



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0101551
(43) 공개일자 2019년09월02일

(51) 국제특허분류(Int. Cl.)

G06N 3/04 (2006.01)

(52) CPC특허분류

G06N 3/0436 (2013.01)

G06N 3/08 (2013.01)

(21) 출원번호 10-2018-0021651

(22) 출원일자 2018년02월23일

심사청구일자 2018년02월23일

(71) 출원인

창원대학교 산학협력단

경상남도 창원시 의창구 창원대학교 20(퇴촌동)

(72) 발명자

차정원

경상남도 창원시 성산구 동산로 115, 105동 1404호 (상남동, 대동아파트)

박다솔

경상남도 김해시 능동로 117, 414동 1101호 (부곡동, 석봉마을부영아파트)

신창욱

경상남도 김해시 가락로 301, 가동 306호 (구산동, 유신주택)

(74) 대리인

김종선, 이형석

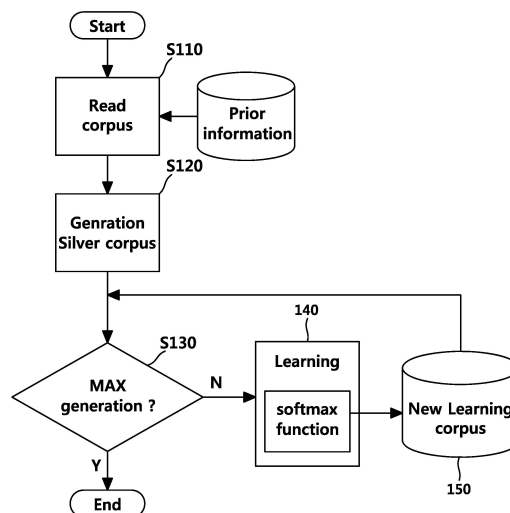
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘을 사용한 분류 방법

(57) 요약

퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘을 사용한 분류 방법이 개시된다. 상기 방법은 데이터베이스로부터 분류를 목적으로 하는 학습 코퍼스를 읽어오는 단계와, 상기 학습 코퍼스에 대해 사전 정보를 이용하여 분류 값에 해당하는 레이블을 생성하는 단계와, 생성된 레이블을 상기 학습 코퍼스에 부착하여 실버 코퍼스를 생성하는 단계와, 상기 실버 코퍼스에 대해 자연어 처리를 위한 CNN(convolutional neural network) 학습 모델을 이용하여 학습을 수행하여 새로운 학습 코퍼스를 생성하는 단계를 포함한다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 1711060015
 부처명 과학기술정보통신부
 연구관리전문기관 한국연구재단
 연구사업명 차세대정보·컴퓨팅기술개발
 연구과제명 설득적 인터랙션에 기반한 멀티모달 대화 모델 개발
 기 여 율 1/2
 주관기관 창원대학교
 연구기간 2017.11.01 ~ 2018.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1345264468
 부처명 교육부
 연구관리전문기관 한국연구재단
 연구사업명 개인기초연구(교육부)
 연구과제명 인공지능 대화 시스템에서 텍스트로부터 감정 인식
 기 여 율 1/2
 주관기관 창원대학교
 연구기간 2017.06.01 ~ 2020.05.31

명세서

청구범위

청구항 1

컴퓨터로 구현되는 방법에 있어서,

데이터베이스로부터 분류를 목적으로 하는 학습 코퍼스를 읽어오는 단계;

상기 학습 코퍼스에 대해 사전 정보를 이용하여 분류 값에 해당하는 레이블을 생성하는 단계;

생성된 레이블을 상기 학습 코퍼스에 부착하여 실버 코퍼스를 생성하는 단계; 및

상기 실버 코퍼스에 대해 자연어 처리를 위한 CNN(convolutional neural network) 학습 모델을 이용하여 학습을 수행하여 새로운 학습 코퍼스를 생성하는 단계;를 포함하는 방법.

청구항 2

제1항에 있어서,

상기 분류 값은 '0' 또는 '1'의 이진논리로 표현되는 것을 특징으로 하는 방법.

청구항 3

제1항에 있어서,

상기 새로운 학습 코퍼스를 생성하는 단계는,

학습 결과, 상기 실버 코퍼스가 어느 레이블에 속하는지를 확률 분포로 표현한 퍼지 범주 표현을 생성하는 것을 특징으로 하는 방법.

청구항 4

제1항에 있어서,

상기 새로운 학습 코퍼스를 생성하는 단계는,

학습 결과, 상기 실버 코퍼스가 어느 레이블에 속하는지를 확률 분포로 표현하기 위해 각 레이블에 대한 퍼지 범주 표현 값을 생성하는 단계; 및

생성된 퍼지 범주 표현 값들과 상기 학습 코퍼스를 상기 새로운 학습 코퍼스로서 생성하는 단계;를 포함하는 방법.

청구항 5

제4항에 있어서,

상기 새로운 학습 코퍼스에 대해 반복적으로 학습을 수행하는 단계; 및

최종 생성된 퍼지 범주 표현 값들 중에서 가장 큰 값에 해당하는 레이블을 상기 학습 코퍼스에 부착하여 골드 코퍼스를 생성하는 단계;를 더 포함하는 방법.

청구항 6

제 5항에 있어서,

상기 반복적으로 학습을 수행하는 단계는,

상기 반복적으로 학습이 수행됨에 따라 상기 퍼지 범주 표현 값들이 업데이트되는 것을 특징으로 하는 방법.

청구항 7

컴퓨터로 구현되는 시스템에 있어서,

컴퓨터가 관독 가능한 명령을 실행하도록 구현되는 적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

데이터베이스로부터 분류를 목적으로하는 학습 코퍼스를 읽어오고, 읽어온 학습 코퍼스에 대해 사전 정보를 이용하여 분류 값에 해당하는 레이블을 생성하고, 생성된 레이블을 상기 학습 코퍼스에 부착하여 실버 코퍼스를 생성하는 실버 코퍼스 생성부; 및

상기 실버 코퍼스에 대해 자연어 처리를 위한 CNN(convolutional neural network) 학습 모델을 이용하여 학습을 수행하여 새로운 학습 코퍼스를 생성하는 새로운 학습 코퍼스 생성부;를 포함하는 시스템.

발명의 설명

기술 분야

[0001] 본 발명의 개념에 따른 실시 예는 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘에 관한 것으로, 특히 학습 코퍼스(corpus)에 대해 학습을 거쳐서 다수의 카테고리로 분류하는 방법에 관한 것이다.

배경 기술

[0002] 청소년의 따돌림 문제가 심각한 사회문제로 대두되고 스마트폰 보급과 인터넷 사용이 일상화됨에 따라 학교 폭력은 사이버상의 따돌림과 괴롭힘으로까지 확대되고 있으며 사이버불링의 발생률이 증가하고 있다.

[0003] '사이버불링'은 사이버 범죄의 다른 말로, 인터넷과 관련된 기술상에서 다른 사람에게 피해를 입히는 모든 행위를 의미한다. 혐오 발언과 같은 언어 폭력으로 인한 사이버불링의 비중이 매우 높고, 사이버불링의 피해자는 자살을 선택할 정도로 심각한 물리적, 정신적 피해를 받는다.

[0004] 블랙리스트를 이용하여 혐오 단어를 매칭할 수 있지만, 단순한 혐오 단어 매칭 방법은 혐오 발언의 미묘하고 교묘한 표현을 찾아내지 못한다는 문제점이 존재한다.

[0005] 준지도 학습(semi-supervised learning)은 레이블이 부착되지 않은 대량의 데이터와 레이블이 부착된 소량의 데이터를 이용한다. 대부분의 준지도 학습 과정은 초기에 레이블이 부착된 데이터만으로 학습한 뒤 레이블이 부착되지 않은 데이터에 점차적으로 레이블을 부착해 나가며 레이블이 부착된 데이터와 레이블이 부착되지 않은 데이터를 동시에 학습한다.

[0006] 한국 등록특허 제10-1417647호는 사용자에게 모음 약화에 대한 피드백을 제공함으로써 외국어 학습의 효과를 증대시키는 발명으로써, 모음 약화가 레이블링(labeling)된 텍스트 모음 약화 말뭉치 DB로부터 텍스트 특징 정보를 추출하고, 상기 추출한 텍스트 특징 정보를 훈련 데이터로 한 기계 학습을 통하여 모음 약화가 발생할 확률 분포 및 가중치 정보를 저장하는 모음 약화 예측 모델 DB와, 상기 모음 약화가 있을 확률 및 가중치를 가중 합(weighted sum)하여 텍스트에 대한 모음 약화를 예측하는 발명이 개시되어 있다.

[0007] 그러나, 상기 선행기술문헌은 여전히 텍스트에 대한 모음 약화 예측에 대해 여전히 신뢰하기 어려운 문제점이 있다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 한국 등록특허 제10-1417647호(2014.07.01)

발명의 내용

해결하려는 과제

[0009] 본 발명은 상기와 같은 문제점을 해결하기 위하여 안출된 것으로서, 본 발명은 레이블이 부착되어 있는 학습 코퍼스에 대해 학습을 진행하여 레이블의 확률 분포인 퍼지 범주 표현을 생성하고, 생성된 퍼지 범주 표현을 학습

하여 학습 코퍼스를 분류할 수 있는 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘을 제공하는 것을 목적으로 한다.

과제의 해결 수단

- [0010] 상기와 같은 목적을 달성하기 위한 본 발명의 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘을 사용한 분류 방법은, 데이터베이스로부터 분류를 목적으로 하는 학습 코퍼스를 읽어오는 단계와, 상기 학습 코퍼스에 대해 사전 정보를 이용하여 분류 값에 해당하는 레이블을 생성하는 단계와, 생성된 레이블을 상기 학습 코퍼스에 부착하여 실버 코퍼스를 생성하는 단계와, 상기 실버 코퍼스에 대해 자연어 처리를 위한 CNN(convolutional neural network) 학습 모델을 이용하여 학습을 수행하여 새로운 학습 코퍼스를 생성하는 단계를 포함한다.
- [0011] 상기 분류 값은 '0' 또는 '1'의 이진논리로 표현된다.
- [0012] 상기 새로운 학습 코퍼스를 생성하는 단계는, 학습 결과, 상기 실버 코퍼스가 어느 레이블에 속하는지를 확률 분포로 표현한 퍼지 범주 표현을 생성한다.
- [0013] 상기 새로운 학습 코퍼스를 생성하는 단계는, 학습 결과, 상기 실버 코퍼스가 어느 레이블에 속하는지를 확률 분포로 표현하기 위해 각 레이블에 대한 퍼지 범주 표현 값을 생성하는 단계와, 생성된 퍼지 범주 표현 값들과 상기 학습 코퍼스를 상기 새로운 학습 코퍼스로서 생성하는 단계를 포함한다.
- [0014] 상기 방법은 상기 새로운 학습 코퍼스에 대해 반복적으로 학습을 수행하는 단계와, 최종 생성된 퍼지 범주 표현 값들 중에서 가장 큰 값에 해당하는 레이블을 상기 학습 코퍼스에 부착하여 골드 코퍼스를 생성하는 단계를 더 포함한다.
- [0015] 상기 반복적으로 학습을 수행하는 단계는, 상기 반복적으로 학습이 수행됨에 따라 상기 퍼지 범주 표현 값들이 업데이트된다.
- [0016] 본 발명의 컴퓨터로 구현되는 시스템은, 컴퓨터가 관독 가능한 명령을 실행하도록 구현되는 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 데이터베이스로부터 분류를 목적으로 하는 학습 코퍼스를 읽어오고, 읽어온 학습 코퍼스에 대해 사전 정보를 이용하여 분류 값에 해당하는 레이블을 생성하고, 생성된 레이블을 상기 학습 코퍼스에 부착하여 실버 코퍼스를 생성하는 실버 코퍼스 생성부와, 상기 실버 코퍼스에 대해 자연어 처리를 위한 CNN(convolutional neural network) 학습 모델을 이용하여 학습을 수행하여 새로운 학습 코퍼스를 생성하는 새로운 학습 코퍼스 생성부를 포함한다.

발명의 효과

- [0017] 상기한 바와 같은 본 발명의 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘은 사람이 직접 학습 코퍼스에 대한 레이블을 생성할 필요 없이 사전 정보를 이용하여 학습 코퍼스에 대한 레이블을 생성하므로 시간과 비용을 줄일 수 있고, 세대(generation)를 거쳐 상기 학습 코퍼스에 대해 학습을 진행하여 레이블의 확률 분포인 퍼지 범주 표현을 생성하고, 생성된 퍼지 범주 표현을 학습하여 학습 코퍼스를 재생성함으로써 사전 매칭에 비해 신뢰도 높은 분류 결과를 얻을 수 있는 효과가 있다.

도면의 간단한 설명

- [0018] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.
- 도 1은 본 발명의 실시 예에 따른 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘을 설명하기 위한 순서도이다.
- 도 2는 본 발명의 실시 예에 따른 실버 코퍼스의 실시 예를 나타낸다.
- 도 3은 본 발명의 실시 예에 따른 학습 모델의 실시 예를 설명하기 위한 개념도이다.
- 도 4는 본 발명의 실시 예에 따른 실버 코퍼스에 대해 학습을 진행하여 생성된 퍼지 범주 표현을 나타내는 개념도이다.
- 도 5는 본 발명의 실시 예에 따른 세대를 생성할 때마다 퍼지 범주 표현 값이 업데이트 되는 실시 예를 나타낸다.
- 도 6은 본 발명의 실시 예에 따른 분류 방법을 수행하기 위해 컴퓨터로 구현되는 시스템을 나타낸다.

발명을 실시하기 위한 구체적인 내용

- [0019] 본 발명은 분류를 목적으로 하는 코퍼스(corpus)를 입력받아 학습을 통해 상기 코퍼스에 대한 분류 단위인 레이블을 확률 분포로 표현하는 것이다. 학습 데이터로는 Google, Twitter, Yahoo, Facebook 등 소셜 네트워크 서비스(social networking service(SNS))와 웹페이지 댓글 등을 크롤링하여 대량의 코퍼스를 제작할 수 있다. 코퍼스의 단위는 문장, 문단, 또는 문서 등으로 다양하게 나타날 수 있다. 본 발명은 코퍼스의 종류나 분류 종류의 수에 관계 없이 동작 가능하다.
- [0020] 또한, 문장은 오타, 띄어쓰기 오류 등을 포함하는 비정형 데이터이므로, 자연어 처리를 바로 적용하기 위해 띄어쓰기를 제거한 문장을 이용할 수 있다.
- [0021] 소셜 네트워크 환경은 복수의 전자 기기들, 복수의 서버들, 및 네트워크를 포함할 수 있고, 복수의 전자 기기들은 컴퓨터 장치로 구현되는 고정형 단말 또는 이동형 단말일 수 있다. 예컨대, 복수의 전자 기기들은 스마트폰(smart phone), 휴대폰, 컴퓨터, 노트북, 태블릿 PC 등일 수 있다.
- [0022] 복수의 서버들 각각은 복수의 전자기기들과 무선 또는 유선 통신 방식을 이용하여 네트워크를 통해 통신하여 명령, 코드, 파일, 콘텐츠, 서비스 등을 제공하는 컴퓨터 장치 또는 복수의 컴퓨터 장치들로 구현될 수 있다.
- [0023] 따라서, 본 발명의 실시 예에 따른 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘은 소셜 네트워크 환경에서의 적어도 하나의 서버에 의해 수행될 수 있다. 상기 서버는 실버 코퍼스 생성부와 새로운 학습 코퍼스 생성부를 포함하며, 이하에서는 본 발명의 실시 예 및 도면을 참조하여, 본 발명을 더욱 상세히 설명한다.
- [0025] 도 1은 본 발명의 실시 예에 따른 퍼지 범주 표현을 이용한 확률 레이블 부착 알고리즘을 설명하기 위한 순서도이다. 도 1을 참조하면, 실버 코퍼스 생성부가 분류를 목적으로 하는 학습 코퍼스를 데이터베이스로부터 읽어오고(S110), 읽어온 학습 코퍼스에 대해 사전 정보를 이용하여 분류 값에 해당하는 레이블을 생성한다. 예컨대, 상기 사전 정보는 비속어 사전, 혐오 단어 사전 등을 의미할 수 있다.
- [0026] 실버 코퍼스 생성부는 생성된 레이블을 상기 학습 코퍼스에 부착하여 실버 코퍼스(silver corpus)를 생성한다. 즉, 사전 정보 매칭으로 생성된 레이블이 부착된 학습 코퍼스를 실버 코퍼스라 한다.
- [0027] 초기 레이블이 부착된 학습 코퍼스를 사람이 직접 생성할 필요 없이 사전 정보를 이용하여 자동으로 생성함으로써 시간과 비용을 줄일 수 있는 효과가 있다.
- [0029] 도 2는 본 발명의 실시 예에 따른 실버 코퍼스의 실시 예를 나타낸다. 도 2를 참조하면, 실버 코퍼스의 분류 값은 분류 클래스를 의미하는 상수나 기호로 나타낼 수 있다. 분류는 복수의 카테고리로 구성될 수 있다.
- [0030] 예컨대, 사전에는 단어 1, 단어 2, 단어 3, 및 단어 4를 포함하고 있다고 가정한다. 학습 코퍼스 문장에 단어 1, 단어 2, 단어 3, 및 단어 4가 하나도 포함되어 있지 않으면 상기 학습 코퍼스 문장을 '0'으로 분류하고, 학습 코퍼스 문장에 단어 1, 단어 2, 단어 3, 및 단어 4 중에서 적어도 하나가 포함되어 있으면 상기 학습 코퍼스 문장을 '0'으로 분류한다.
- [0031] 따라서, 사전 정보 매칭으로 생성된 레이블은 '0' 또는 '1'의 이진 논리로 표현된다. 그러나, 이러한 이진 논리의 레이블은 100% 신뢰하기 어려우므로 학습 코퍼스 문장이 어느 레이블에 속하는지 여부를 확률 분포로 표현하기 위해 퍼지 범주 표현을 생성할 필요가 있다.
- [0033] 다시 도 1을 참조하면, 새로운 학습 코퍼스 생성부는 설정된 파라미터 값으로 반복 루프를 진행하여 퍼지 범주 표현을 생성한다. 퍼지 범주 표현(또는 퍼지 표현)은 명제 혹은 집합에서 참 또는 거짓과 같이 뜻이 명확한 것을 다루는 문제가 아닌 애매모호한 기준을 다루기 위한 것이다.
- [0034] 본 명세서에서는 파라미터 값으로서 세대(generation)라는 개념을 이용한다. 본 명세서에서의 세대란 유전 알고리즘(genetic algorithm)의 세대와 유사한 주기를 의미하며, 구체적으로 학습 코퍼스에 대해 학습을 진행하여 퍼지 범주 표현 값과 학습 코퍼스 문장으로 학습 코퍼스가 재생성되는 주기를 의미한다. 따라서, 새로운 학습 코퍼스 생성부는 설정된 파라미터 값인 세대 값에 대해서 반복 루프를 진행한다.

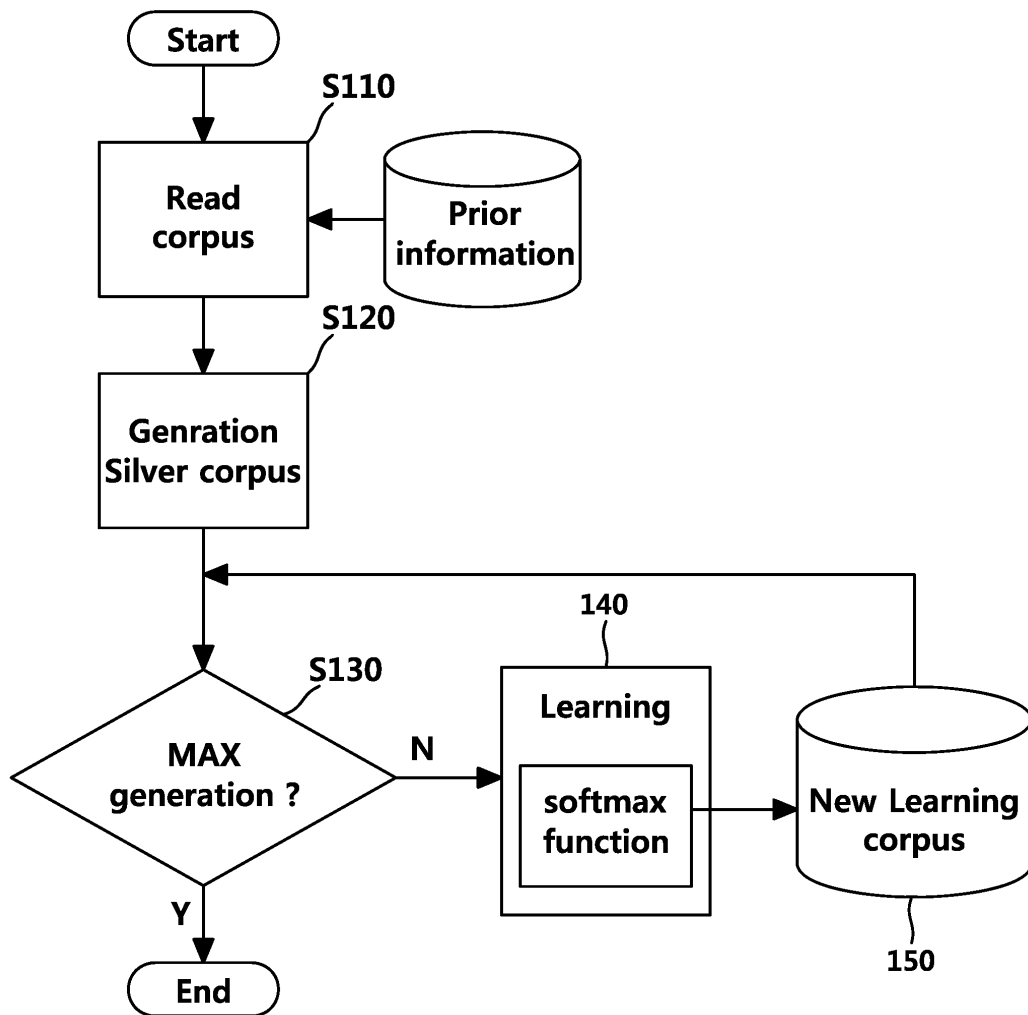
- [0035] 먼저 현재 세대 값이 설정된 세대 값인지 여부를 판단한다(S130). 현재 세대 값이 설정된 세대 값이 아니면(S130의 N), 현재 학습 코퍼스를 기반으로 학습을 수행하여(S140) 새로운 학습 코퍼스를 생성한다(S150).
- [0036] 예컨대, 실버 코퍼스에 대해 자연어 처리를 위한 학습 모델을 이용하여 학습을 수행한다. 학습 결과, 상기 실버 코퍼스가 어느 레이블에 속하는지를 확률 분포로 표현하기 위해 각 레이블에 대한 퍼지 범주 표현 값을 생성한다. 생성된 퍼지 범주 표현 값들과 상기 학습 코퍼스를 상기 새로운 학습 코퍼스로서 생성한다.
- [0037] 현재 세대 값이 설정된 세대 값과 같아질 때까지 상기 새로운 학습 코퍼스에 대해 반복적으로 학습을 수행한다. 상기 반복적으로 학습이 수행됨에 따라 상기 퍼지 범주 표현 값들이 업데이트된다.
- [0038] 즉, 각 세대마다 해당 학습 코퍼스의 문장이 각 레이블에 속할 확률인 퍼지 범주 표현 값을 생성함으로써 퍼지 범주 표현을 생성할 수 있다. 따라서, 각 세대마다 이전 세대에서 생성된 퍼지 범주 표현과 학습 코퍼스의 문장을 입력으로 하여 학습을 진행한다.
- [0039] 학습 모델로서 컨볼루션 신경망(convolutional neural networks(CNN))을 이용할 수 있다. 학습 코퍼스의 문장이 어느 레이블에 속하는지 여부를 확률 분포로 예측하기 위해서 활성화 함수(activation function)를 사용한다. 예컨대, 활성화 함수 중에서 소프트맥스(softmax)를 사용할 수 있으나 이에 한정되는 것은 아니다. CNN 학습에서 소프트맥스 함수의 입력 값을 각 학습 코퍼스의 문장에 부착된 레이블에 대한 퍼지 범주 표현 값으로 설정한다.
- [0041] 도 3은 본 발명의 실시 예에 따른 학습 모델의 실시 예를 설명하기 위한 개념도이다. 도 3을 참조하면, CNN 모델은 두 가지 연산 층인 컨볼루션 레이어(convolution layer), 풀링 레이어(pooling layer), 및 최종적으로 완전 연결 레이어(fully-connected layer)를 통해 분류를 수행하는 모델이다.
- [0042] 크기가 다른 다양한 필터를 사용하여 컨볼루션 레이어에서 여러 개의 특징 맵(feature map)을 생성한다. 특징 맵의 값들을 활성화 함수를 거친 후 max-over-time pooling 연산 단계를 거친다. max-over-time pooling은 각 필터의 여러 개 값 중에서 가장 큰 값을 선택하는 기법이다. 그 연산 값들이 완전 연결 레이어를 거쳐 출력이 결정되는 모델이다.
- [0043] 상기 CNN 구조에 추가로 문장의 정보를 얻기 위해 CNN에 입력하였던 문장을 음절 단위로 Bidirectional-GRU에 입력하여 문장의 분산 표상(distributed representation)에 해당하는 벡터를 구한다. 그렇게 구해진 문장 벡터는 max-over-time pooling이 수행된 후의 CNN 특징 벡터(feature vector)에 합쳐져 최종 분류 값을 결정하기 위한 softmax 분류기에 입력된다.
- [0045] 도 4는 본 발명의 실시 예에 따른 학습 코퍼스에 대해 학습을 진행하여 생성된 퍼지 범주 표현을 나타내는 개념도이다. 도 4를 참조하면, 2개의 범주로 분류한다고 가정하자. 학습 코퍼스의 문장이 '분류 카테고리 1'에 속할 확률이 '0.000110383'이고, '분류 카테고리 2'에 속할 확률이 '0.99989'이다.
- [0046] 이진 분류이므로 '분류 카테고리 1'과 '분류 카테고리 2' 각각에 대응하는 퍼지 범주 표현 값의 합은 1이다.
- [0047] 따라서, 퍼지 범주 표현을 이용하여 학습 코퍼스의 분류 단위가 어느 레이블에 속하는지를 확률 분포로서 표현할 수 있다.
- [0049] 도 5는 본 발명의 실시 예에 따른 세대를 생성할 때마다 퍼지 범주 표현 값이 업데이트 되는 실시 예를 나타낸다. 도 5를 참조하면, 세대 값을 '3'으로 설정하고, 학습 코퍼스의 문장(예컨대, '원하는 대로 이기정의 나이스 샷')이 '골프' 도메인에 포함되는 기사 제목인지, '양궁' 도메인에 포함되는 기사 제목인지를 분류한다고 가정한다. 또한, '샷'에 해당하는 단어는 골프 사전에 포함되어 있다고 가정한다.
- [0050] 세대 값이 '0'일 경우(generation 0), 실버 코퍼스가 생성됨을 의미하고, 퍼지 범주 표현 값 중 'E'는 부동 소수점 10진수의 가수부를 의미한다. '샷'에 해당하는 단어는 골프 사전에 포함되어 있다고 가정했으므로, 세대 값이 '0'일 경우 '골프' 카테고리에 '1'이라는 레이블이 부착되고, '양궁' 카테고리에 '0'이라는 레이블이 부착된다.
- [0051] 세대 값이 '1'일 경우(generation 1), 실버 코퍼스에 대해 학습을 진행하여 해당 카테고리에 해당하는 레이블은

학습 코퍼스가 해당 카테고리에 속할 확률에 대응하는 퍼지 범주 표현 값으로 변경된다.

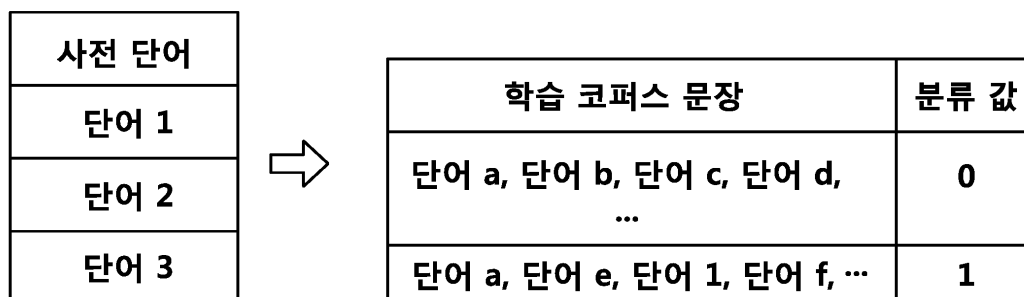
- [0052] 예컨대, '원하는 대로 이기정의 나이스 샷'이 '골프' 도메인에 포함되는 기사 제목일 확률이 '0.999931'이고, '양궁' 도메인에 포함되는 기사 제목일 확률이 '6.85877E-05'이다.
- [0053] 세대 값이 '2'일 경우(generation 2), 이전 세대(generation 1)에서 생성된 퍼지 범주 표현과 학습 코퍼스 문장을 입력으로 하여 학습을 진행한다. 세대가 생성될수록 해당 퍼지 범주 표현 값이 업데이트된다.
- [0054] 예컨대, '원하는 대로 이기정의 나이스 샷'이 '골프' 도메인에 포함되는 기사 제목일 확률이 '0.999827'이고, '양궁' 도메인에 포함되는 기사 제목일 확률이 '0.000172864'이다.
- [0055] 세대 값이 '3'일 경우(generation 3), 이전 세대(generation 2)에서 생성된 퍼지 범주 표현과 학습 코퍼스 문장을 입력으로 하여 학습을 진행한다. 예컨대, '원하는 대로 이기정의 나이스 샷'이 '골프' 도메인에 포함되는 기사 제목일 확률이 '0.999618'이고, '양궁' 도메인에 포함되는 기사 제목일 확률이 '0.000381843'이다.
- [0057] 다시 도 1을 참조하면, 현재 세대 값이 설정된 세대 값과 같을 때(S130의 Y), 새로운 학습 코퍼스 생성부는 최종 생성된 퍼지 범주 표현 값들 중에서 가장 큰 값에 해당하는 레이블을 학습 코퍼스에 부착하여 골드 코퍼스(gold corpus)를 생성한다.
- [0058] 예컨대, 도 4에 도시된 바와 같이 최종 생성된 퍼지 범주 표현 값들(예컨대, '0.999618' 및 '0.000381843') 중에서 가장 큰 값(예컨대, '0.999618')에 해당하는 레이블(예컨대, '골프')을 학습 코퍼스(예컨대, '원하는 대로 이기정의 나이스 샷')에 부착하여 골드 코퍼스를 생성할 수 있다.
- [0059] 실버 코퍼스는 정답이 부착되었지만 오류가 일부 포함된 완전하지 않은 코퍼스를 의미하는 반면, 골드 코퍼스는 정답이 부착된 오류가 없는 코퍼스를 의미한다. 따라서, 사전 매칭에 비해 신뢰도 높은 분류 결과를 얻을 수 있는 효과가 있다.
- [0061] 도 6은 본 발명의 실시 예에 따른 분류 방법을 수행하기 위해 컴퓨터로 구현되는 시스템을 나타낸다. 시스템(10)은 컴퓨터가 판독 가능한 명령을 실행하도록 구현되는 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 실버 코퍼스 생성부(100)와 새로운 학습 코퍼스 생성부(200)를 포함한다.
- [0062] 실버 코퍼스 생성부(100)는 데이터베이스로부터 분류를 목적으로하는 학습 코퍼스를 읽어오고, 읽어온 학습 코퍼스에 대해 사전 정보를 이용하여 분류 값에 해당하는 레이블을 생성하고, 생성된 레이블을 상기 학습 코퍼스에 부착하여 실버 코퍼스를 생성한다.
- [0063] 새로운 학습 코퍼스 생성부(200)는 상기 실버 코퍼스에 대해 자연어 처리를 위한 CNN(convolutional neural network) 학습 모델을 이용하여 학습을 수행하여 새로운 학습 코퍼스를 생성한다.
- [0065] 본 발명은 도면에 도시된 일 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

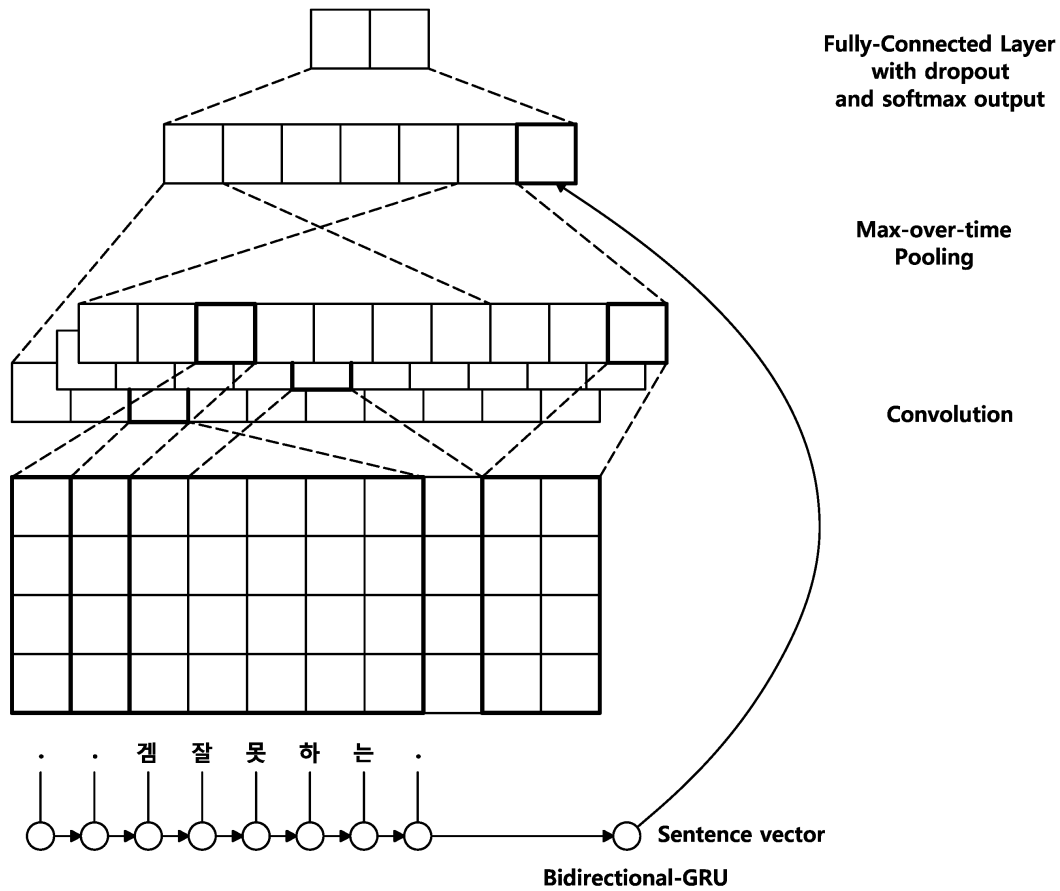
도면1



도면2



도면3



도면4

분류 카테고리 1	분류 카테고리 2	학습 코퍼스 문장
0.000110383	0.99989	단어 a, 단어 b, 단어 1, 단어 c, ...

도면5

generation	골프	양궁	학습 코퍼스 문장
0	1	0	'원하는 대로' 이기정의 나이스 샷
1	0.999931	6.85877E-05	'원하는 대로' 이기정의 나이스 샷
2	0.999827	0.000172864	'원하는 대로' 이기정의 나이스 샷
3	0.999618	0.000381843	'원하는 대로' 이기정의 나이스 샷

도면6

