(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0069058
(43) 공개일자 2025년05월19일

(51) 국제특허분류(Int. Cl.)
G06F 16/33 (2025.01) *G06F 16/332* (2025.01)
G06F 16/338 (2019.01) *G06F 16/36* (2019.01)
G06N 5/04 (2023.01)
(52) CPC특허분류
G06F 16/3347 (2019.01)
G06F 16/3329 (2025.01)
(21) 출원번호 10-2023-0155301
(22) 출원일자 2023년11월10일
심사청구일자 2023년11월10일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애기능생활관 311호 NLP&AI연구실
손준영
서울특별시 동대문구 무학로49길 42, 302호
(74) 대리인
김등용

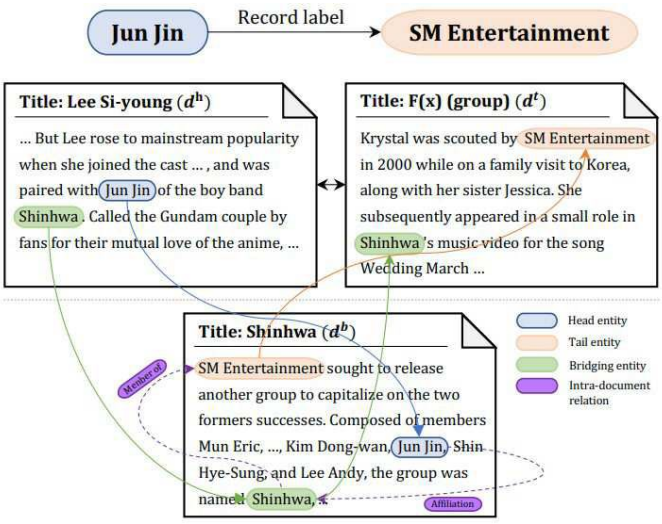
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **추론 경로 확장을 통한 교차 문서 관계 추출 장치 및 방법**

(57) 요 약

교차 문서 관계 추출(Cross-Document Relation Extraction)을 위한 추론 경로(reasoning path) 확장 장치 및 방법이 개시된다. 상기 추론 경로 확장 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 추론 경로 확장 방법으로, 머리 개체(head entity), 꼬리 개체(tail entity), 머리 문서(head document), 및 꼬리 문서(tail document)를 수신하는 단계, 후보 다리 개체(bridging entity)를 도출하는 단계, 상기 후보 다리 개체와 연관된 문서를 검색하는 단계, 상기 연관된 문서를 필터링하는 단계, 상기 연관된 문서에 대한 연관성 점수를 계산하는 단계, 및 상기 연관성 점수에 기초하여, 다리 개체와 상기 다리 개체에 대응하는 연관된 문서를 결정하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/338 (2019.01)
G06F 16/36 (2019.01)
G06N 5/04 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193711
과제번호 2018-0-01405-006
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345365402
과제번호 2021R1A6A1A03045425
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 Human-inspired AI 연구소
기 여 율 1/2
과제수행기관명 고려대학교
연구기간 2023.03.01 ~ 2024.02.29

명 세 서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는, 교차 문서 관계 추출(Cross-Document Relation Extraction)을 위한 추론 경로(reasoning path) 확장 방법에 있어서,

머리 개체(head entity), 꼬리 개체(tail entity), 머리 문서(head document), 및 꼬리 문서(tail document)를 수신하는 단계;

후보 다리 개체(bridging entity)를 도출하는 단계;

상기 후보 다리 개체와 연관된 문서를 검색하는 단계;

상기 연관된 문서를 필터링하는 단계;

상기 연관된 문서에 대한 연관성 점수를 계산하는 단계; 및

상기 연관성 점수에 기초하여, 다리 개체와 상기 다리 개체에 대응하는 연관된 문서를 결정하는 단계;

추론 경로 확장 방법.

청구항 2

제1항에 있어서,

상기 후보 다리 개체는 상기 머리 문서에 포함되어 있는 개체에 대한 집합과 상기 꼬리 문서에 포함되어 있는 개체에 대한 집합의 교집합인,

추론 경로 확장 방법.

청구항 3

제2항에 있어서,

상기 연관된 문서는 상기 후보 다리 개체에 대한 임베딩과 미리 정해진 값 이상의 내적값을 갖는 임베딩을 가진 문서인,

추론 경로 확장 방법.

청구항 4

제3항에 있어서,

상기 필터링하는 단계는, 연관된 문서 중에서 상기 후보 다리 개체를 포함하지 않는 문서를 배제하는,

추론 경로 확장 방법.

청구항 5

제4항에 있어서,

상기 연관성 점수를 계산하는 단계는, 상기 머리 문서(doc^{head}) 내에서 상기 후보 다리 개체(e^{cand})가 출현한 빈도, 상기 연관된 문서(doc^{cand}) 내에서 상기 머리 개체(head)가 출현한 빈도, 상기 꼬리 문서(doc^{tail}) 내에서 상기 후보 다리 개체가 출현한 빈도, 및 상기 연관된 문서 내에서 상기 꼬리 개체(tail)가 출현한 빈도 중 적어도 하나에 기초하여 계산되는,

추론 경로 확장 방법.

청구항 6

제5항에 있어서,

상기 연관성 점수는 수학식에 의해 계산되고,

상기 수학식은

$I(e^{caned}) = (N(e^{cand}, doc^{head}) + N(head, doc^{cand}))*(N(e^{cand}, doc^{tail}) + N(tail, doc^{cand}))$ 인,

추론 경로 확장 방법.

청구항 7

제6항에 있어서,

상기 머리 개체, 상기 꼬리 개체, 상기 머리 문서, 상기 꼬리 문서, 상기 다리 개체, 및 상기 다리 개체에 대응하는 문서를, 교차 문서 관계 추출 모델에 입력하여, 상기 머리 개체와 상기 꼬리 개체 사이의 관계를 추출하는 단계를 더 포함하는,

추론 경로 확장 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 문서 간 관계 추출 기법에 관한 것이며, 보다 상세하게는 임의의 개체 쌍에 대한 관계를 추론하기 위하여 해당 개체 쌍 내의 각 개체와 관련된 두 문서로 구성된 추론 경로를 확장하는 방법에 관한 것이다. 추론 경로는 주체 개체와 객체 개체 및 각각 관련된 주체 문서와 객체 문서로 구성되나, 본 발명의 방법을 적용하여 주체 문서와 객체 문서 사이에 간선 문서를 검색 및 추가하여 교차 문서 관계를 추출하는 방법에 관한 것이다.

배 경 기 술

[0002] 교차 문서 관계 추출(Cross-Document Relation Extraction)은 단일 문서 관계 추출과는 다른 작업(task)이다. 교차 문서 관계 추출은 두 개체 사이의 관계를 단일 문서를 보고 추론할 수 없으며, 두 개 이상의 문서를 추론하여 관계를 추출하는 작업이다. 일반적으로 현실 세계의 데이터는 여러 문서에 걸쳐서 분산되어 있으며, 단일 문서에 원하는 모든 정보가 담겨있을 가능성이 낮다. 실제로 방대한 지식 베이스인 Wikipedia(Wikidata)에 저장된 관계 정보를 분석하면 57.6%의 지식 정보가 단일 문서에 나타나지 않는다. 따라서, 이 관계 추출 작업은 현실 세계의 관계 추출 문제를 해결하기 위해 중요한 작업이다.

[0003] 교차 문서 관계 추출 작업의 대표적인 데이터셋인 CodRED는 인공지능 모델의 문서 간 관계 추출 학습을 위해서 각 개체 쌍마다 두 문서로 이루어진 추론 경로의 집합을 제공한다. 각 추론 경로는 주체 개체에 대한 문서인 주체 문서와 객체 개체에 대한 문서인 객체 문서의 두 종류의 문서로 구성된다. 주체 문서와 객체 문서는 두 문서에 공통으로 존재하면서 관계적인 단서를 제공하는 임의의 다리 개체에 의해서 연결된다. 이 다리 개체는 두 개체(주체 개체와 객체 개체) 사이에 잠재적으로 중요한 단서를 제공하는 핵심적인 요소이다.

[0004] 본 발명에서는, 교차 문서 관계 추출을 위해 질의(query)가 주어졌을 때 의미적 관련도가 높은 문서 후보군을 반환하는 밀집 검색(Dense Retrieval)을 이용한다. 모델은 BERT 기반의 사전학습 언어 모델을 2개 활용하여 질의를 인코딩하는 질의 모델과 문서를 인코딩하는 문서 모델로 구성될 수 있다. 질의가 주어지면 질의 모델을 활용하여 질의 벡터를 생성하고, 사전에 문서 모델로 구축한 문서 벡터 집합에서 MIPS(Maximum Inner Product Search, 최대 내적 검색)를 활용하여 가장 의미적으로 관련성이 높은 문서를 반환할 수 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 대한민국 공개특허 제2022-0083414호 (2022.06.20. 공개)

(특허문헌 0002) 대한민국 공개특허 제2020-0066972호 (2020.06.11. 공개)

(특허문헌 0003) 대한민국 공개특허 제2023-0044914호 (2023.04.04. 공개)

비특허문헌

[0006] (비특허문헌 0001) Yao, Yuan, et al. "CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

(비특허문헌 0002) Wang, Fengqi, et al. "Entity-centered Cross-document Relation Extraction." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022.

(비특허문헌 0003) Lu, Keming, et al. "Multi-hop Evidence Retrieval for Cross-document Relation Extraction." Proceedings of the Findings of the Association for Computational Linguistics. 2023.

(비특허문헌 0004) Karpukhin, Vladimir, et al. "Dense Passage Retrieval for Open-Domain Question Answering." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

발명의 내용

해결하려는 과제

[0007] 본 발명에서는 문서 간 관계 추출 작업에서 입력으로 사용되는 추론 경로를 활용하여 주체 개체와 객체 개체 사이의 다리 개체를 식별하고, 식별된 다리 개체와 밀집 검색(Dense Retrieval) 기술을 활용하여 다리 문서 후보군을 검색한다. 다리 문서 후보군에서 가장 유사도가 높은 다리 문서 집합을 구축하기 위해, 개체 기반 문서 필터링/스코어링 기술을 적용하고, 최종 다리 문서를 활용하여 추론 경로를 (주체 문서 - 객체 문서)에서 (주체 문서 - 다리 문서 - 객체 문서)의 형태로 확장한다. 이후 확장된 추론 경로를 인공지능 모델에 입력하여 관계를 추출하고자 한다.

[0008] 즉, 본 발명이 해결하고자 하는 과제는 추론 경로 확장을 이용한 교차 문서 관계 추출 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0009] 본 발명의 일 실시예에 따른 추론 경로(reasoning path) 확장 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는, 교차 문서 관계 추출(Cross-Document Relation Extraction)을 위한 추론 경로 확장 방법으로, 머리 개체(head entity), 꼬리 개체(tail entity), 머리 문서(head document), 및 꼬리 문서(tail document)를 수신하는 단계, 후보 다리 개체(bridging entity)를 도출하는 단계, 상기 후보 다리 개체와 연관된 문서를 검색하는 단계, 상기 연관된 문서를 필터링하는 단계, 상기 연관된 문서에 대한 연관성 점수를 계산하는 단계, 및 상기 연관성 점수에 기초하여, 다리 개체와 상기 다리 개체에 대응하는 연관된 문서를 결정하는 단계를 포함한다.

발명의 효과

[0010] 본 발명의 실시예에 따른, 추론 경로 확장을 이용한 교차 문서 관계 추출 장치 및 방법에 의할 경우, 확장된 추론 경로를 통해 두 개체 사이의 관계를 보다 정확하게 추출할 수 있는 효과가 있다.

[0011] 또한, 확장된 추론 경로는 두 개체 사이에 추론한 관계와 관련된 추가적인 정보를 포함하므로, 두 개체 사이에 추론된 관계에 대한 설명성을 보충해 줄 수 있다.

도면의 간단한 설명

[0012] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 따른 추론 경로 확장 방법을 설명하기 위한 개념도이다.

도 2는 본 발명의 일 실시예에 따른 추론 경로 확장 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0014] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0015] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0016] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0017] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0018] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0019] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0020] 관계 추출(Relation Extraction, RE) 태스크는 주어진 텍스트 내에서 두 엔티티(개체) 간의 관계를 예측하는 것을 목표로 하며, 지식 베이스 구축(Knowledge Base Construction, KBC) 및 질의 응답(Question Answering, QA)과 같은 응용 작업에 필수적인 기반이 된다. 문장-수준(sentece-leve)과 문서-수준(document-lever) RE는 상당한 성능을 달성했지만, 여전히 여러 문서들 사이에서의 관계를 식별하는 것은 어려운 과제이다. 엔티티 간의 관계는 여러 문서로부터 추론될 수 있기 때문에, Yao 등은 CodRED(Cross-Document Relation Extraction)을 제안하였다(Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. Codred: A cross-document relation extraction dataset for acquiring knowledge in the wild. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4452-4472.). 이는 두 엔티티(즉, 헤드(head)와 테일(tail)) 간의 관계를 여러 문서를 통해 제공한다.

[0021] CodRED 모델을 구현하기 위해, Wang 등은 엔티티 기반 문맥 필터(entity-based document-context filter)를 사용하여 다리 엔티티(bridging entities)를 캡처하는 방법을 개선하였고, 추론 경로(reasoning paths) 사이의 강화된 연결 인식(connection awareness)을 위해 교차 문서 엔티티 관계 어텐션 모듈(cross-document entity relatio attention module)을 활용한다. 반면에, Lu 등은 명시적으로 다리 엔티티를 고려하지 않고 밀집 검색

(dense retrieval)을 사용하여 증거 경로(evidence path)를 추출하고 순위를 매기는 데 중점을 두었다. 마찬가지로, 이전 연구들은 주로 다리 엔티티를 암묵적으로 모델링하려고만 시도하며, 다리 엔티티가 제공할 수 있는 잠재적으로 가지고 있는 추가 정보를 활용하지 않는다.

[0022] 그러나, 추론 경로의 탐험(exploration)에 대한 명확한 가이드를 제공할 수 있는, 다리 엔티티와 관련된 명시적인 정보를 활용하는 것은, 올바른 관계를 식별하는 데 중요하다. 여러 문서를 이용하여 엔티티 간의 관계를 추론할 때, 인간은 무작위 정보를 따르지 않을 수 있다. 대신, 인간은 문서 간의 다리 엔티티에서 명시적으로 연결된 정보를 본능적으로 활용한다. 이 정보는 경로를 꼬리 엔티티에 연결하고 관계에 대한 구조적 단서를 제공한다.

[0023] 본 발명에서는, PILOT(Path wIth expLOraTion)이라 명명될 수 있는 교차 문서 관계 추출 방법을 제안하다. 제안 기법에 의할 경우, 먼저 문서 내 관계(intra-document relation)와 같은 구조적 엔티티 정보를 사용하여 다리 엔티티를 찾고, 그와 관련된 문서를 검색한다. 그 이후에는, 엔티티-인지 스코어 함수(entity-aware scoring function)를 사용하여 다시 순위를 매기고, 높은 점수를 획득한 문서를 헤드(head)와 테일(tail)을 연결하는 디딤돌로 활용한다.

[0024] 제안하는 기법은 헤드와 테일 엔티티 사이의 추론 경로를 구축하기 위해 가장 중요한 다리 엔티티를 활용하는 데 중점을 둔다. 우선, 구조적인 정보(즉, 문서내 관계(intra-document relations))와 밀집 검색기(dense retriever)를 사용하여 후보 다리 엔티티와 그에 상응하는 문서의 집합을 앙상블한다. 다음으로, 엔티티-기반 필터링과 스코어링 기법을 통해 후보 중에서 가장 관련성 높은 것을 선택할 수 있다. 마지막으로, 최종 다리 엔티티를 사용하여 확장된 추론 경로를 구축할 수 있다. 이와 관련하여, 도 1에는 본 발명의 일 실시예에 따른 추론 경로 확장 방법을 설명하기 위한 개념도가 도시되어 있다.

[0025] **태스크 정의(Task Definition).** 교차 문서 관계 추출(cross-document RE)에서, 헤드 엔티티(또는 머리 엔티티)와 테일 엔티티(또는 꼬리 엔티티)는 각각 h와 t로 표기된다. $P = \{p_i\}_{i=1}^{N}$는 추론 경로 집합(set of reasoning paths)을 나타내고, 각 추론 경로 p_i는, h와 t를 포함한, 두 개의 문서 (d_i^h, d_i^t)를 포함한다. 태스크는 두 엔티티 사이의 관계 r∈R을 추론한 것을 목표로 한다. 여기서, R은 미리 정의된 관계 집합(relation set)이다. 특정 엔티티가 d_i^h와 d_i^t 모두에서 언급된다면, 이러한 엔티티는 다리 엔티티로 고려될 수 있다. 문서들 사이의 관계를 정립하는 잠재적인 다리 엔티티들(potential bridging entities)이 존재한다. 본 명세서에서는, 두 문서 사이의 다리 엔티티의 집합을 $E^b = \{e_i^b\}_{i=1}^{M}$로 나타낸다.

[0026] **엔티티-인지 검색기(Entity-aware Retriever).** 엔티티(개체) e가 쿼리로 주어졌을 때 관련된 문서들 $D_e = \{d_e\}_{i=1}^{L}$을 검색하기 위하여, 밀집 검색기(Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769-6781, Online. Association for Computational Linguistics.)를 학습할 수 있다. 먼저, 검색기를 훈련시키기 위한 데이터셋을 구축하기 위해, CodRED와 위키데이터(Wikidata)로부터 대응되는 위키피디아 페이지(Wikipedia page)를 가진 엔티티들을 수집한다. 그런 다음, 엔티티 e의 위키피디아 페이지를 긍정예(positive example) p^+로 간주하고 Karpukhin 등이 제시한 기법(Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769-6781, Online. Association for Computational Linguistics.)을 따르는 BM25로부터 추출된 위키피디아 페이지를 부정예(negative examples) $\{p_j^-\}_{j=}^{J}$로 간주한다. 이때, 단어 유사도는 높으나, 실제로 문맥적 관련성은 없는 문서 집합이 부정예가 될 수 있다. 학습 목표는, 수학식 1과 같이, 대조 학습(Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In In ternational conference on machine learning, pages 1597-1607. PMLR.)을 따른다.

[0027] [수학식 1]

$$l(e,p^+,\{p_j\}_{j=1}^J) = -\log \frac{\exp^{sim(e,p_i^+)}}{\exp^{sim(e,p_i^+)} + \sum_{j=1}^J \exp^{sim(e,p_j^-)}}$$

[0028]

[0029] 개체 e와 정답 문서 p^+, 및 오답 문서 집합 $\{p^-\}$이 주어지면, 수학식 1과 같은 대조 학습 규칙에 따라 엔티티-인지 검색기가 학습될 수 있다.

[0030] 엔티티-인지 검색기는 BERT 기반 사전학습 언어 모델로 이루어진 질의 모델(제1 모델)과 문서 모델(제2 모델)로 구성될 수 있다. 질의 모델은 개체 이름을 질의 벡터로 반환(또는 변환)하는 BERT 기반 인코더이며, 문서 모델은 개체에 대한 문서의 텍스트를 문서 벡터로 반환(또는 변환)하는 BERT 기반 인코더이다. 실시예에 따르면, 문서 모델은 모든 문서 집합에 대하여 사전에 문서 벡터를 계산한 뒤 데이터베이스에 색인한다. 이후 질의가 입력되면 질의에 대한 벡터만을 계산한 뒤, 데이터베이스의 문서 벡터 집합에 대하여 MIPS(Maximum Inner Product Search) 기술을 활용하여 관련도가 높은 문서 집합을 반환한다.

[0031] **Bridging Candidate Construction.** 두 문서 사이의 후보 다리 엔티티 집합을 구축하기 위해, 공유 엔티티 집합(shared entity set) $E^{cand} = E^h \cap E^t$을 이용한다. 여기서 (E^h, E^t)는 각각 (d^h, d^t) 내에 존재하는 모든 엔티티들의 집합이다. 밀집 검색기(dense retriever)로, $e^{cand} \in E^{cand}$을 쿼리로 이용하여, 문서 집합 $D^{cand} = \{d^{cand}\}_{i=1}^{|D^{cand}|}$을 검색할 수 있다.

[0032] **Entity-based Filtering.** 문서 집합을 검색한 후, e^{cand}와 보다 더 밀접한 문서들을 보존하기 위하여 문서들을 필터링할 수 있다. 구체적으로, e^{cand}가 d^{cand} 내에서 언급되지 않는다면, d^{cand}를 제거(또는 배제)할 수 있다. 추가적으로, 위키데이터로부터의 구조적인 정보를 이용하여 e^{cand}의 유효성을 검사할 수 있다. 위키데이터에 따라 e^{cand}가 h 및 t와 연결되어 있다면 다리 엔티티 후보로 이용할 수 있고, 그렇지 않으면 필터링할 수 있다.

[0033] **Bridging Entity Scoring.** 헤드 및 테일 엔티티와 연결된 후보 엔티티들의 연관성을 정량화하기 위해, 주어진 문서에 기초하여 각 엔티티에 대한 (연관성) 스코어를 산출할 수 있다. 구체적으로, e^{cand}와 문서 d에 대하여, d 내의 엔티티들을 검색하고 e^{cand}의 출현 빈도(the number of occurrences of e^{cand})를, 수학식 2와 같이, 카운트할 수 있다.

[0034] [수학식 2]

$$N(e^{cand}, d) = \sum_{e \in E^d} \delta(e^{cand}, e)$$

[0035]

$$\delta(e_1, e_2) = \begin{cases} 1 \text{ if } e_1 = e_2 \\ 0 \text{ otherwise} \end{cases}$$

[0036]

$$I(e^{caned}) = (N(e^{cand}, doc^{head}) + N(head, doc^{cand}))*(N(e^{cand}, doc^{tail}) + N(tail, doc^{cand}))$$

[0037]

[0038] 수학식 2에서, $\delta(e_1,e_2)$는, $e_1 = e_2$이면 1을, 그렇지 않으면 0을 반환한다. 이와 같은 연관성 스코어($I(e^{cand})$)에 기초하여, 상위-k(k는 임의의 자연수) 개의 다리 엔티티 $\hat{E}^b$를 추출하고, 이를 이용하여 추론 경로를 구축(또는 확장)할 수 있다. 즉, 위 수식의 연관성 점수에 기반하여 다리 개체 후보군의 순위를 내림차순으로 정렬하여 최종적으로 두 개체(head, tail)와 대응하는 기존의 문서 쌍(doc_{head}, doc_{tail})에 대하여 다리 개체 및 다리 개체 문서 집합 E^{bridge}, D^{bridge}를 획득할 수 있다.

[0039] **경로 구축(Path Construction)**

[0040] 마지막으로, 기존의 추론 경로 $P = \{p_i\}_{i=1}^{|P|}$와 $\hat{E}_i^b = \{\hat{e}_{ij}^b\}_{j=1}^{|\hat{E}_i^b|}$를 이용함으로써, 확장된 추론 경로 P′ 를 구축할

수 있다. P′ 를 구축하기 위한 구체적인 수학식은 아래 수학식 4와 같다.

[0041] [수학식 4]

[0042] $$P' = \bigcup_{i=1}^{|P|} \bigcup_{j=1}^{|\hat{E}_i^b|} \{(d_i^h, d^{\hat{e}_{ij}^b}, d_i^t) \mid (d_i^h, d_i^t) \in P\}$$

[0043] 수학식 4에서, $d^{\hat{e}_{ij}^b}$는, $\hat{e}_{ij}^b$를 쿼리로 이용하여, 엔티티-인지 검색기에 의해 검색된 문서를 나타낸다. 즉, 기존에 주어지는 두 문서 쌍을 이어주는 다리 문서 및 다리 개체를 활용하여 새로운 추론 경로를 구축할 수 있다. 이 추론 경로를 활용하여 기존 관계 추출 모델에 입력하면 모델링적인 개선을 하지 않더라도 성능이 크게 향상되는 결과를 얻을 수 있다. 표 1은 성능 평가 결과를 나타낸다.

[0044] [표 1]

[0045]

Method	Development				Test	
	AUC	F1	P@500	P@1000	AUC	F1
End-to-End (Yao et al., 2021)	47.94	51.26	62.80	51.00	47.46	51.02
+ PILOT	53.23	56.12	70.86	55.57	54.31	57.33
+ PILOT (- Wikidata)	52.98	55.72	72.72	55.60	53.37	57.52
ECRIM (Wang et al., 2022)	60.91	61.12	78.89	60.17	60.67	62.48
+ PILOT	64.45	63.03	80.24	62.34	62.90	63.84
+ PILOT (- Wikidata)	63.74	62.51	80.04	61.99	62.87	62.41

[0046] 도 2는 본 발명의 일 실시예에 따른 교차 문서 관계 추출을 위한 추론 경로 확장 방법을 설명하기 위한 흐름도이다.

[0047] 도 2를 참조하면, 교차문서 관계 추출 방법, 추론 경로 생성 방법 등으로 명명될 수도 있는 추론 경로 확장 방법은 적어로 프로세서(Processor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 즉, 추론 경로 확장 방법을 이루는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server), 랩탑 컴퓨터, 태블릿 PC 등을 포함할 수 있다. 또한, 컴퓨팅 장치는 물리적으로 분리된 복수의 장치로 구현될 수도 있다.

[0048] 우선, 검색 모델이 생성된다(S110). 검색 모델은 개체(또는 개체 이름)을 입력받아 입력된 개체와 관련된 문서를 출력(또는 추론)하는 모델을 의미할 수 있다. 검색 모델은 BERT 기반 사전학습 언어 모델로써, 임의의 학습 데이터를 이용하여, 대조 학습을 통해 생성될 수 있다. 또한, 검색 모델은 질의 모델과 문서 모델을 포함할 수 있다. 다만, 컴퓨팅 장치에 미리 저장되어 있는 문서들에 대한 벡터(문서 벡터)가 생성이 완료된 경우, 이후의 과정에서 문서 모델은 활용되지 않을 수도 있다.

[0049] 다음으로, 교차 문서 관계 추출은 위해, 머리 개체, 꼬리 개체, 머리 문서, 및 꼬리 문서가 수신된다(S120). 실시예에 따라, 머리 개체, 꼬리 개체, 머리 문서, 및 꼬리 문서는 미리 저장되어 있을 수도 있다.

[0050] 후보 다리 개체가 도출된다(S130). 후보 다리 개체는 주어진 두 문서(머리 문서와 꼬리 문서)에 공통적으로 포함되어 있는 개체를 의미할 수 있다. 즉, 후보 다리 개체는 머리 문서에 포함되어 있는 개체들의 집합과 꼬리 문서에 포함되어 있는 개체들의 집합의 교집합을 의미할 수 있다.

[0051] 후보 다리 개체 각각에 대하여, 연관성이 있는 문서들이 검색된다(S140). 연관성 있는 문서는 MIPS(Maximum Inner Product Search)를 통해 검색될 수 있다. 즉, 후보 다리 개체에 대한 벡터와 높은 내적값, 가장 높은 내적값, 또는 미리 정해진 값 이상의 내적값을 갖는 적어도 하나의 문서가 연관성 있는 문서로 검색될 수 있다. 이때, 검색 대상 문서는 컴퓨팅 장치에 미리 저장되어 있거나, 유무선 통신망을 통하여 수신될 수 있다.

[0052] 검색된 문서들에 대한 필터링이 수행될 수 있다(S150). 필터링 동작은 후보 다리 개체의 포함 여부로 수행될 수 있다. 즉, 후보 다리 개체를 포함하지 않는 문서는 필터링되어 배제될 수 있다.

[0053] 필터링 이후, 문서에 대한 연관성 점수가 계산된다(S160). 연관성 점수는 머리 문서에서 후보 다리 개체가 출현한 빈도, 후보 문서에서 머리 개체가 출현한 빈도, 꼬리 문서에서 후보 다리 개체가 출현한 빈도, 및 후보 문서에서 꼬리 개체가 출현한 빈도 중 적어도 하나에 기초하여 산출될 수 있다. 예시적인 산출식은 앞서 기재한 수학식과 같다.

[0054] 다리 개체 집합 및 다리 개체 문서 집합이 획득된다(S170). 다리 개체 집합은 연관성 점수가 가장 높은 적어도 하나의 후보 다리 개체를 의미하며, 다리 개체 문서 집합은 선택된 다리 개체에 대응하는 문서를 의미한다. 이를 통하여, 교차 문서 관계 추출은 위한 확장된 추론 경로가 생성된다.

[0055] 교차 문서 관계 추출 동작이 수행된다(S180). 교차 문서 관계 추출은 기존의 교차 문서 관계 추출 모델을 이용하여 수행될 수 있다. 주어진 머리 개체, 머리 문서, 꼬리 개체, 및 꼬리 문서 이외에 적어도 하나의 다리 개체와 다리 개체에 대응하는 문서를 교차 문서 관계 추출 모델에 입력함으로써, 교차 문서 관계가 추출될 수 있다.

[0056] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

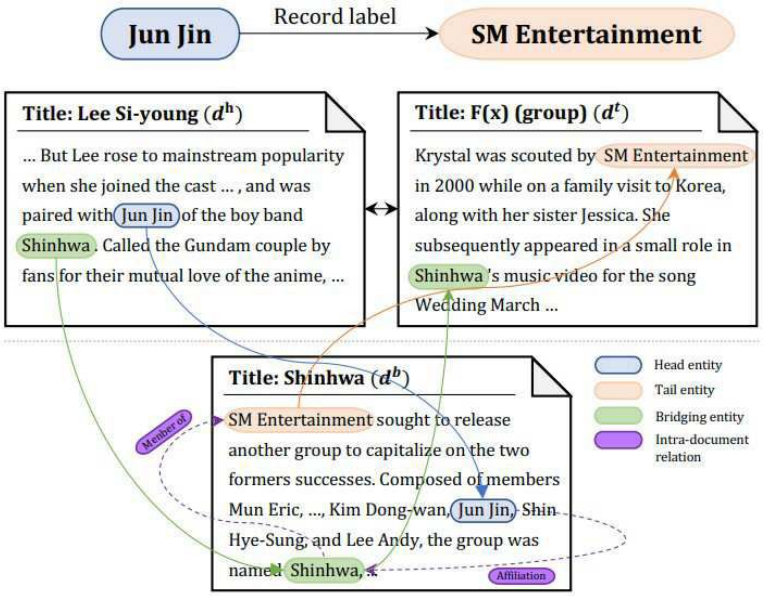
[0057] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0058] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

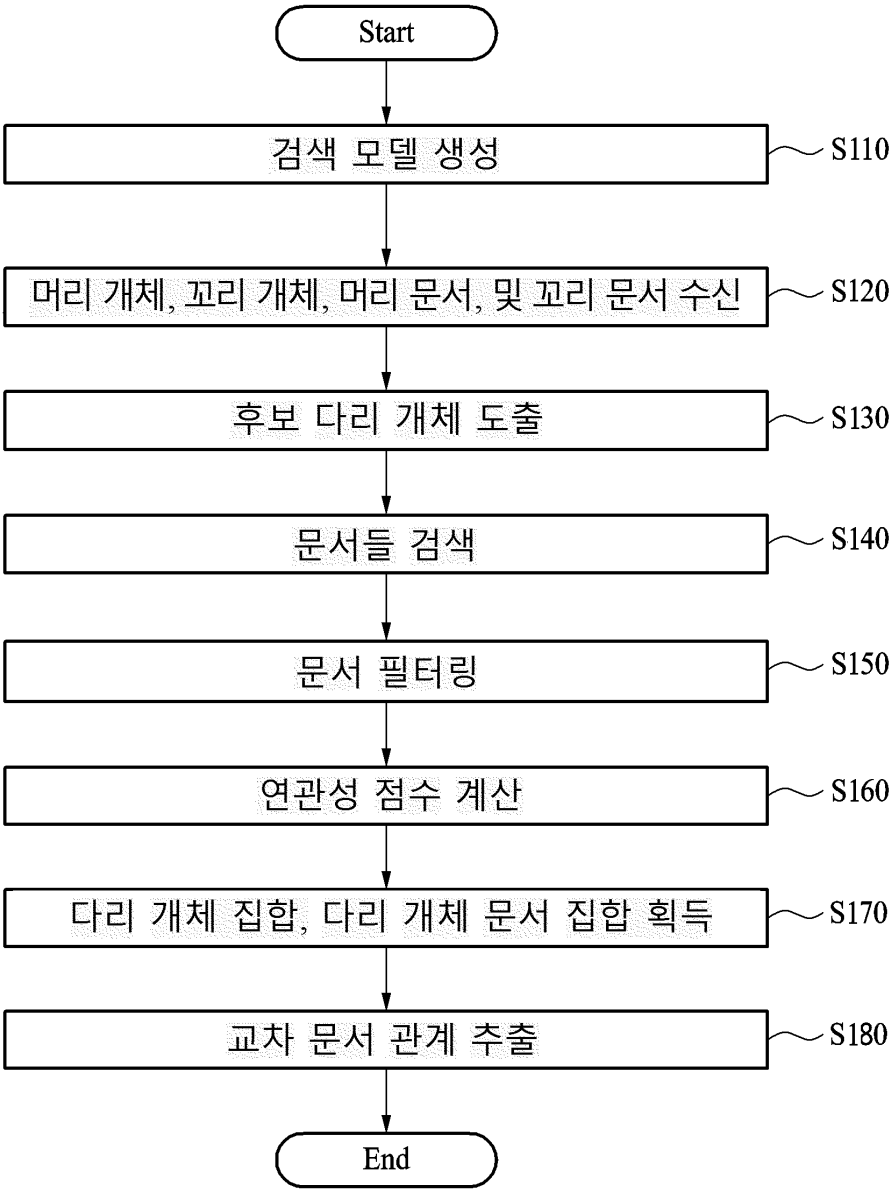
[0059] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

도면1


Jun Jin
Record label
SM Entertainment
Title: Lee Si-young (d^h)
... But Lee rose to mainstream popularity when she joined the cast ... , and was paired with Jun Jin of the boy band Shinhwa. Called the Gundam couple by fans for their mutual love of the anime, ...
Title: F(x) (group) (d^t)
Krystal was scouted by SM Entertainment in 2000 while on a family visit to Korea, along with her sister Jessica. She subsequently appeared in a small role in Shinhwa 's music video for the song Wedding March ...
Title: Shinhwa (d^b)
SM Entertainment sought to release another group to capitalize on the two formers successes. Composed of members Mun Eric, ..., Kim Dong-wan, Jun Jin, Shin Hye-Sung, and Lee Andy, the group was named Shinhwa, ...
Member of
Affiliation
Head entity
Tail entity
Bridging entity
Intra-document relation

도면2


Start
검색 모델 생성
S110
머리 개체, 꼬리 개체, 머리 문서, 및 꼬리 문서 수신
S120
후보 다리 개체 도출
S130
문서들 검색
S140
문서 필터링
S150
연관성 점수 계산
S160
다리 개체 집합, 다리 개체 문서 집합 획득
S170
교차 문서 관계 추출
S180
End