



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년11월20일
(11) 등록번호 10-1799681
(24) 등록일자 2017년11월14일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) G06F 15/18 (2017.01)
G06N 99/00 (2010.01)
(52) CPC특허분류
G06F 17/2785 (2013.01)
G06F 15/18 (2013.01)
(21) 출원번호 10-2016-0074761
(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)
신준철
울산광역시 남구 신북로31번길 9-1, 208호 (무거동)
이주상
울산광역시 중구 옥교3길 120, B동 301호 (학성동, 옥성맨션)
(74) 대리인
김중선, 이형석

전체 청구항 수 : 총 16 항

심사관 : 경연정

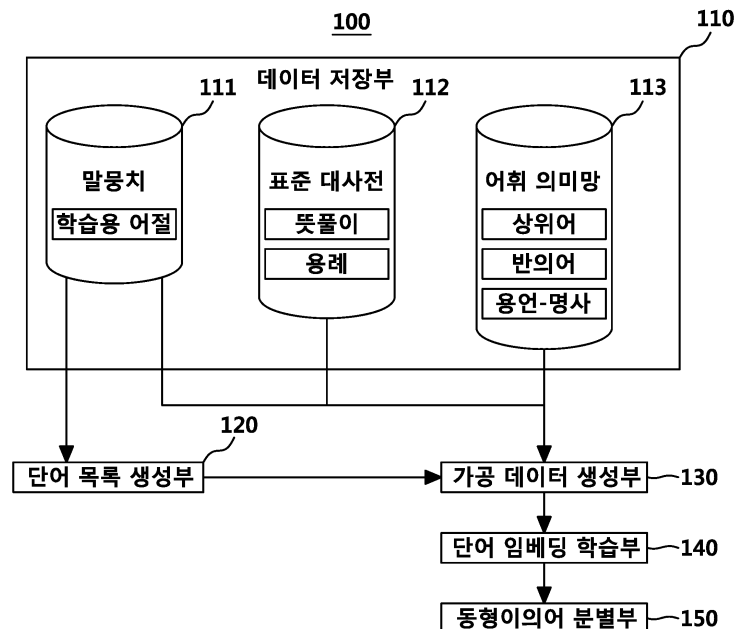
(54) 발명의 명칭 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법

(57) 요약

본 발명은 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법에 관한 것으로서, 더욱 상세하게는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치와 단어 의미 데이터가 포함된 표준 대사전 및 어휘 의미망을 저장하는 데이터 저장부; 상기 말뭉치로부터 학습할 단어 목록을 생성하는 단어 목록 생성부; 상기 생성된 단

(뒷면에 계속)

대표도 - 도1



어 목록에 포함된 학습할 단어의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환하고, 상기 생성된 학습할 단어 목록, 상기 변환된 말뭉치 및 상기 단어 의미 데이터를 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 가공 데이터 생성부; 상기 생성된 가공 데이터를 가지고 입출력층 및 프로잭션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단어 임베딩 학습부; 및 상기 생성된 단어 벡터를 이용하여 동형이의어와 인접 어절 간의 유사도를 비교하고 상기 비교 결과에 따라 동형이의어를 분별하는 동형이의어 분별부를 포함한다.

(52) CPC특허분류

G06F 17/2795 (2013.01)

G06N 99/005 (2013.01)

(56) 선행기술조사문헌

US20090276414 A1*

KR101508070 B1*

KR1020110017129 A

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

'어휘별 중의성 제거 규칙과 통계정보를 이용한 한국어 품사 태깅', 한국콘텐츠학회 v.9 no.2, 안광모 외 2인 (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

적어도 하나 이상의 학습용 어절이 포함된 말뭉치와 단어 의미 데이터가 포함된 표준 대사전 및 어휘 의미망을 저장하는 데이터 저장부;

상기 말뭉치로부터 학습할 단어 목록을 생성하는 단어 목록 생성부;

상기 생성된 단어 목록에 포함된 학습할 단어의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환하고, 상기 생성된 학습할 단어 목록, 상기 변환된 말뭉치 및 상기 단어 의미 데이터를 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 가공 데이터 생성부;

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 학습부; 및

상기 생성된 단어 벡터를 이용하여 동형이의어와 인접 어절 간의 유사도를 비교하고 상기 비교 결과에 따라 동형이의어를 분별하는 동형이의어 분별부를 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 가공 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 2

제1항에 있어서,

상기 데이터 저장부는

적어도 하나 이상의 학습용 어절이 포함된 말뭉치와, 용례 및 뜻풀이가 단어 의미 데이터로 포함된 표준 대사전과, 용언-명사 관계 정보, 상위어 및 반의어가 단어 의미 데이터로 포함된 어휘 의미망을 저장하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 3

제1항에 있어서,

상기 가공 데이터 생성부는

적어도 하나 이상의 학습용 어절이 포함된 말뭉치, 표준 대사전에 포함된 용례 및 뜻풀이, 및 어휘 의미망에 포함된 용언-명사 관계 정보 중에서 적어도 하나를 말뭉치 형태로 변환하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 4

제1항에 있어서,

상기 가공 데이터 생성부는

상기 변환된 말뭉치에서 상기 학습할 단어와 인접한 어절 및 상기 인접한 어절의 상위어를 상기 학습할 단어의 인접 어절로 처리하거나, 용언-명사 관계 정보를 상기 학습할 단어의 인접 어절로 처리하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 5

제1항에 있어서,

상기 단어 임베딩 학습부는

상기 가공 데이터 중에서 상기 학습할 단어, 인접 어절, 용언-명사 관계 정보 및 반의어가 각각 가공된 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 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 6

삭제

청구항 7

제1항에 있어서,

상기 단어 임베딩 학습부는

단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 8

제1항에 있어서,

상기 단어 임베딩 학습부는

단어 임베딩 학습을 통해 학습할 단어와 인접하고 조사 또는 어미를 제외한 실질형태소의 단어 벡터를 생성하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 9

제1항에 있어서,

상기 동형이의어 분별부는

상기 생성된 실질형태소의 단어 벡터를 이용하여 분별할 동형이의어의 실질형태소와 인접 어절의 실질형태소 간의 유사도를 비교하여 동형이의어를 분별하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치.

청구항 10

동형이의어 분별 장치에 의해 수행되는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법에 있어서,

적어도 하나 이상의 학습용 어절이 포함된 말뭉치로부터 학습할 단어 목록을 생성하는 단계;

상기 생성된 단어 목록에 포함된 학습할 단어의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환하는 단계;

상기 생성된 학습할 단어 목록, 상기 변환된 말뭉치 및 상기 단어 의미 데이터를 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 단계;

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

상기 생성된 단어 벡터를 이용하여 동형이의어와 인접 어절 간의 유사도를 비교하고 상기 비교 결과에 따라 동형이의어를 분별하는 단계를 포함하고,

상기 단어 벡터를 생성하는 단계는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로파게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행하고, 백 프로파게이션 과정에서 학습할 단어의 가공 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

청구항 11

제10항에 있어서,

상기 말뭉치로 변환하는 단계는

용례 및 뜻풀이가 포함된 표준 대사전과, 용언-명사 관계 정보, 상위어 및 반의어가 포함된 어휘 의미망에서의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

청구항 12

제10항에 있어서,

상기 말뭉치로 변환하는 단계는

적어도 하나 이상의 학습용 어절이 포함된 말뭉치, 표준 대사전에 포함된 용례 및 뜻풀이, 및 어휘 의미망에 포함된 용언-명사 관계 정보 중에서 적어도 하나를 말뭉치 형태로 변환하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

청구항 13

제10항에 있어서,

상기 가공 데이터를 생성하는 단계는

상기 변환된 말뭉치에서 상기 학습할 단어와 인접한 어절 및 상기 인접한 어절의 상위어를 상기 학습할 단어의 인접 어절로 처리하거나, 용언-명사 관계 정보를 상기 학습할 단어의 인접 어절로 처리하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

청구항 14

제10항에 있어서,

상기 단어 벡터를 생성하는 단계는

상기 가공 데이터 중에서 상기 학습할 단어, 인접 어절, 용언-명사 관계 정보 및 반의어가 각각 가공된 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 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

청구항 15

삭제

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 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

청구항 17

제10항에 있어서,

상기 단어 벡터를 생성하는 단계는

단어 임베딩 학습을 통해 학습할 단어와 인접하고 조사 또는 어미를 제외한 실질형태소의 단어 벡터를 생성하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

청구항 18

제10항에 있어서,

상기 동형이의어를 분별하는 단계는

상기 생성된 실질형태소의 단어 벡터를 이용하여 분별할 동형이의어의 실질형태소와 인접 어절의 실질형태소 간의 유사도를 비교하여 동형이의어를 분별하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법.

발명의 설명

기술 분야

[0001] 본 발명은 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법에 관한 것으로서, 더욱 상세하게는 다양한 자원(예컨대, 말뭉치, 표준 대사전 및 어휘 의미망)으로부터의 학습할 단어 목록, 변환된 말뭉치 및 단어 의미 데이터를 이용한 단어 임베딩 학습을 통해 학습하고 동형이의어와 인접 어절 간의 유사도를 비교하여 동형이의어를 분별함으로써, 미학습 패턴에 대해서 동형이의어를 용이하게 분별할 수 있는, 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법에 관한 것이다.

배경 기술

[0002] 정보기술이 발달하면서 자연어를 자동으로 처리하려는 기술 또한 같이 발달하고 있다. 이런 기술들은 문서분류나 기계번역 포탈검색 등에 응용되고 있다. 한국어는 교착어로 형태소 분석이 어려운 편이지만 최근 들어 학습 말뭉치를 이용한 기계학습 방법으로 형태소 복원과 품사 태깅이 높은 정확률(예컨대, 약 98%)을 보이고 있다.

[0003] 그러나 동형이의어 분별 정확률은 아직은 다소 낮은 편으로(약 96.5%), 실제로 기계번역에서 단어의 의미를 완전히 잘못 번역하는 경우를 쉽게 찾아볼 수 있다. 이런 오류를 줄이려면 동형이의어 분별 정확률을 향상시키는 것에 집중해야 한다.

[0004] 한국어 동형이의어 분별을 위해 문맥정보를 이용하는 방법으로는 크게 말뭉치학습 방법과 어휘망 학습 방법이 존재한다. 전자는 통계학습 방식이고 후자는 지식기반 방식에 속한다. 말뭉치를 학습하는 방법은 세종말뭉치처럼 대용량의 태그 부착 말뭉치를 학습하여서 인접어절에 대한 패턴을 학습하는 것이다. 이 방식은 인접 어절 전체나 일부를 그대로 저장하는 것으로, 어절 전체, 일부 형태소, 품사 또는 음절 등을 그대로 저장한다.

[0005] 그러나 종래의 말뭉치학습 방식은 직접적으로 학습한 적이 없다면 아무런 효과를 발휘할 수 없다. 예로 "사과는 열매다."를 분석한다고 가정하면, '사과(apple)'가 '열매'와 인접한 경우가 말뭉치에 존재해야 이 방식이 효과를 발휘한다. 하지만, 세종말뭉치에 그런 문장은 없으며, 만약 "사과는 식물이다." 또는 "사과는 과일이다."처럼 비슷한 문장이 말뭉치에 존재한다고 하더라도 '열매'가 직접적으로 나타나지 않았다면 아무런 효과가 없다. 다만, 개별 형태소 단위로 '사과'가 애플(apple)로 태깅된 비율을 측정하여 사용할 수는 있다. 이 방법으로는 베이스 라인만큼의 정확률만 기대할 수 있다.

[0006] 이렇게 종래의 말뭉치학습 방법으로는 미학습 패턴을 전혀 처리할 수 없기 때문에 최근에 들어서 워드넷(WordNet), 한국어 어휘의미망(Korlex, Korean Wordnet) 또는 한국어 어휘지도(UWordMap)와 같은 어휘망을 사용하는 연구가 이뤄지고 있다. 이 방법들은 어휘망에서 상하위 관계나 용언-명사 관계에 등록되지 않은 경우에 효과를 발휘할 수 없다는 점에서는 말뭉치학습 방식에서와 유사한 문제점이 있다. 하지만, 재현율은 부분적으로 훨씬 더 높다고 평가할 수 있다. 왜냐하면, 대부분의 잘 알려진 명사는 상하위망에 등록되어 있으며, 용언-명사의 관계는 상하위망에서 최소상계노드만을 이용하여 작성되기 때문이다. 무작위로 특정 노드 하나를 선택한다고 가정할 때, 그 노드의 하위에는 수백 또는 수천 개의 하위노드가 포함되어 있을 것으로 기대할 수 있다. 그러나 용언-명사 관계망 자체가 많이 빈약하기 때문에 재현율의 문제는 여전히 남아 있다. 이 문제를 완화시키려면 큰 비용을 들여서 어휘망을 계속 보완해야 한다.

[0007] 최근에는 워드(단어) 임베딩(word embedding)의 유용성이 알려지면서 자연어처리나 자연어이해의 다양한 분야에 이것을 적용하려는 시도가 이뤄지고 있다. 대표적으로 알려진 워드 임베딩 모델로는 워드투벡(Word2Vec)이 있다. 본래 단어 임베딩은 그 자체로 유사한 단어끼리 묶어주거나, 단어 간의 의미적 유사도 또는 연관성의 정도를 계산할 수 있게 해준다.

[0008] 한편, 의미처리시스템 개발을 위해서는 동형이의어 분별 기술은 필수적이다. 최근까지의 연구로는 말뭉치 학습 기반의 방법이 비교적 정확한 결과를 보여주고 있다. 종래의 말뭉치학습 방식은 인접 어절의 내용 중의 일부(음절 또는 형태소)를 그대로 기억하여 동형이의어를 분별하는 것이다.

[0009] 그러나 종래의 말뭉치학습 방식은 학습한 적이 없는 패턴에 대해서는 정확률은 낮다.

선행기술문헌

특허문헌

[0010] (특허문헌 0001) 대한민국 등록특허공보 제10-1358614호 (2014.01.27. 등록)

발명의 내용

해결하려는 과제

[0011] 본 발명의 실시 예들은 다양한 자원(예컨대, 말뭉치, 표준 대사전 및 어휘 의미망)으로부터의 학습할 단어 목록, 변환된 말뭉치 및 단어 의미 데이터를 이용한 단어 임베딩 학습을 통해 학습하고 동형이의어와 인접 어절 간의 유사도를 비교하여 동형이의어를 분별함으로써, 미학습 패턴에 대해서 동형이의어를 정확하게 분별할 수 있는, 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법을 제공하고자 한다.

[0012] 본 발명의 실시 예는 종래의 단어 임베딩이 아닌 말뭉치에서 실질형태소의 의미적 정보를 이용한 단어 임베딩 학습을 수행하기 때문에 적은 학습 데이터로도 단어 임베딩 학습이 가능하며 학습 시간도 적게 걸릴 수 있는, 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법을 제공하고자 한다.

[0013] 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 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법을 제공하고자 한다.

[0014] 또한, 본 발명의 실시 예는 종래의 위치 기반의 단어 임베딩이 아닌 실질형태소와 인접한 어절과 그 인접한 어절의 단어 의미 데이터(예컨대, 상위어)를 인접 어절로 처리하여 단어 벡터를 생성함으로써, 단어들의 관계를 코사인 유사도를 통해 볼 수 있는, 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법을 제공하고자 한다.

과제의 해결 수단

[0015] 본 발명의 제1 측면에 따르면, 적어도 하나 이상의 학습용 어절이 포함된 말뭉치와 단어 의미 데이터가 포함된 표준 대사전 및 어휘 의미망을 저장하는 데이터 저장부; 상기 말뭉치로부터 학습할 단어 목록을 생성하는 단어 목록 생성부; 상기 생성된 단어 목록에 포함된 학습할 단어의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환하고, 상기 생성된 학습할 단어 목록, 상기 변환된 말뭉치 및 상기 단어 의미 데이터를 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 가공 데이터 생성부; 상기 생성된 가공 데이터를 가지고 입력력층 및 프로젝트층 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단어 임베딩 학습부; 및 상기 생성된 단어 벡터를 이용하여 동형이의어와 인접 어절 간의 유사도를 비교하고 상기 비교 결과에 따라 동형이의어를 분별하는 동형이의어 분별부를 포함하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치가 제공될 수 있다.

[0016] 상기 데이터 저장부는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치와, 용례 및 뜻풀이가 단어 의미 데이터로 포함된 표준 대사전과, 용언-명사 관계 정보, 상위어 및 반의어가 단어 의미 데이터로 포함된 어휘 의미망을 저장할 수 있다.

[0017] 상기 가공 데이터 생성부는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치, 표준 대사전에 포함된 용례 및 뜻풀이, 및 어휘 의미망에 포함된 용언-명사 관계 정보 중에서 적어도 하나를 말뭉치 형태로 변환할 수 있다.

[0018] 상기 가공 데이터 생성부는 상기 변환된 말뭉치에서 상기 학습할 단어와 인접한 어절 및 상기 인접한 어절의 상위어를 상기 학습할 단어의 인접 어절로 처리하거나, 용언-명사 관계 정보를 상기 학습할 단어의 인접 어절로 처리할 수 있다.

[0019] 상기 단어 임베딩 학습부는 상기 가공 데이터 중에서 상기 학습할 단어, 인접 어절, 용언-명사 관계 정보 및 반의어가 각각 가공된 가공 데이터를 단어 임베딩에서의 스킵-그램(Skip-Gram)의 입력층 및 출력층이 합쳐진 하나

의 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.

- [0020] 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 가공 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경할 수 있다.
- [0021] 상기 단어 임베딩 학습부는 단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습할 수 있다.
- [0022] 상기 단어 임베딩 학습부는 단어 임베딩 학습을 통해 학습할 단어와 인접하고 조사 또는 어미를 제외한 실질형태소의 단어 벡터를 생성할 수 있다.
- [0023] 상기 동형이의어 분별부는 상기 생성된 실질형태소의 단어 벡터를 이용하여 분별할 동형이의어의 실질형태소와 인접 어절의 실질형태소 간의 유사도를 비교하여 동형이의어를 분별할 수 있다.
- [0024] 한편, 본 발명의 제2 측면에 따르면, 적어도 하나 이상의 학습용 어절이 포함된 말뭉치로부터 학습할 단어 목록을 생성하는 단계; 상기 생성된 단어 목록에 포함된 학습할 단어의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환하는 단계; 상기 생성된 학습할 단어 목록, 상기 변환된 말뭉치 및 상기 단어 의미 데이터를 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성하는 단계; 상기 생성된 가공 데이터를 가지고 입출력층 및 프로젝션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성하는 단계; 및 상기 생성된 단어 벡터를 이용하여 동형이의어와 인접 어절 간의 유사도를 비교하고 상기 비교 결과에 따라 동형이의어를 분별하는 단계를 포함하는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법이 제공될 수 있다.
- [0025] 상기 말뭉치로 변환하는 단계는 용례 및 뜻풀이가 포함된 표준 대사전과, 용언-명사 관계 정보, 상위어 및 반의어가 포함된 어휘 의미망에서의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환할 수 있다.
- [0026] 상기 말뭉치로 변환하는 단계는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치, 표준 대사에 포함된 용례 및 뜻풀이, 및 어휘 의미망에 포함된 용언-명사 관계 정보 중에서 적어도 하나를 말뭉치 형태로 변환할 수 있다.
- [0027] 상기 가공 데이터를 생성하는 단계는 상기 변환된 말뭉치에서 상기 학습할 단어와 인접한 어절 및 상기 인접한 어절의 상위어를 상기 학습할 단어의 인접 어절로 처리하거나, 용언-명사 관계 정보를 상기 학습할 단어의 인접 어절로 처리할 수 있다.
- [0028] 상기 단어 벡터를 생성하는 단계는 상기 가공 데이터 중에서 상기 학습할 단어, 인접 어절, 용언-명사 관계 정보 및 반의어가 각각 가공된 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.
- [0029] 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 가공 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경할 수 있다.
- [0030] 상기 단어 벡터를 생성하는 단계는 단어 임베딩에서의 네거티브-샘플링(Negative-Sampling)을 이용해 가공 데이터 이외의 오답을 학습할 단어에 학습할 수 있다.
- [0031] 상기 단어 벡터를 생성하는 단계는 단어 임베딩 학습을 통해 학습할 단어와 인접하고 조사 또는 어미를 제외한 실질형태소의 단어 벡터를 생성할 수 있다.
- [0032] 상기 동형이의어를 분별하는 단계는 상기 생성된 실질형태소의 단어 벡터를 이용하여 분별할 동형이의어의 실질형태소와 인접 어절의 실질형태소 간의 유사도를 비교하여 동형이의어를 분별할 수 있다.

발명의 효과

- [0033] 본 발명의 실시 예들은 다양한 자원(예컨대, 말뭉치, 표준 대사전 및 어휘 의미망)으로부터의 학습할 단어 목록, 변환된 말뭉치 및 단어 의미 데이터를 이용한 단어 임베딩 학습을 통해 학습하고 동형이의어와 인접 어절

간의 유사도를 비교하여 동형이의어를 분별함으로써, 미학습 패턴에 대해서 동형이의어를 정확하게 분별할 수 있다.

[0034] 본 발명의 실시 예는 종래의 단어 임베딩이 아닌 말뭉치에서 실질형태소의 의미적 정보를 이용한 단어 임베딩 학습을 수행하기 때문에 적은 학습 데이터로도 단어 임베딩 학습이 가능하며 학습 시간도 적게 걸릴 수 있다.

[0035] 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 있다.

[0036] 또한, 본 발명의 실시 예는 종래의 위치 기반의 단어 임베딩이 아닌 실질형태소와 인접한 어절과 그 인접한 어절의 단어 의미 데이터(예컨대, 상위어)를 인접 어절로 처리하여 단어 벡터를 생성함으로써, 단어들의 관계를 코사인 유사도를 통해 볼 수 있다.

도면의 간단한 설명

[0037] 도 1은 본 발명의 실시 예에 따른 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 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 동형이의어 분별 장치에서의 변형된 스킵-그램 모델인 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 동형이의어 분별 장치에 의해 수행되는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법에 대한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0038] 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 설명한다. 본 발명에 따른 동작 및 작용을 이해하는 데 필요한 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.

[0039] 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.

[0040] 도 1은 본 발명의 실시 예에 따른 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치의 구성도이다.

[0041] 도 1에 도시된 바와 같이, 본 발명의 실시 예에 따른 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치(100)는 데이터 저장부(110), 단어 목록 생성부(120), 가공 데이터 생성부(130), 단어 임베딩 학습부(140) 및 동형이의어 분별부(150)를 포함한다.

[0042] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 단어 임베딩의 변형된 스킵-그램(Skip-Gram) 모델인 피쳐 미러(Feature Mirror) 모델을 이용해서 다양한 자원들을 학습하고 실질형태소의 단어 벡터를 생성한다.

[0043] 이를 위해, 동형이의어 분별 장치(100)는 우선 말뭉치(111)로부터 학습할 단어 목록을 생성한다. 또한, 동형이의어 분별 장치(100)는 학습용 어절이 포함된 말뭉치(111), 표준 대사전(112)의 용례를 말뭉치의 형태로 변환하여 학습한다. 또한, 동형이의어 분별 장치(100)는 어휘 의미망(113)의 용언-명사 관계를 말뭉치의 형태로 변환하여 학습한다. 추가로 동형이의어 분별 장치(100)는 어휘 의미망(113)의 상위어와 표준 대사전(112)의 뜻풀이를 이용하여 학습할 수 있다. 여기서, 어휘 의미망(113)의 반의어는 네거티브-샘플링(Negative-Sampling) 용도로 이용한다. 이용되는 모든 학습 자원들은 동형이의어 수준에서 태그가 부착되어 있다.

[0044] 동형이의어 분별 장치(100)는 이러한 단어 임베딩 학습을 통해 학습한 결과물로 형태소의 단어 벡터를

생성한다. 그리고 동형이의어 분별 장치(100)는 형태소의 단어 벡터를 이용하여 두 형태소의 유사도를 계산하여 동형이의어를 분별할 수 있다.

- [0045] 여기서, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 말뭉치(111), 표준 대사전(112) 및 어휘 의미망(113)이 구축될 수 있는 한국어뿐만 아니라 영어(WordNet), 중국어(HowNet) 등의 모든 언어에 적용 가능하다.
- [0046] 이하, 도 1의 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치(100)의 각 구성요소들의 구체적인 구성 및 동작을 설명한다.
- [0047] 데이터 저장부(110)는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치(111)와 단어 의미 데이터가 포함된 표준 대사전(112) 및 어휘 의미망(113)을 저장한다. 데이터 저장부(110)는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치(111)와, 용례 및 뜻풀이가 단어 의미 데이터로 포함된 표준 대사전(112)과, 용언-명사 관계 정보, 상위어 및 반의어가 단어 의미 데이터로 포함된 어휘 의미망(113)을 저장한다. 예컨대, 말뭉치(111)는 기설정된 말뭉치로서, 세종말뭉치가 적용될 수 있다.
- [0048] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 학습할 단어의 종류와 단어 의미 데이터로 말뭉치(111), 표준 대사전(112) 및 어휘 의미망(113)을 이용한다. 여기서, 어휘 의미망(113)은 어휘 지도로 지칭되고 한다. 어휘 의미망(113)은 표준 사전을 기반으로 명사, 용언, 부사 어휘들이 의미제약으로 상호 연결된 어휘 의미망을 의미한다.
- [0049] 단어 목록 생성부(120)는 데이터 저장부(110)에서 저장된 말뭉치(111)로부터 학습할 단어 목록을 생성한다.
- [0050] 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.
- [0051] 여기서, 가공 데이터 생성부(130)는 단어 목록 생성부(120)에서 생성된 학습할 단어 목록에 포함된 학습할 단어의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환한다.
- [0052] 구체적으로 살펴보면, 가공 데이터 생성부(130)는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치(111), 상기 표준 대사전(112)에 포함된 용례 및 상기 어휘 의미망(113)에 포함된 용언-명사 관계 정보 중에서 적어도 하나를 말뭉치 형태로 변환한다.
- [0053] 그리고 가공 데이터 생성부(130)는 변환된 말뭉치에서 상기 학습할 단어와 인접한 어절 및 상기 인접한 어절의 상위어를 상기 학습할 단어의 인접 어절로 처리하거나, 용언-명사 관계 정보를 상기 학습할 단어의 인접 어절로 처리할 수 있다.
- [0054] 그리고 가공 데이터 생성부(130)는 그 생성된 학습할 단어 목록, 변환된 말뭉치 및 단어 의미 데이터를 단어 임베딩 학습에 맞게 가공하여 가공 데이터를 생성한다.
- [0055] 단어 임베딩 학습부(140)는 가공 데이터 생성부(130)에서 생성된 가공 데이터를 가지고 입출력층 및 프로젝션층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성한다. 이때, 단어 임베딩 학습부(140)는 단어 임베딩 학습을 통해 학습할 단어와 인접하고 조사 또는 어미를 제외한 실질형태소의 단어 벡터를 생성할 수 있다.
- [0056] 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.
- [0057] 구체적으로 살펴보면, 단어 임베딩 학습부(140)는 가공 데이터 중에서 학습할 단어, 인접 어절, 용언-명사 관계 정보 및 반의어가 각각 가공된 가공 데이터를 단어 임베딩에서의 스킵-그램(Skip-Gram)의 입력층 및 출력층이 합쳐진 하나의 입출력층에 위치시킨다. 그리고 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.
- [0058] 이때, 단어 임베딩 학습부(140)는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로퍼게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행한다. 단어 임베딩 학습부(140)는 백 프로퍼게이션 과정에서 학습할 단어의 가공 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경한다.

- [0059] 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.
- [0060] 한편, 동형이의어 분별부(150)는 단어 임베딩 학습부(140)에서 생성된 단어 벡터를 이용하여 동형이의어와 인접 어절 간의 유사도를 비교하고 그 비교 결과에 따라 동형이의어를 분별한다.
- [0061] 여기서, 동형이의어 분별부(150)는 실질형태소의 단어 벡터를 이용한다. 즉, 동형이의어 분별부(150)는 단어 임베딩 학습부(140)에서 생성된 실질형태소의 단어 벡터를 이용하여 분별할 동형이의어의 실질형태소와 인접 어절의 실질형태소 간의 유사도를 비교하여 동형이의어를 분별할 수 있다.
- [0062] 이와 같이, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 말뭉치뿐만 아니라 용례, 뜻풀이 또한 단순 말뭉치로 취급하여 학습하게 된다. 동형이의어 분별 장치(100)는 변환된 말뭉치에서 학습할 단어의 바로 다음 어절이나 형태소에 나타날 단어를 가공 데이터로 생성하게 된다. 또한, 동형이의어 분별 장치(100)는 상위어, 반의어, 용언-명사 관계를 가공 데이터로 이용한다.
- [0063] 한편, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)에서는 말뭉치(111)와 표준 대사전(112)의 용례를 묶어서 하나의 말뭉치로 취급한다. 종래의 워드투벡이나 기타 알려진 단어 임베딩 모델들은 영어와 영어 말뭉치를 대상으로 한다.
- [0064] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)에서는 영어뿐만 아니라 다른 언어(예컨대, 교착어인 한국어 등)에 적용하기 위해서 각 언어의 특성에 맞게 변경하여 단어 임베딩 학습을 수행한다.
- [0065] 일례로, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)가 한국어에 적용되는 경우, 한국어에서는 '을', '를'같은 조사나 '하다', '였다'와 같은 어미들은 단어 벡터를 가지지 않게 한다. 즉, 단어 임베딩 학습부(140)는 일반명사(NNG), 동사(VV), 형용사(VA) 등의 실질형태소에 대해서 단어 벡터를 생성한다. 그리고 단어 임베딩 학습부(140)는 변환된 말뭉치에서 실질형태소만 남기고 바로 인접하는 실질형태소를 단어 임베딩 학습에 이용한다. 또한, 단어 임베딩 학습부(140)는 인접 형태소를 기설정된 횟수(예컨대, 1회)로 학습할 때마다 네거티브-샘플링을 수행할 수 있다. 단어 임베딩 학습부(140)는 무작위로 선택된 동형이의어의 실질형태소와 동일한 형태소의 다른 동형이의어에 대해서 네거티브-샘플링을 수행한다. 예컨대, 단어 임베딩 학습부(140)는 '사과_05/NNG'에 대해서 학습할 때, 이 형태소는 '사과'이며 품사가 일반명사(NNG)인 다른 형태소(사과_08/NNG 등)'와 네거티브-샘플링을 수행한다.
- [0066] 한편, 다양한 정보들이 있는 어휘 의미망(113)을 학습하는 과정에 대해서 살펴보면 다음과 같다. 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 이 중에서 상위어, 반의어, 용언-명사 관계 정보를 이용한다. 가공 데이터 생성부(130)는 변환된 말뭉치를 가공할 때 인접 어절에 대해 처리하면서 동시에 그 어절의 상위어도 인접한 것과 동일하게 처리한다. 그리고 단어 임베딩 학습부(140)는 반의어에 대해서 네거티브-샘플링을 수행한다.
- [0067] 그리고 동형이의어 분별 장치(100)는 변환된 말뭉치 전체에 대해 이런 과정이 1회 수행되고 나면 용언-명사 관계 정보를 학습할 수 있다. 동형이의어 분별 장치(100)는 용언-명사 관계망에 있는 용언과 명사는 서로 인접한 것으로 처리하여 학습할 수 있다.
- [0068] 도 2는 본 발명의 실시 예에 따른 입출력층 및 미러층으로 이루어진 학습 모델의 설명도이다.
- [0069] 도 2에 도시된 바와 같이, 본 발명의 실시 예에 따른 동형이의어 분별 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.
- [0070] 그리고 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 각 단어를 50차원의 단어 벡터로 표현한다.
- [0071] 도 2에는 본 발명의 실시 예에서 이용한 학습 모델이 나타나 있다. 동형이의어 분별 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, 동형이의어 분별 장치(100)는 실제 가중치 값을 변경하는 경우, 학습할 단어 A

와 연결된 가중치(204)만 변경하게 되고, 단어 B와 연결된 가중치(203)를 변경하지 않는다.

- [0072] 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, 동형이의어 분별 장치(100)는 뜻풀이와 상위어에 대해 출력(Output) 값으로 1로 설정해 학습하고, 반의어의 경우에만 출력 값을 0으로 설정해 학습한다.
- [0073] 또한, 동형이의어 분별 장치(100)는 네거티브-샘플링을 이용해 학습 데이터 이외의 오답을 각 단어에 학습시켜 학습의 정확률을 높일 수 있다.
- [0074] 이하, 도 3 및 도 4를 참조하여 종래의 단어 임베딩의 스킵-그램 모델과 본 발명의 실시 예에 따른 피쳐 미리 모델과의 차이점을 살펴보기로 한다.
- [0075] 도 3은 종래의 단어 임베딩에서의 스킵-그램 모델에 대한 설명도이다.
- [0076] 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.
- [0077] 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.
- [0078] 도 4는 본 발명의 실시 예에 따른 동형이의어 분별 장치에서의 피쳐 미리 모델에 대한 설명도이다.
- [0079] 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.
- [0080] 본 발명의 실시 예에 따른 피쳐 미리 모델에서의 입출력층에는 학습할 단어($X(t)$), 학습할 단어($X(t)$)의 단어 의미 데이터가 원-핫 형태로 가공되어 들어가게 된다. 여기서, 학습할 단어($X(t)$)의 단어 의미 데이터에는 학습할 단어($X(t)$)와 인접한 인접 어절($X(t+1)$), 용언-명사 관계(V-N relation) 정보, 반의어(antonym)가 포함될 수 있다. 예컨대, 단어 의미 데이터에는 단어 뜻풀이(Word definition), 상위어(hypernym) 및 반의어(antonym) 등이 포함될 수 있다.
- [0081] 단어 의미 임베딩 장치(100)는 단어 임베딩 학습의 피드포워드(feedforward) 과정과 백 프로파게이션(back propagation) 과정을 통해 단어 임베딩 학습을 수행한다. 단어 의미 임베딩 장치(100)는 백 프로파게이션 과정에서 학습할 단어의 단어 의미 데이터와 연결된 가중치 값을 변경하지 않고, 학습할 단어와 연결된 가중치 값을 변경한다.
- [0082] 즉, 본 발명의 실시 예에 따른 동형이의어 분별 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.
- [0083] 이하, 도 5를 참조하여 종래의 스킵-그램 모델과 본 발명의 실시 예에 따른 스킵-그램 모델을 이용한 학습 과정을 살펴보기로 한다.
- [0084] 도 5는 종래의 스킵-그램 모델을 이용한 형태소 단위 학습 과정에 대한 예시도이다.
- [0085] 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. 종래의 단어 임베딩 방식은 주변의 실질 형태소뿐만 아니라 인접한 조사나 어미들도 학습에 이용한다.
- [0086] 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.
- [0087] 한편, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 워드투벡(Word2Vec)의 변형 모델 즉, 스킵-그램

을 변형한 모델을 이용한다. 학습된 단어들을 단어 벡터 값을 가지고 동형이의어의 중의성을 해소하기 위한 계산을 하기와 같이 수행한다.

[0088] 실질형태소의 단어 벡터를 이용한 유사도 계산 과정은 다음과 같다.

[0089] 동형이의어 분별 장치(100)는 "사과를 먹다"라는 문장에서 동형이의어인 사과와 중의성을 해소하기 위해서는 "사과"에 해당하는 모든 단어 벡터 값을 찾는다. 여기서, 모든 단어 벡터 값에는 사과(05)(사과나무의 열매), 사과(08)(자기 잘못을 인정하고 용서를 빚) 등이 포함된다.

[0090] 그리고 동형이의어 분별 장치(100)는 먹다(음식을 배 속으로 들이키다)의 단어 벡터와 하기의 [수학식 1]과 같은 코사인 유사도(Cosine Similarity)를 이용하여 값을 구한다.

수학식 1

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

[0091]

[0092] 여기서, A 및 B는 두 개의 단어 A 및 B 각각의 단어 벡터를 나타낸다.

[0093] 이후, 동형이의어 분별 장치(100)는 가장 높은 값을 가진 사과의 벡터를 "사과를 먹다"라는 문장에서 사용된 "사과"라는 단어의 의미로 결정한다.

[0094] 예를 들어서, “배가 썩는다”에서 ‘배’라는 동형이의어의 중의성을 해소하는 과정을 살펴보기로 한다.

[0095] 이 문장 자체는 세종말뭉치에 없기 때문에 미학습 패턴에 속한다. 정확히 말하자면, 열매라는 의미의 '배'와 '썩다'라는 표현이 인접하는 경우가 세종말뭉치에 없다. 종래의 학습 방법들로는 이러한 미학습 패턴이 발생하면 정확률이 현저히 낮아진다.

[0096] 하지만, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)에서 생성된 단어 벡터를 이용하는 경우에는 정확률의 하락 정도가 적다. 왜냐하면, 열매 '배'의 단어 벡터 방향이 '썩다'와 인접하는 경우가 있는 '열매'나 '사과', '음식'의 벡터와 비슷하기 때문이다.

[0097] 비슷해지도록 학습되는 근거는 어휘 의미망(113)에서 '배'의 상위어에 '열매'가 나타나기 때문이다. 그 외에도 '배'의 인접어가 '음식'의 인접어(예컨대, 먹다, 삼키다 등)와 비슷하기 때문에 단어 벡터가 유사해지도록 학습되는 것이다. '사과'나 '음식' 등의 단어 벡터는 동사 '썩다'와 인접하는 경우가 있다. 인접하는 단어끼리는 그들의 벡터가 유사해지게 된다. 결국, 열매 '배'의 단어 벡터가 '썩다'와 유사해지게 된다.

[0098] 그리고 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 '배'의 다른 의미들(예컨대, 선박, 복부 등)의 단어 벡터가 서로 방향으로 멀어지도록 네거티브-샘플링하는 과정을 수행한다. 그래서 선박을 뜻하는 '배'의 단어 벡터는 열매 '배'의 단어 벡터와 크게 다르게 된다. 결국, 선박을 뜻하는 '배'의 단어 벡터는 '썩다' 벡터와 유사도가 낮아지게 된다.

[0099] 즉, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 단어 임베딩 학습을 통해 생성된 단어 벡터를 동형이의어 분별 과정에 이용함으로써, 미학습 패턴에 강건한 경향이 있다. 또한, 동형이의어 분별 장치(100)는 어휘 의미망(113)을 추가로 학습하고, 의미번호만 다른 동형이의어끼리 네거티브-샘플링을 한다는 점(예를 들어, 열매 '배'는 선박 '배'와 네거티브-샘플링) 때문에 더욱 정확하게 동형이의어를 분별할 수 있다.

[0100] 한편, 피쳐 미리 모델을 이용하여 "사과(먹는 사과)"라는 학습할 단어를 학습하는 과정을 살펴보기로 한다.

[0101] 동형이의어 분별 장치(100)는 사과와 뜻풀이인 사과나무의 열매에서 명사와 용언을 추출한 "사과나무", "열매"라는 단어 의미 데이터들 가공 데이터로 이용한다. 그리고 동형이의어 분별 장치(100)는 "사과"라는 학습할 단어의 상위어인 "과일"을 사과와 가공 데이터로 추가한다. 여기서, 학습할 단어 "사과"의 반의어가 없기 때문에 반의어를 가공 데이터는 추가하지 않는다. 그렇지만, 반의어가 있는 경우에는 반의어를 추가한다.

[0102] "사과"라는 학습할 단어의 가공 데이터로 "사과나무", "열매", "과일"을 찾은 후 오른쪽에 학습 모델을 이용하

여 단어 임베딩 학습을 수행한다.

[0103] 일례로, "사과"라는 학습할 단어와 뜻풀이의 "사과나무"라는 단어를 이용한 단어 임베딩 학습 과정을 하기의 [수학식 2] 및 [수학식 3]을 참조하여 살펴보기로 한다.

수학식 2

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

[0104]

[0105] 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 3

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

[0106]

[0107] 상기의 [수학식 2]는 사과 및 사과나무 간의 출력값을 계산하는 수식을 나타낸다.

[0108] 상기의 [수학식 2]에 따라 사과의 벡터와 사과나무의 벡터를 곱한 후 모두를 더한 값에 상기의 [수학식 3]에 나타난 시그모이드 함수(Sigmoid Function)를 이용하면 출력 값이 된다.

수학식 4

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

[0109]

[0110] 여기서, E 는 오차 값, $d(\text{사과나무})$ 는 가공 데이터 상의 정답 값을 나타낸다.

[0111] 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.

[0112] 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 연결된 가중치에만 변화량을 적용하고, 사과나무와 연결된 가중치에는 변화량을 적용하지 않는다.

[0113] 도 6은 본 발명의 실시 예에 따른 동형이의어 분별 장치에 의해 수행되는 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 방법에 대한 흐름도이다.

[0114] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 적어도 하나 이상의 학습용 어절이 포함된 말뭉치로부터 학습할 단어 목록을 생성한다(S101).

[0115] 그리고 동형이의어 분별 장치(100)는 생성된 단어 목록에 포함된 학습할 단어의 단어 의미 데이터 중에서 변환 가능한 단어 의미 데이터를 말뭉치로 변환한다(S102). 이때, 동형이의어 분별 장치(100)는 단어 의미 데이터를 데이터 저장부(110)에서 저장된 말뭉치(111), 표준 대사전(112) 및 어휘 의미망(113)으로부터 가져온다.

[0116] 이어서, 동형이의어 분별 장치(100)는 생성된 학습할 단어 목록, 변환된 말뭉치 및 단어 의미 데이터를 단어 임

배딩 학습에 맞게 가공하여 가공 데이터를 생성한다(S103).

[0117] 이후, 동형이의어 분별 장치(100)는 생성된 가공 데이터를 가지고 입출력층 및 프로젝션 층으로 이루어진 학습 모델을 이용한 단어 임베딩 학습을 통해 학습할 단어를 학습하여 단어 벡터를 생성한다(S104).

[0118] 그리고 동형이의어 분별 장치(100)는 생성된 단어 벡터를 이용하여 동형이의어와 인접 어절 간의 유사도를 비교한다(S105).

[0119] 이어서, 동형이의어 분별 장치(100)는 S105 과정에서의 비교 결과에 따라 동형이의어를 분별한다(S106).

[0120] 한편, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)에 의해 실험된 실험 과정 및 결과를 하기의 [표 1]을 참조하여 살펴보기로 한다.

[0121] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 학습한 적이 없는 패턴을 처리하기 위해서 동형이의어 형태소를 단어 벡터로 벡터화하고, 그 동형이의어와 인접 어절의 실질형태소 간의 유사도를 비교한다.

[0122] 여기서, 단어를 벡터화하는 과정에서 학습용 말뭉치(111), 표준 대사전(112) 및 어휘 의미망(113)이 이용된다. 하기 미학습 패턴들에 대한 소규모의 실험에서 충분히 의미가 있는 결과가 하기의 [표 1]과 같이 나타난다.

표 1

[0123]

분별할 명사	인접 형태소	정답
사과	열매_01/NNG	사과_05/NNG
배	씨/VV	배_03/NNG
면	삶/VV	면_08/NNG

[0124] [표 1]은 간이 실험용 테스트세트의 일부를 나타낸다.

[0125] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)가 의미가 있는지를 간단하게 확인하기 위해 간이 실험 환경을 구성하였다.

[0126] 먼저, 실험을 위해, 동형이의어가 여러 가지이면서 각각의 동형이의어가 고르게 나타나는 즉, 베이스라인이 비교적 낮은 명사를 선택한다(예컨대, 사과).

[0127] 그리고 '사과'와 인접할 수 있지만 학습용 말뭉치(111)에서는 인접한 적이 없는 형태소(예컨대, 열매)를 정한다. [표 1]은 실험에 사용한 테스트세트에서 3개만 예로 표시한 것이다.

[0128] 일례로, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 '사과'와 '열매_01/NNG'를 입력받게 되고, '사과'의 동형이의어 중에서 '열매'와 가장 유사한 것이 '사과_05/NNG'로 계산되면 정답을 맞춘 것이다.

[0129] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 총 24개 중에 18개를 맞추었다. 비록 테스트세트가 아주 작지만 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)가 의미가 있음을 알 수 있다.

[0130] 이와 같이, 본 발명의 실시 예에서는 한국어뿐만 아니라 영어, 중국어 등 모든 언어에 적용가능한 동형이의어 분별을 위해서 단어 임베딩 학습을 통해 학습하고 단어 벡터들 간의 유사도 비교를 통해 동형이의어를 분별한다.

[0131] 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 단어 임베딩 학습을 위해 말뭉치(111)뿐만 아니라 표준 대사전(112), 어휘 의미망(113)을 이용한다. 상기의 [표 1]과 같이 미학습 패턴만으로 간단한 실험을 하였으며 의미 있는 결과를 보였다.

[0132] 미학습패턴은 종래의 방법으로는 매우 낮은 베이스라인 수준의 정확률만을 기대할 수 있다. 따라서 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 종래의 동형이의어 분별기가 재현하지 못하는 패턴에 대해서 후속 보완 모듈로도 이용될 수 있다. 또한, 후속 보완 모듈로 이용된다면 전체적인 정확률을 향상시킬 수 있다. 즉, 동형이의어 분별 장치(100)가 적용되면, 미학습패턴에 대해서도 안정적으로 작동하는 동형이의어 분별기가 제공될 수 있다. 이와 같이, 본 발명의 실시 예에 따른 동형이의어 분별 장치(100)는 종래의 동형이의어 분별기(UTagger)에 통합될 수 있다.

[0133] 또한, 동형이의어 분별 장치(100)는 의존명사, 부사, 접미사, 접두사 등에 대해서도 단어 벡터를 생성할 수 있다. 이들은 일반명사나 용언과는 다른 성격을 가지고 있기 때문에 동형이의어 분별 장치(100)는 그 성격에 따라

변경하여 단어 임베딩 학습을 수행할 수 있다.

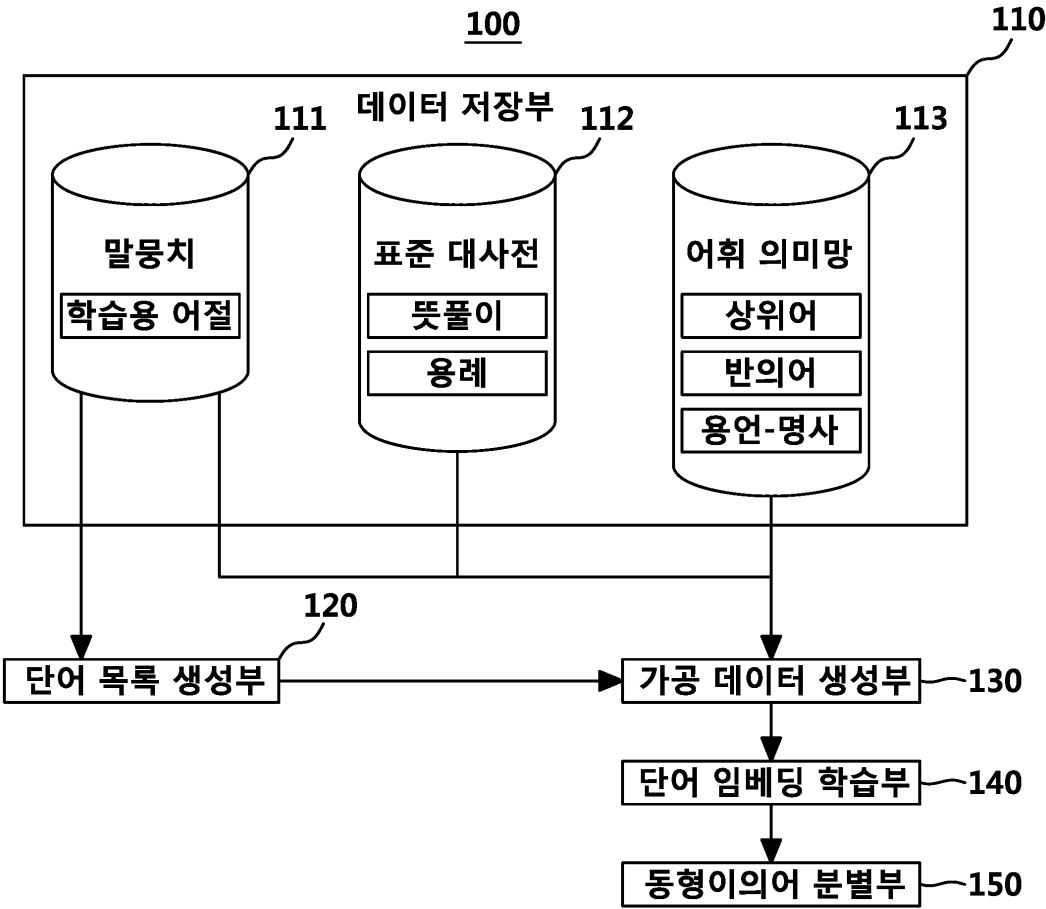
[0134] 이상에서 설명한 실시 예들은 그 일 예로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명에 개시된 실시 예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시 예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

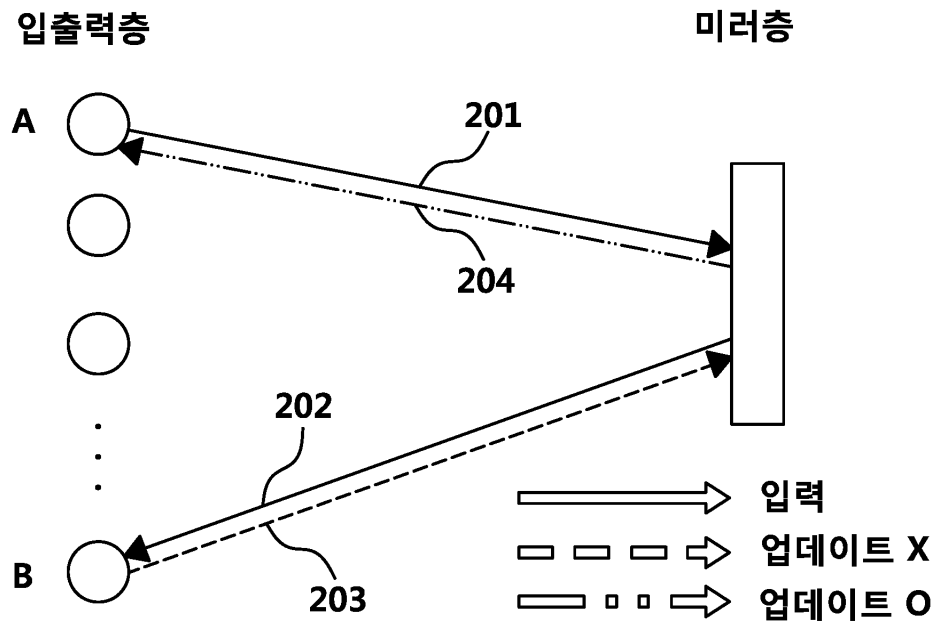
[0135] 100: 동형이의어 분별 장치
110: 데이터 저장부
111: 말뭉치
112: 표준 대사전
113: 어휘 의미망
120: 단어 목록 생성부
130: 가공 데이터 생성부
140: 단어 임베딩 학습부
150: 동형이의어 분별부

도면

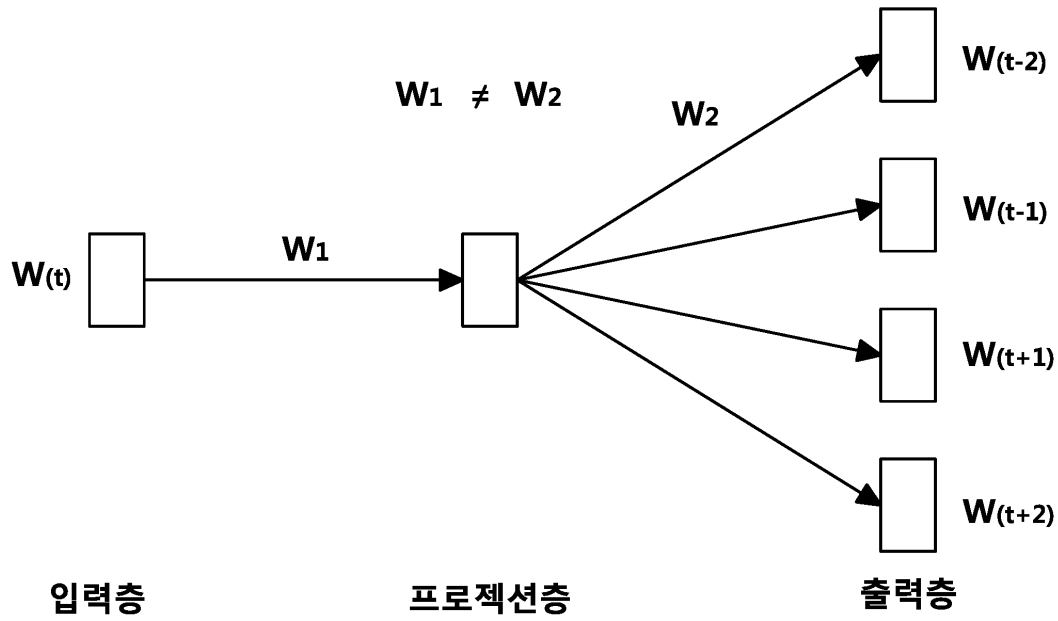
도면1



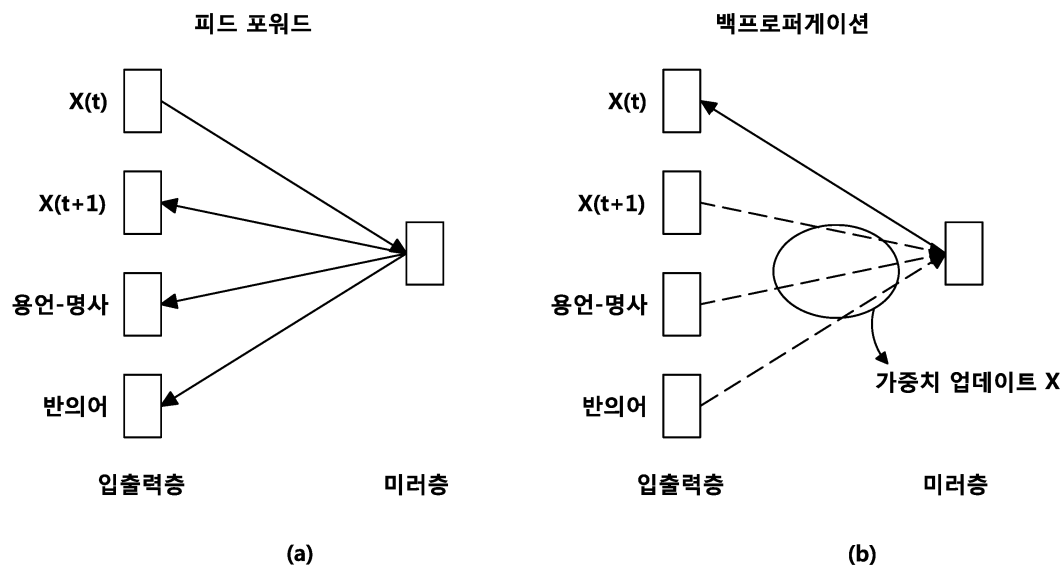
도면2



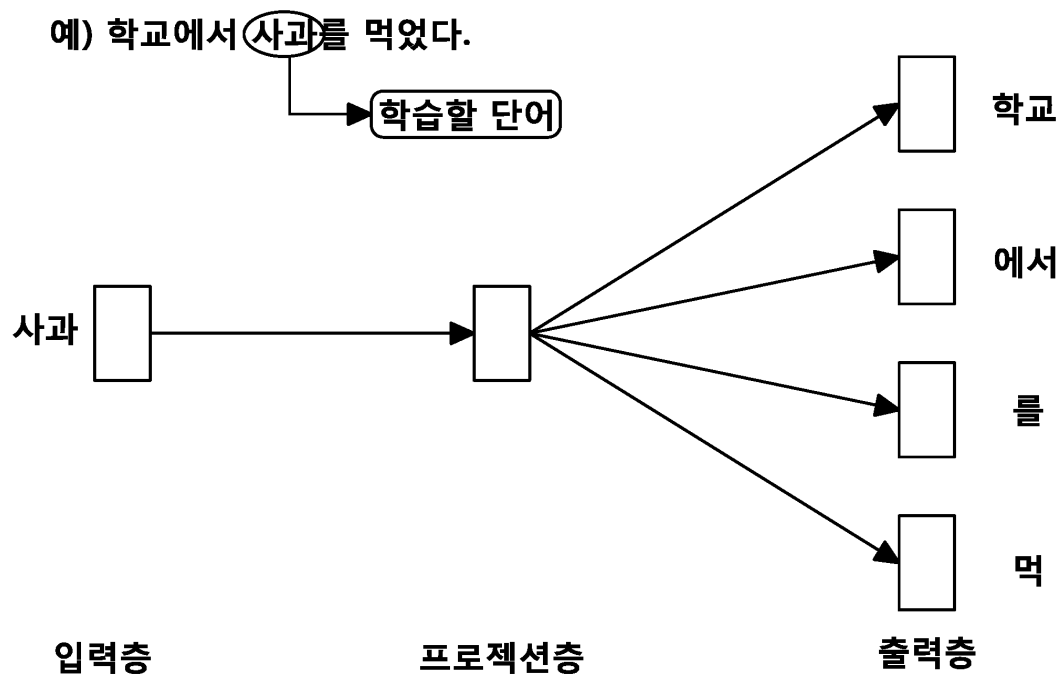
도면3



도면4



도면5



도면6

