



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0102327
(43) 공개일자 2022년07월20일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06F 40/117 (2020.01)
G06N 3/04 (2006.01) G06N 5/02 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06F 40/117 (2020.01)
(21) 출원번호 10-2021-0004560
(22) 출원일자 2021년01월13일
심사청구일자 2021년01월13일

(71) 출원인
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
정상근
서울특별시 동작구 장승배기로16길 134 쌍용예가
아파트 102동 801호
황태욱
대전광역시 유성구 유성대로 757-5 동잠빌라 204
호
(74) 대리인
이은철, 김재문

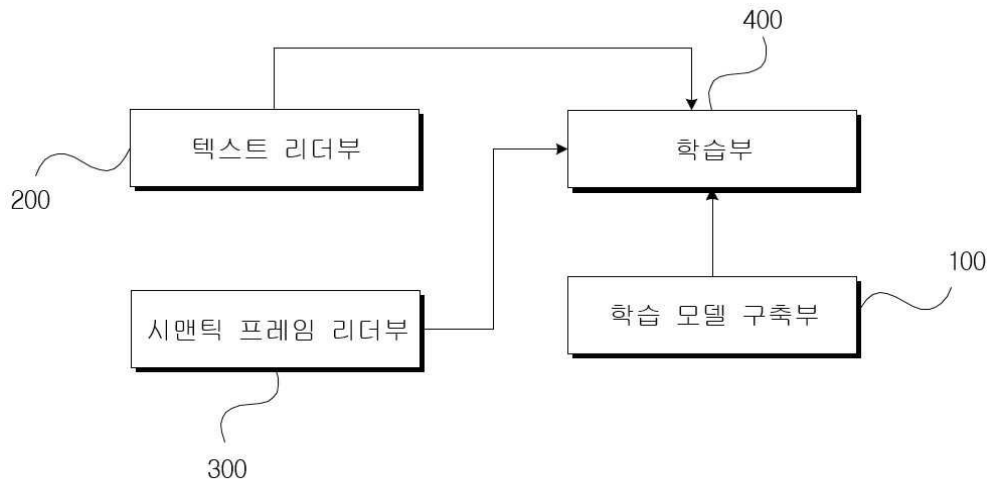
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템 및 방법

(57) 요약

본 기술은 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템 및 방법이 개시된다. 이러한 기술에 대한 구체적인 구현 예는 의미들의 의미 태그, 개체 태그, 및 개체값 중 적어도 하나를 포함하는 구조 정보의 손실없이 저차원의 시각화 학습 모델을 구축하고 구축된 시각화 학습 모델을 통해 회귀 기반 학습으로 입력된 말뭉치에 대한 학습을 수행함에 따라 학습된 의미 벡터와 정답 의미 벡터에 대한 정량화가 가능하고 저차원의 시각화 학습 모델의 훈련의 효율성을 향상시킬 수 있고, 또한, 수집된 말뭉치에 대한 시각화 학습 모델을 통해 학습 결과를 직관적 분석이 가능하고, 학습 결과에 따라 훈련에 필요한 학습 변수를 제어함에 따라 학습 성능을 더욱 향상시킬 수 있다.

대표도 - 도1



(52) CPC특허분류

G06F 40/205 (2020.01)

G06F 40/30 (2020.01)

G06N 3/04 (2013.01)

G06N 5/02 (2019.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711108241
과제번호	2019R1F1A1060601
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	설명 가능한 딥러닝 자연어 처리 기술 개발
기 여 율	1/1
과제수행기관명	충남대학교
연구기간	2020.03.01 ~ 2021.02.28

명세서

청구범위

청구항 1

다수의 수집된 말뭉치 각각에 대해 의미들의 구조 정보를 생성하고, 생성된 각 구조 정보에 대해 3차원 좌표 형태로 변환한 다음 3차원 좌표 형태의 각 구조 정보를 시각화하여 시각화 학습 모델을 구축하는 학습 모델 구축부;

입력된 말뭉치에 대해 분리된 토큰 각각의 임베딩값으로 전처리한 다음 기 구축된 선학습 BERT(Bidirectional Encoder Representation from Transformers) 언어 모델을 토대로 학습 수행하여 문장 벡터를 도출하는 텍스트 리더부;

입력된 말뭉치에 대해 의미들의 구조 정보를 생성하고 생성된 각 구조 정보 기반에 대해 BERT 언어 모델을 토대로 학습 수행하여 다수의 의미 벡터를 도출하는 시맨틱 프레임 리더부; 및

상기 시맨틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델 기반 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 유사 의미 벡터와 문장 벡터 간의 정해진 거리함수를 토대로 정답 의미 벡터를 도출하는 학습부를 포함하는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템.

청구항 2

제1항에 있어서, 상기 구조 정보는,

문장의 주제인 의미 태그, 문장 요소의 이름인 개체 태그, 및 개체 태그의 실제값인 개체값 중 적어도 하나인 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템.

청구항 3

제2항에 있어서, 상기 학습 모델 구축부는,

수집된 다수의 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하는 구조 정보 생성모듈;

생성된 구조 정보 중 의미 태그와의 거리가 일정한 다수의 개체 태그를 분포시키고, 각 개체 태그와의 거리가 일정한 개체값을 분포시켜 각 말뭉치에 대해 시각화하는 시각화 모듈; 및

수집된 다수의 말뭉치 각각에 대한 시각화 결과로 시각화 학습모델을 구축하는 모델 구축모듈을 포함하는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템.

청구항 4

제3항에 있어서, 상기 텍스트 리더부는,

입력된 말뭉치에 대해 토큰으로 분리한 다음 분리된 토큰에 대한 임베딩값을 연산하고, 연산된 임베딩값에 대해 BERT 언어 모델을 토대로 학습 수행하여 완전 결합 계층의 출력값으로 문장 벡터를 출력하며, 상기 문장 벡터는 3차원 벡터값으로 도출되도록 구비되는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템.

청구항 5

제4항에 있어서, 상기 시맨틱 프레임 리더부는

입력된 말뭉치에 대해 의미들의 각 구조 정보를 생성하고, 생성된 구조 정보 중 문장 태그에 대해 BERT 언어 모델 기반으로 학습 수행하여 완전 결합 계층의 출력값으로 3차원 벡터 형태의 의미 벡터를 생성하도록 구비되는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템.

청구항 6

제5항에 있어서, 상기 학습부는,

입력된 말뭉치에 대한 시맨틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델을 토대로 학습 수행하여 다수의 유사 의미 벡터를 추정하는 유사 의미 벡터 추정모듈; 및

추정된 다수의 유사 의미 벡터와 상기 문장 벡터 간의 정해진 거리 함수를 토대로 문장 벡터와 다수의 유사 의미 벡터 중 문장 벡터와 가장 인접된 거리의 의미 벡터를 정답 의미 벡터로 출력하는 정답 의미 벡터 도출모듈을 포함하도록 구비되는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템.

청구항 7

제6항에 있어서, 상기 학습부는,

상기 임의의 문장 벡터 t 에 대해 정답 의미 벡터의 좌표값과 문장 벡터 간의 유클리디안 거리를 도출하고, 상기 의미 태그의 좌표 및 개체 태그의 좌표 중 하나를 원점으로 도출된 정답 의미 벡터의 좌표값과 문장 벡터 간의 코사인 함수의 거리를 도출하며, 도출된 유클리디안 거리 및 코사인 함수의 거리의 합으로 문장 벡터에 대한 손실 비용을 도출하는 제1 손실 비용 연산모듈;

상기 임의의 의미 벡터 s 에서 정답 의미 벡터의 좌표값과 유사 의미 벡터 간의 유클리디안 거리를 도출하고, 의미 태그의 좌표 및 개체 태그의 좌표 중 하나를 원점으로 도출된 정답 의미 벡터의 좌표값과 유사 의미 벡터 간의 코사인 함수의 거리를 도출하며, 도출된 유클리디안 거리 및 코사인 함수의 거리의 합으로 정답 의미 벡터에 대한 손실 비용을 도출하는 제2 손실 비용 연산모듈; 및

상기 문장 벡터의 손실 비용과 정답 의미 벡터의 손실 비용의 합으로 최종 손실 비용을 산출하고 산출된 최종 손실 비용으로 시각화 학습 모델의 학습 변수를 제어하는 학습 성능 제어모듈;을 포함하는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템.

청구항 8

다수의 수집된 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하고, 생성된 각 구조 정보에 대해 3차원 좌표 형태로 변환한 다음 3차원 좌표 형태의 각 구조 정보를 시각화하여 시각화 학습 모델을 구축하는 학습 모델 구축단계;

입력된 말뭉치에 대해 분리된 토큰 각각의 임베딩값으로 전처리한 다음 기 구축된 선학습 BERT(Bidirectional Encoder Representation from Transformers) 언어 모델을 토대로 학습 수행하여 문장 벡터를 도출하는 텍스트 리더단계;

입력된 말뭉치에 대해 의미들의 구조 정보를 생성하고 생성된 각 구조 정보 기반에 대해 BERT 언어 모델을 토대로 학습 수행하여 다수의 의미 벡터를 도출하는 시맨틱 프레임 리더단계; 및

상기 시맨틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델 기반의 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 유사 의미 벡터와 문장 벡터 간의 정해진 거리함수를 토대로 정답 의미 벡터를 도출하는 학습단계를 포함하는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 방법.

청구항 9

제8항에 있어서, 상기 학습 모델 구축 단계는

수집된 다수의 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하는 단계;

생성된 구조 정보 중 의미 태그와의 거리가 일정한 다수의 개체 태그를 분포시키고, 각 개체 태그와의 거리가 일정한 개체값을 분포시킴에 따라 수집된 다수의 말뭉치 각각에 대한 시각화를 수행하는 단계; 및

수집된 다수의 말뭉치 각각에 대한 시각화 결과로 시각화 학습모델을 구축하는 단계를 포함하는 것을 특징으로 하는 딥러닝 학습 프레임워크 기반의 자연어 처리 방법.

청구항 10

제8항 또는 제9항의 딥러닝 학습 프레임워크 기반의 자연어 처리 방법을 실행하기 위한 프로그램이 기록되고 컴퓨터에서 판단 가능한 기록매체.

발명의 설명

기술 분야

- [0001] 본 발명은 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템 및 방법에 관한 것으로서, 더욱 상세하게는 의미들의 의미 태그, 개체 태그, 및 개체값 중 적어도 하나를 포함하는 구조 정보를 저차원의 시각화 학습 모델을 구축함에 따라, 시각화 학습 모델을 통해 회귀 기반 학습으로 입력된 말뭉치에 대한 학습을 수행할 수 있도록 한 기술에 관한 것이다.

배경 기술

- [0002] 자연어 이해(Natural Language Understanding)란 인간이 발화하는 언어 현상을 기계적으로 분석해 컴퓨터가 이해할 수 있는 형태로 구조화하는 것을 말한다. 자연어 이해를 위해, 추출된 정보, 문서 요약 등을 기계가 이해할 수 있는 의미들이 모델링되어야 한다.
- [0003] 선행문헌 1 및 2는 수집된 자연어 문장과 이에 따른 의미들을 하나의 벡터 공간 상에서 연관성을 기반으로 의미 벡터에 대한 학습을 수행한다. 그러나, 학습 결과를 도출된 의미 벡터값은 질의자의 의도와 상이할 수 있고 학습을 위한 모델링에서 분포 나 위치 등을 제어할 수 없다.
- [0004] 또한, 최근에는 학습 위한 의미들의 모델링의 부하 용량을 감소하여 경량의 디바이스에 적용하기 위해 차원 축소 기술을 사용하여 저차원의 시각화를 시도하였다. 그러나, 저차원 시각화 알고리즘 수행 시 정보 손실이 발생하고 이에 학습 성능이 저하되는 한계에 도달하였다.
- [0005] 이에 본 출원인은 저차원 시각화 학습 모델을 통해 전처리된 수집된 자연어 문장의 의도태크, 객체태그, 및 객체값의 구조 정보에 대해 학습 수행함에 따라 정보 손실 없이 도출된 정답 의미 벡터의 정확도를 향상시킬 수 있는 방안을 제안하고자 한다.

선행기술문헌

비특허문헌

- [0006] (비특허문헌 0001) 1. 김현지, 정상근. "선학습 BERT를 활용한 의미벡터 학습 방법." 한국정보과학회 학술발표논문집 (2019): 1427-1429. 2019
- (비특허문헌 0002) 2. Jung, Sangkeun. "Semantic vector learning for natural language understanding." Computer Speech & Language 56 (2019): 130-145.

발명의 내용

해결하려는 과제

- [0007] 따라서, 본 발명은 학습된 정답 의미 벡터에 대한 학습 성능에 대한 정량화가 가능하고 저차원의 시각화 학습 모델의 훈련의 효율성을 향상시킬 수 있는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템 및 방법을 제공하고자 함에 있다.
- [0008] 또한 본 발명은 수집된 말뭉치에 대한 시각화 학습 모델을 통해 학습 결과를 직관적 분석이 가능하고, 학습 결과에 따라 훈련에 필요한 학습 변수를 제어함에 따라 학습 성능을 더욱 향상시킬 수 있는 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템 및 방법을 제공하고자 한다.
- [0009] 본 발명의 목적은 이상에서 언급한 목적으로 제한되지 않으며, 언급되지 않은 본 발명의 다른 목적 및 장점들은 하기의 설명에 의해서 이해될 수 있으며, 본 발명의 실시예에 의해 보다 분명하게 알게 될 것이다. 또한, 본 발명의 목적 및 장점들은 특허청구 범위에 나타난 수단 및 그 조합에 의해 실현될 수 있음을 쉽게 알 수 있을 것이다.

과제의 해결 수단

- [0010] 본 발명의 일 실시예에 따른 딥러닝 학습 프레임워크 기반의 자연어 처리 시스템은,
- [0011] 다수의 수집된 말뭉치 각각에 대해 의미들의 구조 정보를 생성하고, 생성된 각 구조 정보에 대해 3차원 좌표 형태로 변환한 다음 3차원 좌표 형태의 각 구조 정보를 시각화하여 시각화 학습 모델을 구축하는 학습 모델 구축부;
- [0012] 입력된 말뭉치에 대해 분리된 토큰 각각의 임베딩값으로 전처리한 다음 기 구축된 선학습 BERT(Bidirectional Encoder Representation from Transformers) 언어 모델을 토대로 학습 수행하여 문장 벡터를 도출하는 텍스트 리더부;
- [0013] 입력된 말뭉치에 대해 의미들의 구조 정보를 생성하고 생성된 각 구조 정보 기반에 대해 BERT 언어 모델을 토대로 학습 수행하여 다수의 의미 벡터를 도출하는 시맨틱 프레임 리더부; 및
- [0014] 상기 시맨틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델 기반 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 유사 의미 벡터와 문장 벡터 간의 정해진 거리함수를 토대로 정답 의미 벡터를 도출하는 학습부를 포함하는 것을 일 특징으로 한다.
- [0015] 바람직하게 상기 구조 정보는,
- [0016] 문장의 주제인 의미 태그, 문장 요소의 이름인 개체 태그, 및 개체 태그의 실제값인 개체값 중 적어도 하나일 수 있다.
- [0017] 바람직하게 상기 학습 모델 구축부는,
- [0018] 수집된 다수의 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하는 구조 정보 생성모듈;
- [0019] 생성된 구조 정보 중 의미 태그와의 거리가 일정한 다수의 개체 태그를 분포시키고, 각 개체 태그와의 거리가 일정한 개체값을 분포시켜 각 말뭉치에 대해 시각화하는 시각화 모듈; 및
- [0020] 수집된 다수의 말뭉치 각각에 대한 시각화 결과로 시각화 학습모델을 구축하는 모델 구축모듈을 포함할 수 있다.
- [0021] 바람직하게 상기 텍스트 리더부는,
- [0022] 입력된 말뭉치에 대해 토큰으로 분리한 다음 분리된 토큰에 대한 임베딩값을 연산하고, 연산된 임베딩값에 대해 BERT 언어 모델을 토대로 학습 수행하여 완전 결합 계층의 출력값으로 문장 벡터를 출력하며, 상기 문장 벡터는 3차원 벡터값으로 도출되도록 구비될 수 있다.
- [0023] 바람직하게 상기 시맨틱 프레임 리더부는
- [0024] 입력된 말뭉치에 대해 의미들의 각 구조 정보를 생성하고, 생성된 구조 정보 중 문장 태그에 대해 BERT 언어 모델 기반으로 학습 수행하여 완전 결합 계층의 출력값으로 3차원 의미 벡터값을 생성하도록 구비될 수 있다.
- [0025] 바람직하게 상기 학습부는,
- [0026] 입력된 말뭉치에 대한 시맨틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델을 토대로 학습 수행하여 다수의 유사 의미 벡터를 추정하는 유사 의미 벡터 추정모듈; 및
- [0027] 추정된 다수의 유사 의미 벡터와 상기 문장 벡터 간의 정해진 거리 함수를 토대로 문장 벡터와 다수의 유사 의미 벡터 중 문장 벡터와 가장 인접된 거리의 의미 벡터를 정답 의미 벡터로 출력하는 정답 의미 벡터 도출모듈을 포함하도록 구비될 수 있다.
- [0028] 바람직하게 상기 학습부는,
- [0029] 상기 임의의 문장 벡터 t에 대해 정답 의미 벡터의 좌표값과 문장 벡터 간의 유클리디안 거리를 도출하고, 상기 의미 태그의 좌표 및 개체 태그의 좌표 중 하나를 원점으로 도출된 정답 의미 벡터의 좌표값과 문장 벡터 간의 코사인 함수의 거리를 도출하며, 도출된 유클리디안 거리 및 코사인 함수의 거리의 합으로 문장 벡터에 대한 손실 비용을 도출하는 제1 손실 비용 연산모듈;
- [0030] 상기 임의의 의미 벡터 s에서 정답 의미 벡터의 좌표값과 유사 의미 벡터 간의 유클리디안 거리를 도출하고, 의

미 태그의 좌표 및 개체 태그의 좌표 중 하나를 원점으로 도출된 정답 의미 벡터의 좌표값과 유사 의미 벡터 간의 코사인 함수의 거리를 도출하며, 도출된 유클리디안 거리 및 코사인 함수의 거리의 합으로 정답 의미 벡터에 대한 손실 비용을 도출하는 제2 손실 비용 연산모듈; 및

- [0031] 상기 문장 벡터의 손실 비용과 정답 의미 벡터의 손실 비용의 합으로 최종 손실 비용을 산출하고 산출된 최종 손실 비용으로 시각화 학습 모델의 학습 변수를 제어하는 학습 성능 제어모듈;을 포함할 수 있다.
- [0032] 본 발명의 다른 실시 양태에 의거한, 딥러닝 학습 프레임워크 기반의 자연어 처리 방법은
- [0033] 다수의 수집된 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하고, 생성된 각 구조 정보에 대해 3차원 좌표 형태로 변환한 다음 3차원 좌표 형태의 각 구조 정보를 시각화하여 시각화 학습 모델을 구축하는 학습 모델 구축단계;
- [0034] 입력된 말뭉치에 대해 분리된 토큰 각각의 임베딩값으로 전처리한 다음 기 구축된 선학습 BERT(Bidirectional Encoder Representation from Transformers) 언어 모델을 토대로 학습 수행하여 문장 벡터를 도출하는 텍스트 리더단계;
- [0035] 입력된 말뭉치에 대해 의미들의 구조 정보를 생성하고 생성된 각 구조 정보 기반에 대해 BERT 언어 모델을 토대로 학습 수행하여 다수의 의미 벡터를 도출하는 시멘틱 프레임 리더단계; 및
- [0036] 상기 시멘틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델 기반의 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 유사 의미 벡터와 문장 벡터 간의 정해진 거리함수를 토대로 정답 의미 벡터를 도출하는 학습단계를 포함하는 것을 일 특징으로 한다.
- [0037] 바람직하게 상기 학습 모델 구축 단계는
- [0038] 수집된 다수의 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하는 단계;
- [0039] 생성된 구조 정보 중 의미 태그와의 거리가 일정한 다수의 개체 태그를 분포시키고, 각 개체 태그와의 거리가 일정한 개체값을 분포시킴에 따라 수집된 다수의 말뭉치 각각에 대한 시각화를 수행하는 단계; 및
- [0040] 수집된 다수의 말뭉치 각각에 대한 시각화 결과로 시각화 학습모델을 구축하는 단계를 포함할 수 있다.

발명의 효과

- [0041] 일 실시 예에 따르면, 의미들의 의미 태그, 개체 태그, 및 개체값 중 적어도 하나를 포함하는 구조 정보의 손실 없이 저차원의 시각화 학습 모델을 구축하고 구축된 시각화 학습 모델을 통해 회귀 기반 학습으로 입력된 말뭉치에 대한 학습을 수행함에 따라 학습된 의미 벡터와 정답 의미 벡터에 대한 정량화가 가능하고 저차원의 시각화 학습 모델의 훈련의 효율성을 향상시킬 수 있다.
- [0042] 또한, 수집된 말뭉치에 대한 시각화 학습 모델을 통해 학습 결과를 직관적 분석이 가능하고, 학습 결과에 따라 훈련에 필요한 학습 변수를 제어함에 따라 학습 성능을 더욱 향상시킬 수 있다.

도면의 간단한 설명

- [0043] 본 명세서에서 첨부되는 다음의 도면들은 본 발명의 바람직한 실시 예를 예시하는 것이며, 후술하는 발명의 상세한 설명과 함께 본 발명의 기술사상을 더욱 이해시키는 역할을 하는 것이므로, 본 발명은 그러한 도면에 기재된 사항에만 한정되어 해석되어서는 아니된다.

도 1은 일 실시예의 자연어 처리 시스템의 구성도이다.

도 2는 일 실시예의 시스템의 학습 모델 구축부의 세부 구성도이다.

도 3은 일 실시예의 구조 정보를 보인 예시도이다.

도 4는 일 실시예의 시각화 학습 모델을 보인 예시도이다.

도 5는 일 실시예의 시스템의 학습부의 세부 구성도이다.

도 6은 일 실시예의 입력된 말뭉치의 학습 성능을 보인 예시도들이다.

발명을 실시하기 위한 구체적인 내용

- [0044] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 실시될 수 있다. 따라서, 실시예들은 특정한 개시형태로 한정되는 것이 아니며, 본 명세서의 범위는 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.
- [0045] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.
- [0046] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.
- [0047] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0048] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술분야에서 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.
- [0049] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시 예를 설명함으로써, 본 발명을 상세히 설명한다.
- [0050] 도 1은 일 실시예의 딥러닝 학습 프레임워크를 이용한 자연어 처리 시스템의 구성도이고, 도 2는 도 1에 도시된 학습 모델 구축부의 세부 구성도이며, 도 3은 일 실시예에 적용되는 구조 정보를 보인 예시도이고, 도 4는 도 3의 구조 정보의 시각화 학습 모델을 보인 예시도이며, 도 5는 도 1에 도시된 학습부의 세부 구성도이다.
- [0051] 도 1 내지 도 5를 참조하면, 일 실시 예에 따른 딥러닝 학습 프레임워크를 이용한 자연어 처리 시스템은 수집된 다수의 각 말뭉치에 대해 통계학적 분석 결과를 토대로 3차원 벡터 형태의 의미 벡터를 생성하고 생성된 각 말뭉치에 대한 의미 벡터를 모델링하여 시각화 학습 모델을 구축하며, 입력된 말뭉치에 대해 문장 벡터 및 다수의 의미 벡터에 대해 시각화 학습 모델에 의거 다수의 유사 의미 벡터를 도출하고 도출된 다수의 유사 의미 벡터와 문장 벡터 간의 거리 함수를 토대로 정답 의미 벡터를 도출하는 구성을 갖추며, 이에 시스템은, 학습 모델 구축부(100), 텍스트 리더부(200), 시맨틱 프레임 리더부(300), 및 학습부(400)를 포함할 수 있다.
- [0052] 여기서, 학습 모델 구축부(100)는 다수의 수집된 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하고, 생성된 각 구조 정보에 대해 3차원 좌표 형태로 변환한 다음 3차원 좌표 형태의 각 구조 정보를 시각화하여 시각화 학습 모델을 구축하며, 도 2를 참조하면, 학습 모델 구축부(100)는 구조 정보 생성모듈(110), 시각화 모듈(120), 및 모델 구축모듈(130)로 구비된다.
- [0053] 구조 정보 생성모듈(110)은 수집된 각각의 물뭉치를 분리하여 의미들의 구조 정보를 생성하며 구조 정보는 의미 태그, 개체 태그, 및 개체값 중 적어도 하나이다. 예를 들어 “오늘 비 와?”에 대한 말뭉치에 대한 의미들의 문장은 “weather.rain(day.p=오늘)” 이고, 의미 태그는 'weather.rain'이며, 개체 태그는 'day.p'이고, 개체값은 '오늘' 이다. 즉, 의미 태그는 예시된 바와 같이, 입력된 말뭉치의 주제이고, 개체 태그는 문장 요소의 이름이며, 개체값은 개체 태그의 실제값이다.
- [0054] 그리고 시각화 모듈(120)은 수집된 각 말뭉치 각각의 의미 태그, 개체 태그, 및 개체값에 대한 통계학적 분석에 의거 3차원 벡터 형태로 시각화할 수 있다. 이때 의미 태그 중심으로 일정한 거리에 개체 태그를 배열하고 각 개체 태그를 중심으로 일정한 거리에 개체값을 배열하여 3차원 벡터 형태로 의미 태그, 개체 태그, 및 개체값을 시각화한다. 일례로, 도 3을 참조하면, “weather.rain”의 의미 태그를 포함하는 다수의 개체 태그는 의미 태그를 중심으로 일정한 거리 및 각도로 모두 배열되고, 각 개체 태그를 포함하는 다수의 개체값은 각 개체 태그를 중심으로 일정한 거리 및 일정 각도로 모두 배열된다. “weather” 말뭉치에 대한 의도 태그는 시각화되며, 이때 일정한 거리는 개체 태그의 수로 정해진다.
- [0055] 그리고 모델 구축모듈(130)은 시각화된 3차원 벡터 형태로 의미 태그, 개체 태그, 및 개체값 각각에 대해 좌표값으로 시각화 학습 모델을 구축한다. 즉, 도 4를 참조하면, “weather” 의미 태그에 대해 (x, y, z) 좌표값은 (30.62, -49.06, 484.45)이며, 시각화된 각 의미 태그의 좌표값으로 시각화 학습 모델이 구축된다.

- [0056] 한편, 텍스트 리더부(200)는 입력된 말뭉치에 대해 토큰으로 분리한 다음 분리된 토큰에 대한 임베딩값을 연산하고, 연산된 임베딩값에 대해 BERT(Bidirectional Encoder Representations from Transformation) 언어 모델을 토대로 학습 수행하여 완전 결합 계층의 출력값으로 문장 벡터를 출력하며, 상기 문장 벡터는 3차원 벡터값으로 도출한다. 도출된 문장 벡터는 학습부(400)로 전달된다.
- [0057] 여기서, 입력된 말뭉치에 대해 BERT 언어 모델을 통해 문장 벡터를 도출하는 일련의 과정은 일반적인 말뭉치에 대한 문장 벡터를 도출하는 과정과 동일 또는 유사하므로 이에 대한 구체적인 설명은 생략한다. 이때 선학습 BERT 언어 모델은 한국어 말뭉치의 경우 BERT multilingual 선학습 모델이고, 영어 말뭉치의 경우 BERT base uncased 선학습 모델이다.
- [0058] 그리고, 시맨틱 프레임 리더부(300)는 입력된 말뭉치에 대해 의미들의 각 구조 정보를 생성하고, 생성된 구조 정보 중 의미 태그에 대해 BERT 언어 모델 기반으로 학습 수행하여 완전 결합 계층의 출력값으로 3차원 의미 벡터값을 생성한다. 이에 다수의 의미 벡터는 학습부(400)로 전달된다. 여기서, 입력된 말뭉치에 대한 구조 정보를 생성하는 일련의 과정은 전술한 바와 같고, 생성된 구조 정보 중 의미 태그에 대한 학습 수행 결과로 다수의 의미 벡터를 도출하는 일련의 과정은 일반적인 의미 태그에 대해 BERT 언어 모델 기반으로 학습 수행하여 완전 결합 계층의 출력값으로 3차원 의미 벡터값을 생성하는 과정과 동일 또는 유사하다. 이때 선학습 BERT 언어 모델은 BERT multilingual 선학습 모델이다.
- [0059] 한편 학습부(400)는 시맨틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델 기반 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 유사 의미 벡터와 문장 벡터 간의 정해진 거리함수를 토대로 정답 의미 벡터를 도출하도록 구비되며, 이에 학습부(400)는 도 4에 도시된 바와 같이, 유사 의미 벡터 추정모듈(410) 및 정답 의미 벡터 도출모듈(420)을 포함할 수 있다.
- [0060] 유사 의미 벡터 추정모듈(410)은 입력된 말뭉치에 대한 시맨틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델을 토대로 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 다수의 유사 의미 벡터는 정답 의미 벡터 도출모듈(420)로 전달된다.
- [0061] 정답 의미 벡터 도출모듈(420)은 추정된 각 유사 의미 벡터와 문장 벡터 간에 정의된 거리 함수로 정답 의미 벡터를 도출한다. 일례로 유사 의미 벡터와 문장 벡터 간에 정의된 거리는 코사인 함수로 정해진다. 정답 의미 벡터는 유사 의미 벡터와 문장 벡터 간의 거리가 가장 작은 유사 의미 벡터로 도출된다.
- [0062] 이러한 정답 의미 벡터를 반영하여 입력된 말뭉치에 대한 응답 말뭉치를 생성함에 따라 자연스러운 대화가 유지될 수 있다.
- [0063] 이에 일 실시예는 구조 정보의 손실없이 저차원 시각화 학습 모델을 구축함에 따라 학습된 의미 벡터와 정답 의미 벡터에 대한 정량화가 가능하고 저차원의 시각화 학습 모델의 훈련의 효율성을 향상시킬 수 있다.
- [0064] 한편 학습부(400)는 학습 결과로 도출된 정답 의미 벡터와 실측된 정답 의미 벡터 간의 손실 비용을 토대로 시각화 학습 모델에 대한 회귀 기반 학습 성능을 제어하는 구성을 갖추며, 이에 학습부(400)는 제1 손실비용 연산모듈(430), 제2 손실비용 연산모듈(440), 및 학습 성능 제어모듈(450)을 포함할 수 있다.
- [0065] 여기서, 제1 손실비용 연산모듈(430)은 임의의 문장 벡터 t 에 대해 정답 의미 벡터의 좌표값과 문장 벡터 v_t 간의 유클리디안 거리 L_{euc}^t 를 도출하고, 상기 의미 태그의 좌표 및 개체 태그의 좌표 중 하나를 원점으로 도출된 정답 의미 벡터 v_s 의 좌표값과 문장 벡터 간의 코사인 함수의 거리 L_{cos}^t 를 도출하며, 도출된 유클리디안 거리 L_{euc}^t 및 코사인 함수의 거리 L_{cos}^t 의 합으로 임의의 문장 벡터 t 에 대한 손실 비용 L^t 을 도출한다.
- [0066] 그리고, 제2 손실비용 연산모듈(440)은 임의의 의미 벡터 s 에서 정답 의미 벡터 v_s 의 좌표값과 유사 의미 벡터 간의 유클리디안 거리 L_{euc}^s 를 도출하고, 의미 태그의 좌표 및 개체 태그의 좌표 중 하나를 원점으로 도출된 정답 의미 벡터의 좌표값과 유사 의미 벡터 간의 코사인 함수의 거리 L_{cos}^s 를 도출하며, 도출된 유클리디안 거리 L_{euc}^s 및 코사인 함수의 거리 L_{cos}^s 의 합으로 정답 의미 벡터에 대한 손실 비용 L^s 을 도출한다.
- [0067] 그리고 학습 성능 제어모듈(450)은 문장 벡터 t 에 대한 손실 비용 L^t 와 의미 벡터 s 에 대한 손실 비용 L^s 에 대한 합으로 총 손실 비용 L 을 도출하고 도출된 총 손실 비용 L 을 토대로 회귀 기반 학습 변수, 예를 들어 시그모

이드(sigmoid) 함수를 이용한 활성 함수의 종속 변수, 배치(batch) 사이즈, 및 데이터 에포치(epoch) 등을 조절됨에 따라 학습 성능이 향상된다.

[0068] 이에 일 실시예는, 수집된 말뭉치에 대한 시각화 학습 모델을 통해 학습 결과를 손실 비용을 토대로 직관적 분석이 가능하고, 학습 결과에 따라 훈련에 필요한 학습 변수를 제어함에 따라 학습 성능을 더욱 향상시킬 수 있다.

[0069] 한국어 말뭉치의 일반적인 언어 체계인 weather, 네비게이션 언어 체계인 Navi, 영어 말뭉치의 항공 언어 체계인 ATIS, 일반적인 언어 체계인 snips 각각에 “오늘 비와?”의 말뭉치가 입력되었다고 가정하자.

[0070] 입력된 “오늘 비와?”의 말뭉치에 대한 구조 정보 각각에 대한 통계학적 분석 결과는 다음 표 1에 도시된 바와 같다.

[0071] [표 1]

Domain	Weather	Navi	ATIS	Snips
# of train	6993	5601	4478	12658
# of test	2998	2402	882	685
# of 의도태그	19	8	21	7
# of 개체태그	14	4	79	39
# of 개체태그 조합	80	6	765	549

[0072]

[0073] 표 1을 참조하면 “weather”의 의도 태그와 관련된 통계학적 분석 결과는 각 언어 체계 별로 19, 14, 21, 7 개가 검색됨을 확인할 수 있고, 각 개체태그와 관련된 통계학적 분석 결과는 각 언어 체계 별로 14, 4, 79, 39 개임을 알 수 있다.

[0074] 이러한 각 구조 정보에 대한 통계학적 분석 결과를 토대로 각 의미 태그를 중심으로 개체 태그 및 개체값에 대해 3차원 벡터 형태로 시각화를 수행하여 시각화 학습 모델이 구축된다.

[0075] 구축된 시각화 학습 모델을 토대로 입력된 말뭉치에 대해 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 유사 의미 벡터의 좌표와 가장 인접된 정답 의미 좌표 k 개의 의도 태그와 실측된 정답 의도 태그와의 일치율은 하기 표 2에 도시된 바와 같다.

[0076] [표 2]

	Weather			Navi			ATIS			Snips		
	J[4]	K[1]	Text	J[4]	K[1]	Text	J[4]	K[1]	Text	J[4]	K[1]	Text
k=1	0.987	0.991	0.961	0.992	0.997	0.997	0.941	0.980	0.918	0.972	0.981	0.988
k=5	0.985	0.998	0.959	0.995	0.988	0.997	0.907	0.980	0.917	0.973	0.981	0.988
k=10	0.984	0.998	0.954	0.995	0.988	0.997	0.935	0.980	0.904	0.972	0.981	0.988

[0077]

[0078] 여기서, k는 정답 의미 태그의 수이다. 표 2를 참조하면, ATIS 및 weather 언어체계에서는 학습 성능이 상대적으로 낮고 Navi와 Snips의 언어 체계에서는 학습 성능이 상대적으로 향상됨을 알 수 있다.

[0079] 도 6은 일 실시예에 의거 구축된 시각화 학습 모델과 각 언어 체계 별 학습 결과에 대한 그래프를 보인 예시도들로서, 도 7을 참조하면, 각 언어 체계 별로 도출된 정답 의미 벡터의 좌표값과 실측된 정답 의미 벡터 간의 거리 연산의 결과가 0.15 미만으로 예측된 데이터는 파란색 점으로 나타내고, 각 언어 체계 별로 도출된 정답 의미 벡터의 좌표값과 실측된 정답 의미 벡터 간의 거리 연산의 결과가 0.15 이상으로 예측된 데이터는 붉은 색으로 표시된다. 이에 snips 및 Navi는 파란색 점 근처에 밀집됨을 확인할 수 있다.

[0080] 본 발명의 다른 실시 양태에 의거한, 딥러닝 학습 프레임워크 기반의 자연어 처리 방법은 다수의 수집된 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하고, 생성된 각 구조 정보에 대해 3차원 좌표 형태로 변환한 다음 3차원 좌표 형태의 각 구조 정보를 시각화하여 시각화 학습 모델을 구축하는 학습 모델 구축단계; 입력된 말뭉치에 대해 분리된 토큰 각각의 임베딩값으로 전처리한 다음 기 구축된 선학습 BERT(Bidirectional Encoder Representation from Transformers) 언어 모델을 토대로 학습 수행하여 문장 벡터를 도출하는 텍스트 리더단계; 입력된 말뭉치에 대해 의미들의 구조 정보를 생성하고 생성된 각 구조 정보 기반에 대해 BERT 언어 모델을 토대로 학습 수행하여 다수의 의미 벡터를 도출하는 시멘틱 프레임 리더단계; 및 상기 시멘틱 프레임 리더부의 다수의 의미 벡터에 대해 시각화 학습 모델 기반의 학습 수행하여 다수의 유사 의미 벡터를 추정하고 추정된 유사 의미 벡터와 문장 벡터 간의 정해진 거리함수를 토대로 정답 의미 벡터를 도출하는 학습단계를 포함하고, 여기서, 상기 학습 모델 구축 단계는 수집된 다수의 말뭉치 각각에 대해 의미들의 각 구조 정보를 생성하는 단계; 생성된 구조 정보 중 의미 태그와의 거리가 일정한 다수의 개체 태그를 분포시키고, 각 개체 태그와의 거리가

일정한 개체값을 분포시킴에 따라 수집된 다수의 말뭉치 각각에 대한 시각화를 수행하는 단계; 및 수집된 다수의 말뭉치 각각에 대한 시각화 결과로 시각화 학습모델을 구축하는 단계를 포함할 수 있고, 상기의 딥러닝 학습 프레임워크를 이용하여 자연어를 처리하는 각 단계는, 전술한 학습 모델 구축부(100), 텍스트 리더부(200), 시맨틱 프레임 리더부(300), 및 학습부(400)에서 수행되는 기능으로 자세한 원용은 생략한다.

[0081] 이에 일 실시예는 의미들의 의미 태그, 개체 태그, 및 개체값 중 적어도 하나를 포함하는 구조 정보의 손실없이 저차원의 시각화 학습 모델을 구축하고 구축된 시각화 학습 모델을 통해 회귀 기반 학습으로 입력된 말뭉치에 대한 학습을 수행함에 따라 학습된 의미 벡터와 정답 의미 벡터에 대한 정량화가 가능하고 저차원의 시각화 학습 모델의 훈련의 효율성을 향상시킬 수 있다.

[0082] 또한, 일 실시예는 수집된 말뭉치에 대한 시각화 학습 모델을 통해 학습 결과를 직관적 분석이 가능하고, 학습 결과에 따라 훈련에 필요한 학습 변수를 제어함에 따라 학습 성능을 더욱 향상시킬 수 있다.

[0083] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 컨트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate 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(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 기술분야에서 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 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0084] 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, 구성요소(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.

[0085] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 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 광매체(magneto-optical media), 및 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.

[0086] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

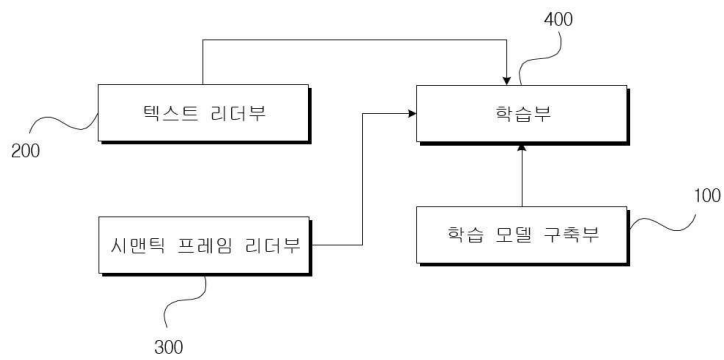
부호의 설명

[0087]

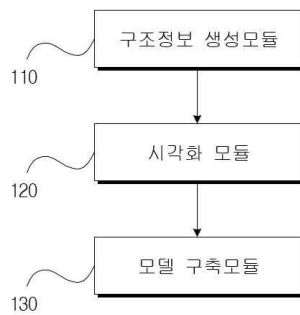
- 100 : 학습 모델 구축부
- 110 : 구조 정보 생성모듈
- 120 : 시각화 모듈
- 130 : 모델 구축모듈
- 200 : 텍스트 리더부
- 300 : 시맨틱 프레임 리더부
- 400 : 학습부

도면

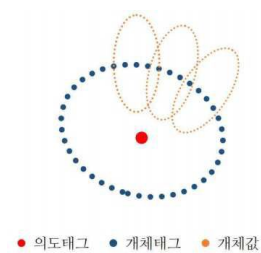
도면1



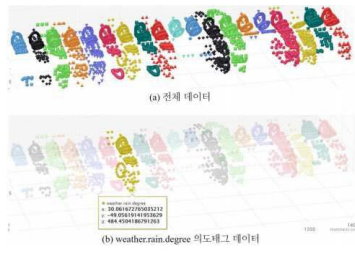
도면2



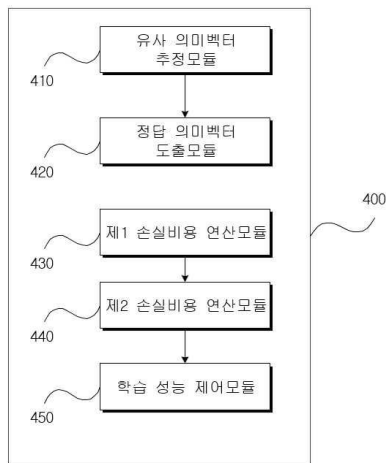
도면3



도면4



도면5



도면6

