(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0083094
(43) 공개일자 2022년06월20일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 20/00* (2019.01)
G06N 5/02 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06F 40/151 (2020.01)
(21) 출원번호 10-2020-0172972
(22) 출원일자 2020년12월11일
심사청구일자 2020년12월11일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
최원익
인천광역시 연수구 해송로30번길 20, 403동 2301호(송도동, 송도 웰카운티 4단지)
이상원
세종특별자치시 보람로 15, 904동 2301호(도담동, 도램마을9단지)
(74) 대리인
양성보

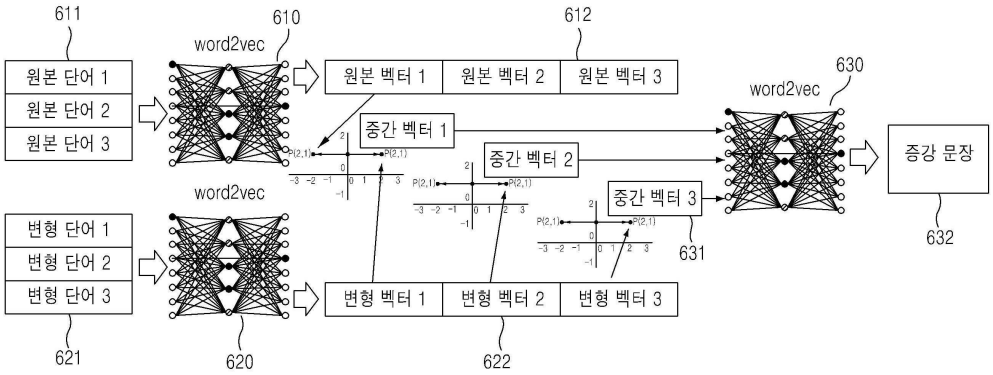
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 **번역 기반 문장 데이터 변형과 Word2Vec 보정을 이용한 문장 데이터 증강 방법 및 장치**

(57) 요 약

번역 기반 문장 데이터 변형과 Word2Vec 보정을 이용한 문장 데이터 증강 방법 및 장치가 제시된다. 일 실시예에 따른 컴퓨터 장치를 통해 구현되는 Word2Vec 보정을 이용한 문장 데이터 증강 방법은, 원본 문장 및 상기 원본 문장의 변형 문장의 단어를 미리 학습된 Word2Vec에 입력하여 벡터로 변환하여, 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터를 획득하는 단계; 상기 Word2Vec의 벡터 공간에서 변환된 상기 원본 문장 단어의 벡터 및 상기 변형 문장 단어의 벡터의 중간 벡터에서 가장 가까운 벡터를 선택하는 단계; 및 선택된 상기 벡터를 상기 Word2Vec을 통해 단어로 변환하고, 상기 단어를 순서대로 나열하여 증강 문장을 생성하는 단계를 포함하여 이루어질 수 있다.

대 표 도



(52) CPC특허분류
G06F 40/279 (2020.01)
G06F 40/56 (2020.01)
G06F 40/58 (2020.01)
G06N 20/00 (2021.08)
G06N 5/022 (2019.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711120431
과제번호	2017M3C8A8091768
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	기초 . 원천기술 개발사업
연구과제명	[Ezbaro] 디지털 컴패니언 기반의 고령자 모니터링 및 대응 소프트웨어 기술 개발
기 여 율	1/1
과제수행기관명	인하대학교
연구기간	2020.07.20 ~ 2021.06.19

명 세 서

청구범위

청구항 1

컴퓨터 장치를 통해 구현되는 Word2Vec 보정을 이용한 문장 데이터 증강 방법에 있어서,

원본 문장 및 상기 원본 문장의 변형 문장의 단어를 미리 학습된 Word2Vec에 입력하여 벡터로 변환하여, 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터를 획득하는 단계;

상기 Word2Vec의 벡터 공간에서 변환된 상기 원본 문장 단어의 벡터 및 상기 변형 문장 단어의 벡터의 중간 벡터에서 가장 가까운 벡터를 선택하는 단계; 및

선택된 상기 벡터를 상기 Word2Vec을 통해 단어로 변환하고, 상기 단어를 순서대로 나열하여 증강 문장을 생성하는 단계

를 포함하는, 문장 데이터 증강 방법.

청구항 2

제1항에 있어서,

상기 원본 문장을 번역기에 입력하여 상기 변형 문장을 생성하는 단계

를 더 포함하는, 문장 데이터 증강 방법.

청구항 3

제1항에 있어서,

상기 원본 문장을 상기 Word2Vec에 입력하여 단어를 벡터로 변환하도록 학습하는 단계

를 더 포함하는, 문장 분류 데이터 증강 방법.

청구항 4

Word2Vec 보정을 이용한 문장 데이터 증강 장치에 있어서,

원본 문장 및 상기 원본 문장의 변형 문장의 단어를 미리 학습된 Word2Vec에 입력하여 벡터로 변환하여, 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터를 획득하는 벡터화부;

상기 Word2Vec의 벡터 공간에서 변환된 상기 원본 문장 단어의 벡터 및 상기 변형 문장 단어의 벡터의 중간 벡터에서 가장 가까운 벡터를 선택하는 벡터 선택부; 및

선택된 상기 벡터를 상기 Word2Vec을 통해 단어로 변환하고, 상기 단어를 순서대로 나열하여 증강 문장을 생성하는 증강 문장 생성부

를 포함하는, 문장 데이터 증강 장치.

청구항 5

제1항에 있어서,

상기 원본 문장을 번역기에 입력하여 상기 변형 문장을 생성하는 변형 문장 생성부

를 더 포함하는, 문장 데이터 증강 장치.

발명의 설명

기 술 분 야

[0001] 아래의 실시예들은 문장 데이터 증강 방법 및 장치에 관한 것으로, 더욱 상세하게는 번역 기반 문장 데이터 변

형과 Word2Vec 보정을 이용한 문장 데이터 증강 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근에는 빅데이터(big data) 시대라고 불릴 만큼 데이터가 중요한 시대가 되었다. 많은 양의 데이터를 모아 인공지능을 학습시켜 인류가 더 편안한 생활을 할 수 있도록 한다. 하지만 데이터를 얻기 힘든 여러 가지 경우가 있다. 적은 데이터로 인공지능을 학습시키기엔 분명 한계가 있으며, 데이터가 부족한 경우 확보한 데이터를 더욱 효율적으로 사용할 수 있어야 한다.

[0003] 이러한 방법 중 하나는 데이터 증강이다. 확보한 원본 데이터를 데이터 증강 기법을 통해 늘리고 학습하여 더 좋은 인공지능을 만들 수 있다.

[0004] 한국공개특허 10-2020-0071877호는 이러한 자율 증강형 반복 학습을 이용한 정보 추출 방법 및 시스템에 관한 기술을 기재하고 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 한국공개특허 10-2020-0071877호

발명의 내용

해결하려는 과제

[0006] 실시예들은 번역 기반 문장 데이터 변형과 Word2Vec 보정을 이용한 문장 데이터 증강 방법 및 장치에 관하여 기술하며, 보다 구체적으로 문장 데이터를 여러 번의 번역을 통해 원본 데이터와 의미는 비슷하지만 단어나 형태가 다른 변형 문장을 만들고, Word2Vec을 이용하여 보정하는 방식을 통해 데이터 증강시키는 기술을 제공한다.

[0007] 실시예들은 번역기가 반복적으로 실행되며 원본 문장이 심하게 왜곡된 변형 문장이 생성되었더라도 Word2Vec을 이용하여 원본 문장과 변형 문장의 중간 벡터로 증강 문장을 생성함으로써, 보다 나은 문장을 생성할 수 있는 번역 기반 문장 데이터 변형과 Word2Vec 보정을 이용한 문장 데이터 증강 방법 및 장치를 제공하는데 있다.

과제의 해결 수단

[0008] 일 실시예에 따른 컴퓨터 장치를 통해 구현되는 Word2Vec 보정을 이용한 문장 데이터 증강 방법은, 원본 문장 및 상기 원본 문장의 변형 문장의 단어를 미리 학습된 Word2Vec에 입력하여 벡터로 변환하여, 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터를 획득하는 단계; 상기 Word2Vec의 벡터 공간에서 변환된 상기 원본 문장 단어의 벡터 및 상기 변형 문장 단어의 벡터의 중간 벡터에서 가장 가까운 벡터를 선택하는 단계; 및 선택된 상기 벡터를 상기 Word2Vec을 통해 단어로 변환하고, 상기 단어를 순서대로 나열하여 증강 문장을 생성하는 단계를 포함하여 이루어질 수 있다.

[0009] 상기 원본 문장을 번역기에 입력하여 상기 변형 문장을 생성하는 단계를 더 포함할 수 있다.

[0010] 상기 원본 문장을 상기 Word2Vec에 입력하여 단어를 벡터로 변환하도록 학습하는 단계를 더 포함할 수 있다.

[0011] 다른 실시예에 따른 Word2Vec 보정을 이용한 문장 데이터 증강 장치는, 원본 문장 및 상기 원본 문장의 변형 문장의 단어를 미리 학습된 Word2Vec에 입력하여 벡터로 변환하여, 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터를 획득하는 벡터화부; 상기 Word2Vec의 벡터 공간에서 변환된 상기 원본 문장 단어의 벡터 및 상기 변형 문장 단어의 벡터의 중간 벡터에서 가장 가까운 벡터를 선택하는 벡터 선택부; 및 선택된 상기 벡터를 상기 Word2Vec을 통해 단어로 변환하고, 상기 단어를 순서대로 나열하여 증강 문장을 생성하는 증강 문장 생성부를 포함하여 이루어질 수 있다.

[0012] 상기 원본 문장을 번역기에 입력하여 상기 변형 문장을 생성하는 변형 문장 생성부를 더 포함할 수 있다.

발명의 효과

[0013] 실시예들에 따르면 문장 데이터를 여러 번의 번역을 통해 원본 데이터와 의미는 비슷하지만 단어나 형태가 다른 변형 문장을 만들고, Word2Vec을 이용하여 보정하는 방식을 통해 데이터 증강시키는 번역 기반 문장 데이터 변형과 Word2Vec 보정을 이용한 문장 데이터 증강 방법 및 장치를 제공할 수 있다.

[0014] 실시예들에 따르면 번역기가 반복적으로 실행되며 원본 문장이 심하게 왜곡된 변형 문장이 생성되었더라도 Word2Vec을 이용하여 원본 문장과 변형 문장의 중간 벡터로 증강 문장을 생성함으로써, 보다 나은 문장을 생성할 수 있는 번역 기반 문장 데이터 변형과 Word2Vec 보정을 이용한 문장 데이터 증강 방법 및 장치를 제공할 수 있다.

도면의 간단한 설명

[0015] 도 1은 일 실시예에 따른 Word2Vec의 동작 예시를 나타내는 도면이다.

도 2는 일 실시예에 따른 구글 번역기를 활용한 변형 문장 생성을 설명하기 위한 도면이다.

도 3은 일 실시예에 따른 컴퓨터 장치의 내부 구성의 일례를 설명하기 위한 블록도이다.

도 4는 일 실시예에 따른 문장 데이터 증강 장치를 나타내는 블록도이다.

도 5는 일 실시예에 따른 문장 데이터 증강 방법을 나타내는 흐름도이다.

도 6은 일 실시예에 따른 원본 문장과 변형 단어의 중간 단어로 증강 문장 생성의 예시를 나타내는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0016] 이하, 첨부된 도면을 참조하여 실시예들을 설명한다. 그러나, 기술되는 실시예들은 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 이하 설명되는 실시예들에 의하여 한정되는 것은 아니다. 또한, 여러 실시예들은 당해 기술분야에서 평균적인 지식을 가진 자에게 본 발명을 더욱 완전하게 설명하기 위해서 제공되는 것이다. 도면에서 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0018] 아래의 실시예들은 번역 기반 문장 데이터 변형과 Word2Vec 보정을 이용한 문장 데이터 증강 기술을 제공한다. 실시예들은 문장 데이터를 여러 번의 번역을 통해 원본 데이터와 의미는 비슷하지만 단어나 형태가 다른 변형 문장을 만들고 Word2Vec을 이용하여 보정하는 방식을 통해 데이터 증강시킬 수 있다.

[0020] 도 1은 일 실시예에 따른 Word2Vec의 동작 예시를 나타내는 도면이다.

[0021] 컴퓨터가 자연어를 처리하기 위해서는 자연어를 계산 가능한 형태로 바꾸어 주어야 한다. 자연어를 계산 가능한 벡터 형태로 바꾸는 기법을 임베딩이라고 한다. 임베딩에는 말뭉치(corpus)의 의미, 문법 정보가 응축되어 있다. 벡터이기 때문에 사칙연산이 가능하며, 단어/문서 관련도 역시 계산 가능하다. 여기에서는 임베딩의 여러 기법 중 하나인 Word2Vec를 사용하여 단어의 유사도를 파악하고 문장 증강에 사용할 수 있다.

[0022] Word2Vec에는 CBOW(Continuous Bag of Words)와 Skip-Gram 두 가지 방식이 있다. CBOW는 주변에 있는 단어들을 가지고, 중간에 있는 단어들을 예측하는 방법이고, Skip-Gram은 중간에 있는 단어로 주변 단어들을 예측하는 방법이다. 이 기법에서는 CBOW를 사용 하여 주변에 있는 단어로 중간에 있는 단어를 예측하여 벡터화 한다. 먼저 중심 단어를 예측하기 위해서 앞, 뒤로 몇 개의 단어를 볼지를 결정하여 윈도우(window)를 만들어 중심 단어와 주변 단어를 바꿔가며 학습을 위한 데이터를 만든다.

[0023] 도 1을 참조하면, 문자 메시지가 입력(110)으로 들어오면 미리 저장된 한국어 조사와 띄어쓰기를 기준으로 단어들을 나누고 윈도우의 크기에 따라 묶을 수 있다. 각각의 윈도우는 One-hot 인코딩(120)을 수행하여 프로젝션 레이어(projection layer)(130)에 넣어 줄 수 있다. 이것은 프로젝션 레이어(projection layer)(130)를 통과하며 룩업 테이블 연산을 한 뒤, 각 윈도우의 중간에 있는 단어의 유사도 값이 출력(140)으로 나오게 된다. 예시의 경우 '에서'와 '에'는 각각 장소 뒤에 붙는 조사와 시간 뒤에 붙는 조사이지만 같은 조사이므로 비슷한 유사도 값을 가질 수 있다.

[0024] 여기에서는 번역이 완료된 문장과 원본 문장의 차이를 계산하여야 하므로 Word2Vec을 미리 학습시켜야 한다.

이 때, 하나의 예시로써 학습에는 한국어 데이터를 가장 쉽고 빠르게 확보할 수 있는 한국어 위키백과의 데이터를 사용할 수 있다. 학습된 Word2Vec은 단어를 벡터로 바꿀 수 있게 된다.

[0026] 도 2는 일 실시예에 따른 구글 번역기를 활용한 변형 문장 생성을 설명하기 위한 도면이다.

[0027] 본 실시예에서는 번역기의 하나의 예시로써, 현재 가장 많은 데이터와 신뢰성, 그리고 지원언어를 가지는 구글 번역기를 사용하여 변형 문장을 만들고, 앞서 설명한 미리 학습된 Word2Vec을 사용하여 변형 문장을 보정함으로써, 데이터를 증강시킬 수 있다. 도 2에 도시된 바와 같이, 원본 문장(210)을 랜덤 언어로 번역하여 다른 언어 문장(220)으로 나타낼 수 있고, 랜덤 언어로 번역하는 것을 n번 반복(230)하여 수행한 후 획득한 다른 언어 문장(240)을 한국어(또는 한글)로 번역함으로써, 변형 문장(250)을 획득할 수 있다.

[0028] 보다 구체적으로, 만약 원본 문장(210) 데이터가 한국어인 경우, 구글 번역기를 활용하여 원본 문장(210) 데이터를 원본 문장(210)의 언어(한국어)가 아닌 다른 언어로 번역하여 다른 언어 문장(220)으로 나타낼 수 있다. 여기서, 다른 언어란 구글 번역기가 지원하는 109개 언어 중 한국어를 제외한 108개 언어에서 랜덤으로 선택되는 언어를 의미한다. 이 때, 이러한 랜덤 언어로 번역하는 과정을 n번 반복(230) 수행할 수 있다. n번의 번역을 마치면 마지막으로 한국어로 번역할 수 있다. n번 번역 후, 나오는 다른 언어 문장(240)은 원본 문장(210)과 형태나 단어는 다르지만, 높은 확률로 의미는 비슷한 문장이 된다. 본 실시예에서는 이것을 변형 문장(250)이라 한다.

[0029] 하지만, 이러한 변형 문장(250)은 종종 기존의 문장에서 많이 벗어난 형태를 가지거나 기존과 다른 의미를 가지는 문장이 된다. 이러한 좋지 않은 변형 문장(250)을 학습 데이터로 사용하면 오히려 성능이 떨어지는 경우가 생길 수 있다. 이것을 보완하기 위하여 딥러닝을 활용하여 좋지 않은 변형 문장들(250)을 제거할 수 있다.

[0031] 도 3은 일 실시예에 따른 컴퓨터 장치의 내부 구성의 일례를 설명하기 위한 블록도이다. 예를 들어, 일 실시예들에 따른 문장 데이터 증강 장치가 도 3의 컴퓨터 장치(300)을 통해 구현될 수 있다. 도 3에 도시한 바와 같이, 컴퓨터 장치(300)은 문장 데이터 증강 방법을 실행하기 위한 구성요소로서 프로세서(310), 메모리(320), 영구 저장 장치(330), 버스(340), 입출력 인터페이스(350) 및 네트워크 인터페이스(360)를 포함할 수 있다.

[0032] 프로세서(310)는 명령어들의 임의의 시퀀스를 처리할 수 있는 임의의 장치를 포함하거나 그의 일부일 수 있다. 프로세서(310)는 예를 들어 컴퓨터 프로세서, 이동 장치 또는 다른 전자 장치 내의 프로세서 및/또는 디지털 프로세서를 포함할 수 있다. 프로세서(310)는 예를 들어, 서버 컴퓨팅 디바이스, 서버 컴퓨터, 일련의 서버 컴퓨터들, 서버 팜, 클라우드 컴퓨터, 컨텐츠 플랫폼, 이동 컴퓨팅 장치, 스마트폰, 태블릿, 셋톱 박스, 미디어 플레이어 등에 포함될 수 있다. 프로세서(310)는 버스(340)를 통해 메모리(320)에 접속될 수 있다.

[0033] 메모리(320)는 컴퓨터 장치(300)에 의해 사용되거나 그에 의해 출력되는 정보를 저장하기 위한 휘발성 메모리, 영구, 가상 또는 기타 메모리를 포함할 수 있다. 메모리(320)는 예를 들어 랜덤 액세스 메모리(RAM: random access memory) 및/또는 동적 RAM(DRAM: dynamic RAM)을 포함할 수 있다. 메모리(320)는 컴퓨터 장치(300)의 상태 정보와 같은 임의의 정보를 저장하는 데 사용될 수 있다. 메모리(320)는 예를 들어 문장 데이터 증강을 위한 명령어들을 포함하는 컴퓨터 장치(300)의 명령어들을 저장하는 데에도 사용될 수 있다. 컴퓨터 장치(300)은 필요에 따라 또는 적절한 경우에 하나 이상의 프로세서(310)를 포함할 수 있다.

[0034] 버스(340)는 컴퓨터 장치(300)의 다양한 컴포넌트들 사이의 상호작용을 가능하게 하는 통신 기반 구조를 포함할 수 있다. 버스(340)는 예를 들어 컴퓨터 장치(300)의 컴포넌트들 사이에, 예를 들어 프로세서(310)와 메모리(320) 사이에 데이터를 운반할 수 있다. 버스(340)는 컴퓨터 장치(300)의 컴포넌트들 간의 무선 및/또는 유선 통신 매체를 포함할 수 있으며, 병렬, 직렬 또는 다른 토폴로지 배열들을 포함할 수 있다.

[0035] 영구 저장 장치(330)는 (예를 들어, 메모리(320)에 비해) 소정의 연장된 기간 동안 데이터를 저장하기 위해 컴퓨터 장치(300)에 의해 사용되는 바와 같은 메모리 또는 다른 영구 저장 장치와 같은 컴포넌트들을 포함할 수 있다. 영구 저장 장치(330)는 컴퓨터 장치(300) 내의 프로세서(310)에 의해 사용되는 바와 같은 비휘발성 메인 메모리를 포함할 수 있다. 영구 저장 장치(330)는 예를 들어 플래시 메모리, 하드 디스크, 광 디스크 또는 다른 컴퓨터 판독 가능 매체를 포함할 수 있다.

[0036] 입출력 인터페이스(350)는 키보드, 마우스, 음성 명령 입력, 디스플레이 또는 다른 입력 또는 출력 장치에 대한 인터페이스들을 포함할 수 있다. 구성 명령들 및/또는 문장 데이터 증강을 위한 정보가 입출력 인터페이스

(350)를 통해 수신될 수 있다.

[0037] 네트워크 인터페이스(360)는 근거리 네트워크 또는 인터넷과 같은 네트워크들에 대한 하나 이상의 인터페이스를 포함할 수 있다. 네트워크 인터페이스(360)는 유선 또는 무선 접속들에 대한 인터페이스들을 포함할 수 있다. 구성 명령들 및/또는 문장 데이터 증강을 위한 정보는 네트워크 인터페이스(360)를 통해 수신될 수 있다.

[0038] 또한, 다른 실시예들에서 컴퓨터 장치(300)은 도 3의 구성요소들보다 더 많은 구성요소들을 포함할 수도 있다. 그러나, 대부분의 종래기술적 구성요소들을 명확하게 도시할 필요성은 없다. 예를 들어, 컴퓨터 장치(300)은 상술한 입출력 인터페이스(350)와 연결되는 입출력 장치들 중 적어도 일부를 포함하도록 구현되거나 또는 트랜시버(transceiver), GPS(Global Positioning System) 모듈, 카메라, 각종 센서, 데이터베이스 등과 같은 다른 구성요소들을 더 포함할 수도 있다.

[0040] 도 4는 일 실시예에 따른 문장 데이터 증강 장치를 나타내는 블록도이다.

[0041] 도 4를 참조하면, 도 3의 일 실시예에 따른 컴퓨터 장치의 프로세서(310)가 포함할 수 있는 구성요소의 예를 도시한 도면이다. 여기서, 컴퓨터 장치(300)의 프로세서(310)는 일 실시예에 따른 문장 데이터 증강 장치(400)를 포함할 수 있다. 일 실시예에 따른 문장 데이터 증강 장치(400)는 벡터화부(430), 벡터 선택부(440) 및 증강 문장 생성부(450)을 포함하여 이루어질 수 있고, 실시예에 따라 문장 데이터 증강 장치(400)는 변형 문장 생성부(410) 및 학습부(420)를 더 포함할 수 있다.

[0042] 프로세서(310) 및 프로세서(310)의 구성요소들은 도 5의 문장 데이터 증강 방법이 포함하는 단계들(S110 내지 S150)을 수행할 수 있다. 예를 들어, 프로세서(310) 및 프로세서(310)의 구성요소들은 메모리(320)가 포함하는 운영체제의 코드와 상술한 적어도 하나의 프로그램 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다. 여기서, 적어도 하나의 프로그램 코드는 문장 데이터 증강 방법을 처리하기 위해 구현된 프로그램의 코드에 대응될 수 있다.

[0043] 문장 데이터 증강 방법은 도시된 순서대로 발생하지 않을 수 있으며, 단계들 중 일부가 생략되거나 추가의 과정이 더 포함될 수 있다.

[0045] 도 5는 일 실시예에 따른 문장 데이터 증강 방법을 나타내는 흐름도이다.

[0046] 도 5를 참조하면, 일 실시예에 따른 컴퓨터 장치를 통해 구현되는 Word2Vec 보정을 이용한 문장 데이터 증강 방법은, 원본 문장 및 원본 문장의 변형 문장의 단어를 미리 학습된 Word2Vec에 입력하여 벡터로 변환하여, 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터를 획득하는 단계(S130), Word2Vec의 벡터 공간에서 변환된 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터의 중간 벡터에서 가장 가까운 벡터를 선택하는 단계(S140), 및 선택된 벡터를 Word2Vec을 통해 단어로 변환하고, 단어를 순서대로 나열하여 증강 문장을 생성하는 단계(S150)를 포함하여 이루어질 수 있다.

[0047] 또한, 원본 문장을 번역기에 입력하여 변형 문장을 생성하는 단계(S110)를 더 포함할 수 있다.

[0048] 또한, 원본 문장을 Word2Vec에 입력하여 단어를 벡터로 변환하도록 학습하는 단계(S120)를 더 포함할 수 있다.

[0049] 아래에서 일 실시예에 따른 문장 데이터 증강 방법의 각 단계를 도 4에서 설명한 문장 데이터 증강 장치(400)를 예를 들어 보다 구체적으로 설명한다.

[0051] 단계(S110)에서, 변형 문장 생성부(410)는 원본 문장을 번역기에 입력하여 변형 문장을 생성할 수 있다.

[0052] 변형 문장 생성부(410)는 원본 문장을 번역기에 입력하여 n번의 번역 과정을 반복 수행하여 변형 문장을 생성할 수 있다. 보다 구체적으로, 도 2에서 설명한 바와 같이, 변형 문장 생성부(410)는 원본 문장을 기 설정된 언어로 n번 반복하여 번역하여 다른 언어 문장으로 나타낼 수 있고, 획득한 다른 언어 문장을 다시 한국어로 번역함으로써 변형 문장을 생성할 수 있다. 이 때, 변형 문장은 원본 문장이 심하게 왜곡된 형태로 생성될 수 있다.

[0053] 이에 따라 번역기로 만들어낸 변형 문장 중에서 학습에 방해되는 데이터를 제거하고 학습에 도움이 되는 좋은

품질의 데이터만 남기기 위하여 Word2Vec을 통해 최종적인 증강 문장을 생성하여 학습 등에 사용할 수 있다.

[0054] 단계(S120)에서, 학습부(420)는 원본 문장을 Word2Vec에 입력하여 단어를 벡터로 변환하도록 학습할 수 있다.

[0055] 여기서, Word2Vec은 도 1에서 설명한 바와 같이 원본 문장이 입력되면 미리 설정된 한국어 조사와 띄어쓰기를 기준으로 단어들을 나누고 윈도우의 크기에 따라 묶을 수 있다. 각각의 윈도우는 One-hot 인코딩을 수행하여 프로젝션 레이어(projection layer)에 넣어 줄 수 있고, 이것은 프로젝션 레이어(projection layer)를 통과 하며 룩업 테이블 연산을 한 뒤, 각 윈도우의 중간에 있는 단어의 유사도 값이 출력된다. 이에 따라 학습된 Word2Vec은 단어를 벡터로 바꿀 수 있게 된다.

[0056] 단계(S130)에서, 벡터화부(430)는 원본 문장 및 원본 문장의 변형 문장의 단어를 미리 학습된 Word2Vec에 입력하여 벡터로 변환하여, 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터를 획득할 수 있다.

[0057] 단계(S140)에서, 벡터 선택부(440)는 Word2Vec의 벡터 공간에서 변환된 원본 문장 단어의 벡터 및 변형 문장 단어의 벡터의 중간 벡터에서 가장 가까운 벡터를 선택할 수 있다. 이 때, 선택된 벡터는 원본 벡터, 변형 벡터와 가장 유사한 단어의 벡터가 될 수 있다.

[0058] 단계(S150)에서, 증강 문장 생성부(450)는 선택된 벡터를 Word2Vec을 통해 단어로 변환하고, 단어를 순서대로 나열하여 증강 문장을 생성할 수 있다. 이에 따라 번역기가 반복적으로 실행되며 원본 문장이 심하게 왜곡된 변형 문장이 생성되었더라도 원본 문장과 변형 문장의 중간 벡터로 증강 문장을 생성함으로써 보다 나은 문장을 생성할 수 있다.

[0060] 도 6은 일 실시예에 따른 원본 문장과 변형 단어의 중간 단어로 증강 문장 생성의 예시를 나타내는 도면이다.

[0061] 도 6에 도시된 바와 같이, 미리 학습된 Word2Vec(610, 620)으로 원본 문장의 단어(611)와 변형 문장의 단어(621)를 벡터로 변환할 수 있다. 보다 구체적으로, 원본 문장의 단어(611)를 미리 학습된 Word2Vec(610)에 입력하여 원본 문장 단어의 벡터(612)를 획득할 수 있다. 또한, 변형 문장의 단어(621)를 미리 학습된 Word2Vec(620)에 입력하여 변형 문장 단어의 벡터(622)를 획득할 수 있다. 이 때, 원본 문장의 벡터 변환을 위한 학습된 Word2Vec(610)와 변형 문장의 벡터 변환을 위한 학습된 Word2Vec(620)은 동일한 Word2Vec일 수 있다. 한편, 원본 문장의 변형 문장은 번역기 등을 통해 생성할 수 있다.

[0062] Word2Vec(610, 620)의 벡터 공간에서 원본 문장 단어의 벡터(611)와 변형 문장 단어의 벡터(621)의 중간 벡터(631)에서 가장 가까운 벡터를 선택할 수 있다. 이 때, 선택된 벡터는 원본 벡터, 변형 벡터와 가장 유사한 단어의 벡터가 될 것이다. 선택된 벡터는 다시 Word2Vec(630)을 통해 단어로 변환되고, 그 단어를 순서대로 나열하여 최종적인 증강 문장(632)을 생성할 수 있다. 여기서, Word2Vec는 하나의 Word2Vec가 사용될 수 있고, 필요에 따라 복수개의 Word2Vec가 사용될 수 있다.

[0063] 이러한 방식은 만약 도 2에서 번역기가 반복적으로 실행되며 원본 문장이 심하게 왜곡된 변형 문장이 생성되었더라도 도 6의 과정을 거치며 원본 문장과 변형 문장의 중간 벡터로 증강 문장을 생성함으로써 더 나은 문장을 만들 수 있다.

[0064] 본 실시예들에 따른 방식으로 만들어지는 증강 문장은 원본 문장과 비슷한 의미를 가지지만, 단어나 형태가 다른 문장이 된다. 증강 문장은 기계학습, 딥러닝 등의 학습 데이터로 사용되거나 자동 안내 멘트 생성 등 자연어 생성이 필요한 다양한 분야에서 활용 가능하다.

[0066] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 컨트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설

명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0067] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

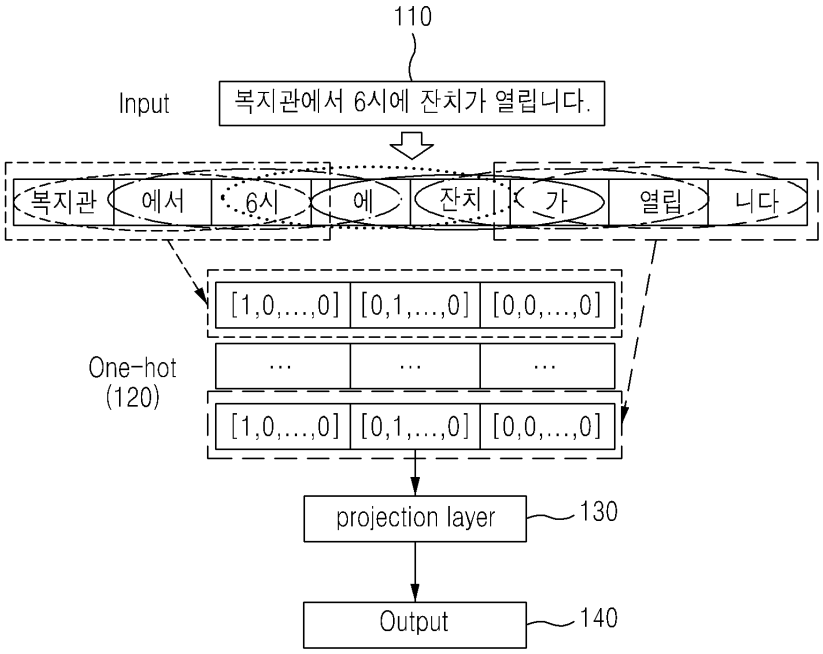
[0068] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0069] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

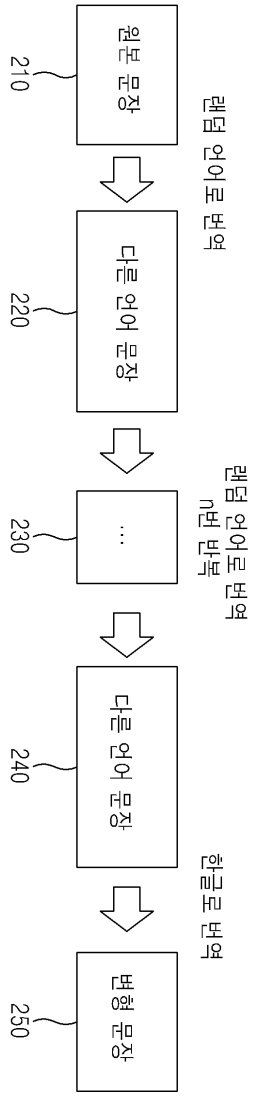
[0070] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

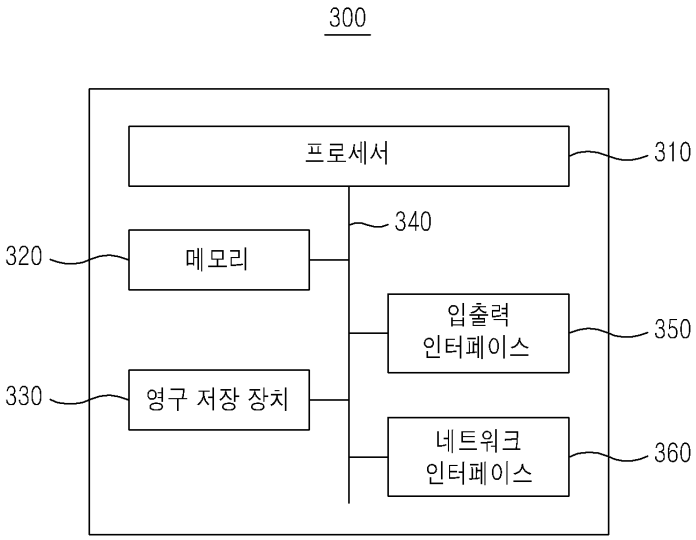
도면1


110
Input
복지관에서 6시에 잔치가 열립니다.
복지관
에서
6시
에
잔치
가
열립
니다
[1,0,...,0]
[0,1,...,0]
[0,0,...,0]
One-hot
(120)
...
...
...
[1,0,...,0]
[0,1,...,0]
[0,0,...,0]
projection layer
130
Output
140

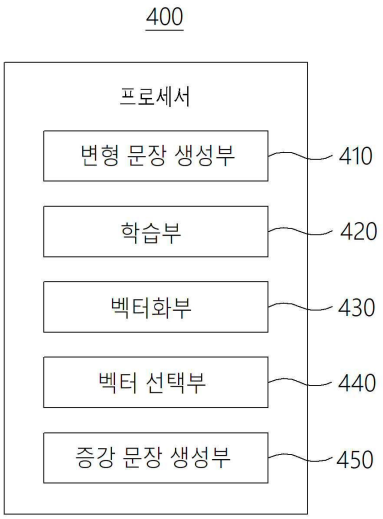
도면2


원본 문장
210
래덤 언어로 번역
다른 언어 문장
220
래덤 언어로 번역
...
230
래덤 언어로 번역
다른 언어 문장
240
한글로 번역
변형 문장
250

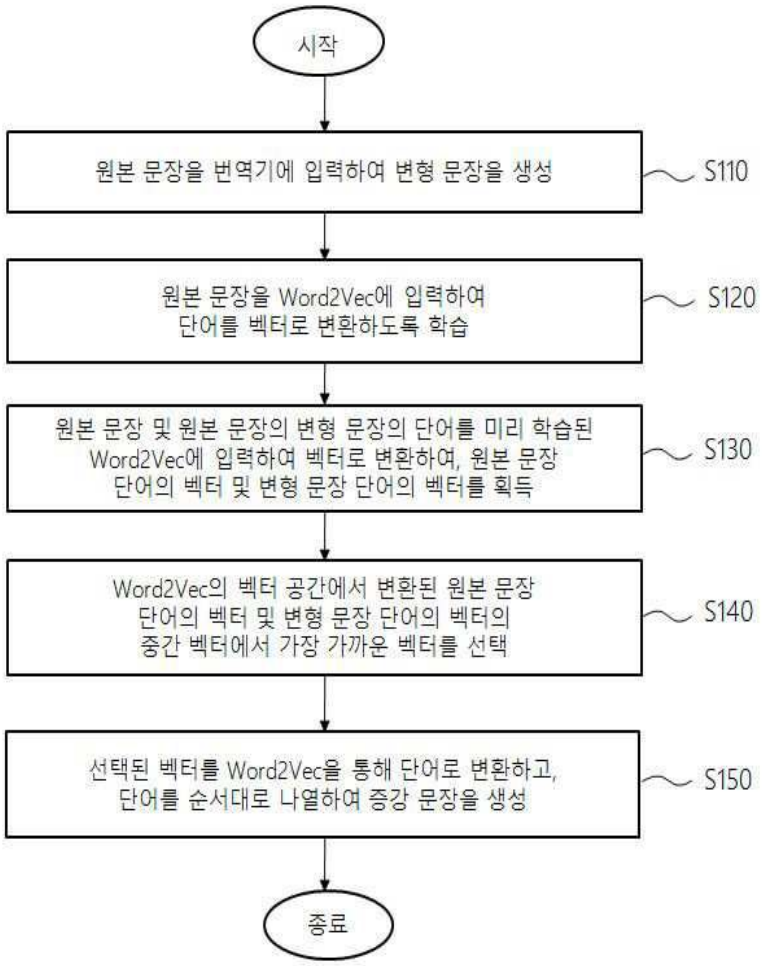
도면3


300
프로세서
310
340
320
메모리
입출력
인터페이스
350
330
영구 저장 장치
네트워크
인터페이스
360

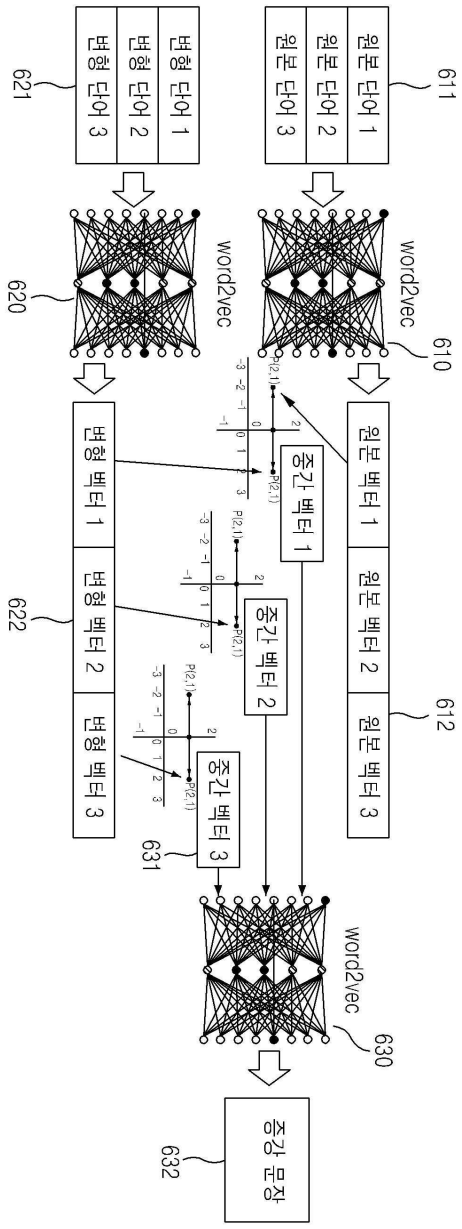
도면4


400
프로세서
변형 문장 생성부
410
학습부
420
벡터화부
430
벡터 선택부
440
증강 문장 생성부
450

도면5


시작
원본 문장을 번역기에 입력하여 변형 문장을 생성
S110
원본 문장을 Word2Vec에 입력하여
단어를 벡터로 변환하도록 학습
S120
원본 문장 및 원본 문장의 변형 문장의 단어를 미리 학습된
Word2Vec에 입력하여 벡터로 변환하여, 원본 문장
단어의 벡터 및 변형 문장 단어의 벡터를 획득
S130
Word2Vec의 벡터 공간에서 변환된 원본 문장
단어의 벡터 및 변형 문장 단어의 벡터의
중간 벡터에서 가장 가까운 벡터를 선택
S140
선택된 벡터를 Word2Vec을 통해 단어로 변환하고,
단어를 순서대로 나열하여 증강 문장을 생성
S150
종료

도면6


원본 단어 1
원본 단어 2
원본 단어 3
611
변형 단어 1
변형 단어 2
변형 단어 3
621
word2vec
610
620
원본 벡터 1
원본 벡터 2
원본 벡터 3
612
변형 벡터 1
변형 벡터 2
변형 벡터 3
622
중간 벡터 1
중간 벡터 2
중간 벡터 3
631
word2vec
630
증강 문장
632