



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0144647
(43) 공개일자 2022년10월27일

(51) 국제특허분류(Int. Cl.)
G06F 40/237 (2020.01) G06F 17/16 (2006.01)
G06F 7/544 (2017.01) G06N 20/00 (2019.01)
(52) CPC특허분류
G06F 40/237 (2022.01)
G06F 17/16 (2013.01)
(21) 출원번호 10-2021-0051220
(22) 출원일자 2021년04월20일
심사청구일자 2021년04월20일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암
동5가)
(72) 발명자
이상근
서울특별시 강남구 선릉로 221(도곡동, 도곡렉슬
아파트)
김도현
서울특별시 도봉구 도봉로136길 28,501동 901호(
창동, 북한산 아이파크)
권오준
서울특별시 마포구 월드컵북로 98,501호(성산동,
성산동 성일 APT)
(74) 대리인
윤귀상

전체 청구항 수 : 총 15 항

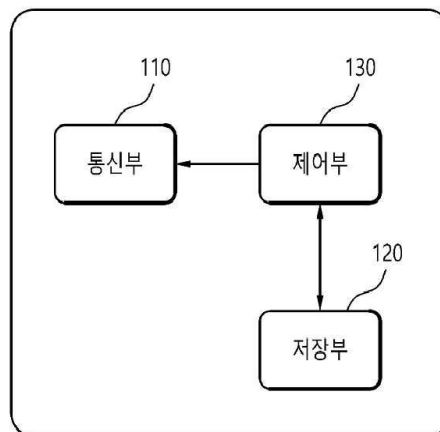
(54) 발명의 명칭 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법, 이를 수행하기 위한 기록
매체 및 시스템

(57) 요약

본 발명의 어휘 밖 단어표현 처리하기 위한 시스템에 의해 수행되는 어휘 밖 단어표현 처리 방법은, 전체 학습 데이터의 기설정된 비율 내에 해당하는 수의 오타를 생성하여 어휘 데이터셋에 저장하는 단계; 어휘 밖 단어가 입력되면, 한글 단어임베딩 모델을 이용하여 단어임베딩을 생성하는 단계; 및 상기 어휘 밖 단어에 대한 단어 표현을 학습하는 단계를 포함할 수 있다. 이에 의해 한글 임베딩 학습 시 인위적으로 만든 오타들에 대한 특징을 추출하여 단어임베딩을 생성하고, 이를 기존의 단어 임베딩과 같아지도록 학습하면서 오타가 많은 시스템에서도 효과적으로 작동할 수 있게 된다.

대 표 도 - 도1

100



(52) CPC특허분류

G06F 7/5443 (2013.01)

G06N 20/00 (2021.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711127793
과제번호	2021R1A2C3010430
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)(R&D)
연구과제명	미플랫폼: 자기지도학습과 연합학습으로 강화되는 작은 인공지능 플랫폼
기 여 율	1/1
과제수행기관명	고려대학교
연구기간	2021.03.01 ~ 2022.02.28

명세서

청구범위

청구항 1

어휘 밖 단어표현 처리하기 위한 시스템에 의해 수행되는 어휘 밖 단어표현 처리 방법은,
전체 학습 데이터의 기설정된 비율 내에 해당하는 수의 오타를 생성하여 어휘 데이터셋에 저장하는 단계;
어휘 밖 단어가 입력되면, 한글 단어임베딩 모델을 이용하여 단어임베딩을 생성하는 단계; 및
상기 어휘 밖 단어에 대한 단어 표현을 학습하는 단계를 포함하는 어휘 밖 단어표현 처리 방법.

청구항 2

제1항에 있어서,
상기 오타를 생성하여 어휘 데이터셋에 저장하는 단계는,
오타 대상 단어가 입력되면 한글 텍스트를 자모 단위로 분리하는 단계; 및
분리된 자모를 기설정된 조건에 따라 상기 오타로 생성하는 단계를 포함하는 것을 특징으로 하는 어휘 밖 단어
표현 처리 방법.

청구항 3

제1항에 있어서,
상기 단어임베딩을 생성하는 단계는,
상기 어휘 밖 단어를 자모 단위로 분리하는 단계;
상기 자모 단위에 해당하는 문자 표현을 연결하여 문자임베딩을 생성하는 단계; 및
상기 문자임베딩을 합성곱 연산을 통해 서로 다른 크기의 필터에 대해 적용하여 피쳐 맵을 생성하는 단계를 포
함하는 것을 특징으로 하는 어휘 밖 단어표현 처리 방법.

청구항 4

제3항에 있어서,
상기 단어임베딩을 생성하는 단계는,
상기 피쳐 맵에서 채널 어텐션 메커니즘을 통해 최종 특징 벡터를 생성하는 단계; 및
상기 최종 특징 벡터를 기존 단어 임베딩과 동일한 크기의 벡터로 변환하여 단어임베딩을 생성하는 단계를 더
포함하는 것을 특징으로 하는 어휘 밖 단어표현 처리 방법.

청구항 5

제4항에 있어서,
생성된 피쳐 맵에서 유의미한 특징을 추출하기 위하여 평균풀링 및 최대풀링 연산을 이용하는 것을 특징으로 하
는 어휘 밖 단어표현 처리 방법.

청구항 6

제5항에 있어서,

상기 채널 어텐션 메커니즘을 통해 최종 특징 벡터를 생성하는 단계는,

상기 평균풀링 및 상기 최대풀링 연산을 통해 생성된 피처 맵의 특징 벡터에서 가중치 행렬 및 소프트맥스 함수를 이용해 상기 특징 벡터를 생성하는 것을 특징으로 하는 어휘 밖 단어표현 처리 방법.

청구항 7

제1항에 있어서,

상기 어휘 밖 단어에 대한 단어 표현을 학습하는 단계는,

생성된 단어임베딩과 기학습된 단어임베딩 정보를 매칭하여 학습하는 단계; 및

상기 생성된 단어임베딩과 주변 단어임베딩을 매칭하여 학습하는 단계를 포함하는 것을 특징으로 하는, 어휘 밖 단어표현 처리 방법.

청구항 8

제7항에 있어서,

상기 기학습된 단어임베딩 정보를 매칭하여 학습하는 단계는,

상기 생성된 단어임베딩과 기학습된 단어 임베딩 정보 간의 L2 손실(loss)을 사용하여 매칭하는 것을 특징으로 하는 어휘 밖 단어표현 처리 방법.

청구항 9

제7항에 있어서,

상기 주변단어임베딩을 매칭하여 학습하는 단계는,

목표 단어를 입력한 경우 주변 단어가 나올 로그-확률의 평균을 사용하여 매칭하는 것을 특징으로 하는 어휘 밖 단어표현 처리 방법.

청구항 10

제1항 내지 제9항 중 어느 하나의 항에 따른 어휘 밖 단어표현 처리 방법을 수행하기 위한, 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체.

청구항 11

자연어 처리에서의 단어 처리 방법을 수행하기 위한 프로그램이 기록되고 어휘 데이터셋이 저장되는 저장부; 및

상기 프로그램을 제어하고 어휘 밖 단어가 입력되면 전체 학습 데이터의 기설정된 비율 내에 해당하는 수의 오타를 생성하여 상기 어휘 데이터셋에 저장하며, 어휘 밖 단어가 입력되면 한글 단어임베딩 모델을 이용하여 단어임베딩을 생성하며, 상기 어휘 밖 단어에 대한 단어 표현을 학습하는 제어부를 포함하는 어휘 밖 단어표현 처리 시스템.

청구항 12

제11항에 있어서,
상기 제어부는,
오타 대상 단어가 입력되면 한글 텍스트를 자모 단위로 분리하고, 분리된 자모를 기설정된 조건에 따라 상기 오타로 생성하는 것을 특징으로 하는 어휘 밖 단어표현 처리 시스템.

청구항 13

제11항에 있어서,
상기 한글 단어임베딩 모델은,
상기 어휘 밖 단어를 자모 단위로 분리하고, 상기 자모 단위에 해당하는 문자 표현을 연결하여 문자임베딩을 생성하며, 상기 문자임베딩을 합성곱 연산을 통해 서로 다른 크기의 필터에 대해 적용하여 피쳐 맵을 생성하는 컨볼루션 모듈을 포함하는 것을 특징으로 하는 어휘 밖 단어표현 처리 시스템.

청구항 14

제13항에 있어서,
상기 컨볼루션 모듈은,
상기 피쳐 맵에서 채널 어텐션 메커니즘을 통해 최종 특징 벡터를 생성하고, 상기 최종 특징 벡터를 기존 단어 임베딩과 동일한 크기의 벡터로 변환하여 단어임베딩을 생성하는 것을 특징으로 하는 어휘 밖 단어표현 처리 시스템.

청구항 15

제13항에 있어서,
상기 한글 단어임베딩 모델은,
L2 손실(loss)을 사용하여 생성된 단어임베딩과 기학습된 단어 임베딩 정보 간의 매칭하거나, 목표 단어를 입력한 경우 주변 단어가 나올 로그-확률의 평균을 사용하여 상기 생성된 단어임베딩과 주변 단어임베딩을 매칭하여 학습하는 최적화 모듈을 더 포함하는 것을 특징으로 하는 어휘 밖 단어표현 처리 시스템.

발명의 설명

기술 분야

[0001] 본 발명은 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법, 이를 수행하기 위한 기록 매체 및 시스템에 관한 것으로, 보다 상세하게는 한글 단어를 임베딩된 표현으로 나타낼 때 관측되지 않은 어휘 밖 단어에 대한 임베딩을 학습하는 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법, 이를 수행하기 위한 기록 매체 및 시스템에 관한 것이다.

배경 기술

[0002] 최근 새로운 알고리즘이 개발되고 컴퓨터의 성능이 향상되며 딥러닝을 이용한 자연어처리시스템에 관한 연구가 급속도로 진행되고 있다. 딥러닝 기술을 이용하는 자연어처리시스템에서는 단어를 수치화하기 위하여 단어를 저차원의 벡터로 임베딩하여 사용해야 한다.

[0003] 이때, 자연어처리시스템은 컴퓨터가 인간 언어를 이해할 수 있는 구문적/의미적 표상을 연구하는 것이고, 단어 임베딩 기술은 신경망에 기반한 언어 모델로부터 도출된 기술로 유사한 단어들을 벡터 공간 상에 가깝게 배치하

여 어휘 의미를 표현할 수 있는 기술이다.

[0004] 그러나 시스템의 메모리 문제로 인해 존재하는 모든 단어에 대한 임베딩을 지니고 있기는 불가능하다. 따라서, 시스템이 텍스트에서 학습 시 관측되지 않았던 어휘 밖 단어를 만나는 것은 불가피한 문제이다. 이러한 문제를 피하는 방법으로 시스템에 존재하는 단어들을 문자 단위로 분리한 하위단어의 정보들을 이용해서 어휘 밖 단어의 의미를 추론하는 방법이 제시되고 있다.

[0005] 하지만, 이러한 방법은 영어 등 고립어의 성격을 가진 언어들에서는 효과적인 해결책이 될 수 있지만, 교착어의 성격이 강하고 여러 형태소를 조합하여 단어를 구성하는 한국어에서는 원래 단어의 뜻을 제대로 표현하지 못하는 결과를 만든다.

[0006] 특히, 이러한 문제는 많은 수의 오타와 의도적으로 변형한 단어들이 존재하는 실제 상황에서의 데이터셋에서 더 명확해진다. 따라서 학습 단계에서 인위적으로 어휘 밖 단어들을 만들어서 중요한 하위단어 정보에 집중하고, 그 단어의 임베딩이 원래의 단어 임베딩과 가까워지도록 하는 언어 모델의 개발을 위한 학습 방법이 필요하다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국등록특허 제10-1799681호

발명의 내용

해결하려는 과제

[0008] 본 발명은 상기와 같은 문제를 해결하기 위해 안출된 것으로, 본 발명의 목적은 인위적으로 오타 어휘들을 생성하여 중요한 하위단어 정보에 집중하여 학습함으로써 오타가 많은 환경에 특화된 한글 단어 임베딩을 생성하는 것이다.

과제의 해결 수단

[0009] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 어휘 밖 단어표현 처리하기 위한 시스템에 의해 수행되는 어휘 밖 단어표현 처리 방법은, 전체 학습 데이터의 기설정된 비율 내에 해당하는 수의 오타를 생성하여 어휘 데이터셋에 저장하는 단계; 어휘 밖 단어가 입력되면, 한글 단어임베딩 모델을 이용하여 단어임베딩을 생성하는 단계; 및 상기 어휘 밖 단어에 대한 단어 표현을 학습하는 단계를 포함한다.

[0010] 여기서 상기 오타를 생성하여 어휘 데이터셋에 저장하는 단계는, 오타 대상 단어가 입력되면 한글 텍스트를 자모 단위로 분리하는 단계; 및 분리된 자모를 기설정된 조건에 따라 상기 오타로 생성하는 단계를 포함할 수 있다.

[0011] 그리고, 상기 단어임베딩을 생성하는 단계는, 상기 어휘 밖 단어를 자모 단위로 분리하는 단계; 상기 자모 단위에 해당하는 문자 표현을 연결하여 문자임베딩을 생성하는 단계; 및 상기 문자임베딩을 합성곱 연산을 통해 서로 다른 크기의 필터에 대해 적용하여 피쳐 맵을 생성하는 단계를 포함할 수 있다.

[0012] 또한, 상기 단어임베딩을 생성하는 단계는, 상기 피쳐 맵에서 채널 어텐션 메커니즘을 통해 최종 특징 벡터를 생성하는 단계; 및 상기 최종 특징 벡터를 기존 단어 임베딩과 동일한 크기의 벡터로 변환하여 단어임베딩을 생성하는 단계를 더 포함할 수도 있다.

[0013] 그리고, 생성된 피쳐 맵에서 유의미한 특징을 추출하기 위하여 평균풀링 및 최대풀링 연산을 이용할 수 있다.

[0014] 또한, 상기 채널 어텐션 메커니즘을 통해 최종 특징 벡터를 생성하는 단계는, 상기 평균풀링 및 상기 최대풀링 연산을 통해 생성된 피쳐 맵의 특징 벡터에서 가중치 행렬 및 소프트맥스 함수를 이용해 상기 특징 벡터를 생성할 수 있다.

[0015] 그리고 상기 어휘 밖 단어에 대한 단어 표현을 학습하는 단계는, 생성된 단어임베딩과 기학습된 단어임베딩 정보를 매칭하여 학습하는 단계; 및 상기 생성된 단어임베딩과 주변 단어임베딩을 매칭하여 학습하는 단계를 포함할 수 있다.

- [0016] 또한, 상기 기학습된 단어임베딩 정보를 매칭하여 학습하는 단계는, 상기 생성된 단어임베딩과 기학습된 단어 임베딩 정보 간의 L2 손실(loss)을 사용하여 매칭할 수 있다.
- [0017] 그리고, 상기 주변단어임베딩을 매칭하여 학습하는 단계는, 목표 단어를 입력한 경우 주변 단어가 나올 로그-확률의 평균을 사용하여 매칭할 수 있다.
- [0018] 또한, 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 어휘 밖 단어표현 처리 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 컴퓨터로 판독 가능한 기록 매체일 수 있다.
- [0019] 한편, 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 어휘 밖 단어표현 처리 시스템은, 자연어 처리에서의 단어 처리 방법을 수행하기 위한 프로그램이 기록되고 어휘 데이터셋이 저장되는 저장부; 상기 프로그램을 제어하고 어휘 밖 단어가 입력되면 전체 학습 데이터의 기설정된 비율 내에 해당하는 수의 오타를 생성하여 상기 어휘 데이터셋에 저장하며, 어휘 밖 단어가 입력되면 한글 단어임베딩 모델을 이용하여 단어임베딩을 생성하며, 상기 어휘 밖 단어에 대한 단어 표현을 학습하는 제어부;를 포함한다.
- [0020] 여기서 상기 제어부는, 오타 대상 단어가 입력되면 한글 텍스트를 자모 단위로 분리하고, 분리된 자모를 기설정된 조건에 따라 상기 오타로 생성할 수 있다.
- [0021] 또한, 상기 한글 단어임베딩 모델은, 상기 어휘 밖 단어를 자모 단위로 분리하고, 상기 자모 단위에 해당하는 문자 표현을 연결하여 문자임베딩을 생성하며, 상기 문자임베딩을 합성곱 연산을 통해 서로 다른 크기의 필터에 대해 적용하여 피쳐 맵을 생성하는 컨볼루션 모듈을 포함할 수 있다.
- [0022] 그리고, 상기 컨볼루션 모듈은, 상기 피쳐 맵에서 채널 어텐션 메커니즘을 통해 최종 특징 벡터를 생성하고, 상기 최종 특징 벡터를 기존 단어 임베딩과 동일한 크기의 벡터로 변환하여 단어임베딩을 생성할 수 있다.
- [0023] 또한, 상기 한글 단어임베딩 모델은, L2 손실(loss)을 사용하여 생성된 단어임베딩과 기학습된 단어 임베딩 정보 간의 매칭하거나, 목표 단어를 입력한 경우 주변 단어가 나올 로그-확률의 평균을 사용하여 상기 생성된 단어임베딩과 주변 단어임베딩을 매칭하여 학습하는 최적화 모듈을 더 포함할 수 있다.

발명의 효과

- [0024] 상술한 본 발명의 일측면에 따르면, 한글 임베딩 학습 시 인위적으로 만든 오타들에 대한 특징을 추출하여 단어 임베딩을 생성하고, 이를 기존의 단어 임베딩과 같아지도록 학습하면서 오타가 많은 시스템에서도 효과적으로 작동할 수 있게 된다.
- [0025] 또한 한글 단어 임베딩 생성 시 단어의 지역적, 형태소적인 특징을 도출하는 데에 적합한 합성곱 신경망과 채널 어텐션 메커니즘을 적용하므로 기존의 방법론보다 오타를 추론함에 있어 향상된 한글 단어 임베딩을 생성할 수 있게 된다.

도면의 간단한 설명

- [0026] 도 1은 본 발명의 일 실시예에 따른 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 시스템의 구성을 설명하기 위한 도면,
- 도 2는 본 발명의 일 실시예에 따른 시스템에서 오타를 생성하는 방법을 설명하기 위한 도면,
- 도 3는 본 발명의 일 실시예에 따른 어휘 밖 한글 단어임베딩 모델을 설명하는 도면,
- 도 4 내지 도 6은 도 2의 일부 구성요소인 컨볼루션 모듈을 설명하기 위한 도면,
- 도 7은 본 발명의 일 실시예에 따른 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법 중 한글 단어임베딩 모델의 학습 과정을 설명하는 흐름도, 그리고,
- 도 8은 본 발명의 일 실시예에 따른 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법 중 한글 단어임베딩 모델의 추론 과정을 설명하는 흐름도이다.

발명을 실시하기 위한 구체적인 내용

- [0027] 후술하는 본 발명에 대한 상세한 설명은, 본 발명이 실시될 수 있는 특정 실시예를 예시로서 도시하는 첨부 도면을 참조한다. 이들 실시예는 당업자가 본 발명을 실시할 수 있기에 충분하도록 상세히 설명된다. 본 발명의 다양한 실시예는 서로 다르지만 상호 배타적일 필요는 없음이 이해되어야 한다. 예를 들어, 여기에 기재되어 있

는 특정 형상, 구조 및 특성은 일 실시예와 관련하여 본 발명의 정신 및 범위를 벗어나지 않으면서 다른 실시예로 구현될 수 있다. 또한, 각각의 개시된 실시예 내의 개별 구성요소의 위치 또는 배치는 본 발명의 정신 및 범위를 벗어나지 않으면서 변경될 수 있음이 이해되어야 한다. 따라서, 후술하는 상세한 설명은 한정적인 의미로서 취하려는 것이 아니며, 본 발명의 범위는, 적절하게 설명된다면, 그 청구항들이 주장하는 것과 균등한 모든 범위와 더불어 첨부된 청구항에 의해서만 한정된다. 도면에서 유사한 참조부호는 여러 측면에 걸쳐서 동일하거나 유사한 기능을 지칭한다.

- [0028] 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: WorldWide Web), 유무선 데이터 통신망, 전화망, 유무선 텔레비전 통신망 등을 포함한다.
- [0029] 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.
- [0030] 이하에서는 도면들을 참조하여 본 발명의 바람직한 실시예들을 보다 상세하게 설명하기로 한다.
- [0031] 도 1은 본 발명의 일 실시예에 따른 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 시스템(100)의 구성을 설명하기 위한 도면, 도 2는 본 발명의 일 실시예에 따른 시스템(100)에서 오타를 생성하는 방법을 설명하기 위한 도면, 도 3는 본 발명의 일 실시예에 따른 한글 단어임베딩 모델(200)을 설명하는 도면, 그리고, 도 4 내지 도 6은 도 2의 일부 구성요소인 컨볼루션 모듈(210)을 설명하기 위한 도면이다.
- [0032] 도 1에 도시된 바와 같이, 어휘 밖 단어표현을 처리하기 위한 시스템(100)은, 통신부(110), 저장부(120) 및 제어부(130)를 포함하여 구성된다.
- [0033] 통신부(110)는 통신망과 연동하여 시스템(100)과 사용자 단말기 간의 송수신 신호를 패킷 데이터 형태로 제공하는 데 필요한 통신 인터페이스를 제공한다. 나아가, 통신부(110)는 사용자 단말기로부터 데이터 요청을 수신하고, 이에 대한 응답으로서 데이터를 송신할 수 있다. 이러한 통신부(110)는 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다.
- [0034] 저장부(120)는 한글 단어임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법을 수행하기 위한 프로그램이 기록된다. 또한 제어부(130)가 처리하는 데이터를 일시적 또는 영구적으로 저장하며, 휘발성 저장매체 또는 비휘발성 저장 매체를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.
- [0035] 그리고 저장부(120)는 단어 처리 방법을 수행하면서 누적되는 데이터가 저장된다. 예컨대, 저장부(120)에는 어휘 데이터셋, 한글 단어임베딩 모델, 등록 단어 및 어휘 밖 단어의 임베딩 벡터 등이 저장될 수 있다.
- [0036] 제어부(130)는 한글 단어임베딩에서의 어휘 밖 단어표현을 처리하기 위한 처리 방법을 제공하는 전체 과정을 제어하기 위해 마련된다. 이러한 제어부(130)는 어휘 데이터셋에 다양한 어휘와 각 단어의 이미 학습된 단어임베딩 벡터를 저장한다.
- [0037] 또한 제어부(130)는 어휘 데이터 셋에 어휘 밖 단어표현 처리를 위한 오타를 저장하고, 어휘 데이터셋에 저장된 한글 단어임베딩의 지도 학습을 기반으로 한글 단어표현을 생성하는 한글 단어임베딩 모델을 학습하며, 학습된 한글 단어임베딩 모델에 기초하여 등록 단어뿐만 아니라 어휘 밖 단어에 대한 단어 표현을 생성한 후 어휘밖 단어의 고유한 의미를 추론할 수 있다. 그리고 제어부(130)는 인위적으로 생성해낸 오타가 포함된 단어에 대한 임베딩을 원래의 단어 임베딩에 가깝게 위치하도록 한글 단어임베딩 모델을 학습하여 오타에 강한 한글 단어임베딩 모델을 제시할 수 있게 된다.
- [0038] 그리고 제어부(130)는 어휘 밖 단어표현 처리 방법을 위한 데이터 셋에 어휘 밖 단어표현 처리를 위한 오타를 저장함에 있어서, 한글 오타를 인위적으로 생성할 수 있다. 구체적으로 제어부(130)는 전체 학습 데이터 중 1%에 해당하는 만큼의 오타를 생성하여 이를 학습 데이터에 추가되도록 할 수 있다.
- [0039] 제어부(130)가 오타를 생성하는 것은 통신부 또는 별도의 입력부를 통해 입력된 오타 대상 단어인 한글 텍스트

를 자모 단위로 분리하여 자모의 순서를 바꾸거나, 자모 중 하나를 추가하거나 또는 자모 중 하나를 삭제하는 방식 중 적어도 하나일 수 있다.

[0040] 구체적으로 도 2에 도시된 바와 같이 "단어"라는 한글 텍스트에 대한 오타를 인위적으로 생성하는 것을 예로 들면, 제어부(130)는 “단어”라는 한글의 음절을 자모 단위로 분리한 후, 도 2의 (가)와 같이 2개 자모의 순서를 바꾸는 방식으로 '당너', '다녕', 'ㄴ다어' 등과 같이 오타를 생성하거나, 도 2의 (나)에서와 같이 자모 중 하나를 추가하는 방식으로 '단어ㄱ', '단ㄴ어', 'ㄷ단어' 등과 같이 오타를 생성하거나 도 2의 (다)와 같이 자모 중 하나를 삭제하는 방식으로 '단ㅇ', '다너', 'ㄷㄴ어' 등과 같이 오타를 생성하는 세 가지의 방법으로 오타를 인위적으로 생성할 수 있다. 제어부(130)는 이러한 방법을 통해 생성된 오타를 학습 데이터로 추가하도록 하여 어휘 데이터셋을 저장할 수 있다.

[0041] 한편, 도 3에 도시된 바와 같이, 한글 단어임베딩 모델(200)은 기학습된 한글 단어임베딩의 지도 학습을 기반으로 단어 표현을 생성하는 것을 학습한다. 이러한 한글 단어임베딩 모델(200)은 컨볼루션 모듈(210) 및 최적화 모듈(220)을 포함할 수 있다.

[0042] 컨볼루션 모듈(210)은 자연어 처리분야에서 널리 활용되는 신경망인 합성곱 신경망(Convolutional Neural Network)을 적용하여 자모 표현에 존재하는 지역적인 특징을 추출하고, 특징벡터에서 중요한 부분에 집중하여 최종적인 특징벡터를 추출하여 단어임베딩을 생성한다. 이를 위해 컨볼루션 모듈(210)은 도 4에 도시된 바와 같이 문자 표현에서 각기 다른 특징을 추출하는 필터들을 포함하여 각 필터를 통해 산출된 행렬인 피쳐 맵을 생성할 수 있다. 그리고 생성된 피쳐 맵을 채널 어텐션 메커니즘을 통해 최종적인 특징 벡터를 만들고 기존 시스템에서 사용하는 단어 임베딩과 동일한 크기의 벡터로 만들어주기 위해서 선형변환(linear transform)을 수행할 수 있다.

[0043] 최적화 모듈(220)은 컨볼루션 모듈(210)에서 생성된 단어임베딩을 이미 학습된 단어 임베딩인 기존 단어임베딩과 매칭하고, 생성된 단어임베딩과 주변 단어임베딩을 매칭하여 생성된 단어임베딩이 기존 단어임베딩 및 주변 단어임베딩과 유사해지도록 최적화를 수행한다.

[0044] 도 7은 본 발명의 일 실시예에 따른 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법 중 한글 단어임베딩 모델(200)의 학습 과정을 설명하는 흐름도이다.

[0045] 먼저 본 실시예에 따른 한글 단어임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법은, 학습을 시작하기 이전에 시스템(100)에서 적어도 하나 이상의 단어를 포함하는 어휘 및 기학습된 단어 임베딩 정보를 포함하는 어휘 데이터 셋을 제공할 수 있다.

[0046] 한글 단어임베딩 모델(200)은 어휘 데이터 셋에 기초한 오타가 입력 데이터로 제공되면, 입력 단어를 자모단위로 나뉘도록 자모를 분해한다(S110),

[0047] 한글 단어임베딩 모델(200)은 각 자모에 해당하는 문자 표현을 연결함으로써 단어에 대한 표현을 생성하는 자연어처리시스템의 문자임베딩을 생성한다(S120)

[0048] 한글 단어임베딩 모델(200)은 문자임베딩을 합성곱 연산을 이용해 피쳐 맵을 생성한다(S130). 그리고 한글 단어임베딩 모델(200)은 채널 어텐션을 이용한 단어임베딩을 생성한다(S140). 이후 한글 단어임베딩 모델(200)은 생성된 단어임베딩과 기존 단어임베딩을 매칭한다(S150). 그리고 한글 단어임베딩 모델(200)은 생성된 단어임베딩과 주변 단어임베딩을 매칭하여(S160), 해당 단어에 대한 단어 표현을 학습한다

[0049] 하위 단어의 범위는 어근, 문자, N-그램(gram) 등 다양하지만 한글 단어임베딩 모델(200)에서는 한글의 각 자모 단위를 문자로 사용한다. 따라서 한글 단어임베딩 모델(200)은 어휘에 존재하는 모든 자모를 문자 단위로 구분하고, 각 문자를 나타내기 위한 표현으로 원-핫 인코딩(one-hot encoding)을 사용한다.

[0050] 여기서 원-핫 인코딩은 해당 문자의 인덱스에서 1의 값을 가지며, 그렇지 않은 경우에는 0의 값을 가지므로, 이러한 문자 표현을 연결하여 단어를 구성한다. 또한 합성곱을 수행하기 위해 제로 패딩(zero-padding)을 사용하여 본 실시예에서는 입력 단어 벡터의 길이를 가장 긴 단어의 크기로 고정할 수 있다.

[0051] 한글 단어임베딩 모델(200)의 컨볼루션 모듈(210)은 도4에 도시된 바와 같이 어휘 밖 단어인 '단ㄴ어'가 입력 문자로 입력되면 해당 단어를 자모 단위로 분리한 후 각 자모에 해당하는 문자 표현을 연결하여 단어에 대한 표현을 생성하고, 이렇게 생성된 단어 표현인 문자임베딩에 하기의 수학식[1]을 사용하여 합성곱을 수행하게 된다.

수학식 1

[0052]
$$c_i = ReLU(K^h \cdot w_{i:i+h-1} + b)$$

[0053] 여기서 c_i 는 합성곱 연산이 행해진 후의 단어 표현을 의미하며, $ReLU$ 는 Rectified Linear Unit으로 활성화 함수를 의미하고, K^h 는 합성곱 신경망에서 사용하는 필터(filter)로서 K^h 는 $K^h \in \mathbb{R}^{h \times d}$ 이고, h 는 필터의 크기, 즉 폭을 의미하고, h 는 $h \in \{1, 2, \dots, 10\}$ 이다. 그리고 w 는 입력 단어의 자모가 분해되어 임베딩된 벡터들의 집합으로, $w \in \mathbb{R}^{m \times d}$ 이고, m 은 입력 단어의 최대 길이, 그리고 d 는 사전에 설정되는 임베딩의 크기이다. 이 때 $w_{i:i+h-1}$ 는 문자 w_i 에서 w_{i+h-1} 사이의 표현이고, b 는 편차(bias)를 의미한다.

[0054] 컨볼루션 모듈(210)은 상술한 [수학식1]에 따른 합성곱 연산을 각각 다른 크기의 필터에 대해 적용한 후, 자모의 주요 특징 $\{w_{1:h}, w_{2:h+1}, \dots, w_{m-h+1}\}$ 을 추출하여 하기의 [수학식 2]와 같은 피쳐 맵을 생성할 수 있다.

수학식 2

[0055]
$$c = [c_1, c_2, \dots, c_{m-h+1}]$$

[0056] 이러한 방식으로 컨볼루션 모듈(210)은 n 개의 피쳐 맵을 생성하고, 결합된 피쳐 맵 C 를 얻을 수 있게 되는데, C 는 $C \in \mathbb{R}^{(m-h+1) \times n}$ 이다. 그리고 컨볼루션 모듈(210)은 추출된 피쳐 맵에서 의미있는 특징들을 뽑기 위하여 평균풀링 C_{avg} 및 최대풀링 C_{max} 를 사용하며 이는 하기의 [수학식 3]과 같고, $C_{avg}, C_{max} \in \mathbb{R}^n$ 이다.

수학식 3

[0057]
$$C_{avg} = average(C)$$

[0058]
$$C_{max} = \max(C)$$

[0059] 컨볼루션 모듈(210)은 이상의 과정을 통해 얻은 특징벡터에서 중요한 부분에 집중하여 최종적인 특징벡터를 만들 수 있다. 컨볼루션 모듈(210)은 어텐션 적용 피쳐 맵 C_{att} 를 추출하기 위해 채널 간의 어텐션 점수를 계산하는 채널 어텐션 메커니즘을 사용하며, 이는 [수학식 4]와 같다.

수학식 4

[0060]
$$C_{score} = W_2(W_1(C_{avg} + C_{max}))$$

[0061]
$$C_{att} = \max(ReLU(C \times softmax(C_{score})))$$

[0062] 여기서 C_{score} 는 도 1에 도시된 바와 같이 어텐션 네트워크를 통과한 결과를 정의하고 W_1 과 W_2 는 가중치 행렬이고, $softmax$ 는 소프트맥스 함수를 의미한다. 이상과 같은 수학식을 이용해 컨볼루션 모듈(210)은 각 필터 크기에 대한 최종적인 특징 벡터를 생성할 수 있게 되고, 결과적으로 이상의 방법을 통해 최종적인 특징 벡터 C_{att}^h 를 얻게 되고, C_{att}^h 은 $C_{att}^h \in \mathbb{R}^{10n}$ 이다.

[0063] 그리고 컨볼루션 모듈(210)은 기존 시스템에서의 단어 임베딩과 동일한 크기의 벡터로 만들어 주기 위해서 산출

된 단어 임베딩 벡터에 하기의 [수학식 5]를 이용해 선형변환(linear transform)을 수행하여 최종 단어 표현 w 를 생성한다.

수학식 5

$$w = C_{att}^h \cdot W + b$$

이렇게 생성된 w 는 단어의 하위 구조를 선택적으로 강화하고 문맥 정보를 반영하는 단어 벡터로서 w 는 $w \in \mathbb{R}^d$ 이다. W 와 b 는 선형 변환의 파라미터들이고, W 는 $W \in \mathbb{R}^{10n \times d}$, b 는 $b \in \mathbb{R}$ 이다.

한편, 한글 단어임베딩 모델(200)의 최적화 모듈(220)은 상술한 과정을 통해 최종 산출된 생성된 단어임베딩이 기존 시스템이 가진 해당 단어의 임베딩, 즉 기학습된 단어 임베딩 벡터와 유사해지도록 학습하기 위해 목적 함수로 하기 [수학식 6]과 같은 L2 손실(loss) 함수를 사용하여 매칭시킨다.

수학식 6

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

[수학식 6]에서 L_s 는 생성한 단어임베딩과 기존 시스템에 있는 단어 임베딩 사이의 loss를 의미하고, V_w 는 어휘 데이터 셋을 의미하고, w_v 는 생성된 단어임베딩 벡터이고, $\hat{w}_v$ 는 기학습된 단어임베딩 벡터이다.

또한 본 실시예에 따른 최적화 모듈(220)은 생성된 단어임베딩이 주변 단어임베딩 벡터와 유사해지도록 학습하기 위해 하기 [수학식 7]과 같은 손실(loss)함수를 사용하여 매칭시킨다.

수학식 7

$$L_c = \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

수학식 7에서 L_c 는 목표 단어를 입력으로 넣었을 때 주변 단어가 나올 로그-확률의 평균이고, 생성한 단어임베딩과 주변 단어임베딩 사이의 loss를 의미하고, c 는 주변 단어의 개수, T 는 전체 단어의 개수를 의미한다.

이렇게 생성된 두 개의 손실(loss)를 합하여 하기의 [수학식 8]과 같이 시스템의 최종적인 손실(loss)를 구한다.

수학식 8

$$Loss = (1 - \alpha)L_s + \alpha L_c$$

여기서 α 는 0과 1 사이의 가중치를 의미한다.

최적화 모듈(220)은, 단어에 대해 산출된 단어 임베딩 벡터와 기학습된 단어 임베딩 정보를 매칭함으로써 해당 단어에 대한 단어 표현을 학습한다.

한편, 도 8은 본 발명의 일 실시예에 따른 한글 단어 임베딩에서의 어휘 밖 단어표현을 처리하기 위한 방법 중 한글 단어임베딩 모델(200)을 이용한 추론 과정을 설명하는 흐름도이다.

[0077] 도 8에 도시된 바와 같이 어휘 데이터 셋 내의 사전 학습 단어와 어휘 밖 단어에 대한 단어 표현을 학습한 한글 단어임베딩 모델(200)은 추론할 단어가 입력되면 추론할 단어를 자모 단위로 나뉘도록 자모를 분해한다(S210).

[0078] 그리고 한글 단어임베딩 모델(200)은 각 자모에 해당하는 문자 표현을 연결함으로써 단어에 대한 표현을 생성하는 자연어처리시스템의 문자임베딩을 수행한다(S220). 이후 한글 단어임베딩 모델(200)은 합성곱 연산을 이용한 피쳐 맵을 생성하고(S330) 채널 어텐션을 이용한 단어임베딩을 생성하며(S340). 이후 한글 단어임베딩 모델(200)은 생성된 단어임베딩을 이용해 추론을 진행한다(S250).

[0079] 표 1은 본 발명의 일 실시예에 따른 시스템을 통해 진행된 유추 실험 결과를 나타낸 표로서 성능 평가 척도는 목표 단어와 계산된 단어 사이의 코사인 거리를 활용한 3COSADD 지표를 사용하고, 이를 위한 계산은 하기의 [수학식 9]와 같다.

수학식 9

$$[0080] \quad 1 - \cos(u_a + u_b - u_c, u_d)$$

[0081] 이 때 u_a 와 u_b , 그리고 u_c 와 u_d 는 각각 같은 관계를 가진 단어들의 임베딩을 의미하며, 단어 임베딩이 잘 이루어졌다면 $u_a + u_b - u_c$ 와 u_d 의 위치는 거의 같아지고 코사인 값은 1에 가까워지기 때문에 상술한 수학식 9의 값은 작아지게 된다. 표 1을 통해 생성한 단어표현들 간의 관계가 기존의 방법론들로 만들어진 단어표현들 간의 관계보다 더 정교함을 확인할 수 있다.

표 1

방법	의미론					구문론				
	수도	성별	국적	언어	기타	주격조사	과거형	수동태	동사변형	존댓말
한글 단어임베딩	0.471	0.574	0.528	0.466	0.587	0.478	0.561	0.633	0.672	0.665
한글 하위단어 임베딩 (6문자)	0.446	0.512	0.524	0.364	0.512	0.194	0.417	0.461	0.524	0.362
한글 하위단어 임베딩 (4문자)	0.441	0.517	0.521	0.368	0.521	0.197	0.422	0.468	0.529	0.369
제안 방법-한글 단어임베딩	0.463	0.482	0.513	0.331	0.486	0.328	0.382	0.436	0.389	0.326
제안 방법-한글 하위단어임베딩	0.460	0.488	0.498	0.327	0.491	0.311	0.388	0.421	0.361	0.294

[0082]

[0083] 또한 표 2는 본 발명의 일 실시예에 따른 한글 단어임베딩 모델(200)을 언어 모델링 실험에 적용했을 때의 결과를 나타낸 표로서 성능평가 척도는 perplexity, 즉 문장 중 일부 단어를 가린 후 해당 자리에 들어갈 단어를 유추하는 것으로, 낮을수록 강한 모델을 의미한다. 표 2를 통해 오타율이 높아질수록 기존의 어휘 밖 단어를 처리하는 방법(하위 단어로 분리하는 방법)보다 본 실시예에 따른 어휘 밖 단어를 처리하는 것이 보다 더 효과적이라는 것을 확인할 수 있다.

표 2

방법	오타율			
	0%	10%	20%	30%
한글 단어임베딩	248.3	259.7	269.1	281.4
한글 하위단어임베딩 (6문자)	206.4	218.1	229.5	241.7
한글 하위단어임베딩 (4문자)	210.2	223.1	231.9	246.3
제안 방법-한글 단어임베딩	213.8	220.7	224.7	229.8
제안 방법-한글 하위단어임베딩	211.6	219.4	224.6	228.4

[0084]

[0085] 그리고 표 3은 본 발명의 일 실시예에 따른 한글 단어임베딩 모델(200)을 문장 분류 실험에 적용했을 때의 결과

를 나타낸 표로서, 성능평가 척도는 정확도이다. 표 3을 통해 알 수 있듯이 오타율이 높아질수록 기존의 어휘 밖 단어를 처리하는 방법보다 본 발명에서 제시하는 방법을 통해 어휘 밖 단어를 처리하는 것이 보다 더 효과적이라는 것을 확인할 수 있다.

표 3

방법	오타율				
	0%	10%	20%	30%	40%
한글 단어임베딩	64.1	62.9	60.4	59.3	58.4
한글 하위단어임베딩 (6문자)	68.8	67.3	66.3	64.1	62.3
한글 하위단어임베딩 (4문자)	68.4	66.4	65.2	63.4	61.7
제안 방법- 한글 단어임베딩	68.3	67.6	66.7	65.4	64.2
제안 방법- 한글 하위단어임베딩	68.1	67.2	66.4	65.5	64.1

한편, [표 4]는 어휘 밖 단어를 입력하였을 때 생성된 단어 임베딩에 관한 것으로, 표 4를 통해 알 수 있듯이, word2vec은 오타인 어휘 밖 단어를 처리할 수 없고, FastText 및 본 발명의 한글 단어임베딩 모델인 misK-ft는 단어임베딩을 생성하지만 FastText에 의해 생성된 단어임베딩은 정확하지 않은 것을 알 수 있다. 반면 본 발명의 한글 단어임베딩 모델(200)에서 생성된 단어임베딩은 적절한 의미를 가진 단어와 매핑됨을 확인할 수 있다.

표 4

	In-Vocabulary				Out-Of-Vocabulary			
	컴퓨터	대한민국	대학교	자동차	커뮤니티	중화민국	침례교	자도앗
<i>Word2vec</i>	개인용 그래픽 컴퓨터의	한국 국내	대학교의 대학교에서	승용차 자동차의 자동차가	-	-	-	-
<i>FastText</i>	컴퓨터에 컴퓨터로	대한민국은 대한민국이	대학교는 대학교와 대학교의	자동차를 자동차도 자동차에	커뮤니티 객실 옴	중화민국 대한민국과	침례교 외교 폐교	자 재차 자작
<i>misK-ft</i>	컴퓨터는 컴퓨터는 소프트웨어	한국 대한민국에서 국내	대학교의 대학의 대학교는	자동차에 자동차의 자동차가	코펜하겐 소프트웨어 컴퓨터는	대한민국 중화민국	대학원생 대학원의 학사	자동차 자동차의 광고

또한 이와 같은 본 발명의 한글 단어 임베딩에서의 어휘 밖 단어표현 처리 방법은 다양한 컴퓨터 구성요소를 통하여 수행될 수 있는 프로그램 명령어의 형태로 구현되어 컴퓨터 판독 가능한 기록 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능한 기록 매체는 프로그램 명령어, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다.

상기 컴퓨터 판독 가능한 기록 매체에 기록되는 프로그램 명령어는 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야의 당업자에게 공지되어 사용 가능한 것일 수도 있다.

컴퓨터 판독 가능한 기록 매체의 예에는, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체, CD-ROM, DVD 와 같은 광기록 매체, 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 ROM, RAM, 플래시 메모리 등과 같은 프로그램 명령어를 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다.

프로그램 명령어의 예에는, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드도 포함된다. 상기 하드웨어 장치는 본 발명에 따른 처리를 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

이상에서는 본 발명의 다양한 실시예에 대하여 도시하고 설명하였지만, 본 발명은 상술한 특정의 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 변형실시가 가능한 것은 물론이고, 이러한 변형실시들은 본 발명의 기술적 사상이나 전망으로부터 개별적으로 이해되어져서는 안될 것이다.

부호의 설명

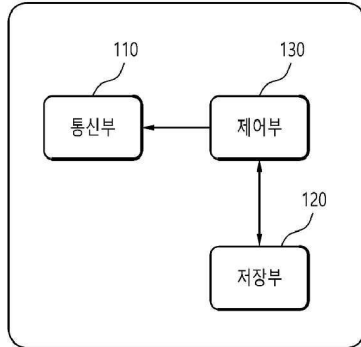
[0094]

100 : 시스템 110 : 통신부
 120 : 저장부 130 : 제어부
 200 : 한글 단어임베딩 모델 210 : 컨볼루션 모듈
 220 : 최적화 모듈

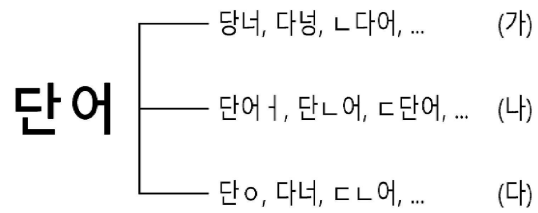
도면

도면1

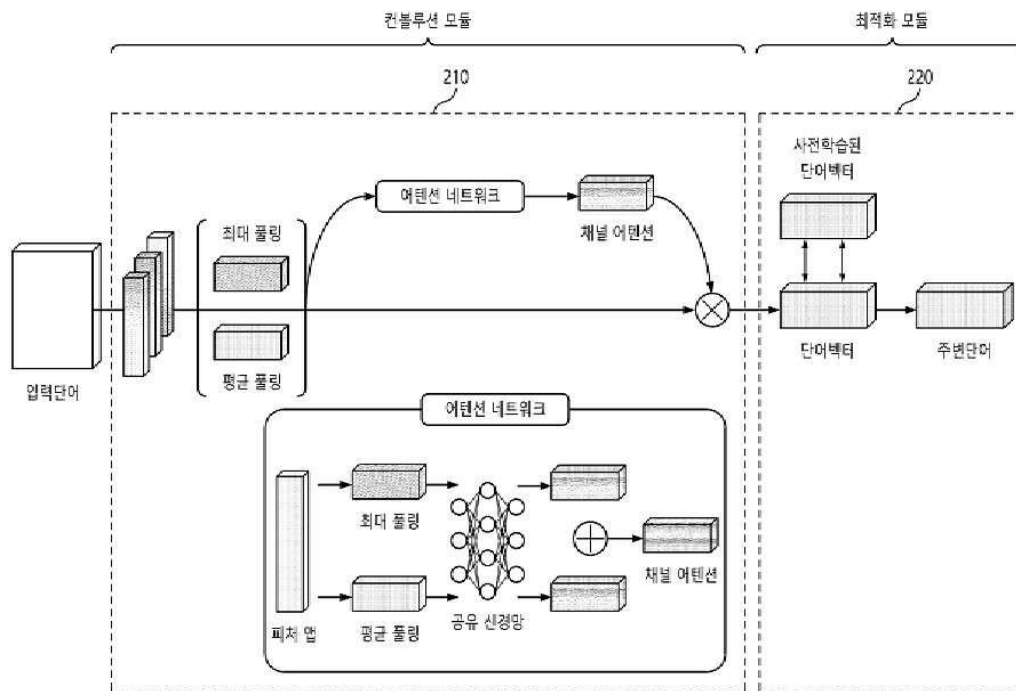
100



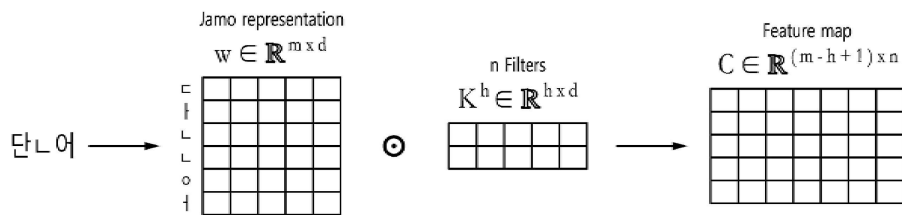
도면2



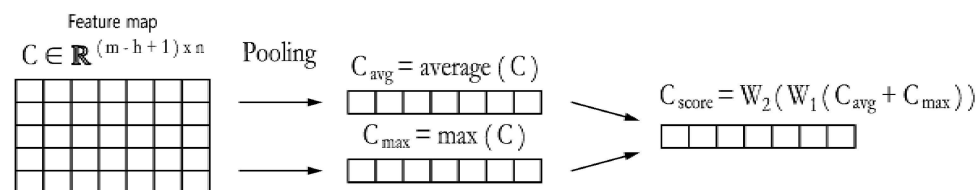
도면3



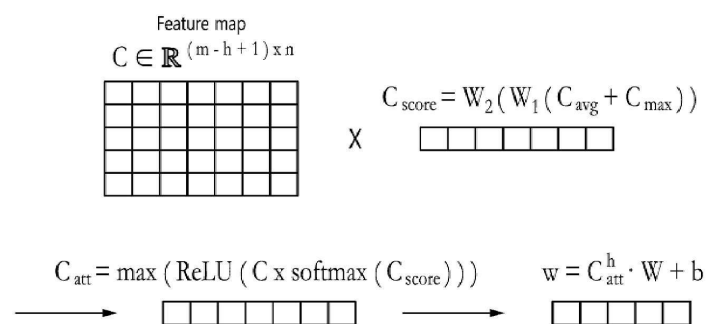
도면4



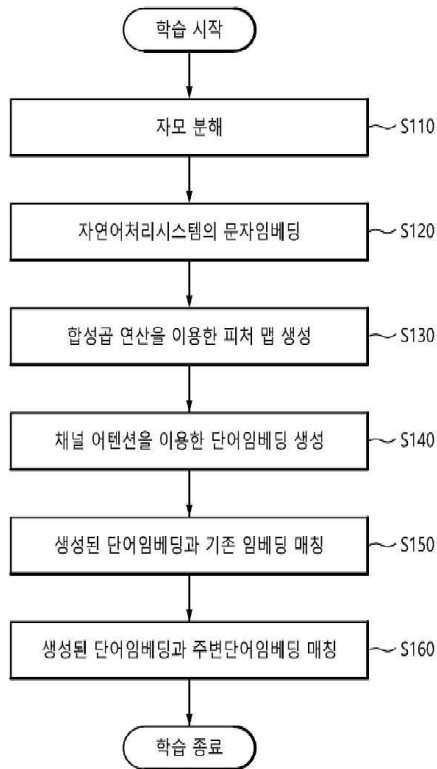
도면5



도면6



도면7



도면8

