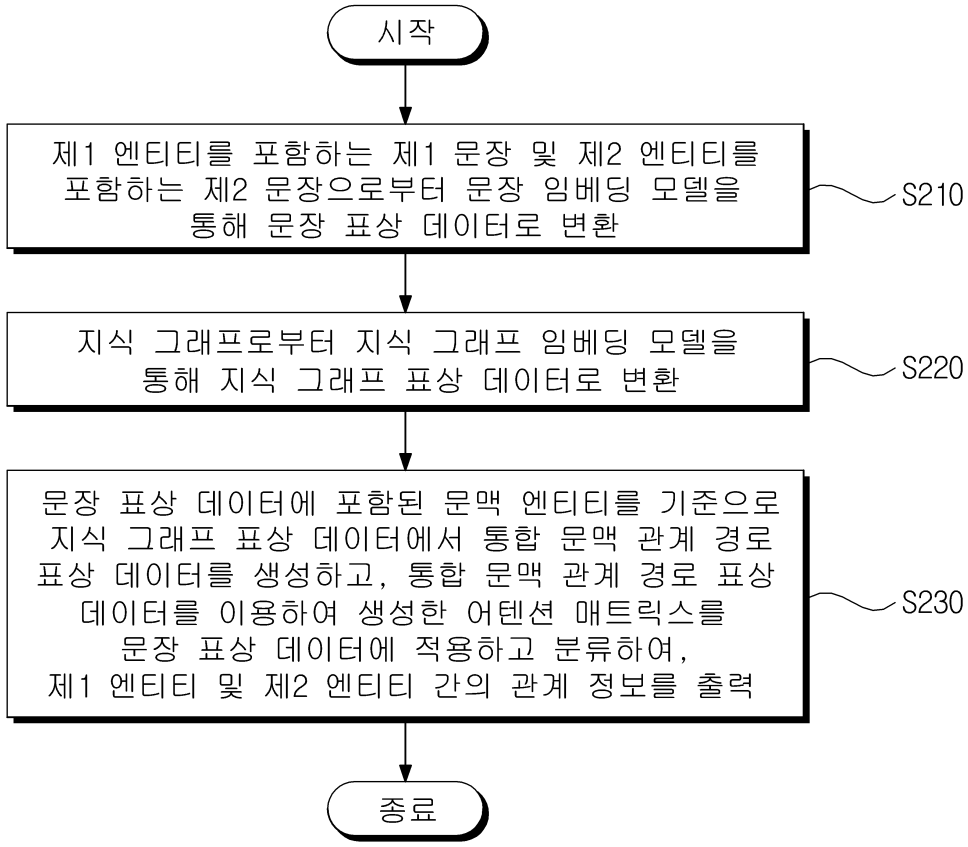
도면3


New Zealand
film_film_country
Siones Wedding(E4)
Knowledge Graph
S1: Victoria Schmidt is a New Zealand theater, film and television actress.
S2: She is most known for her role as Aaliyah in Siones Wedding (2006).
Sentences(paragraph)

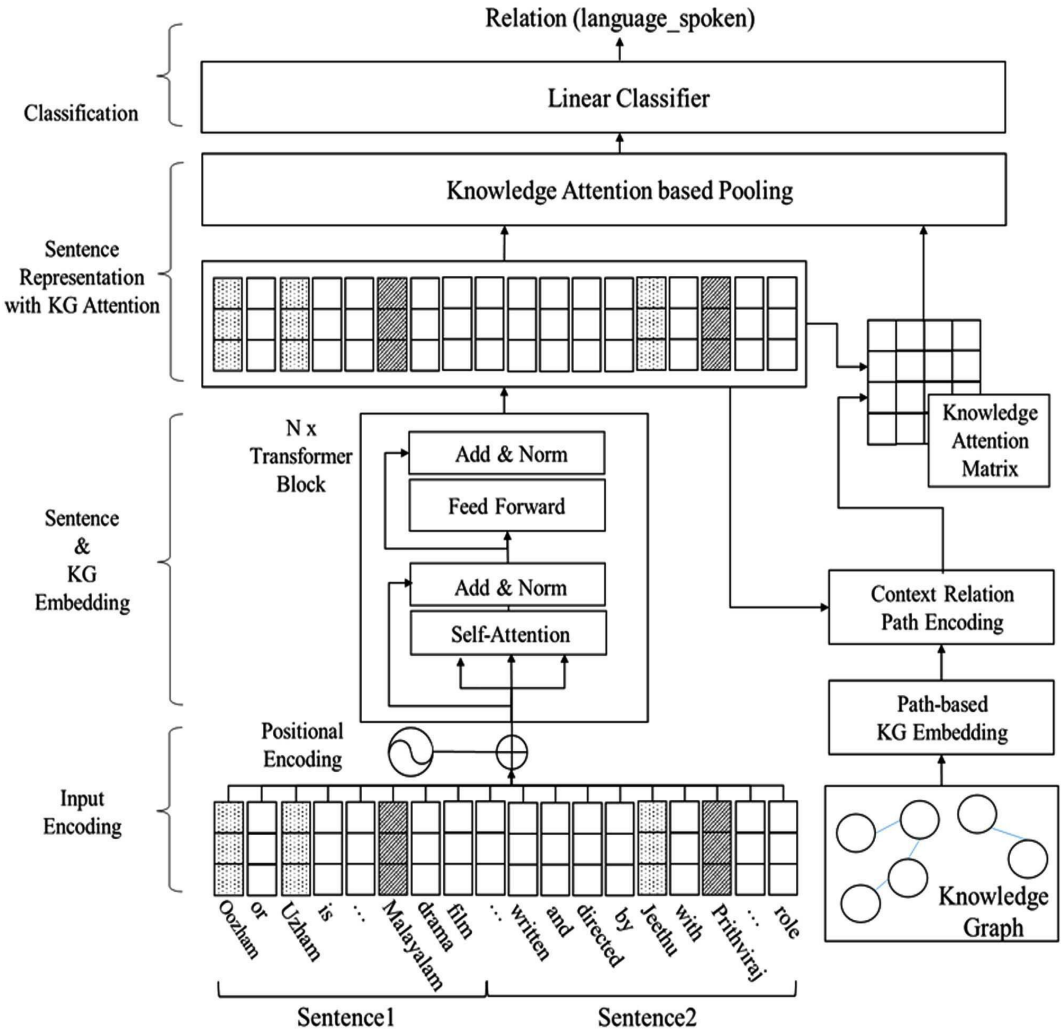
도면4


S1: Oozham (or Uzham) is an upcoming 2016 Malayalam drama film.
S2: The film was written and directed by Jeethu Joseph with Prithviraj Sukumaran in the lead role.
language_spoken (정답)
Sentence1
Sentence2
Attention
India
Oozham
Jeethu Joseph
Malayalam
Ente Katha
Prithviraj
language
country_of_origin
directed_by
original_language
original_language
film_performance
language_spoken
type

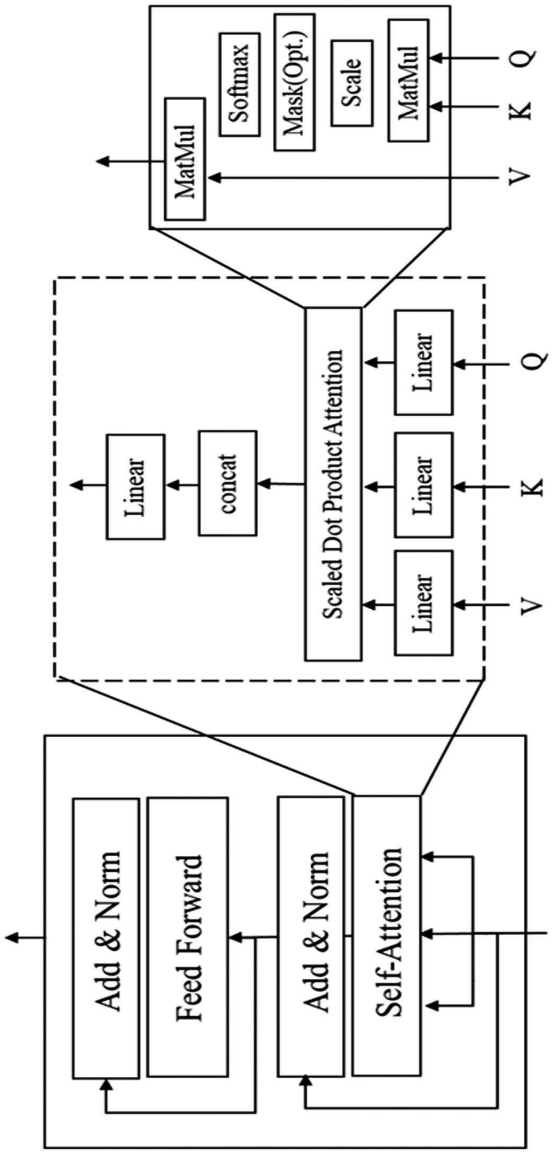
도면5


시작
제1 엔티티를 포함하는 제1 문장 및 제2 엔티티를 포함하는 제2 문장으로부터 문장 임베딩 모델을 통해 문장 표상 데이터로 변환
S210
지식 그래프로부터 지식 그래프 임베딩 모델을 통해 지식 그래프 표상 데이터로 변환
S220
문장 표상 데이터에 포함된 문맥 엔티티를 기준으로 지식 그래프 표상 데이터에서 통합 문맥 관계 경로 표상 데이터를 생성하고, 통합 문맥 관계 경로 표상 데이터를 이용하여 생성한 어텐션 매트릭스를 문장 표상 데이터에 적용하고 분류하여, 제1 엔티티 및 제2 엔티티 간의 관계 정보를 출력
S230
종료

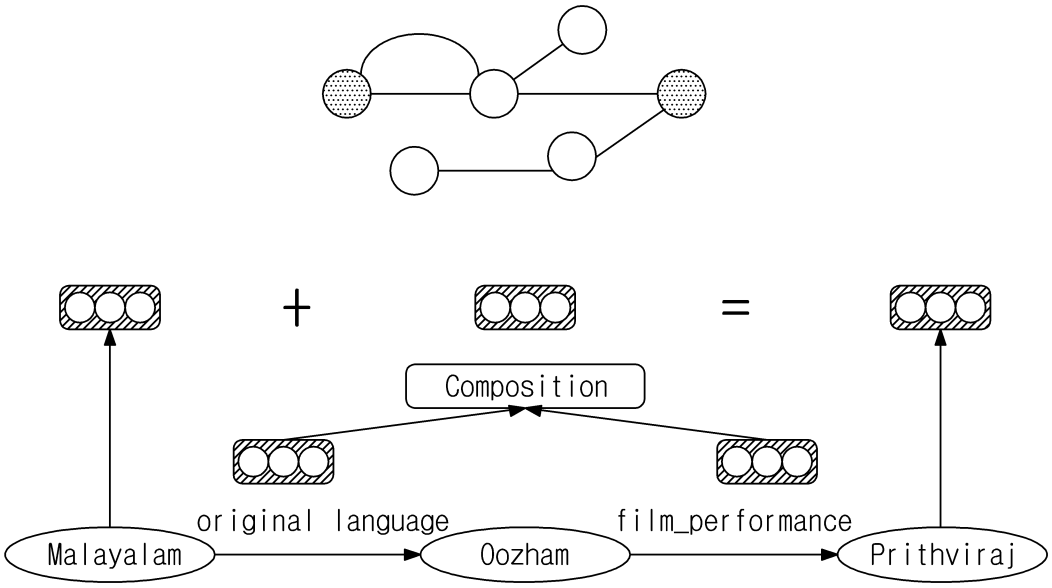
도면6


Relation (language_spoken)
Classification
Linear Classifier
Knowledge Attention based Pooling
Sentence Representation with KG Attention
Knowledge Attention Matrix
N x Transformer Block
Add & Norm
Feed Forward
Add & Norm
Self-Attention
Sentence & KG Embedding
Context Relation Path Encoding
Path-based KG Embedding
Positional Encoding
Input Encoding
Knowledge Graph
Oozham or Uzham is … Malayalam drama film … written and directed by Jeethu with Prithviraj … role
Sentence1
Sentence2

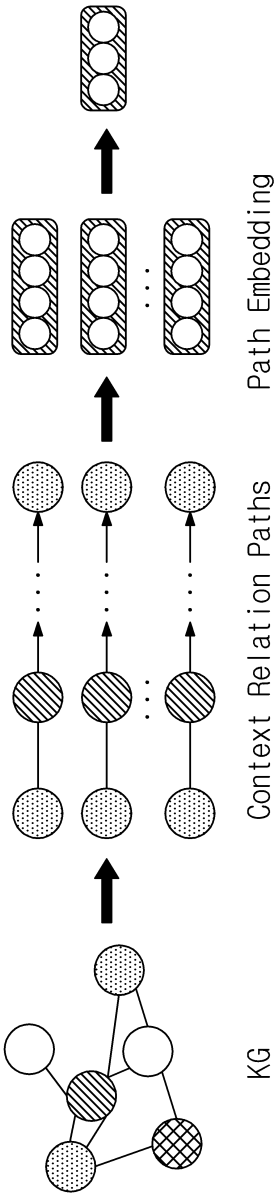
도면7


Add & Norm
Feed Forward
Add & Norm
Self-Attention
Linear
Linear
Linear
Scaled Dot Product Attention
Linear
concat
Q
K
V
MatMul
Softmax
Mask(Opt.)
Scale
MatMul
Q
K
V

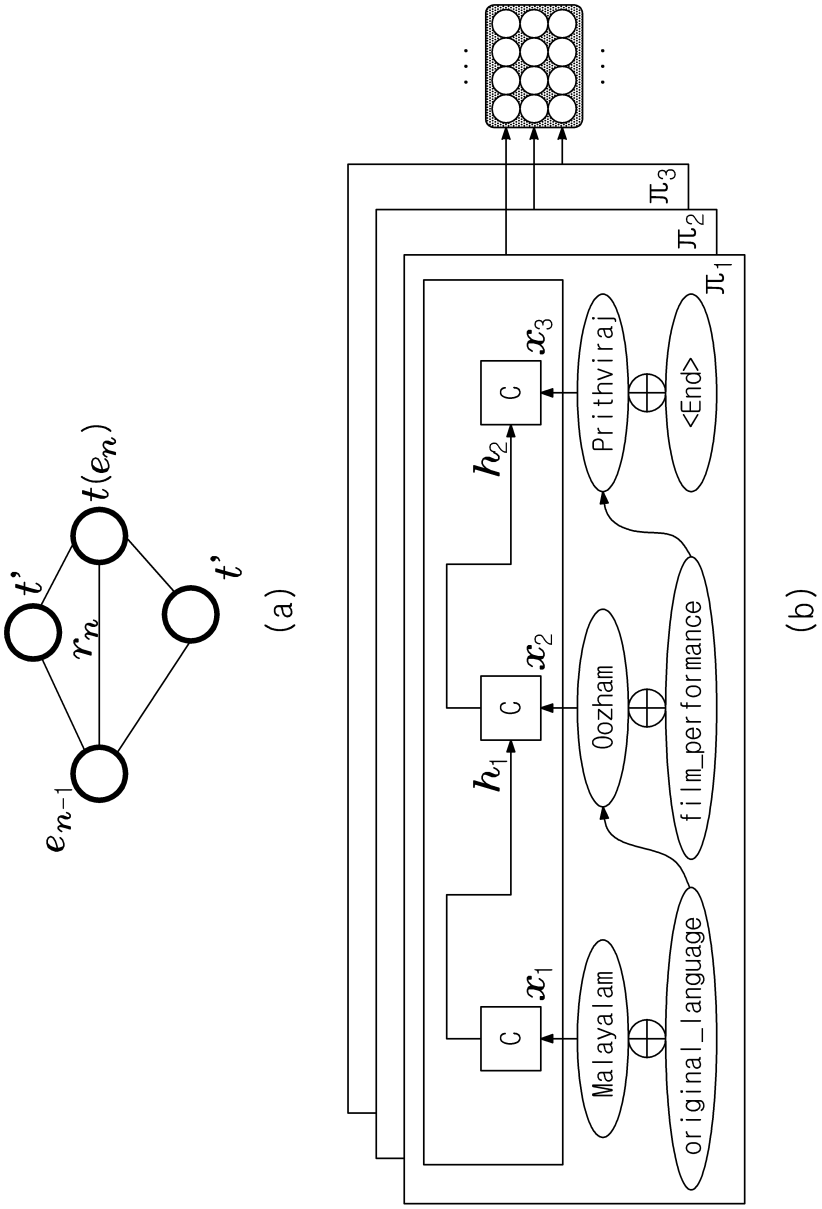
도면8


+
=
Composition
original language
film_performance
Malayalam
Oozham
Prithviraj

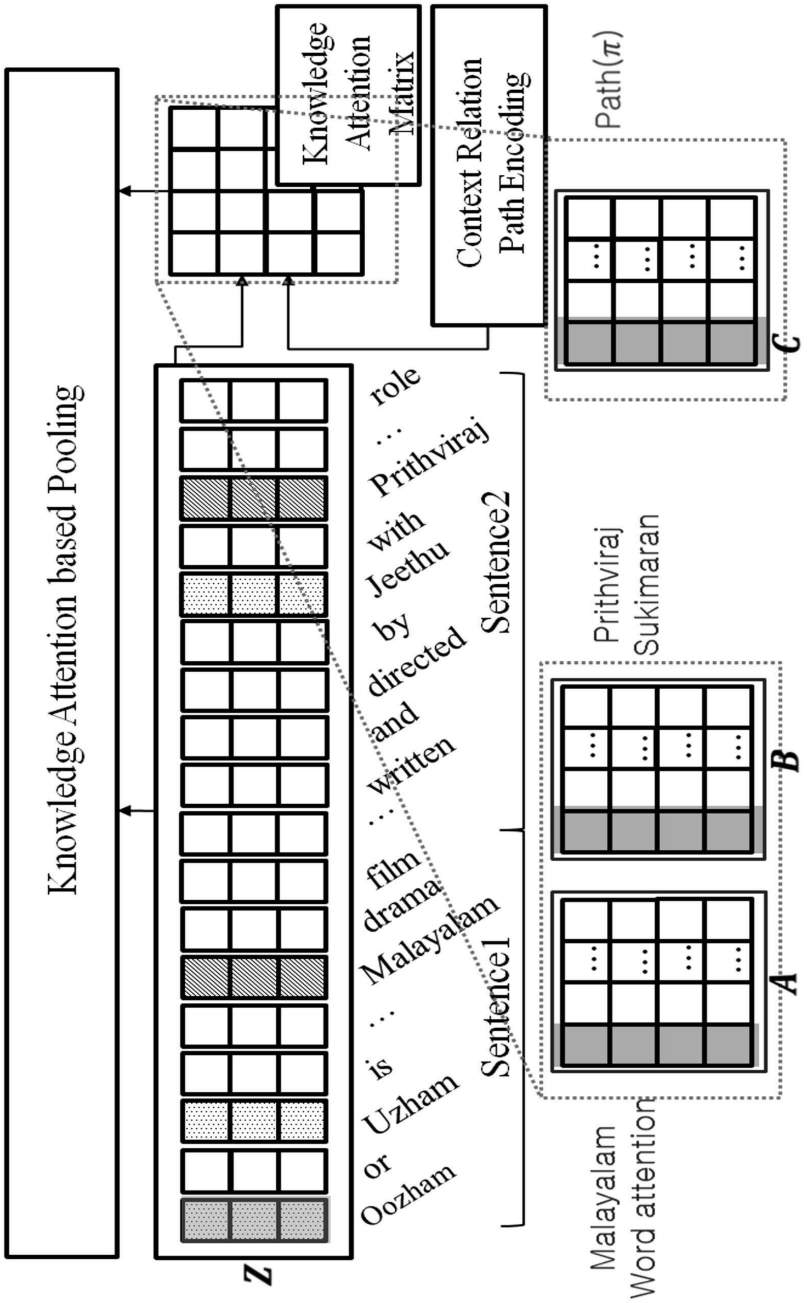
도면9



도면10


t'
t(e_n)
t'
r_n
e_n-1
(a)
...
...
π3
π2
π1
C
x3
h2
C
x2
h1
C
x1
Prithviraj
<End>
Oozham
film_performance
Malayalam
original_language
(b)

도면11


Knowledge Attention based Pooling
Knowledge Attention Matrix
Context Relation Path Encoding
Path(π)
Z
role
…
Prithviraj
with
Jeethu
by
directed
and
written
…
film
drama
Malayalam
…
is
Uzham
or
Oozham
Sentence2
Sentence1
Prithviraj Sukimaran
Malayalam Word attention
A
B
C

도면12

S: **Oozham** (or **Uzham**) is an upcoming 2016 **Malayalam** drama film.