(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0143196
(43) 공개일자 2020년12월23일

(51) 국제특허분류(Int. Cl.)
G06F 16/35 (2019.01) *G06N 5/02* (2006.01)
(52) CPC특허분류
G06F 16/35 (2019.01)
G06N 5/022 (2019.01)
(21) 출원번호 10-2019-0122002
(22) 출원일자 2019년10월02일
심사청구일자 없음
(30) 우선권주장
1020190070119 2019년06월13일 대한민국(KR)

(71) 출원인
한국전자통신연구원
대전광역시 유성구 가정로 218 (가정동)
창원대학교 산학협력단
경상남도 창원시 의창구 창원대학로 20(퇴촌동)
(72) 발명자
방준성
대전광역시 유성구 죽동로 321, 102동 201호(죽동, 금성백조예미지)
차정원
경상남도 창원시 성산구 동산로 115, 105동 1407호(상남동, 대동아파트)
성수진
경상남도 창원시 마산회원구 내서읍 광려천동로 22, 512호(서광맨션)
(74) 대리인
특허법인 무한

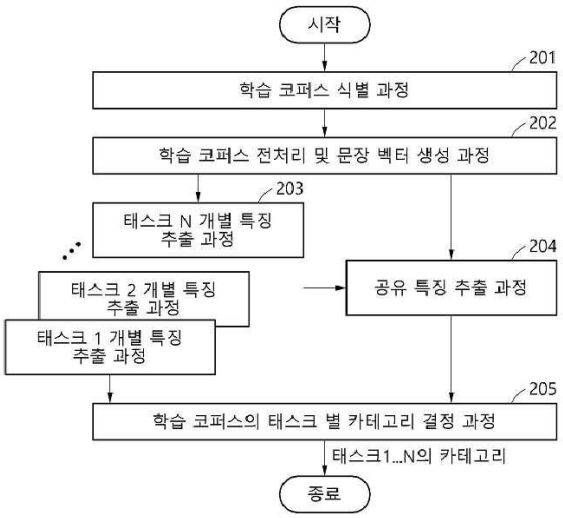
전체 청구항 수 : 총 1 항

(54) 발명의 명칭 **코퍼스 분류 방법 및 시스템**

(57) 요 약

멀티 태스크 학습 방식을 적용하여 코퍼스에 대해서 동시에 여러 태스크에 대한 카테고리를 분류할 수 있는 방법 및 시스템이 개시된다. 코퍼스 분류 방법은 학습 코퍼스를 입력 받는 단계; 학습 코퍼스를 전처리하고 문장 벡터를 생성하는 단계; 상기 전처리된 학습 코퍼스 및 문장 벡터를 이용하여 태스크 별로 개별 특징을 추출하는 단계; (i) 상기 전처리된 학습 코퍼스 및 문장 벡터 및 (ii) 상기 개별 특징 중 적어도 하나 이상을 이용하여 공유 특징을 추출하는 단계; 상기 개별 특징과 상기 공유 특징을 이용하여 상기 학습 코퍼스에 대해 태스크 별로 카테고리를 결정하는 단계를 포함한다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	2018-0-00440
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	2019년 RnD재발견프로젝트
연구과제명	위험상황초기인지를 위한 ICT기반의 범죄위험도예측 및 대응기술개발
기 여 율	1/1
과제수행기관명	한국전자통신연구원
연구기간	2018.04.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

학습 코퍼스를 입력 받는 단계;

학습 코퍼스를 전처리하고 문장 벡터를 생성하는 단계;

상기 전처리된 학습 코퍼스 및 문장 벡터를 이용하여 태스크 별로 개별 특징을 추출하는 단계;

(i) 상기 전처리된 학습 코퍼스 및 문장 벡터 및 (ii) 상기 개별 특징 중 적어도 하나 이상을 이용하여 공유 특징을 추출하는 단계;

상기 개별 특징과 상기 공유 특징을 이용하여 상기 학습 코퍼스에 대해 태스크 별로 카테고리를 결정하는 단계

를 포함하는 코퍼스 분류 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 코퍼스를 분류하는 기술에 관한 것으로, 구체적으로는 멀티 태스크 학습 방식을 적용하여 입력 코퍼스에 대해서 동시에 여러 태스크에 대한 카테고리를 분류할 수 있는 방법 및 시스템에 관한 것이다.

배 경 기 술

[0002] 코퍼스에 대한 뉴럴 네트워크의 학습 데이터량이 증가할수록 학습에 이용할 수 있는 특징이 늘어나기도 하지만 학습을 방해하는 특징(노이즈) 또한 늘어난다는 문제점이 있다. 따라서, 노이즈는 제거하고 유의한 특징을 추출하여 사용하는 것이 중요하며 유의한 특징에 대한 가중치를 학습하는 것이 중요하다.

[0003] 또한, 학습에 사용될 수 있는 데이터를 늘리기 위해서는 사람이 직접 데이터를 생산해야 하기 때문에 많은 시간과 비용이 소모된다. 따라서, 한정된 데이터를 이용하여 학습 효율을 증대하는 기술이 요구된다.

발명의 내용

해결하려는 과제

[0004] 본 발명은 코퍼스를 2개 이상의 태스크에 대해 동시에 분석하고 각 태스크에 대한 카테고리를 결정함으로써 기존 학습 방법에서 얻을 수 없었던 정보를 얻을 수 있고, 분류 정확도를 높일 수 있다.

[0005] 또한, 멀티 태스크 학습 방식을 적용하여 입력 코퍼스에 대해서 동시에 여러 태스크에 대한 카테고리를 분류할 수 있는 방법 및 시스템을 제공할 수 있다.

과제의 해결 수단

[0006] 본 발명의 일실시예에 따른 코퍼스 분류 방법은 학습 코퍼스를 입력 받는 단계; 학습 코퍼스를 전처리하고 문장 벡터를 생성하는 단계; 상기 전처리된 학습 코퍼스 및 문장 벡터를 이용하여 태스크 별로 개별 특징을 추출하는 단계; (i) 상기 전처리된 학습 코퍼스 및 문장 벡터 및 (ii) 상기 개별 특징 중 적어도 하나 이상을 이용하여 공유 특징을 추출하는 단계; 상기 개별 특징과 상기 공유 특징을 이용하여 상기 학습 코퍼스에 대해 태스크 별로 카테고리를 결정하는 단계를 포함한다.

[0007] 상기 학습 코퍼스는 태스크 별로 주어진 카테고리에 대한 정보를 포함하는 코퍼스일 수 있다. 또한, 상기 코퍼스는 자연어로 구성될 수 있다.

[0008] 상기 문장 벡터는 문장을 이루는 토큰 및 매칭되는 임의의 상수 값으로 구성될 수 있다. 상기 토큰은 전처리 작

업의 결과에 따라 문자, 단어, 형태소 등이 될 수 있다.

[0009] 상기 카테고리 결정 단계에 있어서, 상기 개별 특징 및 상기 공유 특징을 이용할 때 접합(Concatenation), 평균 및 다층 퍼셉트론(perceptron) 레이어 중 하나의 연산 방법을 활용하여 카테고리를 결정할 수 있다.

[0010] 상기 카테고리 결정 단계에 있어서, 상기 학습 코퍼스에 대해 태스크 별로 카테고리를 결정할 때, 소프트맥스(softmax) 연산 방법을 활용하여 카테고리를 결정할 수 있다.

[0011] 상기 개별 특징 추출에 사용된 가중치와 상기 공유 특징 추출에 사용된 가중치는 상기 태스크 별로 주어진 카테고리와 상기 결정된 카테고리를 비교한 결과에 따라 업데이트될 수 있다.

[0012] 본 발명의 일실시예에 따른 코퍼스 분류 시스템은 프로세서를 포함하고, 상기 프로세서는 학습 코퍼스를 입력받고, 학습 코퍼스를 전처리하고 문장 벡터를 생성하고, 상기 전처리된 학습 코퍼스 및 문장 벡터를 이용하여 태스크 별로 개별 특징을 추출하고, (i) 상기 전처리된 학습 코퍼스 및 문장 벡터 및 (ii) 상기 개별 특징 중 적어도 하나 이상을 이용하여 공유 특징을 추출하고, 상기 개별 특징과 상기 공유 특징을 이용하여 상기 학습 코퍼스에 대해 태스크 별로 카테고리를 결정한다.

[0013] 상기 학습 코퍼스는 태스크 별로 주어진 카테고리에 대한 정보를 포함하는 코퍼스일 수 있다. 또한, 상기 코퍼스는 자연어로 구성될 수 있다.

[0014] 상기 문장 벡터는 문장을 이루는 토큰 및 매칭되는 임의의 상수 값으로 구성될 수 있다. 상기 토큰은 전처리 작업의 결과에 따라 문자, 단어, 형태소 등이 될 수 있다.

[0015] 상기 프로세서는 상기 개별 특징 및 상기 공유 특징을 이용할 때 접합(Concatenation), 평균 및 다층 퍼셉트론(perceptron) 레이어 중 하나의 연산 방법을 활용하여 카테고리를 결정할 수 있다.

[0016] 상기 프로세서는 상기 학습 코퍼스에 대해 태스크 별로 카테고리를 결정할 때, 소프트맥스(softmax) 연산 방법을 활용하여 카테고리를 결정할 수 있다.

[0017] 상기 개별 특징 추출에 사용된 가중치와 상기 공유 특징 추출에 사용된 가중치는 상기 태스크 별로 주어진 카테고리와 상기 결정된 카테고리를 비교한 결과에 따라 업데이트될 수 있다.

발명의 효과

[0018] 본 발명의 일실시예에 따르면, 코퍼스를 2개 이상의 태스크에 대해 동시에 분석하고 각 태스크에 대한 카테고리를 결정함으로써 기존 학습 방법에서 얻을 수 없었던 정보를 얻을 수 있고, 분류 정확도를 높일 수 있다.

[0019] 또한, 본 발명의 일실시예에 따르면, 멀티 태스크 학습 방식을 적용하여 입력 코퍼스에 대해서 동시에 여러 태스크에 대한 카테고리를 분류할 수 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일실시예에 따른, 코퍼스 분류 시스템을 도시한 도면이다.

도 2는 본 발명의 일실시예에 따른, 코퍼스 분류 방법의 플로우 차트를 도시한 도면이다.

도 3은 본 발명의 일실시예에 따른, 코퍼스 분류하는 예시를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0021] 이하에서, 첨부된 도면을 참조하여 실시예들을 상세하게 설명한다. 그러나, 실시예들에는 다양한 변경이 가해질 수 있어서 특허출원의 권리 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 실시예들에 대한 모든 변경, 균등물 내지 대체물이 권리 범위에 포함되는 것으로 이해되어야 한다.

[0022] 실시예에서 사용한 용어는 단지 설명을 목적으로 사용된 것으로, 한정하려는 의도로 해석되어서는 안된다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0024] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0025] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0026] 도 1은 본 발명의 일실시예에 따른 코퍼스 분류 시스템(101)을 도시한 도면이다.

[0027] 코퍼스 분류 시스템(101)은 학습 코퍼스를 입력 받아 태스크(task) 별로 카테고리를 결정하여 출력한다. 이 때, 학습 코퍼스는 코퍼스(corpus)와 태스크 별로 주어진 카테고리에 대한 정보를 포함할 수 있다.

[0028] 예를 들어, 학습을 위한 코퍼스는 경우 '입력 문장, 태스크 1의 카테고리, 태스크 2의 카테고리, …, 태스크 N의 카테고리'의 형식으로 주어질 수 있다. 다만, 평가를 위한 코퍼스의 경우 입력 문장만 주어진다. 또한, 코퍼스는 자연어로 구성될 수 있다.

[0029] 본 발명에서 태스크란 해당 코퍼스를 분류하는 기준을 의미할 수 있다. 예를 들어, 태스크 1은 해당 코퍼스가 긍정적인 표현인지, 부정적인 표현인지에 대한 것, 태스크 2는 해당 코퍼스가 영화의 어떤 장르인지에 대한 것을 의미할 수 있다.

[0030] 따라서, 본 발명에서는 입력 코퍼스에 대하여 여러 태스크에 대해 동시에 판단함으로써 코퍼스의 카테고리를 분류할 수 있고, 이에 레이블을 부착할 수 있다.

[0032] 도 2는 본 발명의 일실시예에 따른 코퍼스 분류 방법의 플로우 차트를 도시한 도면이다.

[0033] 이하의 단계들은 모두 프로세서에 의해 수행될 수 있다.

[0034] 단계(201)은 학습 코퍼스를 식별하는 과정이다. 학습 코퍼스는 하나의 코퍼스와 복수의 태스크 별 카테고리에 대한 정보를 포함하고 있다. 이 때, 복수의 태스크는 서로 연관이 있는 태스크를 사용할 수 있다. 서로 연관이 있는 태스크란, 태스크 간의 레이블(label)의 빈도 교차표를 구하고 Chi-Squared Test를 수행했을 때 p-value가 2.2e-16 이하 일 때를 의미한다.

[0035] 단계(202)는 학습 코퍼스를 전처리하고 문장 벡터를 생성하는 과정이다. 전처리 작업은 불필요한 특수 문자 제거와 같은 작업을 포함할 수 있다. 문장 벡터는 문장을 이루는 토큰 및 매칭되는 임의의 상수 값으로 구성될 수 있다. 여기서 토큰은 전처리 작업의 결과에 따라 문자, 단어, 형태소 등이 될 수 있다.

[0036] 단계(203)은 태스크 별로 개별 특징을 추출하는 과정이다. 이 과정은 태스크 각각에만 유익한 영향을 가지면서, 해당 태스크를 배제한 다른 태스크에서 개별 특징을 추출함에 있어서 노이즈로 작용할 수 있는 특징들을 추출하는 것이 목표이다.

[0037] 도 2는 태스크가 총 N개일 때를 코퍼스를 분류하는 과정을 도시한 것이므로, 태스크별 개별 특징을 추출 과정도 N개의 과정이 동시에 처리될 수 있다.

[0038] 태스크 별로 개별 특징을 추출하는 과정은 전처리된 코퍼스 및 문장 벡터를 입력받아 컨볼루션 뉴럴 네트워크(Convolution Neural Network)의 컨볼루션 필터 등의 연산 방법을 적용하여 각 태스크의 카테고리와 연관성이 높은 토큰에 가중치를 높게 부여한다. 가중치는 학습 태스크에 포함된 태스크 별로 주어진 카테고리와 단계(205)에서 결정된 카테고리를 비교한 결과에 따라 업데이트될 수 있다.

[0039] 손실 함수는 학습 태스크에 포함된 태스크 별로 주어진 카테고리와 단계(205)에서 결정된 카테고리를 비교 결과로 계산된다. 가중치는 해당 태스크에 대한 카테고리간 비교 결과로 계산되는 손실 함수의 값에 의존하여 업데이트 될 수 있다.

[0040] 단계(204)는 공유 특징을 추출하는 과정이다. 이 과정에서 (i) 전처리된 학습 코퍼스 및 문장 벡터 및 (ii) 상기 개별 특징 중 적어도 하나 이상을 이용할 수 있다. 이 때, 컨볼루션 필터 등의 연산 방법을 적용하여 전체 태스크 분류에 유의한 토큰 부분의 가중치를 높게 부여할 수 있다. 따라서, 여기서 가중치는 태스크 1에서 태스

크 N까지에 대한 카테고리간 비교 결과로 계산되는 손실 함수의 값에 의하여 업데이트된다.

[0041] 단계(205)는 학습 코퍼스에 대한 태스크 별 카테고리를 결정하는 단계이다. 추출된 개별 특징과 공유 특징을 이용하여 상기 학습 코퍼스에 대해 태스크 별로 카테고리를 결정한다.

[0042] 개별 특징 및 공유 특징을 이용할 때 접합(Concatenation), 평균 및 다층 퍼셉트론(perceptron) 레이어 중 하나의 연산 방법을 활용하여 카테고리를 결정할 수 있다. 또한, 위 개별 특징 및 공유 특징이 종합된 결과를 소프트맥스(softmax) 연산 방법을 활용하여 각 태스크 별로 각 카테고리에 대한 값을 도출하고, 그 중 가장 높은 값인 카테고리를 선택함으로써 태스크 별 카테고리를 결정한다.

[0043] 태스크 별 결정된 카테고리는 학습 코퍼스에 포함된 태스크 별 주어진 카테고리와 비교하여 손실 함수의 값을 결정한다. 각각의 태스크에 대해서 비교하여 얻어진 손실 함수의 평균 값을 최종 손실 함수 값으로 사용한다. 이 때, 손실 함수로는 크로스 엔트로피(Cross-Entropy) 등이 사용될 수 있다.

[0044] 업데이트된 가중치 값들은 현재 단계의 최종 손실 함수 값이 이전 단계의 최종 손실 함수 값보다 작을 경우 파일로 저장되고, 다음 학습에서는 저장된 가중치의 값이 업데이트된다. 학습은 단계를 반복하며 최종 손실 함수 값이 사용자가 설정한 횟수만큼 갱신에 실패했을 때 종료된다.

[0045] 학습 이후, 평가 단계에서는 학습 중에 저장된 가중치 중 가장 마지막에 저장된 가중치를 이용한다. 입력 코퍼스는 문장만을 포함하며 카테고리 정보가 주어지지 않는다. 입력된 코퍼스는 코퍼스 식별과정, 코퍼스 전처리 및 문장벡터 생성 과정, 태스크 별 특징 추출 과정, 공유 특징 추출과정을 거쳐 태스크 1의 카테고리, 태스크 2의 카테고리, …, 태스크 N의 카테고리를 동시에 결정할 수 있다.

[0046] 결정된 카테고리에 따라 코퍼스에 레이블을 부착할 수 있고, 레이블에 따라 다수의 코퍼스들을 분류할 수 있다.

[0048] 도 3은 본 발명의 일실시예에 따른 코퍼스 분류하는 예시를 도시한 도면이다.

[0049] 입력 문장은 영화 리뷰 중 한 문장이고, 태스크 1은 긍정/부정 분류, 태스크 2는 영화 장르 분류인 경우의 예시에 대한 것이다.

[0050] 식별된 영화 리뷰 문장(301)이 '통쾌한 액션씬이 빛남!! 음향쪽이 아쉽긴 했지만… 전반적으로 만족스러움' 일 때, 문장 전처리 및 문장 벡터 생성 과정(302)을 거쳐 '!!', '…'와 같은 특수 문자가 제거되고, '통쾌/XR 하/XSA ㄴ/ETM 액션/NNG 씬/NNG 이/JKC 빛나/VV ㅁ/ETN 음향/NNG 쪽/NNB 이/JKS 아쉽/VA 기/ETN ㄴ/JX 하/VX 았/EP 지만/EC 전반/NNG 적/XSN 으로/JKB 만족/NNG 스럽/XSA ㅁ/ETN'와 같이 형태소로 분리된다. 또한, 문장 전처리 결과 이용하여 [3,6,43,68,120, …,]와 같이 문장 내 토큰의 개수만큼의 차원을 가지는 문장 벡터가 생성된다.

[0051] 태스크 별 개별 특징 추출 과정을 통해, 태스크 1(긍정/부정) (303)에 대해서는 '통쾌/XR', '빛나/VV', '아쉽/VA'에 대해 높은 가중치를 부여하고, 태스크 2(장르) (304)에 대해서는 '액션/NNG', '씬/NNG' 등에 높은 가중치가 부여된다.

[0052] 공유 특징 추출 과정(305)을 통해, '통쾌/XR 하/XSA ㄴ/ETM 액션/NNG 씬/NNG'와 같이 태스크 1 및 태스크 2 모두 연관성이 높은 부분에 높은 가중치가 부여된다.

[0053] 코퍼스에 대한 태스크 별 카테고리 결정 과정(306)을 통해, 태스크 1에 대해선 '긍정', 태스크 2에 대해선 '액션/모험' 이라는 카테고리가 결정된다.

[0054] 한편, 본 발명에 따른 방법은 컴퓨터에서 실행될 수 있는 프로그램으로 작성되어 마그네틱 저장매체, 광학적 판독매체, 디지털 저장매체 등 다양한 기록 매체로도 구현될 수 있다.

[0055] 본 명세서에 설명된 각종 기술들의 구현들은 디지털 전자 회로조직으로, 또는 컴퓨터 하드웨어, 펌웨어, 소프트웨어로, 또는 그들의 조합들로 구현될 수 있다. 구현들은 데이터 처리 장치, 예를 들어 프로그램가능 프로세서, 컴퓨터, 또는 다수의 컴퓨터들의 동작에 의한 처리를 위해, 또는 이 동작을 제어하기 위해, 컴퓨터 프로그램 제품, 즉 정보 캐리어, 예를 들어 기계 판독가능 저장 장치(컴퓨터 판독가능 매체) 또는 전파 신호에서 유형적으로 구체화된 컴퓨터 프로그램으로서 구현될 수 있다. 상술한 컴퓨터 프로그램(들)과 같은 컴퓨터 프로그램은 컴파일된 또는 인터프리트된 언어들을 포함하는 임의의 형태의 프로그래밍 언어로 기록될 수 있고, 독립형 프로그램으로서 또는 모듈, 구성요소, 서브루틴, 또는 컴퓨팅 환경에서의 사용에 적절한 다른 유닛으로서 포함하는 임

의의 형태로 전개될 수 있다. 컴퓨터 프로그램은 하나의 사이트에서 하나의 컴퓨터 또는 다수의 컴퓨터들 상에서 처리되도록 또는 다수의 사이트들에 걸쳐 분배되고 통신 네트워크에 의해 상호 연결되도록 전개될 수 있다.

[0056] 컴퓨터 프로그램의 처리에 적절한 프로세서들은 예로서, 범용 및 특수 목적 마이크로프로세서들 둘 다, 및 임의의 종류의 디지털 컴퓨터의 임의의 하나 이상의 프로세서들을 포함한다. 일반적으로, 프로세서는 판독 전용 메모리 또는 랜덤 액세스 메모리 또는 둘 다로부터 명령어들 및 데이터를 수신할 것이다. 컴퓨터의 요소들은 명령어들을 실행하는 적어도 하나의 프로세서 및 명령어들 및 데이터를 저장하는 하나 이상의 메모리 장치들을 포함할 수 있다. 일반적으로, 컴퓨터는 데이터를 저장하는 하나 이상의 대량 저장 장치들, 예를 들어 자기, 자기-광 디스크들, 또는 광 디스크들을 포함할 수 있거나, 이것들로부터 데이터를 수신하거나 이것들에 데이터를 송신하거나 또는 양쪽으로 되도록 결합될 수도 있다. 컴퓨터 프로그램 명령어들 및 데이터를 구체화하는데 적절한 정보 캐리어들은 예로서 반도체 메모리 장치들, 예를 들어, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM(Compact Disk Read Only Memory), DVD(Digital Video Disk)와 같은 광 기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-Optical Media), 롬(ROM, Read Only Memory), 램(RAM, Random Access Memory), 플래시 메모리, EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM) 등을 포함한다. 프로세서 및 메모리는 특수 목적 논리 회로조직에 의해 보충되거나, 이에 포함될 수 있다.

[0057] 또한, 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용매체일 수 있고, 컴퓨터 저장매체 및 전송매체를 모두 포함할 수 있다.

[0058] 본 명세서는 다수의 특정한 구현물의 세부사항들을 포함하지만, 이들은 어떠한 발명이나 청구 가능한 것의 범위에 대해서도 제한적인 것으로서 이해되어서는 안되며, 오히려 특정한 발명의 특정한 실시형태에 특유할 수 있는 특징들에 대한 설명으로서 이해되어야 한다. 개별적인 실시형태의 문맥에서 본 명세서에 기술된 특정한 특징들은 단일 실시형태에서 조합하여 구현될 수도 있다. 반대로, 단일 실시형태의 문맥에서 기술한 다양한 특징들 역시 개별적으로 혹은 어떠한 적절한 하위 조합으로도 복수의 실시형태에서 구현 가능하다. 나아가, 특징들이 특정한 조합으로 동작하고 초기에 그와 같이 청구된 바와 같이 묘사될 수 있지만, 청구된 조합으로부터의 하나 이상의 특징들은 일부 경우에 그 조합으로부터 배제될 수 있으며, 그 청구된 조합은 하위 조합이나 하위 조합의 변형물로 변경될 수 있다.

[0059] 마찬가지로, 특정한 순서로 도면에서 동작들을 묘사하고 있지만, 이는 바람직한 결과를 얻기 위하여 도시된 그 특정한 순서나 순차적인 순서대로 그러한 동작들을 수행하여야 한다거나 모든 도시된 동작들이 수행되어야 하는 것으로 이해되어서는 안 된다. 특정한 경우, 멀티태스킹과 병렬 프로세싱이 유리할 수 있다. 또한, 상술한 실시형태의 다양한 장치 컴포넌트의 분리는 그러한 분리를 모든 실시형태에서 요구하는 것으로 이해되어서는 안되며, 설명한 프로그램 컴포넌트와 장치들은 일반적으로 단일의 소프트웨어 제품으로 함께 통합되거나 다중 소프트웨어 제품에 패키징 될 수 있다는 점을 이해하여야 한다.

[0060] 한편, 본 명세서와 도면에 개시된 본 발명의 실시 예들은 이해를 돕기 위해 특정 예를 제시한 것에 지나지 않으며, 본 발명의 범위를 한정하고자 하는 것은 아니다. 여기에 개시된 실시 예들 이외에도 본 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 자명한 것이다.

부호의 설명

[0061] 101: 코퍼스 분류 시스템

201: 학습 코퍼스 식별 과정

202: 학습 코퍼스 전처리 및 문장 벡터 생성 과정

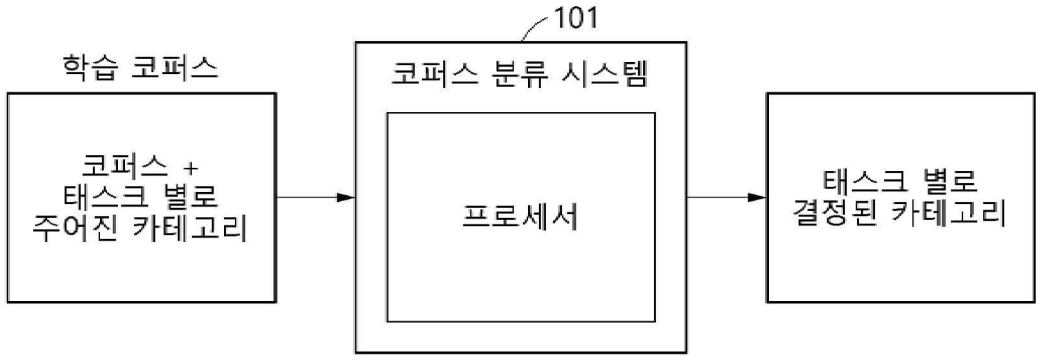
203: 태스크 별 개별 특징 추출 과정

204: 공유 특징 추출 과정

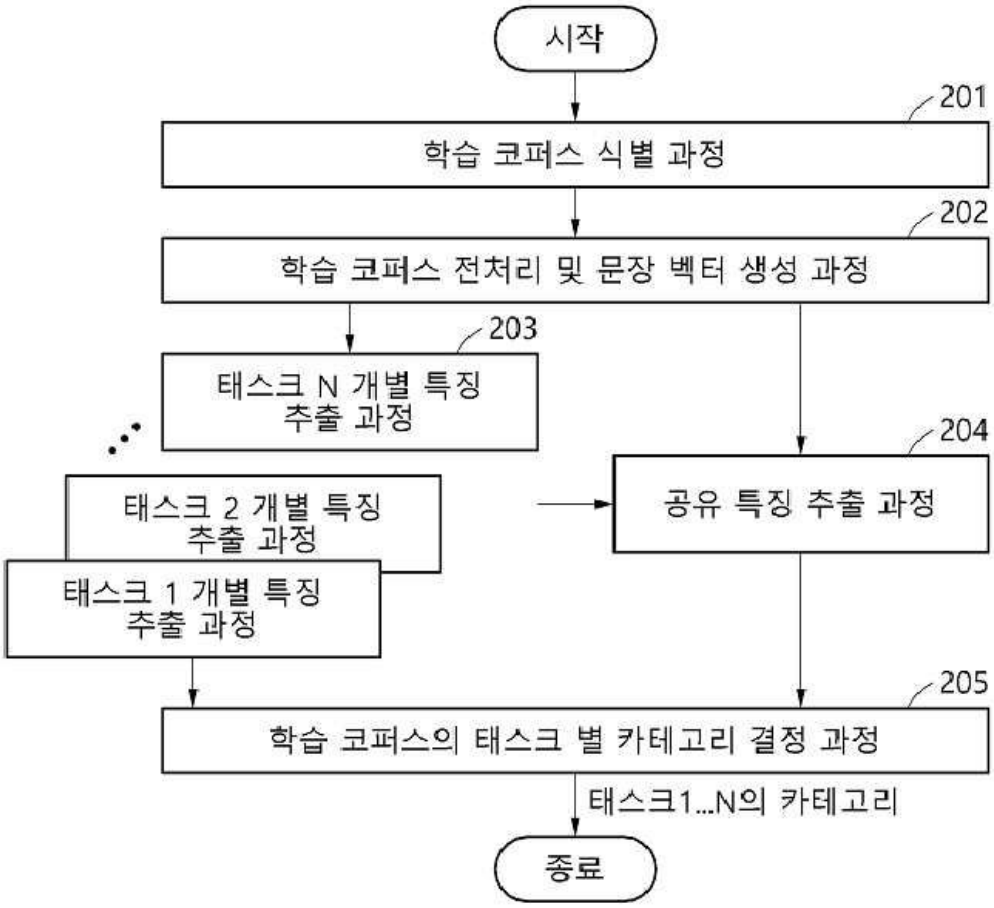
205: 카테고리 결정 과정

도면

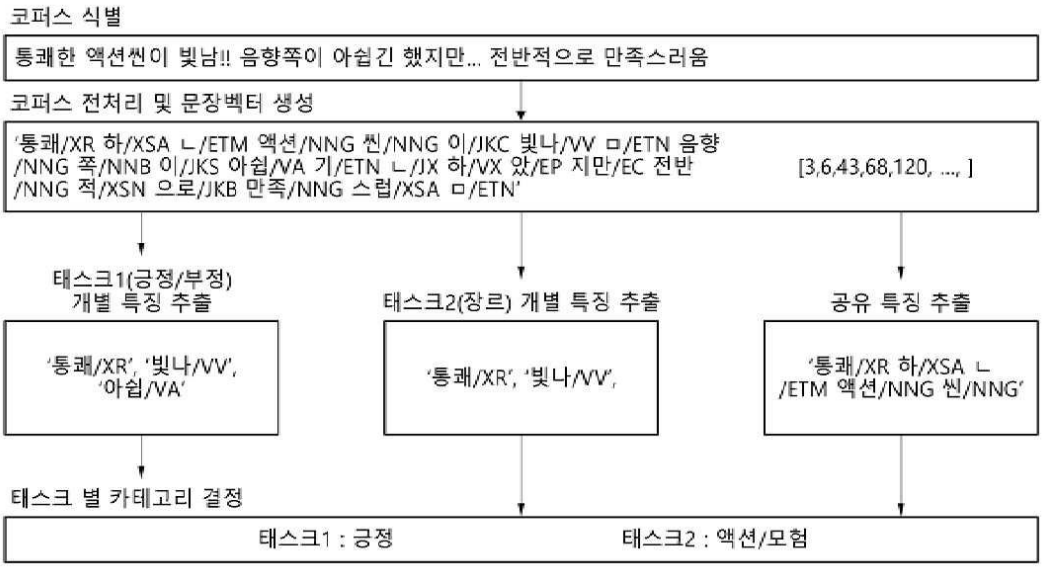
도면1


101
학습 코퍼스
코퍼스 분류 시스템
코퍼스 +
태스크 별로
주어진 카테고리
프로세서
태스크 별로
결정된 카테고리

도면2


시작
201
학습 코퍼스 식별 과정
202
학습 코퍼스 전처리 및 문장 벡터 생성 과정
203
태스크 N 개별 특징
추출 과정
204
태스크 2 개별 특징
추출 과정
공유 특징 추출 과정
태스크 1 개별 특징
추출 과정
205
학습 코퍼스의 태스크 별 카테고리 결정 과정
태스크1...N의 카테고리
종료

도면3


코퍼스 식별
통쾌한 액션씬이 빛남!! 음향쪽이 아쉽긴 했지만... 전반적으로 만족스러움
코퍼스 전처리 및 문장벡터 생성
'통쾌/XR 하/XSA ㄴ/ETM 액션/NNG 씬/NNG 이/JKC 빛나/VV ㅁ/ETN 음향 /NNG 쪽/NNB 이/JKS 아쉽/VA 기/ETN ㄴ/JX 하/VX 았/EP 지만/EC 전반 /NNG 적/XSN 으로/JKB 만족/NNG 스럽/XSA ㅁ/ETN'
[3,6,43,68,120, ...,]
태스크1(긍정/부정) 개별 특징 추출
태스크2(장르) 개별 특징 추출
공유 특징 추출
'통쾌/XR', '빛나/VV', '아쉽/VA'
'통쾌/XR', '빛나/VV',
'통쾌/XR 하/XSA ㄴ /ETM 액션/NNG 씬/NNG'
태스크 별 카테고리 결정
태스크1 : 긍정
태스크2 : 액션/모험