



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년11월15일
(11) 등록번호 10-1797365
(24) 등록일자 2017년11월07일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) G06F 15/18 (2017.01)
G06N 3/08 (2006.01)
(52) CPC특허분류
G06F 17/2785 (2013.01)
G06F 15/18 (2013.01)
(21) 출원번호 10-2016-0074758
(22) 출원일자 2016년06월15일
심사청구일자 2016년06월15일
(56) 선행기술조사문헌
KR1020030002930 A
(뒷면에 계속) 기술이전 희망 : 기술양도

(73) 특허권자
울산대학교 산학협력단
울산광역시 남구 대학로 93(무거동)
(72) 발명자
옥철영
울산광역시 남구 신북로35번길 6, 101동 602호 (무거동, 쌍용스윗닷홈)
이주상
울산광역시 중구 옥교3길 120, B동 301호 (학성동, 옥성맨션)
(74) 대리인
김종선, 이형석

전체 청구항 수 : 총 16 항

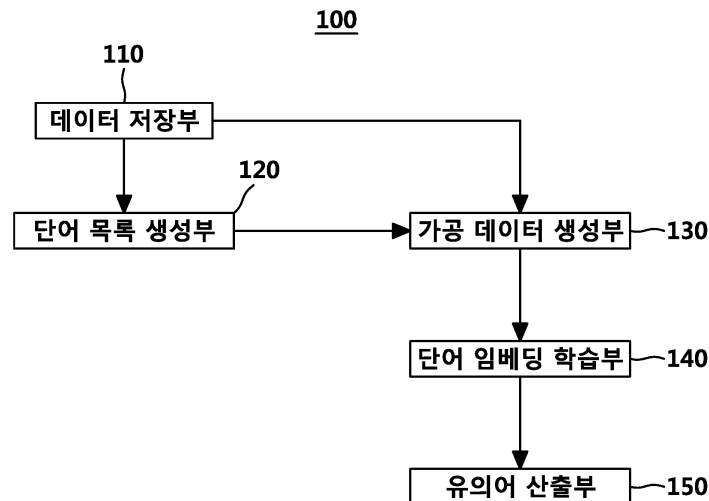
심사관 : 경연정

(54) 발명의 명칭 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법

(57) 요약

본 발명은 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법에 관한 것으로서, 더욱 상세하게는 단어 사전 데이터 및 단어 의미 데이터가 포함된 어휘 의미망을 저장하는 데이터 저장부; 상기 저장된 단어 사전 데이터로부터 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성하는 단어 목록 생성부; 상기 생성된 학습할 단어 목록과 상기 학습할 단어 목록에 포함된 학습할 단어의 단어 의미 데이터를 상기 데이터 저장부로부터 가져와서 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 가공 데이터 생성부; 및 상기 생성된 가공 데이터를 가지고 입출력층 및 프로젝션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단어 임베딩 학습부를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06F 17/2795 (2013.01)

G06N 3/08 (2013.01)

(56) 선행기술조사문헌

US20090276414 A1*

KR101508070 B1*

KR1020110017129 A

'한국어 어휘 의미망에 기반을 둔 비감독 어의 중의성 해소', 부산대학교 석사논문, 김민호 (2009.2)

나승훈 외, 심층학습에 기반한 자연언어처리, 정보과학회지, 33(10), pp.28-39 (2015.10)

*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호 1711035223

부처명 미래창조과학부

연구관리전문기관 정보통신기술진흥센터

연구사업명 SW컴퓨팅산업원천기술개발

연구과제명 (엑소브레인-3세부) Symbolic Approach 기반 인간모사형 자가학습 지능 원천 기술 개발

기 여 율 1/1

주관기관 한국과학기술원

연구기간 2016.03.01 ~ 2017.02.28

명세서

청구범위

청구항 1

단어 사전 데이터 및 단어 의미 데이터가 포함된 어휘 의미망을 저장하는 데이터 저장부;

상기 저장된 단어 사전 데이터로부터 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성하는 단어 목록 생성부;

상기 생성된 학습할 단어 목록과 상기 학습할 단어 목록에 포함된 학습할 단어의 단어 의미 데이터를 상기 데이터 저장부로부터 가져와서 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 가공 데이터 생성부; 및

상기 생성된 가공 데이터를 가지고 입출력층 및 프로젝션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단어 임베딩 학습부를 포함하고,

상기 단어 임베딩 학습부는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로퍼게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행하고, 백 프로퍼게이션 과정에서 학습할 단어의 단어 의미 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 2

제1항에 있어서,

상기 생성된 단어 벡터를 이용한 코사인 유사도를 계산해 입력 단어의 유의어를 산출하는 유의어 산출부를 더 포함하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 3

제2항에 있어서,

상기 유의어 산출부는

제1 및 제2 입력 단어를 입력받고, 제2 입력 단어에 가장 근접한 단어를 코사인 유사도를 이용하여 기설정된 개수의 유의어들을 추출하고, 상기 추출된 유의어들과 제1 입력 단어 간의 코사인 유사도를 계산하여 가장 높은 유사도를 가진 단어를 제1 및 제2 입력 단어의 유의어로 산출하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 4

제2항에 있어서,

상기 유의어 산출부는

입력 단어의 단어 벡터와 기설정된 속성을 가지는 속성 단어의 단어 벡터의 속성 연산 과정을 통해 입력 단어의 속성이 변경된 유의어를 산출하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 5

제1항에 있어서,

상기 단어 목록 생성부는

명사, 용언, 부사 및 의존명사 어휘들이 의미제약으로 상호 연결된 어휘 의미망에서 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 6

제1항에 있어서,

상기 가공 데이터 생성부는

상기 생성된 학습할 단어 목록에 포함된 단어의 뜻풀이, 상위어 및 반의어 중 적어도 하나가 포함된 단어 의미 데이터를 이용하여 학습용 가공 데이터를 생성하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 7

제1항에 있어서,

상기 단어 임베딩 학습부는

단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 8

제1항에 있어서,

상기 단어 임베딩 학습부는

상기 가공 데이터 중에서 상기 학습할 단어 및 상기 학습할 단어의 단어 의미 데이터가 가공된 각각의 가공 데이터를 단어 임베딩에서의 스킵-그램(Skip-Gram)의 입력층 및 출력층이 합쳐진 하나의 입출력층에 위치시키고, 상기 각각의 가공 데이터가 위치한 입출력층 및 미러층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 어휘 의미망을 이용한 단어 의미 임베딩 장치.

청구항 9

삭제

청구항 10

단어 의미 임베딩 장치에 의해 수행되는 어휘 의미망을 이용한 단어 의미 임베딩 방법에 있어서,

어휘 의미망에 포함된 단어 사전 데이터로부터 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성하는 단계;

상기 생성된 학습할 단어 목록과 상기 학습할 단어 목록에 포함된 학습할 단어의 단어 의미 데이터를 가져와서 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 단계; 및

상기 생성된 가공 데이터를 가지고 입출력층 및 프로젝션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단계를 포함하고,

상기 단어 벡터를 생성하는 단계는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로퍼게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행하고, 백 프로퍼게이션 과정에서 학습할 단어의 단어 의미 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 11

제10항에 있어서,

상기 생성된 단어 벡터를 이용한 코사인 유사도를 계산해 입력 단어의 유의어를 산출하는 단계를 더 포함하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 12

제11항에 있어서,

상기 유의어를 산출하는 단계는

제1 및 제2 입력 단어를 입력받는 단계;

제2 입력 단어에 가장 근접한 단어를 코사인 유사도를 이용하여 기설정된 개수의 유의어들을 추출하는 단계; 및
상기 추출된 유의어들과 제1 입력 단어 간의 코사인 유사도를 계산하여 가장 높은 유사도를 가진 단어를 제1 및 제2 입력 단어의 유의어로 산출하는 단계
를 포함하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 13

제11항에 있어서,

상기 유의어를 산출하는 단계는

입력 단어의 단어 벡터와 기설정된 속성을 가지는 속성 단어의 단어 벡터의 속성 연산 과정을 통해 입력 단어의 속성이 변경된 유의어를 산출하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 14

제10항에 있어서,

상기 학습할 단어 목록을 생성하는 단계는

명사, 용언, 부사 및 의존명사 어휘들이 의미제약으로 상호 연결된 어휘 의미망에서 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 15

제10항에 있어서,

상기 가공 데이터를 생성하는 단계는

상기 생성된 학습할 단어 목록에 포함된 단어의 뜻풀이, 상위어 및 반의어 중 적어도 하나가 포함된 단어 의미 데이터를 이용하여 학습용 가공 데이터를 생성하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 16

제10항에 있어서,

상기 단어 벡터를 생성하는 단계는

단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 17

제10항에 있어서,

상기 단어 벡터를 생성하는 단계는

상기 가공 데이터 중에서 상기 학습할 단어 및 상기 학습할 단어의 단어 의미 데이터가 가공된 각각의 가공 데이터를 단어 임베딩에서의 스킵-그램(Skip-Gram)의 입력층 및 출력층이 합쳐진 하나의 입출력층에 위치시키고, 상기 각각의 가공 데이터가 위치한 입출력층 및 미러층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 어휘 의미망을 이용한 단어 의미 임베딩 방법.

청구항 18

삭제

발명의 설명

기술 분야

본 발명은 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법에 관한 것으로서, 더욱 상세하게는 어휘 의미망으로부터 학습할 단어 목록과 학습할 단어의 의미적 정보를 생성하고 그 생성된 학습할 단어 목록과 학습할

[0001]

단어의 의미적 정보를 이용한 단어 임베딩 학습을 통해 적은 단어 데이터로 단어 임베딩 학습이 가능한, 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법에 관한 것이다.

배경 기술

- [0002] 최근에 컴퓨터와 인간의 바둑 대결이 큰 화제가 되면서 일반인들 사이에서 기계 학습에 대한 관심이 증가하고 있다. 기계 학습 중에 딥 러닝(Deep learning)을 이용한 자연어 처리 시스템이 많이 등장하고 있으며, 관련 연구도 활발히 진행되고 있다.
- [0003] 기계 학습을 사용하는 자연어 처리 시스템에서는 문자로 표현된 단어를 숫자로 변환하는 방법이 중요하다. 단어의 표현 방식을 원-핫(One-hot) 방식을 이용하면 단어의 양에 따라 벡터 차원이 결정되기 때문에 고차원의 벡터로 표현되게 된다. 여기서, 원-핫 방식은 해당하는 단어를 1로 나머지를 0으로 표현하는 하는 방식을 나타낸다.
- [0004] 원-핫 방식을 이용하여 수십만 차원으로 단어를 표현하게 되면 딥 러닝을 이용한 자연어 처리에서 입력이 매우 커지는 문제가 발생한다. 이러한 원-핫 방식의 문제점을 해결할 방법으로 단어 임베딩(Word Embedding)이 등장하게 되었다.
- [0005] 단어 임베딩은 단어를 신경망(Neural Network)과 대량의 말뭉치를 이용해 저차원의 벡터로 표현하는 방법이다. 단어 임베딩은 자연어 처리를 딥 러닝에 사용하기 전 사전 학습(pre-training)으로 사용되고 있다. 또한, 단어 임베딩 자체를 이용한 여러 연구 결과들이 나오고 있다.
- [0006] 단어 임베딩은 초기에는 대량의 말뭉치를 이용해 문장에서 학습할 단어에 인접한 단어를 이용해 학습을 한다. 초기 단어 임베딩 방식은 말뭉치에서 인접 단어가 유사한 단어들끼리 벡터가 유사하도록 학습한다. 단어의 의미적인 접근이 아닌 단어의 위치적인 접근을 통해 학습하는 방식이다. 초기에는 기본적인 신경망(Neural Network)을 이용한 NNLM(Neural Network Language Model)이 단어 임베딩으로 사용했다. 이후, 이전 학습을 다음 학습에 사용하는 RNNLM(Recurrent Neural Network Language Model)과 학습 속도를 향상한 워드투벡(Word2Vec)이 등장하게 되었다.

선행기술문헌

특허문헌

- [0007] (특허문헌 0001) 대한민국 등록특허공보 제10-1358614호 (2014.01.27. 등록)

발명의 내용

해결하려는 과제

- [0008] 본 발명의 실시 예들은 어휘 의미망에서 학습할 단어 목록과 학습할 단어의 단어 의미 데이터(예컨대, 사전적인 뜻풀이와 상위어, 반의어 등)를 이용해 가공 데이터를 생성하고, 그 생성된 가공 데이터를 가지고 단어 의미 임베딩의 변형된 스킵-그램(Skip-Gram) 모델인 피쳐 미러(Feature Mirror) 모델과 네거티브-샘플링(Negative-Sampling)을 이용해 학습함으로써, 단어의 의미적 관계와 연관성을 벡터로 표현할 수 있는, 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법을 제공하고자 한다.
- [0009] 본 발명의 실시 예는 종래의 단어 임베딩이 아닌 의미적 정보를 이용한 단어 임베딩 학습을 수행하기 때문에 적은 학습 데이터로도 단어 임베딩 학습이 가능하며 학습 시간도 적게 걸릴 수 있는, 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법을 제공하고자 한다.
- [0010] 또한, 본 발명의 실시 예는 단순히 등장하는 단어 위치적 정보가 아닌 학습할 단어의 의미 기반으로 학습함으로써 저 빈도의 단어 학습에도 효율적이며, 사전에 등장한 단어를 대상으로 하기 때문에 종래의 단어 임베딩보다 많은 단어를 벡터로 표현할 수 있는, 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법을 제공하고자 한다.
- [0011] 또한, 본 발명의 실시 예는 종래의 위치 기반의 단어 임베딩이 아닌 단어 의미 데이터를 이용한 단어 의미 임베딩 학습을 통해 단어 벡터를 생성하기 때문에 단어들의 관계를 코사인 유사도를 통해 볼 수 있고, 종래의 단어 임베딩에서 얻을 수 없는 동의어 관계나 유의어 등을 찾을 수 있는, 어휘 의미망을 이용한 단어 의미 임베딩 장

치 및 방법을 제공하고자 한다.

- [0012] 또한, 본 발명의 실시 예는 학습할 단어 목록을 부사와 의존명사 등까지 확장하여 학습할 수 있으며, 단어의 의미 정보를 추가해 적용할 수 있는, 어휘 의미망을 이용한 단어 의미 임베딩 장치 및 방법을 제공하고자 한다.

과제의 해결 수단

- [0013] 본 발명의 제1 측면에 따르면, 단어 사전 데이터 및 단어 의미 데이터가 포함된 어휘 의미망을 저장하는 데이터 저장부; 상기 저장된 단어 사전 데이터로부터 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성하는 단어 목록 생성부; 상기 생성된 학습할 단어 목록과 상기 학습할 단어 목록에 포함된 학습할 단어의 단어 의미 데이터를 상기 데이터 저장부로부터 가져와서 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 가공 데이터 생성부; 및 상기 생성된 가공 데이터를 가지고 입출력층 및 프로젝션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단어 임베딩 학습부를 포함하는 어휘 의미망을 이용한 단어 의미 임베딩 장치가 제공될 수 있다.
- [0014] 상기 장치는, 상기 생성된 단어 벡터를 이용한 코사인 유사도를 계산해 입력 단어의 유의어를 산출하는 유의어 산출부를 더 포함할 수 있다.
- [0015] 상기 유의어 산출부는 제1 및 제2 입력 단어를 입력받고, 제2 입력 단어에 가장 근접한 단어를 코사인 유사도를 이용하여 기설정된 개수의 유의어들을 추출하고, 상기 추출된 유의어들과 제1 입력 단어 간의 코사인 유사도를 계산하여 가장 높은 유사도를 가진 단어를 제1 및 제2 입력 단어의 유의어로 산출할 수 있다.
- [0016] 상기 유의어 산출부는 입력 단어의 단어 벡터와 기설정된 속성을 가지는 속성 단어의 단어 벡터의 속성 연산 과정을 통해 입력 단어의 속성이 변경된 유의어를 산출할 수 있다.
- [0017] 상기 단어 목록 생성부는 명사, 용언, 부사 및 의존명사 어휘들이 의미제약으로 상호 연결된 어휘 의미망에서 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성할 수 있다.
- [0018] 상기 가공 데이터 생성부는 상기 생성된 학습할 단어 목록에 포함된 단어의 뜻풀이, 상위어 및 반의어 중 적어도 하나가 포함된 단어 의미 데이터를 이용하여 학습용 가공 데이터를 생성할 수 있다.
- [0019] 상기 단어 임베딩 학습부는 단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습할 수 있다.
- [0020] 상기 단어 임베딩 학습부는 상기 가공 데이터 중에서 상기 학습할 단어 및 상기 학습할 단어의 단어 의미 데이터가 가공된 각각의 가공 데이터를 단어 임베딩에서의 스킵-그램(Skip-Gram)의 입력층 및 출력층이 합쳐진 하나의 입출력층에 위치시키고, 상기 각각의 가공 데이터가 위치한 입출력층 및 미러층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성할 수 있다.
- [0021] 상기 단어 임베딩 학습부는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로파게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행하고, 백 프로파게이션 과정에서 학습할 단어의 단어 의미 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경할 수 있다.
- [0022] 한편, 본 발명의 제2 측면에 따르면, 어휘 의미망에 포함된 단어 사전 데이터로부터 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성하는 단계; 상기 생성된 학습할 단어 목록과 상기 학습할 단어 목록에 포함된 학습할 단어의 단어 의미 데이터를 가져와서 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 단계; 및 상기 생성된 가공 데이터를 가지고 입출력층 및 프로젝션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단계를 포함하는 어휘 의미망을 이용한 단어 의미 임베딩 방법이 제공될 수 있다.
- [0023] 상기 방법은, 상기 생성된 단어 벡터를 이용한 코사인 유사도를 계산해 입력 단어의 유의어를 산출하는 단계를 더 포함할 수 있다.
- [0024] 상기 유의어를 산출하는 단계는 제1 및 제2 입력 단어를 입력받는 단계; 제2 입력 단어에 가장 근접한 단어를 코사인 유사도를 이용하여 기설정된 개수의 유의어들을 추출하는 단계; 및 상기 추출된 유의어들과 제1 입력 단어 간의 코사인 유사도를 계산하여 가장 높은 유사도를 가진 단어를 제1 및 제2 입력 단어의 유의어로 산출하는 단계를 포함할 수 있다.
- [0025] 상기 유의어를 산출하는 단계는 입력 단어의 단어 벡터와 기설정된 속성을 가지는 속성 단어의 단어 벡터의 속

성 연산 과정을 통해 입력 단어의 속성이 변경된 유의어를 산출할 수 있다.

- [0026] 상기 학습할 단어 목록을 생성하는 단계는 명사, 용언, 부사 및 의존명사 어휘들이 의미제약으로 상호 연결된 어휘 의미망에서 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성할 수 있다.
- [0027] 상기 가공 데이터를 생성하는 단계는 상기 생성된 학습할 단어 목록에 포함된 단어의 뜻풀이, 상위어 및 반의어 중 적어도 하나가 포함된 단어 의미 데이터를 이용하여 학습용 가공 데이터를 생성할 수 있다.
- [0028] 상기 단어 벡터를 생성하는 단계는 단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습할 수 있다.
- [0029] 상기 단어 벡터를 생성하는 단계는 상기 가공 데이터 중에서 상기 학습할 단어 및 상기 학습할 단어의 단어 의미 데이터가 가공된 각각의 가공 데이터를 단어 임베딩에서의 스킵-그램(Skip-Gram)의 입력층 및 출력층이 합쳐진 하나의 입출력층에 위치시키고, 상기 각각의 가공 데이터가 위치한 입출력층 및 미러층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성할 수 있다.
- [0030] 상기 단어 벡터를 생성하는 단계는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로퍼게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행하고, 백 프로퍼게이션 과정에서 학습할 단어의 단어 의미 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경할 수 있다.

발명의 효과

- [0031] 본 발명의 실시 예는 종래의 단어 임베딩이 아닌 의미적 정보를 이용한 단어 임베딩 학습을 수행하기 때문에 적은 학습 데이터로도 단어 임베딩 학습이 가능하며 학습 시간도 적게 걸릴 수 있다.
- [0032] 또한, 본 발명의 실시 예는 단순히 등장하는 위치적 정보가 아닌 학습할 단어의 뜻풀이 기반으로 학습함으로써 저 빈도의 단어 학습에도 효율적이며, 사전에 등장한 단어를 대상으로 하기 때문에 종래의 단어 임베딩보다 많은 단어를 벡터로 표현할 수 있다.
- [0033] 또한, 본 발명의 실시 예는 종래의 단어 임베딩에 비해 단어 의미 임베딩 학습을 수행하기 때문에 단어들의 관계를 코사인 유사도를 통해 볼 수 있고, 종래의 단어 임베딩에서 얻을 수 없는 동의어 관계나 유의어 등을 찾을 수 있다.
- [0034] 또한, 본 발명의 실시 예는 학습할 단어 목록을 부사와 의존명사 등까지 확장하여 학습할 수 있으며, 단어의 의미적 정보를 추가해 적용할 수 있다.
- [0035] 더 나아가, 본 발명의 실시 예는 단어 의미 임베딩 학습을 통한 단어 벡터를 이용해 사람 수준의 의미 추론이 가능할 수 있다.

도면의 간단한 설명

- [0036] 도 1은 본 발명의 실시 예에 따른 어휘 의미망을 이용한 단어 의미 임베딩 장치의 구성도이다.
- 도 2는 본 발명의 실시 예에 따른 입출력층 및 미러층으로 이루어진 학습 모델의 설명도이다.
- 도 3은 종래의 단어 임베딩에서의 스킵-그램 모델에 대한 설명도이다.
- 도 4는 본 발명의 실시 예에 따른 단어 의미 임베딩 장치에서의 피쳐 미러 모델에 대한 설명도이다.
- 도 5는 종래의 스킵-그램 모델을 이용한 형태소 단위 학습 과정에 대한 예시도이다.
- 도 6은 본 발명의 실시 예에 따른 피쳐 미러 모델을 이용한 단어 임베딩 학습 과정에 대한 예시도이다.
- 도 7은 본 발명의 실시 예에 따른 단어 의미 임베딩 장치에 의해 수행되는 어휘 의미망을 이용한 단어 의미 임베딩 방법에 대한 흐름도이다.
- 도 8은 본 발명의 실시 예에 따른 단어 의미 임베딩 방법에서 단어 목록, 가공 데이터 및 단어 임베딩 학습 과정에 대한 상세 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0037] 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 설명한다. 본 발명에 따른 동작 및 작용을 이해하는 데 필

요한 부분을 중심으로 상세히 설명한다. 본 발명의 실시 예를 설명하면서, 본 발명이 속하는 기술 분야에 익히 알려졌고 본 발명과 직접적으로 관련이 없는 기술 내용에 대해서는 설명을 생략한다. 이는 불필요한 설명을 생략함으로써 본 발명의 요지를 흐리지 않고 더욱 명확히 전달하기 위함이다.

- [0038] 또한, 본 발명의 구성 요소를 설명하는 데 있어서, 동일한 명칭의 구성 요소에 대하여 도면에 따라 다른 참조부호를 부여할 수도 있으며, 서로 다른 도면임에도 동일한 참조부호를 부여할 수도 있다. 그러나 이와 같은 경우라 하더라도 해당 구성 요소가 실시 예에 따라 서로 다른 기능을 갖는다는 것을 의미하거나, 서로 다른 실시 예에서 동일한 기능을 갖는다는 것을 의미하는 것은 아니며, 각각의 구성 요소의 기능은 해당 실시 예에서의 각각의 구성 요소에 대한 설명에 기초하여 판단하여야 할 것이다.
- [0039] 도 1은 본 발명의 실시 예에 따른 어휘 의미망을 이용한 단어 의미 임베딩 장치의 구성도이다.
- [0040] 도 1에 도시된 바와 같이, 본 발명의 실시 예에 따른 어휘 의미망을 이용한 단어 의미 임베딩 장치(100)는 데이터 저장부(110), 단어 목록 생성부(120), 가공 데이터 생성부(130), 단어 임베딩 학습부(140) 및 유의어 산출부(150)를 포함한다.
- [0041] 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 어휘 의미망으로부터 학습할 단어 목록을 생성하고, 그 생성된 학습할 단어 목록에 포함된 단어의 사전적인 뜻풀이와 단어의 상위어, 반의어를 이용해 가공 데이터를 생성한다. 그리고 단어 의미 임베딩 장치(100)는 단어 임베딩에서의 네거티브 샘플링(Negative-Sampling)과 스킵-그램(Skip-Gram) 방식이 변형된 단어 임베딩 학습을 통해 단어의 의미적 관계와 연관성을 단어 벡터로 표현하기 위한 것이다.
- [0042] 여기서, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 단어 사전 데이터 및 단어 의미 데이터가 포함되는 어휘 의미망이 구축될 수 있는 한국어뿐만 아니라 영어(WordNet), 중국어(HowNet) 등의 모든 언어에 적용가능하다.
- [0043] 이하, 도 1의 어휘 의미망을 이용한 단어 의미 임베딩 장치(100)의 각 구성요소들의 구체적인 구성 및 동작을 설명한다.
- [0044] 데이터 저장부(110)는 단어 사전 데이터 및 단어 의미 데이터가 포함된 어휘 의미망을 저장한다. 단어 사전 데이터에는 단어 사전에 수록된 단어들이 포함될 수 있다. 단어 의미 데이터에는 단어 사전에 수록된 단어들의 단어 의미적 정보가 포함될 수 있다. 예컨대, 단어 의미 데이터에는 단어의 뜻풀이(정의), 상위어, 반의어, 용례 등이 포함될 수 있다.
- [0045] 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 학습할 단어의 종류와 단어 의미 데이터로 어휘 의미망(UWordMap)을 이용한다. 여기서, 어휘 의미망은 어휘 지도로 지칭되기도 한다. 어휘 의미망은 표준 사전을 기반으로 명사, 용언, 부사 어휘들이 의미제약으로 상호 연결된 어휘 의미망을 의미한다.
- [0046] 단어 목록 생성부(120)는 데이터 저장부(110)에서 저장된 단어 사전 데이터로부터 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성한다.
- [0047] 여기서, 단어 목록 생성부(120)는 명사, 용언, 부사 및 의존명사 어휘들 중에서 적어도 하나의 어휘들이 상호 연결된 어휘 의미망을 이용하여 학습할 단어 목록을 생성할 수 있다. 예를 들면, 단어 목록 생성부(120)는 학습에 사용할 단어를 구성하되, 어휘 의미망에 존재하는 단어 중 명사와 용언을 대상으로 특정 지역말(예컨대, 북한말), 방언, 의존명사를 제외한 동형이의어 수준으로 총 387,152개의 학습할 단어 목록을 생성하거나 데이터베이스로 구축할 수 있다.
- [0048] 가공 데이터 생성부(130)는 단어 목록 생성부(120)에서 생성된 학습할 단어 목록에 포함된 단어의 단어 의미 데이터를 데이터 저장부(110)로부터 가져와서 단어 임베딩 학습에 맞게 가공하여 학습용 가공 데이터를 생성한다.
- [0049] 여기서, 가공 데이터 생성부(130)는 단어 목록 생성부(120)에서 생성된 학습할 단어 목록에 포함된 단어의 뜻풀이, 상위어 및 반의어 중 적어도 하나가 포함된 단어 의미 데이터를 이용하여 학습용 가공 데이터를 생성한다. 예를 들면, 가공 데이터 생성부(130)는 단어의 뜻풀이에서 단어 사전 목록에 등장하는 명사, 용언을 해당 단어의 가공 데이터로 이용한다. 또한, 가공 데이터 생성부(130)는 명사의 경우에 명사의 상위어와 반의어를 가공 데이터로 이용한다. 그리고 가공 데이터 생성부(130)는 각 단어들을 단어 임베딩 학습에 맞게 원-핫(One-hot) 방식으로 표현된 입력 값과 출력 값으로 가공하여 가공 데이터를 생성한다.
- [0050] 단어 임베딩 학습부(140)는 가공 데이터 생성부(130)에서 생성된 가공 데이터를 기초로 입출력층 및 프로젝션

층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성한다.

- [0051] 여기서, 단어 임베딩 학습부(140)는 단어 임베딩에서의 스킵-그램(Skip-Gram)의 입력층 및 출력층이 합쳐진 하나의 입출력층 및 미러층(mirror layer)으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성한다. 단어 임베딩 학습부(140)는 각각의 가공 데이터가 위치한 입출력층 및 미러층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성한다.
- [0052] 이때, 단어 임베딩 학습부(140)는 단어 임베딩 학습의 백 프로파게이션(back propagation) 과정에서 출력 단어와 연결된 가중치 값을 변경하지 않고, 입력 단어와 연결된 가중치 값을 변경한다.
- [0053] 또한, 단어 임베딩 학습부(140)는 단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습할 수 있다. 단어 임베딩 학습부(140)는 반의어에 대해서 네거티브-샘플링을 수행한다. 단어 임베딩 학습부(140)는 반의어의 경우에만 출력 값을 0으로 설정해 학습한다. 단어 임베딩 학습부(140)는 학습할 단어의 다른 동형이의어에 대해서 네거티브-샘플링을 수행한다. 예를 들면, '사과_05/NNG'에 대해서 학습할 때 이 단어는 '사과'이며 품사가 NNG인 다른 형태소(사과_08/NNG 등)'와 네거티브-샘플링이 수행될 수 있다.
- [0054] 이와 같이, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 어휘 의미망에 있는 단어 사전 데이터를 이용해 단어 목록을 구축한다. 그리고 구축한 단어 목록과 어휘 의미망의 단어 의미 데이터(예컨대, 뜻풀이, 상위어, 반의어를 이용해 가공 데이터를 생성한다. 이후, 단어 의미 임베딩 장치(100)는 그 생성한 가공 데이터를 가지고 단어 임베딩의 스킵-그램을 변형한 학습 모델을 이용해 단어 임베딩 학습을 수행하여 단어 벡터를 생성한다.
- [0055] 한편, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 단어 벡터를 이용하여 유의어를 산출하는 경우에 유의어 산출부(150)를 더 포함할 수 있다.
- [0056] 일례로, 유의어 산출부(150)는 단어 임베딩 학습부(140)에서 생성된 단어 벡터를 이용한 코사인 유사도를 계산해 입력 단어의 유의어를 산출한다.
- [0057] 다른 예로, 유의어 산출부(150)는 제1 및 제2 입력 단어를 입력받는다. 그리고 유의어 산출부(150)는 제2 입력 단어에 가장 근접한 단어를 코사인 유사도를 이용하여 기설정된 개수의 유의어들을 추출하고, 그 추출된 유의어들과 제1 입력 단어 간의 코사인 유사도를 계산하여 가장 높은 유사도를 가진 단어를 제1 및 제2 입력 단어의 유의어로 산출한다.
- [0058] 또 다른 예로, 유의어 산출부(150)는 입력 단어의 단어 벡터와 기설정된 속성을 가지는 속성 단어의 단어 벡터의 속성 연산 과정을 통해 입력 단어의 속성이 변경된 유의어를 산출한다.
- [0059] 이와 같이, 자연어 처리 분야에서 기계학습을 사용하는 경우 단어의 표현이 중요하다. 단어 임베딩은 문자를 컴퓨터가 이해할 수 있는 벡터로 변경하는 방법이다. 단어 임베딩은 인공 신경망을 이용하여 저차원의 벡터로 학습한다.
- [0060] 실험 예로, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)에서는 어휘 의미망을 이용하여 387,152개의 명사, 용언의 단어 목록을 구축하였으며, 사전상의 뜻풀이와 명사의 상위어, 반의어를 이용해 가공 데이터를 생성했다.
- [0061] 그리고 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 단어 임베딩의 스킵-그램 방식을 변형한 학습 모델과 네거티브-샘플링을 이용해 학습한다. 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 종래의 단어 표현 방식보다 적은 데이터를 통해 빠르게 학습이 가능하며 의미적 유사성을 가진 단어들끼리 인접해 있음을 보일 수 있다. 이는 단순한 단어 벡터의 조합이 의미가 있을 수 있다는 것을 보여준다.
- [0062] 도 2는 본 발명의 실시 예에 따른 입출력층 및 미러층으로 이루어진 학습 모델의 설명도이다.
- [0063] 도 2에 도시된 바와 같이, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 각 단어들을 학습하기 위해서 종래의 단어 임베딩 또는 워드투벡(Word2Vec)의 스킵-그램(Skip-Gram)을 변형한 학습 모델과 네거티브-샘플링(Negative-Sampling)을 이용한다. 여기서, 가공 데이터의 입력 값과 결과 값은 원-핫(One-hot) 방식으로 표현되어 이용된다.
- [0064] 그리고 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 각 단어를 50차원의 단어 벡터로 표현한다.

- [0065] 도 2에는 본 발명의 실시 예에서 이용한 학습 모델이 나타나 있다. 단어 의미 임베딩 장치(100)는 이러한 학습 모델을 통해 단어 A를 학습하기 위해 단어 B와 연결된 가중치(Weight) 값을 이용하여 A라는 단어를 201 및 202 간선을 통해 학습한다. 여기서, 단어 의미 임베딩 장치(100)는 실제 가중치 값을 변경하는 경우, 학습할 단어 A와 연결된 가중치(204)만 변경하게 되고, 단어 B와 연결된 가중치 값(203)을 변경하지 않는다.
- [0066] 이러한 학습 모델은 단어 A의 단어 임베딩 결과에 단어 B와 연결된 단어들의 정보도 포함하기 위해서 이용된다. 그리고 단어들 간의 의미적 연쇄효과를 얻기 위해 도 2와 같은 학습 모델이 이용된다. 이때, 단어 의미 임베딩 장치(100)는 뜻풀이와 상위어에 대해 출력(Output) 값으로 1로 설정해 학습하고, 반의어의 경우에만 출력 값을 0으로 설정해 학습한다.
- [0067] 또한, 단어 의미 임베딩 장치(100)는 네거티브-샘플링을 이용해 학습 데이터 이외의 오답을 각 단어에 학습시켜 학습의 정확률을 높일 수 있다.
- [0068] 이하, 도 3 및 도 4를 참조하여 종래의 단어 임베딩의 스킵-그램 모델과 본 발명의 실시 예에 따른 변형된 스킵-그램 모델인 피쳐 미리 모델과의 차이점을 살펴보기로 한다.
- [0069] 도 3은 종래의 단어 임베딩에서의 스킵-그램 모델에 대한 설명도이다.
- [0070] 도 3에 도시된 바와 같이, 종래의 스킵-그램 모델은 입력층(input layer), 프로젝션층(projection layer) 및 출력층(output layer)의 3개의 층(layer)을 이용하여 구성된 인공 신경망이다.
- [0071] 입력층에는 원-핫(One-hot) 형태로 학습할 단어가 들어가게 된다. 출력층에는 입력층에서 사용한 단어의 앞의 두 단어($w(t-2)$, $w(t-1)$), 뒤의 두 단어($w(t+1)$, $w(t+2)$)가 들어가서 학습하게 된다. 도 3에 도시된 스킵-그램 모델에서는 입력층과 프로젝션층 사이의 간선과 프로젝션층 및 출력층 사이의 간선은 서로 다른 값($W_1 \neq W_2$)으로 구성되어 있다.
- [0072] 도 4는 본 발명의 실시 예에 따른 단어 의미 임베딩 장치에서의 피쳐 미리 모델에 대한 설명도이다.
- [0073] 도 4의 (a) 및 (b)에 도시된 바와 같이, 본 발명의 실시 예에 따른 피쳐 미리 모델은 두 개의 층(Layer)으로 이루어진다. 피쳐 미리 모델은 종래의 스킵-그램 모델의 입력층 및 출력층을 합친 하나의 입출력층과 미리층으로 이루어져 있다.
- [0074] 본 발명의 실시 예에 따른 피쳐 미리 모델에서의 입출력층에는 학습할 단어($x(\text{target})$)와 학습할 단어의 단어 의미 데이터가 원-핫 형태로 가공되어 들어가게 된다. 예컨대, 단어 의미 데이터에는 단어 뜻풀이(Word definition), 상위어(hypernym) 및 반의어(antonym) 등이 포함될 수 있다.
- [0075] 단어 의미 임베딩 장치(100)는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로퍼게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행한다. 단어 의미 임베딩 장치(100)는 백 프로퍼게이션 과정에서 학습할 단어의 단어 의미 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경한다.
- [0076] 즉, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 피쳐 미리 모델을 통해 학습을 적용할 때, 학습할 단어 즉, 타겟이 되는 단어($x(\text{target})$)와 연결된 간선(가중치)만을 변화시키고, 학습할 단어의 단어 뜻풀이, 상위어 및 반의어 등과 연결된 간선(가중치)을 변화시키지 않는다.
- [0077] 이하, 도 5 및 도 6을 참조하여 종래의 스킵-그램 모델과 본 발명의 실시 예에 따른 스킵-그램 모델을 이용한 학습 과정을 살펴보기로 한다.
- [0078] 도 5는 종래의 스킵-그램 모델을 이용한 형태소 단위 학습 과정에 대한 예시도이다.
- [0079] 종래의 단어 임베딩 방식은 문장에서 학습할 단어의 주변 단어를 이용해 학습한다. 도 5에 도시된 바와 같이, "학교에서 사과를 먹었다."라는 문장에서 학습할 단어가 "사과"인 경우에 주변 단어 즉, 주변 형태소인 "학교", "에서", "를" 및 "먹"라는 주변 단어가 단어 임베딩 방식에 이용된다. 학습할 단어인 "사과"와는 의미상으로 관련이 없을 수도 있지만 문장에서 학습할 단어와 인접한 단어가 단어 임베딩 방식에 이용되는 것이다.
- [0080] 여기서, 종래의 단어 임베딩 방식에서는 단어의 수나 정확률을 위해서는 대용량의 형태소 의미 번호가 부착된 말뭉치가 필요하게 된다. 말뭉치의 양에 따라 단어 임베딩의 결과에 영향을 주므로, 종래의 단어 임베딩 방식은 많은 말뭉치를 사용할 것을 권장하고 있다.

- [0081] 도 6은 본 발명의 실시 예에 따른 피쳐 미리 모델을 이용한 단어 임베딩 학습 과정에 대한 예시도이다.
- [0082] 피쳐 미리 모델을 이용하여 "사과(먹는 사과)"라는 학습할 단어를 학습하는 과정을 도 6을 참조하여 살펴보기로 한다.
- [0083] 도 6의 (a) 및 (b)에 도시된 바와 같이, 단어 의미 임베딩 장치(100)는 사과의 뜻풀이인 사과나무의 열매에서 명사와 용언을 추출한 "사과나무", "열매"라는 단어 의미 데이터를 가공 데이터로 이용한다. 그리고 단어 의미 임베딩 장치(100)는 "사과"라는 학습할 단어의 상위어인 "과일"을 사과의 가공 데이터로 추가한다. 여기서, 학습할 단어 "사과"의 반의어가 없기 때문에 반의어를 가공 데이터는 추가하지 않는다. 그렇지만, 반의어가 있는 경우에는 반의어를 추가한다.
- [0084] "사과"라는 학습할 단어의 가공 데이터로 "사과나무", "열매", "과일"을 찾은 후 오른쪽에 학습 모델을 이용하여 단어 임베딩 학습을 601 내지 604 과정에 따라 수행한다.
- [0085] 일례로, "사과"라는 학습할 단어와 뜻풀이의 "사과나무"라는 단어를 이용한 단어 임베딩 학습 과정을 하기의 [수학식 1] 및 [수학식 2]를 참조하여 살펴보기로 한다.

수학식 1

$$y(\text{사과}, \text{사과나무}) = \varphi \left[\sum_{n=0}^m (W_{n(\text{사과})} * W_{n(\text{사과나무})}) \right]$$

[0086]

[0087] 여기서, $W_{n(\text{사과})}$ 는 사과의 벡터, $W_{n(\text{사과나무})}$ 는 사과나무의 벡터, $y(\text{사과}, \text{사과나무})$ 는 사과의 벡터와 사과나무의 벡터를 곱한 후 모두를 더한 값을 나타낸다.

수학식 2

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

[0088]

[0089] 상기의 [수학식 1]은 사과 및 사과나무 간의 출력값을 계산하는 수식을 나타낸다.

[0090] 상기의 [수학식 1]에 따라 사과의 벡터와 사과나무의 벡터를 곱한 후 모두를 더한 값에 상기의 [수학식 2]에 나타난 시그모이드 함수(Sigmoid Function)를 이용하면 출력 값이 된다.

수학식 3

$$E = d(\text{사과나무}) - y(\text{사과}, \text{사과나무})$$

[0091]

[0092] 여기서, E 는 오차 값, $d(\text{사과나무})$ 는 가공 데이터 상의 정답 값을 나타낸다.

[0093] 상기의 [수학식 3]과 같이, 오차 값은 $d(\text{사과나무})$ 에서 출력 값인 $y(\text{사과}, \text{사과나무})$ 를 뺀 값으로 나타내진다. 여기서, $d(\text{사과나무})$ 는 가공 데이터 상의 정답 값을 의미한다. 이러한 정답 값은 반의어 학습이 아닌 경우에 값이 1이며 반의어 학습인 경우 0이다.

[0094] 단어 임베딩 학습부(140)는 오차 값(E)을 이용해 $W_{(\text{사과나무})}$, $W_{(\text{사과})}$ 의 변화량을 구한다. 단어 임베딩 학습부(140)는 변화량을 구한 후 $W_{(\text{사과})}$ 에만 변화량을 적용한다. 즉, 단어 임베딩 학습부(140)는 사과와 연결

된 가중치(604)에만 변화량을 적용하고, 사과나무와 연결된 가중치(603)에는 변화량을 적용하지 않는다.

- [0095] 이와 같이, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 단어의 뜻풀이나 상위어, 반의어 같은 의미적 정보만을 이용하여 학습하므로 종래의 학습 모델보다 적은 데이터로도 학습할 수 있다. 예컨대, 단어 의미 임베딩 장치(100)에서는 총 387,152개의 단어를 학습하는데 250 내지 300만 개의 학습데이터가 필요하다. 즉, 단어 의미 임베딩 장치(100)는 적은 가공 데이터를 이용하여 단어 임베딩 학습을 수행할 수 있다.
- [0096] 도 7은 본 발명의 실시 예에 따른 단어 의미 임베딩 장치에 의해 수행되는 어휘 의미망을 이용한 단어 의미 임베딩 방법에 대한 흐름도이다.
- [0097] 본 발명의 실시 예에 따른 어휘 의미망을 이용한 단어 의미 임베딩 장치(100)는 단어 사전 데이터로부터 기설정된 품사를 갖는 어휘들을 추출해 학습할 단어 목록을 생성한다(S101).
- [0098] 그리고 단어 의미 임베딩 장치(100)는 생성된 단어 목록에 포함된 학습할 단어에 대한 단어 의미 데이터를 이용하여 가공 데이터를 생성한다(S102). 이때, 단어 의미 임베딩 장치(100)는 단어 의미 데이터를 데이터 저장부(110)에서 저장된 어휘 의미망으로부터 가져온다.
- [0099] 이후, 단어 의미 임베딩 장치(100)는 생성된 가공 데이터를 이용한 단어 의미 임베딩 학습을 통해 단어 벡터를 생성한다(S103).
- [0100] 그리고 단어 의미 임베딩 장치(100)는 입력 단어를 입력받고, 그 입력된 입력 단어의 단어 벡터를 이용하여 유의어를 산출한다(S104).
- [0101] 도 8은 본 발명의 실시 예에 따른 단어 의미 임베딩 방법에서 단어 목록, 가공 데이터 및 단어 임베딩 학습 과정에 대한 상세 흐름도이다.
- [0102] 도 8에 도시된 바와 같이, 단어 의미 임베딩 장치(100)는 프로그램 메인(810)을 통해 단어 임베딩 학습을 위한 가공 데이터 생성 과정, 단어 임베딩 학습 과정을 요청하고 그 결과를 반환할 수 있다.
- [0103] 우선, 프로그램 메인(810)은 단어 의미 임베딩을 위한 가공 데이터 생성을 단어 목록 생성부(120) 및 가공 데이터 생성부(130)에 요청한다(S201).
- [0104] 그리고 단어 목록 생성부(120)는 데이터 저장부(110)에 단어 사전 데이터 즉, 단어의 사전적 정보를 요청한다(S202).
- [0105] 이어서, 단어 목록 생성부(120)는 데이터 저장부(110)로부터 단어 사전 데이터 즉, 단어의 사전적 정보를 반환받는다(S203).
- [0106] 그리고 단어 목록 생성부(120)는 반환된 단어 사전 데이터에서 기설정된 품사(예컨대, 명사, 용언 등) 어휘들을 추출해 학습할 단어 목록을 생성한다(S204). 여기서, 단어 목록은 각 단어마다 인덱스(index) 번호가 부여되어 있다.
- [0107] 이후, 가공 데이터 생성부(130)는 학습할 단어 목록에 포함된 학습할 단어의 단어 의미 데이터(예컨대, 단어의 뜻풀이, 상위어, 반의어 등)를 데이터 저장부(110)에 요청한다(S205).
- [0108] 이어서, 가공 데이터 생성부(130)는 데이터 저장부(110)로부터 단어 의미 데이터를 반환받는다(S206).
- [0109] 가공 데이터 생성부(130)는 학습할 단어 목록과 단어 의미 데이터(예컨대, 단어 뜻풀이(정의), 상위어, 반의어 등)를 이용하여 가공 데이터를 생성한다(S207).
- [0110] 그리고 가공 데이터 생성부(130)는 생성된 가공 데이터를 프로그램 메인(810)에 반환한다(S208).
- [0111] 이후, 프로그램 메인(810)은 반환된 가공 데이터를 이용하여 단어 임베딩 학습부(140)에 단어 의미 임베딩에 대한 학습을 요청한다(S209).
- [0112] 그리고 단어 임베딩 학습부(140)는 가공 데이터를 기초로 피쳐 미리 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성한다(S210).
- [0113] 이어서, 단어 임베딩 학습부(140)는 단어 의미 임베딩에 대한 학습 결과 즉, 단어 벡터를 프로그램 메인(810)에 반환한다(S211).
- [0114] 한편, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)와 종래 기술들과의 유의어 추천 결과에 대해서 살

퍼보기로 한다. 여기서, 종래 기술들에는 NNLM, Ranking(hinge loss), Ranking(logit loss), word2vec이 포함된다.

[0115] 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 하기의 [수학식 4]와 같은 코사인 유사도(Cosine similarity)를 이용해 계산한다.

수학식 4

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

[0116]

[0117] 여기서, A 및 B는 두 개의 단어 A 및 B 각각의 단어 벡터를 나타낸다.

[0118] 종래의 단어 임베딩 기술은 단순히 문장에서 앞쪽 또는 뒤쪽 단어가 유사한 단어들끼리 비슷한 벡터를 형성하도록 학습이 된다.

[0119] 하지만, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 단어의 의미적 정보인 뜻풀이, 상위어, 반의어 등을 이용하기 때문에 의미상으로 유사한 단어들이 비슷한 벡터를 형성한다.

[0120] 하기의 [표 1]과 [표 2]는 "서울_01/NNP"(한반도 중심부에 있는 도시, 대한민국의 수도)에 대해서 종래 기술들과 본 발명의 실시 예를 이용하여 검색된 유의어 결과이다.

[0121] [표 1]에는 종래 기술들에 따라 "서울_01/NNP"과 유사한 유의어로 검색된 결과가 나타나 있다.

표 1

NNLM[10]	Ranking(hinge loss)[11]	Ranking(logit loss)	word2vec
인천/NNP	일본/NNP	부산/NNP	강남구/NNP
울산/NNP	미국/NNP	대전/NNP	영등포구/NNP
부산/NNP	부산/NNP	중국/NNP	중구/NNP
충북/NNP	중국/NNP	미국/NNP	부산/NNP
광주/NNP	독일/NNP	일본/NNP	서대문구/NNP

[0122]

[0123] 반면, [표 2]에는 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)에 의해 "서울_01/NNP"과 유사한 유의어로 검색된 결과가 나타나 있다.

표 2

본 발명의 유사어
베이징/NNP
라왈핀디/NNP
브뤼셀/NNP
아스마라/NNP
사라예보/NNP

[0124]

[0125] 상기 [표 1] 및 [표 2]에 나타난 바와 같이, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 "서울_01/NNP"를 수도라는 의미적 정보를 강하게 받아서 각 나라의 수도였거나 수도인 도시들의 이름을 유의어로 산출하였다. 하지만, 종래의 단어 임베딩 기술에서는 "서울"이 나타난 문장구조와 유사한 문장구조를 가졌던 단어들이 유의어로 나타나 수도와 도시가 혼합되거나, 도시와 구가 혼합된 결과가 나온다.

[0126] 한편, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)를 이용한 유의어 실험 결과를 살펴보기로 한다.

[0127] 실험 과정에서, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 어휘 의미망을 통해 387,152개의 명사와 용언으로 이루어진 단어 목록을 생성하였다. 그리고 단어 의미 임베딩 장치(100)는 그 생성된 단어 목록과 단어 의미 데이터를 이용한 가공 데이터에 대한 단어 임베딩 학습을 통해 50차원의 벡터로 변환하였다.

[0128] 그리고 3가지의 실험 방법을 통해 단어 의미 임베딩이 어떤 결과를 보여주는지 실험했다.

표 3

입력 단어	유사어 목록
마시다	들이마시다__01, 비음하다__03, 폭음하다__02
자르다__01	사절하다__03, 사단하다, 단취하다__02
과일__01	백과__03, 실과__01, 유실수
휴대전화	핸드폰, 간이무선, 전신회선
번역	수역__07, 역__13, 개역__02

[0129]

[0130] 상기의 [표 3]에는 하나의 단어에 대해 유의어를 검색한 결과가 나타나 있다.

[0131] 상기의 [표 3]은 하나의 단어에 대해 코사인 유사도(Cosine Similarity)를 이용하여 가장 근접한 단어의 목록을 보여주고 있다. 각 단어의 의미상 유사하거나 같은 상위어를 가지는 단어들이 나타난다. 하지만, '과일'의 경우 '유실수'처럼 과일이 열리는 나무의 의미를 가지는 단어가 근접해 나오기도 한다.

표 4

입력 단어1	입력 단어2	유사 단어
여자__02	선생__01	여선생
이탈리아	국수__01	파스타__01
축구__04	경기장	축구장
초보__01	학생	초학자
프랑스	빵	바게트

[0132]

[0133] 상기의 [표 4]에는 두 명사의 유의어를 산출한 결과가 나타나 있다.

[0134] 상기의 [표 4]에는 두 명사를 입력하여 먼저 입력 단어 2에 근접한 단어를 코사인 유사도를 이용하여 200개를 추출한 결과가 나타나 있다. 그리고 추출한 200개와 입력 단어 1 간의 코사인 유사도를 구해서 가장 높은 유사도가 나온 단어의 결과이다. '초보__01'과 '학생' 두 단어의 유의어로 '초심자{어떤 일을 처음 배우는 사람}'를 기대했지만 오히려 잘 쓰이지는 않지만, 더욱 정확한 의미의 '초학자{학문을 처음으로 배우기 시작한 사람}'가 나타났다.

표 5

입력 단어	(-) 속성	(+) 속성	결과
오토바이	동력__02	사람	양치__03, 별사람, 여인__03
임금__01	남자__02	여자__02	여주__02, 여왕, 발런
원자력발전소	원자력	수력__01	자류식발전소, 오르막길, 클라이밍

[0135]

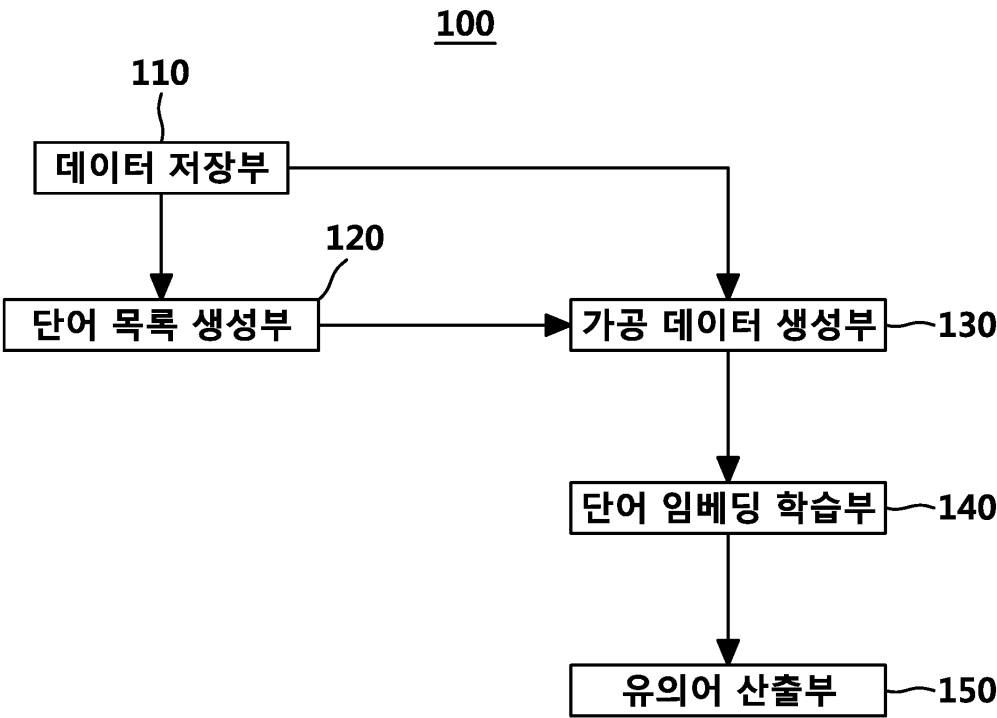
- [0136] 상기의 [표 5]에는 단어의 특정 속성을 빼기(-) 및 더하기(+) 한 결과가 나타나 있다.
- [0137] 상기의 [표 5]는 입력 단어의 벡터에서 단순히 특정 속성 단어의 벡터를 빼고 더한 값에 코사인 유사도를 이용하여 유사도가 높은 단어들을 보여주고 있다.
- [0138] '오토바이'에서 '동력__02'이라는 속성을 빼고 '사람'이라는 속성을 더하면 "인력거"의 뜻을 가진 '양치__03'와 "수레를 만드는 사람"의 뜻풀이를 가진 '여인__03'이라는 단어가 등장한다.
- [0139] 하지만, 사람의 뜻이 강하게 작용되어 '별사람(생김새나 하는 짓, 말 따위가 보통 사람과 다른 이상스러운 사람)'도 나타나게 된다.
- [0140] 이와 같이, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 종래의 말뭉치 기반 단어 임베딩 방식이 아닌 어휘지도 즉, 어휘 의미망에서 단어의 의미적 정보를 이용하여 단어 임베딩 학습을 수행한다.
- [0141] 여기서, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 종래의 단어 임베딩이 아닌 의미적 정보를 이용한 단어 임베딩 학습을 수행하기 때문에 적은 학습 데이터로도 단어 임베딩 학습이 가능하며 학습 시간도 적게 걸리는 장점이 있다.
- [0142] 또한, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 단순히 등장하는 위치적 정보가 아닌 학습할 단어의 뜻풀이 기반으로 인해 저 빈도의 단어 학습에도 효율적이며 사전에 등장한 단어를 대상으로 하기 때문에 종래의 단어 임베딩보다 많은 단어를 벡터로 표현할 수 있다.
- [0143] 또한, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 종래의 단어 임베딩에 비해 단어 의미 임베딩 학습을 수행하기 때문에 단어들의 관계를 코사인 유사도를 통해 볼 수 있다. 단어 의미 임베딩 장치(100)는 종래의 단어 임베딩에서 얻을 수 없는 동의어 관계나 유의어 등을 찾을 수 있다.
- [0144] 또한, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)는 학습할 단어 목록을 부사와 의존명사 등까지 확장하여 학습할 수 있으며, 단어의 의미적 정보를 추가해 적용할 수 있다.
- [0145] 더 나아가, 본 발명의 실시 예에 따른 단어 의미 임베딩 장치(100)에서는 단어 의미 임베딩 학습을 통한 단어 벡터를 이용해 사람 수준의 의미 추론이 가능할 수 있다.
- [0146] 이상에서 설명한 실시 예들은 그 일 예로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명에 개시된 실시 예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시 예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

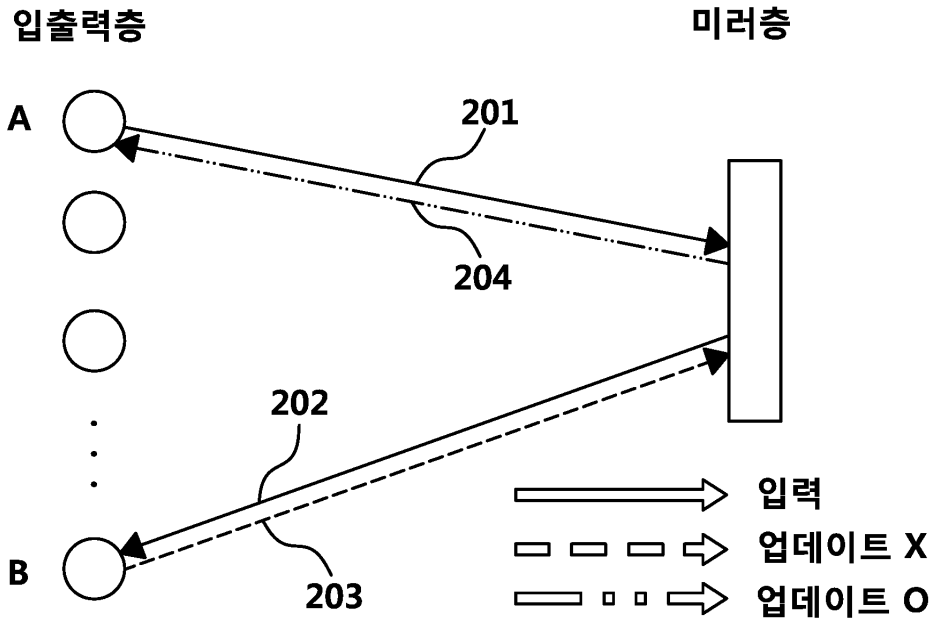
- [0147] 100: 단어 의미 임베딩 장치
- 110: 데이터 저장부
- 120: 단어 목록 생성부
- 130: 가공 데이터 생성부
- 140: 단어 임베딩 학습부
- 150: 유의어 산출부

도면

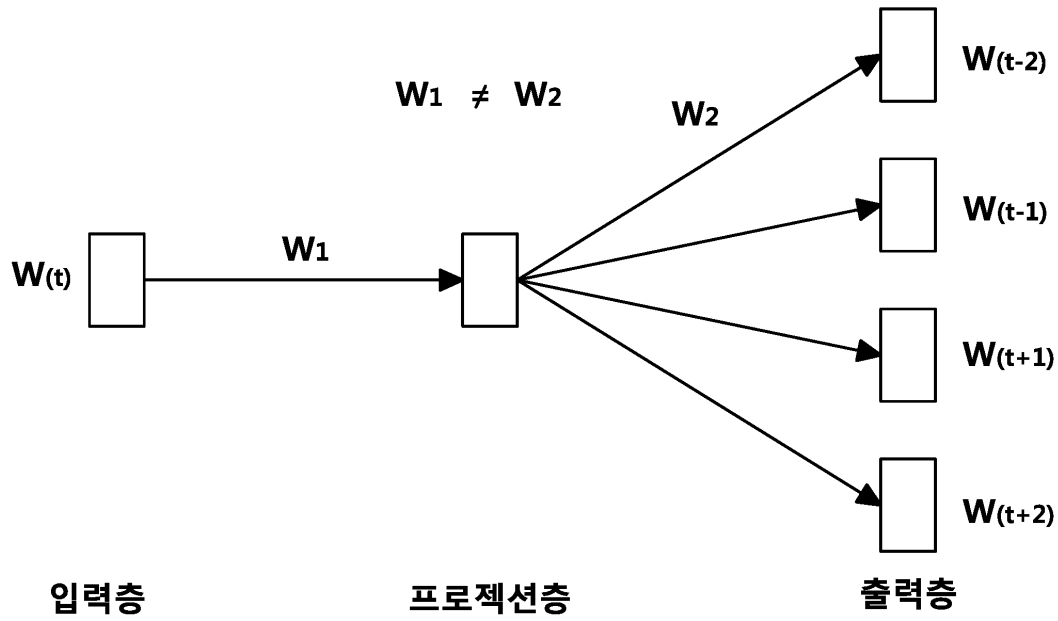
도면1



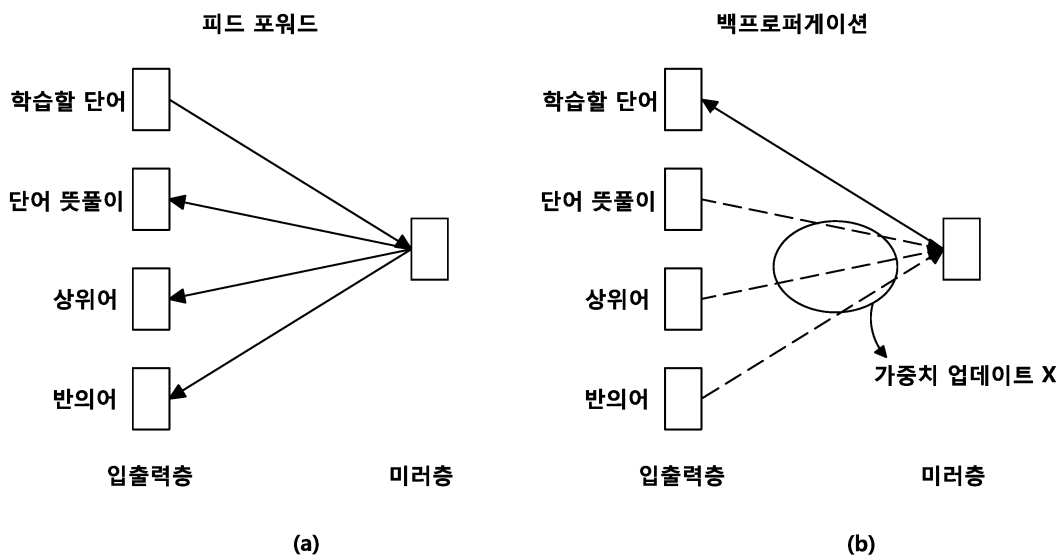
도면2



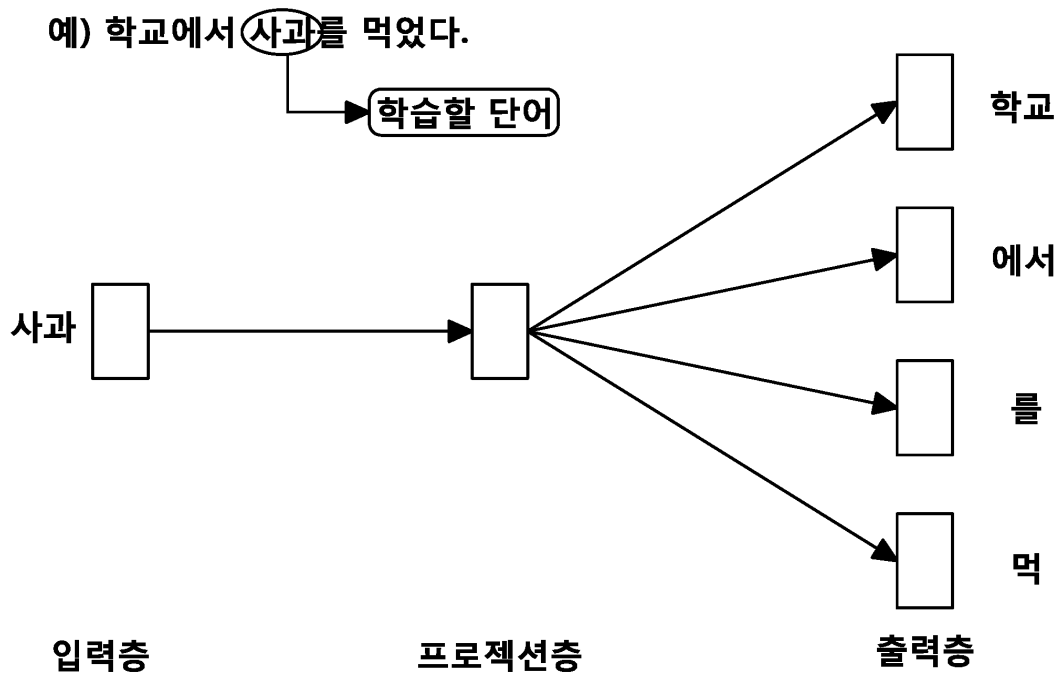
도면3



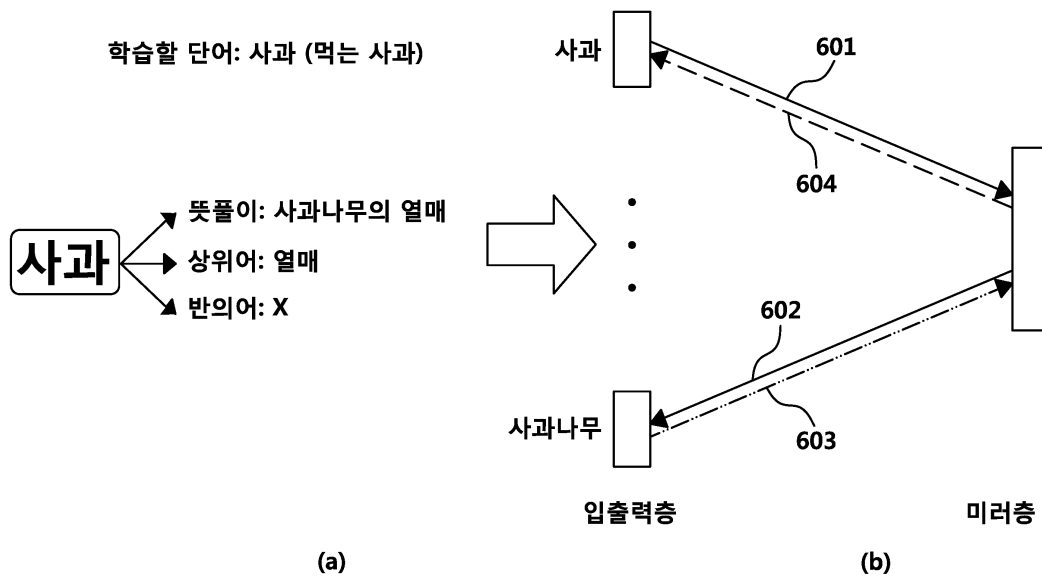
도면4



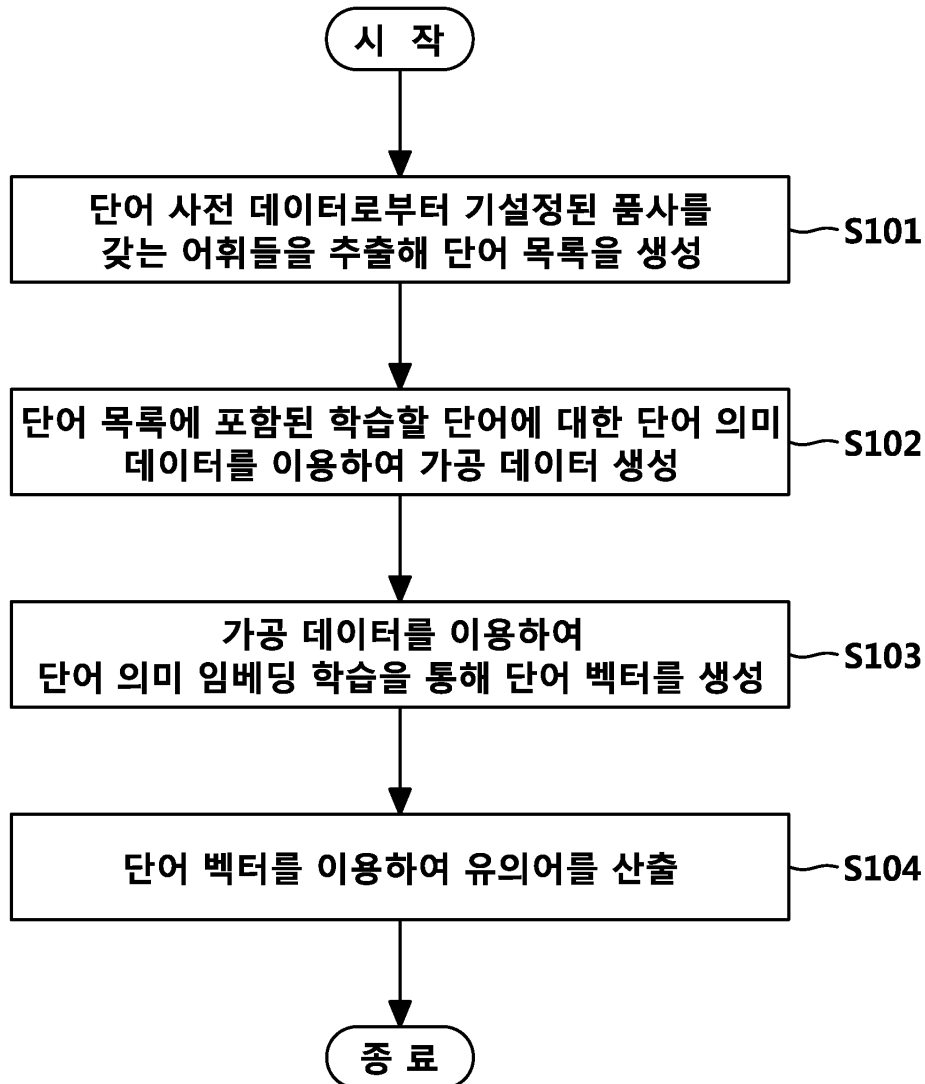
도면5



도면6



도면7



도면8

