(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년06월07일
(11) 등록번호 10-2260646
(24) 등록일자 2021년05월31일

(51) 국제특허분류(Int. Cl.)
G06F 40/40 (2020.01) *G06N 3/08* (2006.01)
(52) CPC특허분류
G06F 40/40 (2020.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2019-0080967
(22) 출원일자 2019년07월04일
심사청구일자 2019년07월04일
(65) 공개번호 10-2020-0040652
(43) 공개일자 2020년04월20일
(30) 우선권주장
1020180120630 2018년10월10일 대한민국(KR)
(56) 선행기술조사문헌
KR1020180008199 A
KR1020180001889 A
Yoon Kim 외, 'Character-Aware Neural Language Models', 2015.12.01.*
Zhenisbek Assylbekov 외, 'Reusing Weights in Subword-aware Neural Laguage Models', 2018.04.25.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이상근
서울특별시 강남구 선릉로 221, 205동 1104호 (도곡동, 도곡렉슬아파트)
김예찬
경기도 남양주시 늘을1로 55-18, 204동 01호 (호평동, 삼우H타운)
(74) 대리인
특허법인엠에이피에스

전체 청구항 수 : 총 17 항 심사관 : 홍경아

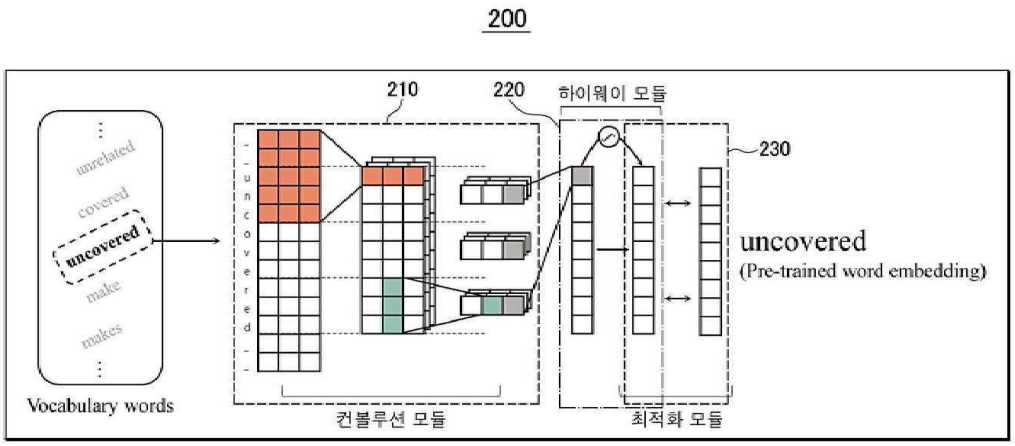
(54) 발명의 명칭 **자연어 처리 시스템 및 자연어 처리에서의 단어 표현 방법**

(57) 요 약

본 발명은 자연어 처리 시스템 및 자연어 처리에서의 단어 표현 방법에 관한 것으로서, 자연어 처리 시스템에 의해 수행되는 자연어 처리에서의 단어 표현 방법에 있어서, a) 적어도 하나 이상의 단어를 포함하는 어휘 및 각 단어에 대해 기학습된 단어 임베딩 정보를 포함하는 어휘 사전 데이터세트를 제공하는 단계; b) 상기 어휘 사전

(뒷면에 계속)

대 표 도



데이터세트에 기초한 어휘가 입력 데이터로 제공되면, 단어 표현 모델을 이용하여 상기 입력 데이터에 존재하는 단어들에 대한 하위 단어(subword) 정보를 추출하고, 상기 하위 단어 정보를 단어 임베딩 벡터를 산출하는 단계; 및 c) 상기 산출된 단어 임베딩 벡터와 해당 단어의 기학습된 단어 임베딩 정보를 매칭함으로써 상기 기학습된 단어 임베딩 정보를 상기 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습하는 단계를 포함하되, 상기 단어 표현 모델은, 상기 하위 단어 정보를 이용하여 하위 단어 특징 벡터들을 산출하는 합성곱 신경망(convolutional neural network) 기반의 컨볼루션 모듈과, 상기 컨볼루션 모듈에서 산출된 하위 단어 특징 벡터들을 적응적으로 결합하여 해당 단어의 단어 임베딩 벡터를 산출하는 하이웨이 네트워크(highway network) 기반의 하이웨이 모듈을 포함하는 것이다.

이 발명을 지원한 국가연구개발사업

과제고유번호	2018R1A2A1A05078380
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	도약연구지원사업
연구과제명	온디바이스 텍스트 인공지능 개발
기 여 율	1/1
과제수행기관명	고려대학교
연구기간	2018.09.01 ~ 2019.02.28

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

자연어 처리 시스템에 의해 수행되는 자연어 처리에서의 단어 표현 방법에 있어서,

a) 적어도 하나 이상의 단어를 포함하는 어휘 및 각 단어에 대해 기학습된 단어 임베딩 정보를 포함하는 어휘 사전 데이터세트를 제공하는 단계;

b) 상기 어휘 사전 데이터세트에 기초한 어휘가 입력 데이터로 제공되면, 단어 표현 모델을 이용하여 상기 입력 데이터에 존재하는 단어들에 대한 하위 단어(subword) 정보를 추출하고, 상기 하위 단어 정보를 단어 임베딩 벡터를 산출하는 단계;

c) 상기 산출된 단어 임베딩 벡터와 해당 단어의 기학습된 단어 임베딩 정보를 매칭함으로써 상기 기학습된 단어 임베딩 정보를 상기 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습하는 단계;

d) 상기 학습된 단어 표현 모델에 미등록 단어(Out of Vocabulary)가 입력 데이터로 제공되면, 상기 미등록 단어에 대해 하위 단어 정보를 추출한 후 상기 추출된 하위 단어 정보를 이용하여 미등록 단어의 단어 임베딩 벡터를 산출하는 단계; 및

e) 상기 산출된 미등록 단어의 단어 임베딩 벡터에 기초한 벡터 연산을 통해 단어 임베딩 벡터 간 유사도를 계산하여 상기 미등록 단어의 이웃 단어를 추출하여 상기 미등록 단어의 고유 의미를 추론하는 단계를 포함하되,

상기 단어 표현 모델은,

상기 하위 단어 정보를 이용하여 하위 단어 특징 벡터들을 산출하는 합성곱 신경망(convolutional neural network) 기반의 컨볼루션 모듈과, 상기 컨볼루션 모듈에서 산출된 하위 단어 특징 벡터들을 적응적으로 결합하여 해당 단어의 단어 임베딩 벡터를 산출하는 하이웨이 네트워크(highway network) 기반의 하이웨이 모듈을 포함하는 것인,

자연어 처리에서의 단어 표현 방법.

청구항 2

제 1 항에 있어서,

상기 b) 단계는,

b-1) 상기 입력 데이터에 포함된 모든 단어에서 하위 단어들을 추출하고, 상기 추출된 하위 단어에 개별 코드를 부여한 후 해당 단어를 구성하는 코드를 연결하여 시퀀스 표현을 생성하는 단계;

b-2) 상기 컨볼루션 모듈에서 상기 시퀀스 표현과의 합성을 통해 해당 단어에 존재하는 하위 단어 정보를 추출하는 단계;

b-3) 상기 추출된 하위 단어 정보에 풀링(pooling) 연산을 적용하여 유의미한 하위 단어 특징들을 추출하는 단계; 및

b-4) 상기 추출된 하위 단어 특징들을 모두 연결하여 합성곱을 통한 하위 단어 특징 벡터를 산출하는 단계를 포함하는 것인, 자연어 처리에서의 단어 표현 방법.

청구항 3

제 2 항에 있어서,

상기 b) 단계는,

상기 하위 단어를 문자(Character)로 설정한 경우, 미등록 단어(Out of Vocabulary)를 포함한 모든 단어에 대해 문자 기반의 시퀀스 표현을 생성하고,

상기 문자 기반의 시퀀스 표현을 이용하여 등록 단어의 단어 임베딩 벡터 및 미등록 단어의 단어 임베딩 벡터를 산출하는 것인, 자연어 처리에서의 단어 표현 방법.

청구항 4

제 3 항에 있어서,

상기 c) 단계는,

상기 등록 단어의 단어 임베딩 벡터와 해당 등록 단어의 기학습된 단어 임베딩 정보를 매칭함으로써 상기 기학습된 단어 임베딩을 상기 등록 단어의 단어 임베딩 벡터로 대체하여 해당 등록 단어에 대한 단어 표현을 학습하고,

상기 미등록 단어의 단어 임베딩 벡터를 이용하여 해당 미등록 단어의 단어 표현을 학습하는 것인, 자연어 처리에서의 단어 표현 방법.

청구항 5

제 2 항에 있어서,

상기 b-1)은 원-핫 인코딩(One-hot encoding) 을 적용하여 하기 수학식 1에 의해 상기 시퀀스 표현을 나타내는 것인, 자연어 처리에서의 단어 표현 방법.

[수학식 1]

$$r = c_1 \oplus c_2 \oplus \cdots \oplus c_n = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}$$

r: 단어 표현, $r \in \mathbb{R}^{|V_c| \times n}$

Vc: 어휘

$\oplus$: 연결 연산자

c_i: 해당 단어에서 i번째 문자를 표현하기 위한 원-핫 인코딩

청구항 6

제 2 항에 있어서,

상기 b-2)는 하기 수학식 2를 통해 상기 시퀀스 표현과 합성하여 상기 하위 단어 정보를 추출하는 것인, 자연어 처리에서의 단어 표현 방법.

[수학식 2]

$$s_i = tanh(F \cdot r_{i:i+h-1} + b)$$

F: 합성곱 필터, $F \in \mathbb{R}^{|V_c| \times h}$

Vc: 어휘

h: F의 폭

$r_{i:i+h-1}$: 문자 c_i에서 c_{i+h-1} 사이의 시퀀스 표현

b: 바이어스

tanh: 합성곱 결과 값들에 대한 비선형 함수

청구항 7

제 2 항에 있어서,

상기 b-3)은 하기 수학식 3에 의한 스트라이드 풀링(stride pooling) 연산을 적용하여 상기 하위 단어 특징 벡터를 추출하고, 상기 하위 단어 특징 벡터들은 접미사, 어근 및 접두어를 포함하는 것인, 자연어 처리에서의 단어 표현 방법.

[수학식 3]

$$e_i = \max\left[s_{k\cdot i:k\cdot(i+1)-1}\right]$$

S: 하위 단어 특징 벡터

k: 스트라이드의 길이

$s_{ki:k(i+1)-1}$: s에 하나의 스트라이드를 갖는 시퀀스 표현

청구항 8

제 1 항에 있어서,

상기 하이웨이 모듈은 하기 수학식 4에 의한 게이트 메커니즘을 사용하여 상기 하위단어 특징 벡터를 라우트하는 것인, 자연어 처리에서의 단어 표현 방법.

[수학식 4]

$$y = T \odot H + C \odot e$$

H: 입력(e)에서의 비선형 변환

T: 변환 게이트

C: 이동 게이트

청구항 9

제 8 항에 있어서,

상기 수학식 4에서 이동 게이트(C)를 (1-T)로 단순화하여 하기 수학식 5로 나타내는 것인, 자연어 처리에서의 단어 표현 방법.

[수학식 5]

$$T = \sigma(W_t \cdot e + b_t)$$
$$H = \tanh(W_h \cdot e + b_h)$$

W_t, W_h: 정방 행렬(square matrices)

b_t, b_h: 바이어스

tanh: 결과 값들에 대한 비선형 함수

청구항 10

제 1 항에 있어서,

상기 단어 표현 모델은 상기 기학습된 단어 임베딩 정보와 동일한 크기의 단어 임베딩 벡터를 생성하기 위해 하기 수학식 6을 통해 상기 산출된 단어 임베딩 벡터에 선형 변환을 수행하는 것인, 자연어 처리에서의 단어 표현 방법.

[수학식 6]

$$w = W \cdot y + b$$

w: 최종 단어 표현

W, b: 선형변환의 매개 변수

y: 결과 벡터

청구항 11

제 1 항에 있어서,

상기 c) 단계는,

상기 산출된 단어 임베딩 벡터와 기학습된 단어 임베딩 정보간의 코사인 유사도, L 1 거리(L1 Distance or Manhattan Distance) 또는 L2 거리(L2 Distance or Euclidean Distance) 중 어느 하나의 유사도 계산 방식을 사용하여 매칭하는 것인, 자연어 처리에서의 단어 표현 방법.

청구항 12

제 11 항에 있어서,

상기 c) 단계는, 하기 수학식 7에 의한 L2 손실(loss) 함수를 이용하여 상기 산출된 단어 임베딩 벡터와 기학습된 단어 임베딩 정보간의 유사도를 산출하는 것인, 자연어 처리에서의 단어 표현 방법.

[수학식 7]

$$E = \sum_{v \in V_w} \|w_v - \hat{w}_v\|^2$$

V_w: 어휘 사전 데이터세트 내의 어휘

w_v: 산출된 단어 임베딩 벡터

$\hat{w}_v$

: 기학습된 단어 임베딩 벡터

청구항 13

삭제

청구항 14

제 1 항에 있어서,

상기 d) 단계는,

상기 하위 단어를 문자(Character)로 설정한 경우, 상기 미등록 단어(Out of Vocabulary)에 대해 문자 기반의 시퀀스 표현을 생성하고, 상기 문자 기반의 시퀀스 표현을 이용하여 미등록 단어의 단어 임베딩 벡터를 산출하는 것인, 자연어 처리에서의 단어 표현 방법.

청구항 15

제 14 항에 있어서,

상기 단어 표현 모델은,

상기 컨볼루션 모듈에서 상기 문자 기반의 시퀀스 표현과의 합성을 통해 상기 미등록 단어에 존재하는 하위 단어 정보를 추출하고, 상기 추출된 하위 단어 정보에 풀링(pooling) 연산을 적용하여 적어도 하나 이상의 하위 단어 특징들을 추출하며, 상기 하위 단어 특징들을 연결하여 하위 단어 특징 벡터를 산출하고,

상기 하이웨이 모듈에서 산출된 상기 미등록 단어의 하위단어 특징 벡터를 게이트 메커니즘을 사용하여 상기 기학습된 단어 임베딩 정보와 연관시키는 것인, 자연어 처리에서의 단어 표현 방법.

청구항 16

제 1 항에 있어서,

상기 e) 단계는, 최근접 이웃 탐색(nearest-neighbor search) 알고리즘을 이용하여 상기 이웃 단어를 추출하고, 상기 단어 임베딩 벡터간 유사도의 내림차순으로 상기 이웃 단어 내의 단어 표현을 정렬하는 것인, 자연어 처리에서의 단어 표현 방법.

청구항 17

어휘에 포함된 단어의 분산된 표현을 위한 자연어 처리 시스템에 있어서,

자연어 처리에서의 단어 표현 방법을 수행하기 위한 프로그램이 기록된 메모리; 및

상기 프로그램을 실행하기 위한 프로세서를 포함하며,

상기 프로세서는, 상기 프로그램의 실행에 의해,

적어도 하나 이상의 단어를 포함하는 어휘 및 각 단어에 대해 기학습된 단어 임베딩 정보를 포함하는 어휘 사전 데이터세트에 기초한 어휘가 입력 데이터로 제공되고, 단어 표현 모델을 이용하여 상기 입력 데이터에 존재하는 단어들에 대한 하위 단어(subword) 정보를 추출하고, 상기 하위 단어 정보를 단어 임베딩 벡터를 산출하며, 상기 산출된 단어 임베딩 벡터와 해당 단어의 기학습된 단어 임베딩 정보를 매칭함으로써 상기 기학습된 단어 임베딩 정보를 상기 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습하고, 상기 학습된 단어 표현 모델에 미등록 단어(Out of Vocabulary)가 입력 데이터로 제공되면, 상기 미등록 단어에 대해 하위 단어 정보를 추출한 후 상기 추출된 하위 단어 정보를 이용하여 미등록 단어의 단어 임베딩 벡터를 산출하고, 상기 산출된 미등록 단어의 단어 임베딩 벡터에 기초한 벡터 연산을 통해 단어 임베딩 벡터 간 유사도를 계산하여 상기 미등록 단어의 이웃 단어를 추출하여 상기 미등록 단어의 고유 의미를 추론하되,

상기 단어 표현 모델은,

상기 하위 단어 정보를 이용하여 하위 단어 특징 벡터들을 산출하는 합성곱 신경망(convolutional neural network) 기반의 컨볼루션 모듈과, 상기 컨볼루션 모듈에서 산출된 하위 단어 특징 벡터들을 적응적으로 결합하여 해당 단어의 단어 임베딩 벡터를 산출하는 하이웨이 네트워크(highway network) 기반의 하이웨이 모듈을 포함하는 것인, 자연어 처리 시스템.

청구항 18

제 17 항에 있어서,

상기 단어 표현 모델은

상기 산출된 단어 임베딩 벡터와 기학습된 단어 임베딩 정보간의 코사인 유사도, L 1 거리(L1 Distance or Manhattan Distance) 또는 L2 거리(L2 Distance or Euclidean Distance) 중 어느 하나의 유사도 계산 방식을 사용하여 매칭하여 상기 기학습된 단어 임베딩 정보를 상기 산출된 단어 임베딩 벡터로 재구성하는 최적화 모듈을 더 포함하는 것인, 자연어 처리 시스템.

청구항 19

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 기학습된 단어 임베딩의 지도 학습을 기반으로 미등록 단어(Out Of Vocabulary, OOV)를 비롯한 모든 단어에 대한 단어 표현을 생성하는 자연어 처리 시스템 및 자연어 처리에서의 단어 표현 방법에 관한 것이다.

배 경 기 술

[0002] 딥러닝 기술이 컴퓨터 비전 시스템의 큰 발전을 가져옴에 따라 딥러닝을 이용한 자연어처리 시스템에 관한 연구도 급속도로 진행되고 있다. 딥러닝 기술을 이용한 자연어처리 시스템은 단어를 수치화하기 위해 단어를 저차원의 벡터로 임베딩하여 사용해야 한다.

[0003] 이때, 자연어 처리(Natural Language Processing)는 컴퓨터가 인간 언어(human or natural language)를 이해할 수 있는 구문적/의미적 표상을 연구하는 것이고, 단어 임베딩 기술은 신경망에 기반한 언어 모델로부터 도출된 기술로 유사한 단어들을 벡터 공간상에 가깝게 배치하여 어휘 의미를 표현할 수 있는 기술이다.

[0004] 언어 모델(Language Model)은 주어진 문장에서 앞선 단어들을 기초로 다음 단어가 나올 확률을 계산해주는 모델이다. 언어 모델은 어떤 문장이 실제로 존재할 확률을 계산해주기 때문에 주어진 문장이 문법적으로 또는 의미적으로 얼마나 적합한지를 결정할 수 있다.

[0005] 언어 모델은 음성 인식, QA(Question Answering), 자연어 처리, 예측 텍스트(predictive text), 번역 및 통역 등의 분야에 적용될 수 있는데, 개방 어휘(Open Vocabulary) 환경에서는 학습 데이터의 어휘에 속하지 않는 단어(Out Of Vocabulary, OOV), 즉 미등록 단어의 처리가 필요하다. OOV를 처리하기 위해, 미등록 단어를 의사 단어(Pesudo word)로 사용하거나, 단어가 사용된 문맥을 이용하여 미등록 단어에 대한 임시 표현을 생성하는 방법을 사용하였다.

[0006] 기존의 미등록 단어의 처리 방법은 미등록 단어가 소량으로 존재하는 경우에 좋은 해결책이 될 수 있지만, 미등록 단어의 수가 많아질 경우에 자연어 처리 시스템이 텍스트를 제대로 분석하지 못하는 결과를 초래하는 문제점이 있다.

[0007] 이러한 문제점은 소셜 미디어 환경에서 더욱 명확하게 드러난다. 소셜 미디어에서 사용자들이 소비하는 단어들은 축약어, 합성어, 신조어, 오타 등 일반적인 단어의 형태와 다르게 사용되고 있고, 이러한 단어들은 자연어처리 시스템이 지니고 있는 단어 임베딩에 존재하지 않을 확률이 매우 높다.

[0008] 따라서, 자연어처리 시스템이 처리해야 할 단어의 수가 많거나 신조어 등이 빈번하게 발생하는 개방 어휘(Open Vocabulary) 환경에서 미등록 단어들 각각에 대한 표현을 생성하고, 미등록 단어가 지니고 있는 고유한 의미를 추론할 수 있는 언어 모델의 개발이 요구된다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 대한민국등록특허 제10-1799681호(발명의 명칭 : 어휘 의미망 및 단어 임베딩을 이용한 동형이의어 분별 장치 및 방법)

발명의 내용

해결하려는 과제

[0010] 본 발명은 전술한 문제점을 해결하기 위하여, 본 발명의 일 실시예에 따라

[0011] 기존의 미등록 단어들에 대한 고유한 정보를 전혀 고려하지 않은 채 주위 단어를 통해 그 의미를 파악하였으나, 미등록 단어뿐만 아니라 모든 단어의 고유한 의미를 추론하기 위해 해당 단어의 하위단어정보를 이용하여 기학습된 단어 임베딩의 지도 학습을 기반으로 고유한 단어 표현을 생성하는 것에 목적이 있다.

[0012] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0013] 상기한 기술적 과제를 달성하기 위한 기술적 수단으로서 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법은, 자연어 처리 시스템에 의해 수행되는 자연어 처리에서의 단어 표현 방법에 있어서, a) 적어도 하나 이상의 단어를 포함하는 어휘 및 각 단어에 대해 기학습된 단어 임베딩 정보를 포함하는 어휘 사전 데이터세

트를 제공하는 단계; b) 상기 어휘 사전 데이터세트에 기초한 어휘가 입력 데이터로 제공되면, 단어 표현 모델을 이용하여 상기 입력 데이터에 존재하는 단어들에 대한 하위 단어(subword) 정보를 추출하고, 상기 하위 단어 정보를 단어 임베딩 벡터를 산출하는 단계; 및 c) 상기 산출된 단어 임베딩 벡터와 해당 단어의 기학습된 단어 임베딩 정보를 매칭함으로써 상기 기학습된 단어 임베딩 정보를 상기 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습하는 단계를 포함하되, 상기 단어 표현 모델은, 상기 하위 단어 정보를 이용하여 하위 단어 특징 벡터들을 산출하는 합성곱 신경망(convolutional neural network) 기반의 컨볼루션 모듈과, 상기 컨볼루션 모듈에서 산출된 하위 단어 특징 벡터들을 적응적으로 결합하여 해당 단어의 단어 임베딩 벡터를 산출하는 하이웨이 네트워크(highway network) 기반의 하이웨이 모듈을 포함하는 것이다.

[0014] 본 발명의 다른 일 실시예에 따른 자연어 처리 시스템은, 어휘에 포함된 단어의 분산된 표현을 위한 자연어 처리 시스템에 있어서, 자연어 처리에서의 단어 표현 방법을 수행하기 위한 프로그램이 기록된 메모리; 및 상기 프로그램을 실행하기 위한 프로세서를 포함하며, 상기 프로세서는, 상기 프로그램의 실행에 의해, 적어도 하나 이상의 단어를 포함하는 어휘 및 각 단어에 대해 기학습된 단어 임베딩 정보를 포함하는 어휘 사전 데이터세트에 기초한 어휘가 입력 데이터로 제공되고, 단어 표현 모델을 이용하여 상기 입력 데이터에 존재하는 단어들에 대한 하위 단어(subword) 정보를 추출하고, 상기 하위 단어 정보를 단어 임베딩 벡터를 산출하며, 상기 산출된 단어 임베딩 벡터와 해당 단어의 기학습된 단어 임베딩 정보를 매칭함으로써 상기 기학습된 단어 임베딩 정보를 상기 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습하되, 상기 단어 표현 모델은, 상기 하위 단어 정보를 이용하여 하위 단어 특징 벡터들을 산출하는 합성곱 신경망(convolutional neural network) 기반의 컨볼루션 모듈과, 상기 컨볼루션 모듈에서 산출된 하위 단어 특징 벡터들을 적응적으로 결합하여 해당 단어의 단어 임베딩 벡터를 산출하는 하이웨이 네트워크(highway network) 기반의 하이웨이 모듈을 포함하는 것이다.

발명의 효과

[0015] 전술한 본 발명의 과제 해결 수단에 의하면, 미등록 단어뿐만 아니라 모든 단어의 하위단어정보를 이용하여 해당 단어가 가지고 있는 고유한 의미를 정확히 추출하고, 미등록 단어가 많은 개방 어휘(Open Vocabulary) 환경에서 효과적으로 동작할 수 있다.

[0016] 또한, 본 발명은 새로운 단어 임베딩을 생성하기 위해 말뭉치, 즉 대형 코퍼스에서 오랜 시간 학습하지 않고, 기존의 자연어 처리 시스템이 가지고 있는 단어 임베딩을 이용하여 미등록 단어를 생성할 수 있기 때문에 단어 임베딩 생성에 있어서의 효율성 및 효과성이 향상될 수 있다.

도면의 간단한 설명

[0017] 도 1은 본 발명의 일 실시예에 따른 단어 표현을 생성하기 위한 자연어 처리 시스템의 구성을 나타낸 도면이다.

도 2는 본 발명의 일 실시예에 따른 단어 표현 모델을 설명하는 도면이다.

도 3은 도 2의 일부 구성요소인 컨볼루션 모듈을 설명하는 도면이다.

도 4는 도 2의 일부 구성요소인 하이웨이 모듈을 설명하는 도면이다.

도 5는 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법 중 단어 표현 모델의 학습 과정을 설명하는 순서도이다.

도 6은 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법 중 단어 표현 모델의 추론 과정을 설명하는 순서도이다.

도 7은 어휘와 미등록 단어를 위한 단어 표현의 최근접 이웃을 설명하는 도면이다.

도 8은 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법과 다른 학습 모델의 비교 평가를 위해 전체 데이터 세트에 대한 단어 유사성 결과를 설명하는 도면이다.

도 9는 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법을 실제 자연어 처리 시스템에 적용한 실험 결과를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0018] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할

수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0019] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0020] 본 명세서에서 '단말'은 휴대성 및 이동성이 보장된 무선 통신 장치일 수 있으며, 예를 들어 스마트폰, 태블릿 PC 또는 노트북 등과 같은 모든 종류의 핸드헬드(Handheld) 기반의 무선 통신 장치일 수 있다. 또한, '단말'은 네트워크를 통해 다른 단말 또는 서버 등에 접속할 수 있는 PC 등의 유선 통신 장치인 것도 가능하다. 또한, 네트워크는 단말들 및 서버들과 같은 각각의 노드 상호 간에 정보 교환이 가능한 연결 구조를 의미하는 것으로, 근거리 통신망(LAN: Local Area Network), 광역 통신망(WAN: Wide Area Network), 인터넷 (WWW: World Wide Web), 유무선 데이터 통신망, 전화망, 유무선 텔레비전 통신망 등을 포함한다.

[0021] 무선 데이터 통신망의 일례에는 3G, 4G, 5G, 3GPP(3rd Generation Partnership Project), LTE(Long Term Evolution), WIMAX(World Interoperability for Microwave Access), 와이파이(Wi-Fi), 블루투스 통신, 적외선 통신, 초음파 통신, 가시광 통신(VLC: Visible Light Communication), 라이파이(LiFi) 등이 포함되나 이에 한정되지는 않는다.

[0022] 이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다. 따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.

[0024] 이하 첨부된 도면을 참고하여 본 발명의 일 실시예를 상세히 설명하기로 한다.

[0025] 도 1은 본 발명의 일 실시예에 따른 단어 표현을 생성하기 위한 자연어 처리 시스템의 구성을 나타낸 도면이고, 도 2는 본 발명의 일 실시예에 따른 단어 표현 모델을 설명하는 도면이며, 도 3은 도 2의 일부 구성요소인 컨볼루션 모듈을 설명하는 도면이고, 도 4는 도 2의 일부 구성요소인 하이웨이 모듈을 설명하는 도면이다.

[0026] 도 1을 참조하면, 자연어 처리 시스템(100)은 통신 모듈(110), 메모리(120), 프로세서(130) 및 데이터베이스(140)를 포함한다.

[0027] 상세히, 통신 모듈(110)은 통신망과 연동하여 자연어 처리 시스템(100)과 사용자 단말 간의 송수신 신호를 패킷 데이터 형태로 제공하는 데 필요한 통신 인터페이스를 제공한다. 나아가, 통신 모듈(110)은 사용자 단말로부터 데이터 요청을 수신하고, 이에 대한 응답으로서 데이터를 송신하는 역할을 수행할 수 있다.

[0028] 여기서, 통신 모듈(110)은 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다.

[0029] 메모리(120)는 자연어 처리에서의 단어 표현 방법을 수행하기 위한 프로그램이 기록된다. 또한, 프로세서(130)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행한다. 여기서, 메모리(120)는 휘발성 저장 매체(volatile storage media) 또는 비휘발성 저장 매체(non-volatile storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0030] 프로세서(130)는 자연어 처리에서의 단어 표현 방법을 제공하는 전체 과정을 제어한다. 즉, 프로세서(130)는 어휘사전 데이터세트에 다양한 어휘와 각 단어의 기학습된 단어 임베딩 벡터를 저장하고, 어휘사전 데이터세트에 저장된 단어 임베딩의 지도 학습을 기반으로 단어 표현을 생성하는 단어 표현 모델을 학습하며, 학습된 단어 표현 모델에 기초하여 등록 단어뿐만 아니라 미등록 단어에 대한 단어표현을 생성한 후 최근접 이웃 탐색을 통해 미등록 단어의 고유한 의미를 추론할 수 있다. 이러한 프로세서(130)가 수행하는 각각의 동작에 대해서는 추후 보다 상세히 살펴보기로 한다.

[0031] 도 2에 도시된 바와 같이, 단어 표현 모델(200)은 기학습된 단어 임베딩의 지도 학습을 기반으로 단어 표현을 생성하는 것을 학습한다. 이러한 단어 표현 모델(200)은 컨볼루션 모듈(210), 하이웨이 모듈(220) 및 최적화 모듈(230)을 포함한다.

[0032] 컨볼루션 모듈(210)은 합성곱 신경망(convolutional neural network)을 통해 문자 기반의 하위 단어 특징을 추출한다. 하이웨이 모듈(220)은 하이웨이 신경망(highway network)을 활용하여 컨볼루션 모듈(210)에서 추출된 하위단어 특징들을 적응적으로 결합하여 단어 임베딩 벡터를 산출한다. 또한, 최적화 모듈(230)은 하이웨이 모듈(220)에서 산출된 단어 임베딩 벡터가 기학습된 단어 임베딩과 유사해지도록 최적화를 수행한다.

[0033] 컨볼루션 모듈(210)은 자연어 처리에서 로컬 특징들(local features)을 추출할 수 있기 때문에 합성곱 신경망을 문자 시퀀스에 적용하여 하위 단어 정보를 추출한다. 도 3에 도시된 바와 같이, 컨볼루션 모듈(210)은 문자 시퀀스에서 각기 다른 특징을 추출하는 필터들을 포함하고, 각 필터를 통해 산출된 행렬인 특징 맵(Feature maps)을 추출하며, 필터들을 통해 특징 맵이 추출되면 해당 특징의 유무의 비선형 값으로 바꿔주기 위해 비선형 함수(tanh, Hyperbolic tangent)를 적용한다.

[0034] 일반적으로 학습된 단어 표현 모델의 깊이가 증가함에 따라 성능이 향상한다. 하지만, 깊이가 증가할수록 최적화가 어려워지며 훈련에 어려움이 따른다. 하이웨이 신경망은 단어 표현 모델을 깊게 만들면서도 정보의 흐름을 통제하고 학습 가능성을 극대화할 수 있도록 해준다.

[0035] 도 4에 도시된 바와 같이, 하이웨이 모듈(220)은 컨볼루션 모듈(210)로부터 수신한 하위단어 특징 벡터들에 대해 input(y)의 값을 가지고, 비선형한 변환(T)과 이동(C)을 추가로 적용한다. 이때, Output(z)이 input(y)에 대하여 얼마나 변환되고 옮겨졌느냐를 표현해주기 때문에 T를 변환 게이트(transform gate), C를 이동 게이트(carry gate)라고 한다.

[0036] 한편, 프로세서(130)는 프로세서(processor)와 같이 데이터를 처리할 수 있는 모든 종류의 장치를 포함할 수 있다. 여기서, '프로세서(processor)'는, 예를 들어 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 이와 같이 하드웨어에 내장된 데이터 처리 장치의 일 예로써, 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0037] 데이터베이스(140)는 자연어 처리에서의 단어 표현 방법을 수행하면서 누적되는 데이터가 저장된다. 예컨대, 데이터베이스(140)에는 어휘사전 데이터세트, 단어 표현 모델, 등록 단어 및 미등록 단어의 임베딩 벡터 등이 저장될 수 있다.

[0038] 도 5는 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법 중 단어 표현 모델의 학습 과정을 설명하는 순서도이다.

[0039] 도 5를 참조하면, 자연어 처리에서의 단어 표현 방법은, 자연어 처리 시스템(100)에서 적어도 하나 이상의 단어를 포함하는 어휘 및 기학습된 단어 임베딩 정보를 포함하는 어휘 사전 데이터세트를 제공한다(S110).

[0040] 단어 표현 모델은 어휘 사전 데이터세트에 기초한 어휘가 입력 데이터로 제공되면(S120), 입력 데이터에 존재하는 단어들에 대한 하위 단어(subword) 정보를 추출하고(S130), 추출된 하위 단어 정보를 이용하여 하위 단어 특징 벡터들을 생성한 후 하위 단어 특징 벡터들을 결합하여 단어 임베딩 벡터를 산출한다(S140).

[0041] 단어 표현 모델은 산출된 단어 임베딩 벡터와 해당 단어의 기학습된 단어 임베딩 정보를 매칭하여(S150), 기학습된 단어 임베딩을 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습한다(S160).

[0042] 하위 단어의 범위는 어근, 문자, N-그램(gram) 등 다양하지만, 단어 표현 모델(200)에서는 하위 단어의 단위를 문자로 사용한다. 따라서, 단어 표현 모델은 어휘에 존재하는 모든 단어를 문자 단위로 구분하고, 각 문자를 나타내기 위한 표현으로 원-핫 인코딩(One-hot encoding)을 사용한다. 여기서, 원-핫 인코딩은 해당 문자의 인덱스에서 1의 값을 가지며, 그렇지 않은 경우에는 0의 값을 가지므로, 이러한 문자 표현을 연결하여 단어를 구성한다.

[0043] 각 단어를 이루고 있는 문자 표현을 연결함으로써 수학식 1과 같은 문자 시퀀스 표현이 생성된다.

[0044] [수학식 1]

$$r = c_1 \oplus c_2 \oplus \cdots \oplus c_n = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}$$

[0045]

[0046] 수학식 1에서, r은 단어에 대한 시퀀스 표현이고, $r \in \mathbb{R}^{|V_c| \times n}$ 이며, Vc는 문자들의 어휘이고, n은 단어의 길이이며, $\oplus$은 연결 연산자이고, c_i은 해당 단어에서 i번째 문자를 표현하기 위한 문자 표현(즉, 원-핫 인코딩)을 각각 나타낸다.

[0047] 이때, 모든 문자들에 동일한 수의 합성곱을 수행하기 위해 r에 제로 패딩(zero-padding)을 삽입하고, 제로 패딩의 수는 h-1/2이 된다.

[0048] 단어 표현 모델(200)의 컨볼루션 모듈(210)은 수학식 1과 같은 패딩된 문자 시퀀스 표현 r에 하기 수학식 2를 사용하여 합성곱을 수행함으로써 하위 단어 정보를 추출한다.

[0049] [수학식 2]

$$s_i = tanh(F \cdot r_{i:i+h-1} + b)$$

[0050]

[0051] 수학식 2에서, F는 합성곱 신경망에서 사용하는 필터로서, $F \in \mathbb{R}^{|V_c| \times h}$ 이고, h는 필터(F)의 폭이며, $r_{i:i+h-1}$는 문자 c_i에서 c_{i+h-1} 사이의 시퀀스 표현이고, b는 바이어스이며, tanh은 합성곱 결과 값들에 대한 비선형 함수를 각각 나타낸다.

[0052] 컨볼루션 모듈(210)은 단어에 존재하는 고유의 하위 단어 정보를 추출한 후, 추출된 하위 단어 정보에 맥스 풀링(Max pooling) 연산을 적용함으로써, 하기 수학식 3과 같이 유의미한 하위 단어 특징들만을 추출한다.

[0053] [수학식 3]

$$e_i = \max\left[s_{k \cdot i:k \cdot (i+1)-1}\right]$$

[0054]

[0055] 수학식 3에서, S는 하위 단어 특징 벡터이고, k는 스트라이드의 길이이며, $s_{ki:k(i+1)-1}$ 는 s에 하나의 스트라이드를 갖는 시퀀스 표현을 각각 나타낸다. 이때, 스트라이드가 적용된 맥스 풀링은 단어의 길이에 따라 길이가 다른 특징을 생성하므로 벡터e의 요소를 단순히 합하여 고정 크기 특징을 생성한다.

[0056] 컨볼루션 모듈(210)은 여러 개의 필터들이 입력 데이터를 지정한 간격으로 순회하면서 합성곱을 계산하는데, 지정된 간격으로 필터를 순회하는 간격을 스트라이드(Stride)라고 한다. 입력 데이터가 여러 채널을 가질 경우, 필터는 각 채널을 순회하며 합성곱을 계산한 후 채널별 특징 맵을 생성하고, 각 채널의 특징 맵을 모두 합산(concatenate)하여 최종 특징 맵으로 반환한다.

[0057] 결과 벡터 s는 제로 패딩으로 인해 입력 단어와 동일한 길이를 갖고, 맥스 폴링을 스트라이드와 함께 사용하여 하위 단어 정보에 접미사, 어근 및 접두어를 포함하도록 한다. 결과적으로, 스트라이드가 적용된 맥스 풀링에서 파생된 각 특징에는 문자 시퀀스에서 하위단어 정보의 요약을 가진다.

[0058] 이와 같이, 단어 표현 모델은 기존에 대형 코퍼스에서 비지도 학습 방식으로 단어 표현을 학습하던 방식과 다르게, 기학습된 단어 임베딩 정보를 포함하는 어휘사전 데이터세트를 이용하여 어휘에 포함된 모든 단어에 대한 단어 표현을 학습한다.

[0059] 단어 표현 모델(200)의 하이웨이 모듈(220)은 하기 수학식 4에 의한 게이트 메커니즘을 사용하여 하위단어 특징 벡터를 라우트하는 것으로서, 개별 필터의 로컬 특징을 적응적으로 결합하여 합성곱 신경망에 유용하다. 즉, 하이웨이 모듈(220)은 컨볼루션 모듈(210)에서 파생된 하위 단어 특징 벡터를 적응적으로 결합하고, 이 하위 단어 특징 벡터가 기학습된 단어 임베딩 정보에 연관되도록 한다.

[0060] [수학식 4]

[0061] $$y = T \odot H + C \odot e$$

[0062] 수학식 4에서, H는 입력(e)에서의 비선형 변환이고, T는 변환 게이트이며, C는 이동 게이트를 각각 나타낸다.

[0063] 상기한 수학식 4에서 이동 게이트(C)를 (1-T)로 단순화하여 하기 수학식 5로 나타낼 수 있다.

[0064] [수학식 5]

[0065] $$T = \sigma(W_t \cdot e + b_t)$$
$$H = \tanh(W_h \cdot e + b_h)$$

[0066] 수학식 5에서, W_t와 W_h는 정방 행렬(square matrices)이고, b_t와 b_h는 바이어스들이며, tanh는 결과 값들에 대한 비선형 함수이다.

[0067] 하이웨이 모듈(220)은 기학습된 단어 임베딩 정보와 동일한 크기의 단어 임베딩 벡터를 생성하기 위해 산출된 단어 임베딩 벡터에 하기 수학식 6을 이용해 선형 변환을 수행한다.

[0068] [수학식 6]

[0069] $$w = W \cdot y + b$$

[0070] 수학식 6에서, w는 단어 표현 모델로부터 파생된 최종 단어 표현이고, W와 b는 선형 변환의 파라미터들이며, y는 결과 벡터이다. 최적화를 위해 수학식 6을 통한 선형 변형을 이용하여 기학습된 단어 임베딩 결과 벡터 y의 크기가 동일해지도록 설정한다.

[0071] 이와 같이, 단어 표현 모델(200)은 컨볼루션 모듈(210)과 하이웨이 모듈(220)을 이용하여 하위 단어 정보를 고려하면서 단어 표현을 생성한다. 또한, 단어 표현 모델(200)의 최적화 모듈(230)은 산출된 단어 임베딩 벡터가 기학습된 단어 임베딩 벡터와 유사해지도록 학습하기 위해 목적 함수로 하기 수학식 7과 같은 제곱 유클리드 거리를 사용한다. 이때, 유클리드 거리는 L2 거리(L2 Distance)라고도 한다.

[0072] [수학식 7]

[0073] $$E = \sum_{v \in V_w} \|w_v - \hat{w}_v\|^2$$

[0074] 수학식 7에서, V_w는 어휘 사전 데이터세트 내의 어휘이고, w_v는 산출된 단어 임베딩 벡터이며, $\hat{w}_v$는 기학습된 단어 임베딩 벡터이다.

[0075] 최적화 모듈(230)은 상기한 수학식 7에 의한 제곱 유클리드 거리 또는 L2 손실(loss) 함수를 이용하여 산출된 단어 임베딩 벡터와 기학습된 단어 임베딩 정보간의 유사도를 산출하는데, 코사인 유사도, L 1 거리(L1 Distance or Manhattan Distance) 등을 이용하여 두 벡터간의 유사도를 계산할 수 있다.

[0076] 최적화 모듈(230)은 단어에 대해 산출된 단어 임베딩 벡터와 기학습된 단어 임베딩 정보를 매칭함으로써 기학습된 단어 임베딩을 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습한다.

[0077] 다시 도 1을 참조하면, 어휘사전 데이터세트에서 'uncovered'라는 단어가 입력 데이터로 제공될 경우, 컨볼루션 모듈(210)은 h=5, 2개의 제로 패딩이 추가된 3개의 필터들과 다른 폭의 필터를 사용하고, 풀링을 위해 스트라이드 폭을 3으로 설정하는 것으로 가정하며, 하이웨이 모듈(220)에서는 단일 하이웨이 네트워크를 사용한다.

[0078] 단어 표현 모델(200)은 'uncovered'의 기학습된 단어 임베딩 정보와 유사하게 단어 표현을 학습하고, 입력 데이터와 어휘적으로 관련되어 있지만 기학습된 단어 임베딩 정보에 포함되지 않은 단어(예를 들어, uncovering)를 나타낼 수 있다. 즉, 'uncovered'와 'uncovering'은 유사한 문자 시퀀스 표현을 공유하기 때문에 단어 표현 모델(200)은 기학습된 단어 임베딩을 문자로 재구성하고, 학습한 단어 표현의 최근접한 이웃 단어들을 제시할 수 있다.

[0079] 이와 같은 방식으로, 어휘 사전 데이터세트 내의 모든 단어에 대한 기학습된 단어 임베딩 정보를 단어 표현 모델에 의해 새롭게 생성된 단어 표현으로 일반화하게 된다. 이때, 단어 표현 모델은 각 단어의 하위 단어를 문자 단위로 구분하기 때문에 미등록 단어를 비롯한 모든 단어에 대한 단어 표현을 생성할 수 있다.

[0080] 도 6은 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법 중 단어 표현 모델의 추론 과정을 설명하는 순서도이고, 도 7은 어휘와 미등록 단어를 위한 단어 표현의 최근접 이웃을 설명하는 도면이다.

[0081] 도 6을 참조하면, 어휘 사전 데이터세트 내의 등록 단어와 미등록단어에 대한 단어 표현을 학습한 단어 표현 모델(200)은 미등록 단어가 입력 데이터로 제공되면(S210), 미등록 단어를 문자 단위로 구분하고, 문자 시퀀스 표현을 생성한 후 컨볼루션 모듈(210)과 하이웨이 모듈(220)을 통해 미등록 단어 임베딩 벡터를 생성한다(S220).

[0082] 단어 표현 모델(200)은 미등록 단어 임베딩 벡터에 대한 최근접 이웃 탐색(nearest-neighbor search)을 통해 미등록 단어의 고유 의미를 추론할 수 있다(S230).

[0083] 표 1은 단어 표현 모델의 추론 과정을 통해 미등록 단어의 단어 표현에 대한 벡터 공간 상 이웃 단어들을 나타내고 있다.

[0084] [표 1]

미등록 단어 (Out-of-Vocabulary)			
bluejacket	vehicals	globalise	computerization
jacket trouser sleeve t-shirt pants	vehicles bicycles cars automobiles trailers	global worldwide globally globalisation globalization	computational computation computerized visualization computations

[0085]

[0086] 표 1에 나타나 있듯이, 단어 표현 모델에 의한 추론 과정은 미등록 단어에 대해 고유 의미를 잘 표현하고 있는 것을 확인할 수 있다. 즉, 단어 표현 모델(200)을 통해 생성된 단어 표현은 단어 변형과 복합어(computerization, bluejacket)를 잘 나타내고, 다른 단어 스타일과 관련된 단어(global-globally-globalization)를 포착하며, 맞춤법 오류 단어(vehicals)에 대한 견고함을 보여주고, 대부분의 미등록 단어(bluejacket-pants, vehicals-bicycles)에서 의미론적으로 관련된 단어를 포착할 수 있다.

[0087] 도 7에 도시된 바와 같이, 단어 표현 모델(GWR)의 성능을 평가하기 위해 최근접 이웃 단어를 탐색하여 정성 분석을 수행하면, word2vec 및 GWR에서 영어로 학습한 단어 표현의 최근접 이웃 단어를 추출하고, 최근접 이웃 단어는 코사인 유사성에 의해 계산된 유사성의 내림차순으로 정렬된다.

[0088] 또한, 어휘 단어의 최근접 이웃 단어는 의미론적 또는 구문론적으로 관련된 단어가 최근접 이웃단어(taxicab-minibus, teachteaching)에 위치하고 있음을 알 수 있다. 더욱이, GWR은 어휘적으로 관련된 단어 표현을 보다 유사하게 만든다. 예를 들어, "connect"의 이웃 단어의 변형어(connects-connected-connecting)가 벡터 공간에서 기학습된 단어 임베딩보다 더 밀접하게 위치함을 나타낸다. 이 렌더링은 다른 단어들(teach-teaching, computes-compute-computable)도 비슷한 추세를 보여주고 있고, 어휘 관련성에 관한 단어의 특성을 만족시킨다. 이는 어휘적으로 관련된 단어가 문자 시퀀스의 많은 부분을 공유하기 때문이다.

[0089] 도 8은 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법과 다른 학습 모델의 비교 평가를 위해 전체 데이터 세트에 대한 단어 유사성 결과를 설명하는 도면이다.

[0090] 단어 유사성은 단어들 간의 코사인 유사도 사이의 상관 계수를 계산하여 측정할 수 있고, 단어 유사성을 통해 각 모델의 능력을 평가할 수 있다.

[0091] 먼저, 아랍어(Ar), 독일어(De), 영어(En), 스페인어(Es), 프랑스어(Fr) 및 러시아어(Ru)의 6개 언어에 대한 단어의 유사성 데이터세트를 수집하고, 6개 언어에 대해 기학습된 word2vec을 사용한다.

[0092] 여기서, word2vec은 단어를 개별 최소 단위(Atomic unit)로 간주하고 윈도우 기반 학습 기술을 채택하는 모델이다. word2vec은 단어 기반 접근 방식이므로 OOV 단어를 표현할 수 없고, OOV 단어의 기본값으로 단어 유사 작업에 영 벡터(null vector)를 사용하고, 언어 모델링 작업을 위해 무작위로 초기화하여 미세 조정한다.

[0093] 한편, FastText 방법은 word2vec의 확장이며, 문자 n-gram을 최소 단위로 간주하며, word2vec의 기술과 유사하게 학습한다. Mimick 방법은 단어를 표현하기 위해 문자에서 단어 임베딩까지의 기능을 학습하는 문자 기반 모델로서, 기학습된 단어 임베딩을 위해 스킵 그램 버전의 word2vec를 사용한다.

[0094] 단어 표현 모델은 생성된 단어 표현(GWR)으로 표시되고, 기학습된 단어 임베딩에 word2vec를 사용한다.

[0095] 도 8에 도시된 바와 같이, 대부분의 언어에 대한 단어의 유사성 데이터 세트에서 단어 기반 방법(word2vec)에 비해 하위 단어 기반 학습 방법(FastText, Mimick, GWR)이 우수한 성능을 보임을 알 수 있다. 이는 하위 단어 정보를 고려하면 단어를 표현할 때 효과적이며 많은 언어에서 유용하다는 것을 나타낸다. 하위 단어 기반 방법 중 GWR은 영어를 제외한 모든 언어에서 Mimick보다 우수한 성능을 나타낸다.

[0096] 도 8에 도시된 단어 유사성 결과를 통해 단어 표현 모델의 컨볼루션 모델이 로컬 기능, 즉 하위 단어 정보를 추출하는데 유용하며, 대규모 코퍼스에서 단어 표현을 학습하는 FastText보다 기학습된 단어 임베딩 정보의 지도 학습을 기반으로 단어표현을 생성하는GWR이 더 우수한 학습 성능을 보여줌을 알 수 있다.

[0097] GWR에서 파생된 미등록 단어(OOV)의 단어 표현이 실제로 성능을 향상시킨다는 것을 확인하기 위해 미등록 단어를 생성하지 않는 모델(GWR$^-$)보다 상당히 우수한 성능을 보여준다. 이로 인해, GWR이 어휘 범위를 효과적으로 확장하고, 미등록 단어의 표현을 매우 잘 생성함을 나타낸다. 또한 GWR$^-$은 word2vec보다 약간 우수한 성능을 나타내는데, 이는 단어 표현에 하위 단어 정보를 고려할 때의 효과를 보여주는 것이다.

[0098] 도 9는 본 발명의 일 실시예에 따른 자연어 처리에서의 단어 표현 방법을 실제 자연어 처리 시스템에 적용한 실험 결과를 나타낸 도면이다.

[0099] 도 9를 참조하면, 자연어 처리 시스템은 GWR의 유용성을 확인하기 위해 외부 모델로 언어 모델링을 수행한다. 자연어처리 시스템의 성능평가 척도는 퍼플렉시티(perplexity)로서 퍼플렉시티가 낮을수록 강한 모델을 의미한다. 여기서, 퍼플렉시티(Perplexity)는 언어 모델을 평가하기 위한 내부 평가 지표로서 보통 PPL이라고 표현하는데, PPL 측정 방법은 외재적 태스크(Extrinsic task)에 대한 성능 평가 방법의 하나이다.

[0100] 체코어(Cs), 독일어(De), 영어(En), 스페인어(Es), 프랑스어(Fr), 러시아어(Ru)의 6개 언어에 대한 데이터 세트를 사용하여 언어 모델링 작업을 수행한다. GWR을 학습하는 데 사용되는 word2vec의 성능을 기반으로 형태적으로 풍부한 언어에서 성능이 현저하게 향상됨을 알 수 있다.

[0101] 예를 들어, 형태학적으로 풍부한 슬라브 언어(Cs, Ru)의 성능은 형태면에서 다른 언어보다 더 복잡한 혼돈 감소(체코 어 및 러시아어의 경우 각각 15 및 22%)를 나타낸다. 이는 GWR이 형태학적으로 풍부한 언어에서 더 유용하고 효과적임을 나타내는 것이다.

[0102] 이와 같이, 단어 표현 모델은 기존에 미등록 단어를 의사 단어로 표현하여 사용하는 미등록 단어 처리 방법에 비해 미등록 단어의 처리 면에서 있어서 효과성이 우수함을 보여준다.

[0103] 한편, 단어 표현 모델은 CNN에서 널리 사용되는 MLP(multi-layer perceptron) 보다는 하이웨이 네트워크에 기반한 하이웨이 모듈(220)을 사용한다. 하이웨이 네트워크가 기본적으로 학습 손실과 관련하여 MLP보다 빠른 수렴을 보여주어 학습을 더 빠르게 수행할 수 있고, 2-레이어 Highway는 1-레이어 Highway보다 더 빠른 수렴을 보여주므로 레이어 스택(stack)에 따른 하이웨이 네트워크의 특성을 이용하여 단어의 의미상 유사성을 향상시키면서 더 깊은 단어 표현 모델을 학습할 수 있다.

[0104] 이와 같이, 본 발명에서는 GWR로 표시된 문자 기반의 단어 표현 방법을 제공하고 있는데, CNN과 하이웨이 네트워크를 사용하는 단어 표현 모델은 하위 단어 정보를 고려하여 미등록 단어의 고품질 표현을 생성할 수 있다. 이러한 단어 표현 모델은 텍스트 분류와 명명된 개체 인식과 같은 다른 영역에도 적용될 수 있다.

[0106] 이상에서 설명한 본 발명의 실시예에 따른 자연어 처리에서의 단어 표현 방법은, 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행 가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 이러한 기록 매체는 컴퓨터 판독 가능 매체를 포함하며, 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독가능 매체는 컴퓨터 저장 매체를 포함하며, 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.

[0107] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0108] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

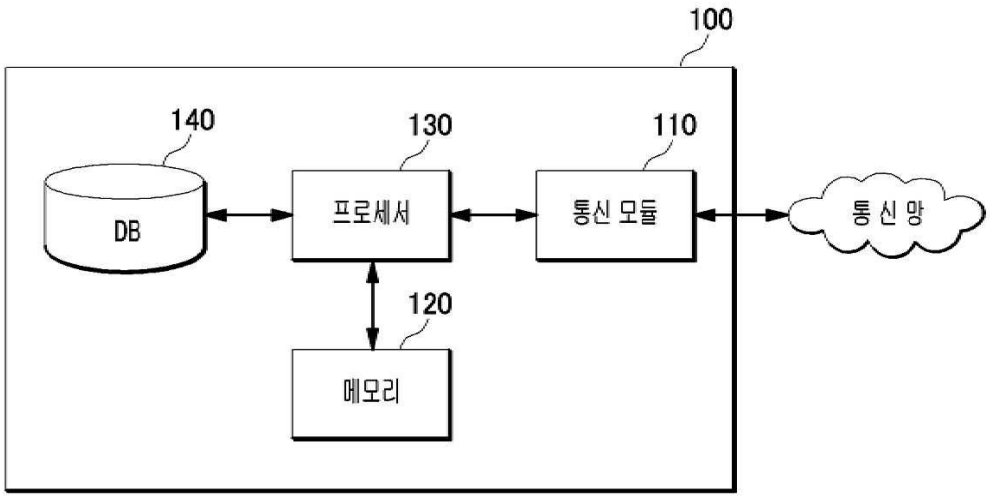
[0109] 100: 자연어 처리 시스템

110: 통신 모듈 120: 메모리

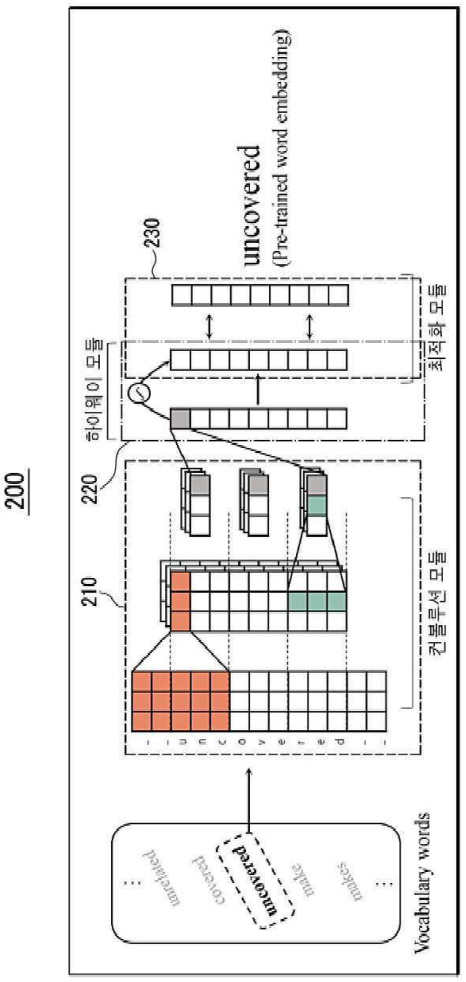
130: 프로세서 140: 데이터베이스

도면

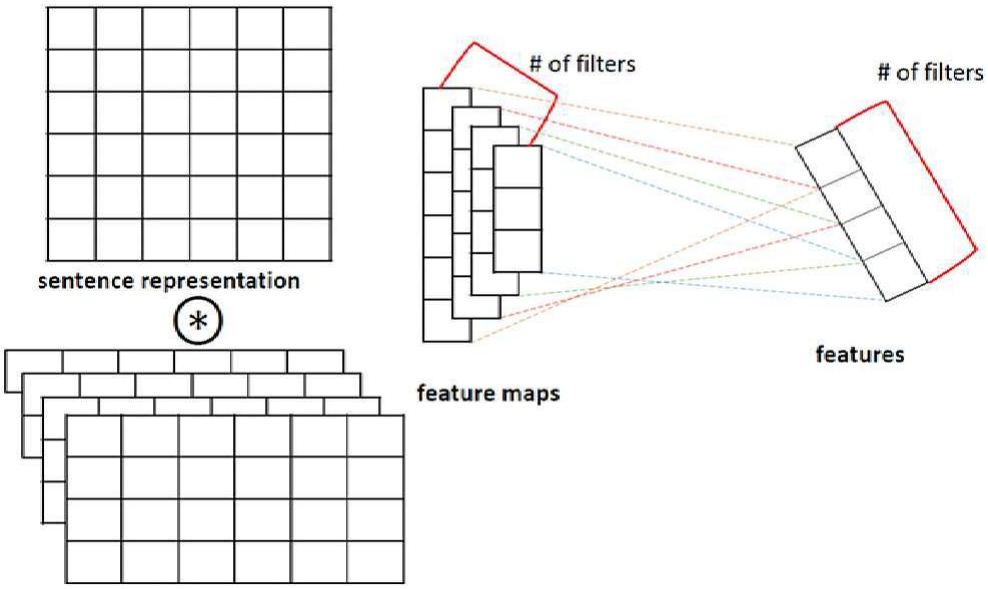
도면1



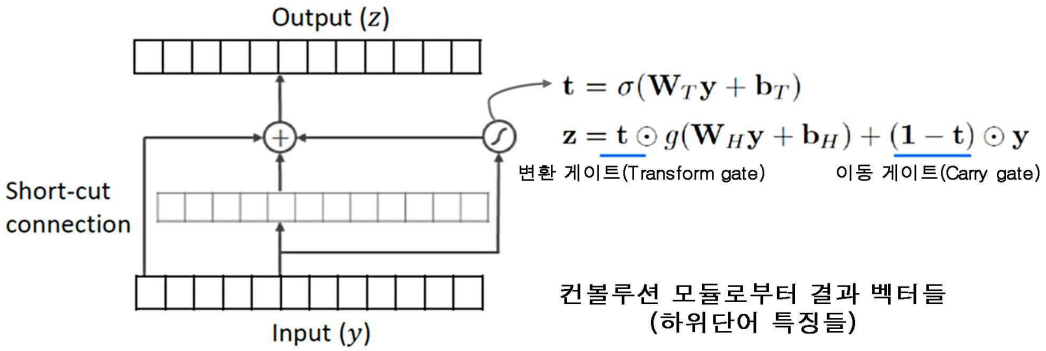
도면2


200
230
uncovered
(Pre-trained word embedding)
하이웨이 모듈
최적화 모듈
220
210
컨볼루션 모듈
Vocabulary words

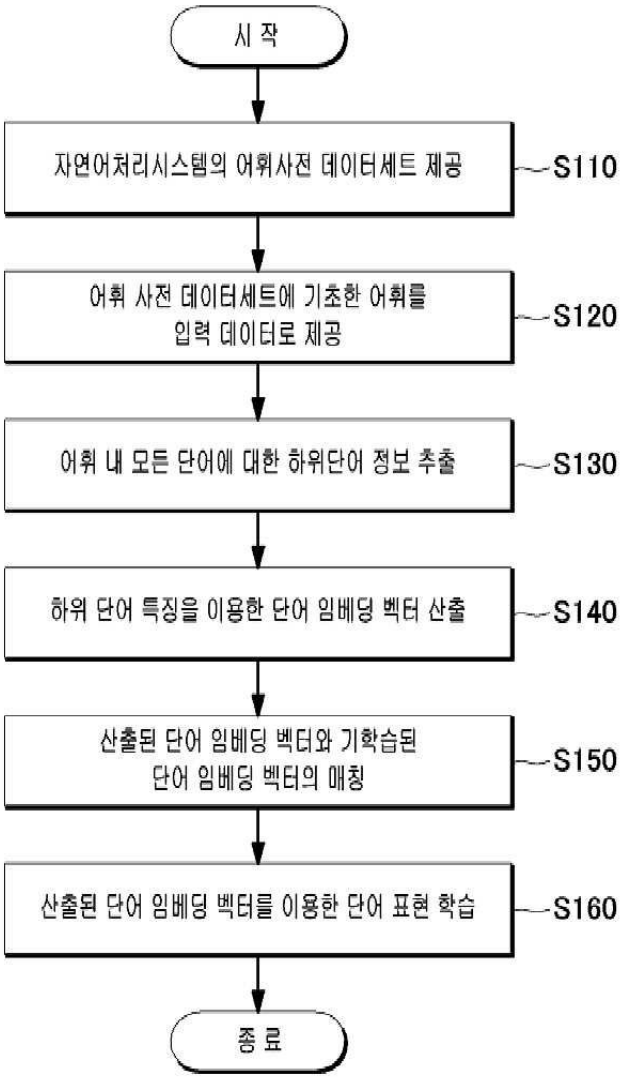
도면3


sentence representation
feature maps
# of filters
# of filters
features

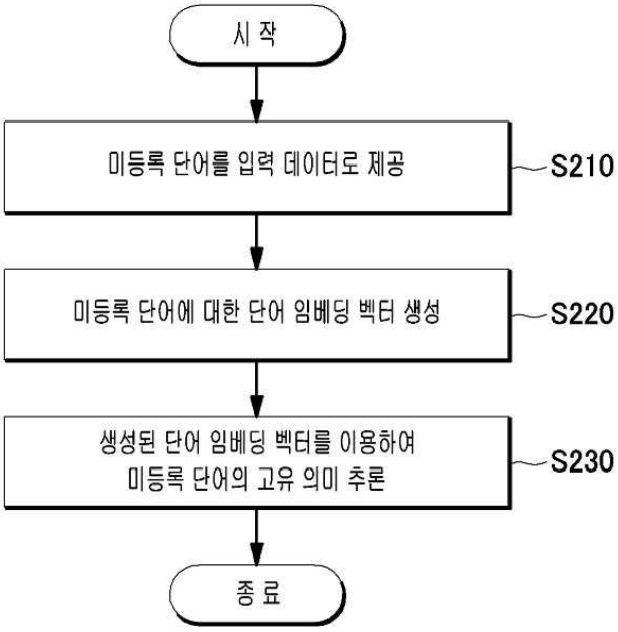
도면4


Output (z)
$\mathbf{t} = \sigma(\mathbf{W}_T\mathbf{y} + \mathbf{b}_T)$
$\mathbf{z} = \mathbf{t} \odot g(\mathbf{W}_H\mathbf{y} + \mathbf{b}_H) + (\mathbf{1} - \mathbf{t}) \odot \mathbf{y}$
변환 게이트(Transform gate)
이동 게이트(Carry gate)
Short-cut connection
Input (y)
컨볼루션 모듈로부터 결과 벡터들
(하위단어 특징들)

도면5


시 작
자연어처리시스템의 어휘사전 데이터세트 제공
S110
어휘 사전 데이터세트에 기초한 어휘를 입력 데이터로 제공
S120
어휘 내 모든 단어에 대한 하위단어 정보 추출
S130
하위 단어 특징을 이용한 단어 임베딩 벡터 산출
S140
산출된 단어 임베딩 벡터와 기학습된 단어 임베딩 벡터의 매칭
S150
산출된 단어 임베딩 벡터를 이용한 단어 표현 학습
S160
종 료

도면6


시 작
미등록 단어를 입력 데이터로 제공
S210
미등록 단어에 대한 단어 임베딩 벡터 생성
S220
생성된 단어 임베딩 벡터를 이용하여
미등록 단어의 고유 의미 추론
S230
종 료

도면7

	In-Vocabulay				Out-of-Vocabulary			
	connect	teach	taxicab	computes	bluejacket	vehicals	globalise	computerization
word2vec	connect	teach	taxicab	computes	-	-	-	-
	connecting	learn	taxi	calculates				
	communicate	instruct	minibus	logarithmic				
	interact	enroll	motorbike	satisfies				
	linking	educate	truck	generates				
GWR	connect	teach	taxicab	computes	jacket	vehicles	global	computational
	connecting	instruct	taxi	compute	trouser	bicycles	worldwide	computation
	connects	learn	limousine	calculates	sleeve	cars	globally	computerized
	connected	teaching	minibus	logarithmic	t-shirt	automobiles	globalisation	visualization
	linking	understand	motorbike	computable	pants	trailers	globalization	computations

도면8

Language	Dataset	word2vec	FastText	Mimick	GWR^{-}	GWR
Ar	WS353	34.9	**52.1**	44.9	36.3	43.8
De	GUR350	34.6	48.7	47.6	35.4	**53.2**
	ZG222	16.2	35.4	36.5	16.4	**38.7**
En	SL999	29.5	27.7	**30.3**	28.6	29.8
	WS353	63.1	64.1	63.7	63.9	**64.6**
Es	MC30	40.6	**72.1**	66.8	41.4	69.8
	WS353	29.5	50.0	40.7	30.8	**50.2**
Fr	RG65	46.8	58.6	51.3	48.1	**60.5**
Ru	HJ	30.3	**60.1**	42.1	31.4	45.5

도면9

언어	기존 단어 임베딩	GWR
체코어	412.7	352.1
독일어	263.8	233.2
영어	259.8	240.7
스페인어	190.7	174.7
프랑스어	205.2	185.6
러시아어	311.5	243.1