



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년12월14일
(11) 등록번호 10-1929509
(24) 등록일자 2018년12월10일

(51) 국제특허분류(Int. Cl.)

G06F 17/27 (2006.01)

(52) CPC특허분류

G06F 17/2755 (2013.01)

(21) 출원번호 10-2018-0111140

(22) 출원일자 2018년09월17일

심사청구일자 2018년09월17일

(56) 선행기술조사문헌

최용석 외, Sequence-to-Sequence 모델 기반으로 한 한국어 형태소 분석의 재순위화 모델, 정보처리학회 논문지, 소프트웨어 및 데이터 공학 제7권 제4호 (2018.04.)*

(뒷면에 계속)

(73) 특허권자

서강대학교 산학협력단

서울특별시 마포구 백범로 35 (신수동, 서강대학교)

(72) 발명자

서정연

서울특별시 서초구 신반포로16길 15-20, 103동 2702호 (반포동, 반포힐스테이트)

홍태석

경기도 고양시 일산동구 일산로 205, 211동 1802호 (마두동, 백마마을2단지아파트)

(뒷면에 계속)

(74) 대리인

유미특허법인

전체 청구항 수 : 총 8 항

심사관 : 경연정

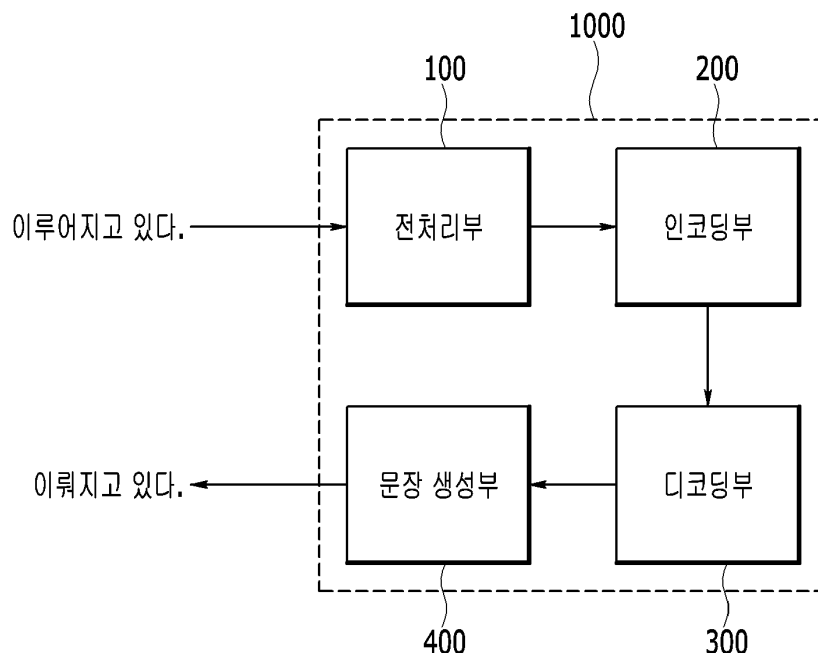
(54) 발명의 명칭 형태소 합성 장치 및 방법

(57) 요약

형태소 합성 장치로서, 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하고, 상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩

(뒷면에 계속)

대표도 - 도1



벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 전처리부, 상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU(Gated Recurrent Unit)에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 인코딩부, 상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 디코딩부, 그리고 상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 문장 생성부를 포함한다.

(72) 발명자

김주애

서울특별시 은평구 갈현로47길 13, 402호 (갈현동)

박영민

경기도 군포시 송부로 291 번안길3-35, 503호 (부곡동, 엠스테이)

서민영

경기도 부천시 경인로 484번길 28

허광호

서울특별시 강서구 공항대로48길 66, 301호

(56) 선행기술조사문헌

박인철, 형태소 합성 기법을 이용한 형태소 패턴 사전의 반자동 구축, 한국산학기술학회 논문지 제12권제11호 (2011.12.11.)*

김도우, Doc2Vec을 활용한 CNN 기반 한국어 신문 기사 분류에 관한 연구, 서강대학교 석사학위 논문 (2017.01.05.)

J. Chung et al., Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, www.arxiv.org (2014.12.11.)

KR101117427 B1

*는 심사관에 의하여 인용된 문헌

이 발명을 지원한 국가연구개발사업

과제고유번호 2017-0-00255

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 정보통신방송기술개발사업

연구과제명 (지능정보-총괄/1세부) 자율지능 디지털 동반자 프레임워크 및 응용 연구개발

기 여 율 1/1

주관기관 전자부품연구원

연구기간 2017-04-01~2017-12-31

공지예외적용 : 있음

명세서

청구범위

청구항 1

형태소 합성 장치로서,

입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하고, 상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 전처리부,

상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU(Gated Recurrent Unit)에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 인코딩부,

상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 디코딩부, 그리고

상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 문장 생성부를 포함하고,

상기 인코딩부는

상기 제1 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제1 은닉 값을 결정하고, 상기 제1 은닉 값과 상기 제1 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 제2 은닉 값을 결정하고, 상기 제2 은닉 값과 상기 제2 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제3 은닉 값을 결정하고, 상기 제3 은닉 값과 상기 제2 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 상기 컨텍스트 벡터를 생성하는 형태소 합성 장치.

청구항 2

제1항에서,

상기 제1 형태소는 상기 입력 문장에서 상기 제2 형태소보다 앞에 위치하는 형태소 합성 장치.

청구항 3

삭제

청구항 4

형태소 합성 장치로서,

입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하고, 상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 전처리부,

상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU(Gated Recurrent Unit)에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 인코딩부,

상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 디코딩부, 그리고

상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 문장 생성부를 포함하고,

상기 디코딩부는

상기 컨텍스트 벡터 및 상기 개시 임베딩 벡터를 상기 제2 GRU에 입력하여 제4 은닉 값을 결정하고 상기 제1 음절을 출력하며, 상기 제4 은닉 값 및 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 상기 제2 음절을 출력하는 형태소 합성 장치.

청구항 5

형태소 합성 장치로서,

입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하고, 상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 전처리부,

상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU(Gated Recurrent Unit)에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 인코딩부,

상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 디코딩부, 그리고

상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 문장 생성부를 포함하고,

상기 문장 생성부는

문장의 끝을 알리는 종료 음절이 출력되는 경우, 상기 종료 음절의 이전에 출력된 상기 제1 음절 및 상기 제2 음절을 순차적으로 연결하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 형태소 합성 장치.

청구항 6

형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 방법으로서,

입력 문장을 수신하고, 상기 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하는 단계,

상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 단계,

상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 단계,

상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 단계, 그리고

상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 단계를 포함하고,

상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 단계는

상기 제1 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제1 은닉 값을 결정하는 단계,

상기 제1 은닉 값과 상기 제1 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 제2 은닉 값을 결정하는 단계,

상기 제2 은닉 값과 상기 제2 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제3 은닉 값을 결정하는 단계, 그리고

상기 제3 은닉 값과 상기 제2 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 상기 컨텍스트 벡터를 생성하는 단계

를 포함하는 문장 생성 방법.

청구항 7

제6항에서,

상기 제1 형태소는 상기 입력 문장에서 상기 제2 형태소보다 앞에 위치하는 문장 생성 방법.

청구항 8

삭제

청구항 9

형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 방법으로서,

입력 문장을 수신하고, 상기 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하는 단계,

상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 단계,

상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 단계,

상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 단계, 그리고

상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 단계를 포함하고,

상기 제1 음절 및 상기 제2 음절을 출력하는 단계는

상기 컨텍스트 벡터 및 상기 개시 임베딩 벡터를 상기 제2 GRU에 입력하여 제4 은닉 값을 결정하고 상기 제1 음절을 출력하는 단계, 그리고

상기 제4 은닉 값 및 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 상기 제2 음절을 출력하는 단계를 포함하는 문장 생성 방법.

청구항 10

형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 방법으로서,

입력 문장을 수신하고, 상기 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하는 단계,

상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 단계,

상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 단계,

상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 단계, 그리고

상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 단계를 포함하고,

상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 단계는

문장의 끝을 알리는 종료 음절이 출력되는 경우, 상기 종료 음절의 이전에 출력된 상기 제1 음절 및 상기 제2 음절을 순차적으로 연결하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 문장 생성 방법.

발명의 설명

기술 분야

[0001] 본 발명은 형태소 합성 장치 및 방법에 관한 것이다.

배경 기술

[0002] 최근 자연어 처리 분야에서 심층 신경망(DNN, Deep Neural Network) 알고리즘을 활용한 연구가 활발히 진행되고 있으며, 특히 문장 생성이 요구되는 기계 번역 또는 자연어 생성 등의 분야에서는 순환 신경망(RNN, Recurrent Neural Network) 알고리즘을 기반으로 하는 시퀀스 투 시퀀스(Seq2Seq, Sequence to Sequence) 모델이 주로 적용되고 있다. Seq2Seq 모델은 시간적 순서가 있는 연속되는 입력에 대해 상응하는 또 다른 연속되는 출력을 생성하는 모델이다.

[0003] 이러한 작업에서 한국어 문장을 출력할 때, 음절, 어절 또는 형태소 단위로 한국어 문장을 출력할 수 있다. 그러나, 음절 단위의 출력은 순환 신경망의 길이가 지나치게 길어지고 표현력이 낮아지고, 어절 단위의 출력은 데이터 희소성이 문제가 되어 높은 성능을 제공하지 못하는 문제가 있으므로, 형태소 단위로 출력하는 것이 단순히 음절 또는 어절 단위로 출력하는 것보다 더 높은 정확도를 제공할 수 있다.

[0004] 한편, 형태소 단위로 한국어 문장을 출력하는 경우에도 출력된 한국어 문장이 최종적인 자연어 문장이 아니기 때문에, 형태소 합성 모델이 필요하다. 기존의 형태소 합성 모델은 말뭉치에서 규칙을 자동 또는 수동으로 구축하였다. 하지만 이러한 규칙 기반 모델은 미리 정의되지 않은 규칙에 대해서는 전혀 대응을 하지 못하는 문제가 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명이 해결하고자 하는 과제는 Seq2Seq 모델을 이용한 형태소 합성 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0006] 본 발명의 일 실시예에 따른 형태소 합성 장치는 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하고, 상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 전처리부, 상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU(Gated Recurrent Unit)에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 인코딩부, 상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 디코딩부, 그리고 상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 문장 생성부를 포함한다.

[0007] 상기 제1 형태소는 상기 입력 문장에서 상기 제2 형태소보다 앞에 위치한다.

[0008] 상기 인코딩부는 상기 제1 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제1 은닉 값을 결정하고, 상기 제1 은닉 값과 상기 제1 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 제2 은닉 값을 결정하고, 상기 제2 은닉 값과 상기 제2 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제3 은닉 값을 결정하고, 상기 제3 은닉 값과 상기 제2 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 상기 컨텍스트 벡터를 생성한다.

[0009] 상기 디코딩부는 상기 컨텍스트 벡터 및 상기 개시 임베딩 벡터를 상기 제2 GRU에 입력하여 제4 은닉 값을 결정하고 상기 제1 음절을 출력하며, 상기 제4 은닉 값 및 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 상기 제2 음절을 출력한다.

[0010] 상기 문장 생성부는 문장의 끝을 알리는 종료 음절이 출력되는 경우, 상기 종료 음절의 이전에 출력된 상기 제1 음절 및 상기 제2 음절을 순차적으로 연결하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성한다.

[0011] 본 발명의 일 실시예에 따른 형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 방법

은 입력 문장을 수신하고, 상기 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 상기 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정하는 단계, 상기 제1 형태소, 상기 제2 형태소, 상기 제1 품사 정보 및 상기 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성하는 단계, 상기 제1 형태소 임베딩 벡터, 상기 제1 품사 정보 임베딩 벡터, 상기 제2 형태소 임베딩 벡터, 상기 제2 품사 정보 임베딩 벡터 순으로 제1 GRU에 입력하여 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 단계, 상기 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 제2 음절을 출력하는 단계, 그리고 상기 제1 음절 및 상기 제2 음절을 이용하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 단계를 포함한다.

[0012] 상기 제1 형태소는 상기 입력 문장에서 상기 제2 형태소보다 앞에 위치한다.

[0013] 상기 입력 문장에 대한 컨텍스트 벡터를 생성하는 단계는 상기 제1 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제1 은닉 값을 결정하는 단계, 상기 제1 은닉 값과 상기 제1 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 제2 은닉 값을 결정하는 단계, 상기 제2 은닉 값과 상기 제2 형태소 임베딩 벡터를 상기 제1 GRU에 입력하여 제3 은닉 값을 결정하는 단계, 그리고 상기 제3 은닉 값과 상기 제2 품사 정보 임베딩 벡터를 상기 제1 GRU에 입력하여 상기 컨텍스트 벡터를 생성하는 단계를 포함한다.

[0014] 상기 제1 음절 및 상기 제2 음절을 출력하는 단계는 상기 컨텍스트 벡터 및 상기 개시 임베딩 벡터를 상기 제2 GRU에 입력하여 제4 은닉 값을 결정하고 상기 제1 음절을 출력하는 단계, 그리고 상기 제4 은닉 값 및 상기 제1 음절의 임베딩 벡터를 상기 제2 GRU에 입력하여 상기 제2 음절을 출력하는 단계를 포함한다.

[0015] 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성하는 단계는 문장의 끝을 알리는 종료 음절이 출력되는 경우, 상기 종료 음절의 이전에 출력된 상기 제1 음절 및 상기 제2 음절을 순차적으로 연결하여 상기 제1 형태소 및 상기 제2 형태소가 합성된 문장을 생성한다.

발명의 효과

[0016] 본 발명에 따르면, 형태소 합성 규칙을 정의하지 않고 심층 신경망 알고리즘 기반의 형태소 합성 모델을 이용하여 한국어 형태소를 합성하는바, 형태소 합성 규칙을 정의하는 번거로움이 없고, 규칙이 정의되지 않은 상황에서도 한국어 형태소 합성이 가능하다.

도면의 간단한 설명

[0017] 도 1은 한 실시예에 따른 형태소 합성 장치를 설명하는 도면이다.

도 2 및 도 3은 한 실시예에 따른 형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 예시적인 방법을 도시한 도면이다.

도 4는 형태소 합성 장치가 시퀀스 투 시퀀스 모델에 기반한 형태소 합성 모델을 학습시키는 방법을 설명하는 도면이다.

도 5는 한 실시예에 따른 형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 방법을 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0018] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 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 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0020] 도 1은 한 실시예에 따른 형태소 합성 장치를 설명하는 도면이다.

[0021] 도 1을 참고하면, 형태소 합성 장치(1000)는 전처리부(100), 인코딩부(200), 디코딩부(300) 및 문장 생성부

(400)를 포함한다.

- [0022] 전처리부(100)는 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 제1 형태소에 대한 제1 품사 정보 및 제2 형태소에 대한 제2 품사 정보를 결정하고, 제1 형태소, 제2 형태소, 제1 품사 정보 및 제2 품사 정보에 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성한다.
- [0023] 구체적으로, 전처리부(100)는 사용자로부터 입력 문장을 수신하고, 형태소 데이터베이스(미도시)를 이용하여 입력 문장을 구성하는 모든 형태소인 제1 형태소 및 제2 형태소들을 추출한다. 이 경우, 제1 형태소는 입력 문장에서 제2 형태소보다 앞에 위치한다.
- [0024] 전처리부(100)는 추출한 형태소들 각각의 품사 정보를 결정한다.
- [0025] 전처리부(100)는 각각의 형태소들에 대해 체언, 용언, 관형사, 부사, 감탄사, 조사, 어미, 접사, 어근, 부호 또는 한글 이외 중 어느 하나로 품사를 결정할 수 있다.
- [0026] 또한, 전처리부(100)는 형태소가 체언인 경우, 일반 명사, 고유 명사, 의존 명사, 수사 또는 대명사 중 어느 하나로 결정할 수 있다.
- [0027] 또한, 전처리부(100)는 형태소가 용언인 경우, 동사, 형용사, 보조 용언, 긍정 지정사 또는 부정 지정사, 부사인 경우 일반 부사 또는 접속 부사 중 어느 하나로 결정할 수 있다.
- [0028] 또한, 전처리부(100)는 형태소가 조사인 경우, 주격 조사, 보격 조사, 관형격 조사, 목적격 조사, 부사격 조사, 호격 조사, 인용격 조사, 보조사 또는 접속 조사 중 어느 하나로 결정할 수 있다.
- [0029] 또한, 전처리부(100)는 형태소가 어말 어미인 경우, 종결 어미, 연결 어미, 명사형 전성 어미 또는 관형형 전성 어미 중 어느 하나로 결정할 수 있다.
- [0030] 또한, 전처리부(100)는 형태소가 접미사인 경우, 명사 파생 접미사, 동사 파생 접미사 또는 형용사 파생 접미사 중 어느 하나로 결정할 수 있다.
- [0031] 또한, 전처리부(100)는 형태소가 부호의 경우, 마침표, 물음표, 느낌표, 쉼표, 가운데점, 콜론, 빗금, 따옴표, 괄호표, 줄표, 줄임표, 붙임표 또는 기타기호 중 어느 하나로 결정할 수 있다.
- [0032] 또한, 전처리부(100)는 형태소가 한글 이외의 경우 외국어, 한자 또는 숫자 중 어느 하나로 결정할 수 있다.
- [0033] 한편, 띄어쓰기로 인한 공란의 경우, 형태소에 해당하지 않으나 이후 형태소들을 합성한 문장에서 표현되어야 하므로 별도로 구분하고 품사 정보는 띄어쓰기로 결정할 수 있다.
- [0034] 형태소 데이터베이스는 말뭉치를 구성하는 문장들에 대해 각각의 문장을 구성하는 형태소들을 정의하고, 정의된 형태소들의 품사 및 특정 품사의 세부 품사를 구분한 정보를 저장한 데이터베이스를 지칭한다.
- [0035] 전처리부(100)는 형태소들 및 형태소들 각각의 품사 정보에 벡터화 알고리즘을 적용하여 임베딩 벡터들을 각각 생성한다. 벡터화 알고리즘을 이용하여 단어를 벡터화하는 방법은 이미 공지된 기술로 본 명세서에서는 자세한 설명을 생략한다.
- [0036] 인코딩부(200)는 제1 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터, 제2 형태소 임베딩 벡터, 제2 품사 정보 임베딩 벡터 순으로 제1 GRU(Gated Recurrent Unit)에 입력하여 입력 문장에 대한 컨텍스트 벡터를 생성한다.
- [0037] 구체적으로, 인코딩부(200)는 제1 형태소 임베딩 벡터를 제1 GRU에 입력하여 제1 은닉 값(hidden state)을 결정하고, 제1 은닉 값과 제1 품사 정보 임베딩 벡터를 제1 GRU에 입력한다.

수학적식 1

$$r_j = \sigma \left([W_r x]_j + [U_r h_{(t-1)}]_j \right)$$

$$z_j = \sigma \left([W_z x]_j + [U_z h_{(t-1)}]_j \right)$$

[0038]

[0039] 수학식 1에서, σ 는 시그모이드(sigmoid) 함수이고, $[\cdot]_j$ 는 벡터의 j 번째 값을 의미한다. x 및 $h_{(t-1)}$ 는 각각 입력과 은닉 값이고, W 및 U 는 학습되는 인자들이다. 인코딩부(200)는 수학식 1을 이용하여 은닉 값을 결정하며, 최종적으로 결정되는 은닉 값은 수학식 2와 같다.

수학식 2

$$h_j^{(t)} = z_j h_j^{(t)} + (1 - z_j) \tilde{h}_j^{(t)}$$

$$\tilde{h}_j^{(t)} = \phi([Wx]_j + [U(r \odot h_{(t-1)})]_j)$$

[0040]

[0041] 인코딩부(200)는 제1 은닉 값과 제1 품사 정보 임베딩 벡터를 제1 GRU에 입력하여 제2 은닉 값을 결정하고, 제2 은닉 값과 제2 형태소 임베딩 벡터를 제1 GRU에 제차 입력하여 제3 은닉 값을 결정한다. 이후, 인코딩부(200)는 제3 은닉 값과 제2 품사 정보 임베딩 벡터를 제1 GRU에 입력하며, 입력 문장을 통해 결정된 임베딩 벡터 중 제2 품사 정보 임베딩 벡터가 마지막 입력이므로, 제3 은닉 값과 제2 품사 정보 임베딩 벡터를 제1 GRU에 입력하여 출력된 벡터가 입력 문장에 대한 컨텍스트 벡터에 해당한다.

[0042] 디코딩부(300)는 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 GRU에 입력하여 제1 음절을 출력하고, 제1 음절의 임베딩 벡터를 제2 GRU에 입력하여 제2 음절을 출력한다.

[0043] 구체적으로, 디코딩부(300)는 인코딩부(200)로부터 컨텍스트 벡터를 수신하고, 컨텍스트 벡터 및 개시 임베딩 벡터를 제2 GRU에 입력하여 제4 은닉 값을 결정하고 제1 음절을 출력한다.

[0044] 이 경우, 디코딩부(300)는 출력 레이어로서, 소프트맥스 분류기(Softmax Classifier)를 이용하여 출력 가능한 음절들에 대한 확률 분포를 출력하고, 가장 높은 확률 분포를 가진 음절을 제1 음절로서 출력할 수 있다.

[0045] 또한, 디코딩부(300)는 제4 은닉 값 및 제1 음절의 임베딩 벡터를 제2 GRU에 입력하여 제2 음절을 출력한다.

[0046] 문장 생성부(400)는 제1 음절 및 제2 음절을 이용하여 제1 형태소 및 제2 형태소가 합성된 문장을 생성한다.

[0047] 구체적으로, 문장 생성부(400)는 문장의 끝을 알리는 종료 음절이 출력되는 경우, 종료 음절 이전에 출력된 제1 음절 및 제2 음절을 순차적으로 연결하여 제1 형태소 및 제2 형태소가 합성된 문장을 생성하고, 합성된 문장을 출력한다.

[0048] 도 2 및 도 3은 한 실시예에 따른 형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 예시적인 방법을 도시한 도면이다.

[0049] 도 2를 참고하면, 사용자로부터 입력 문장 "이루어지고 있다."를 수신한 경우, 전처리부(100)는 입력 문장을 형태소 데이터베이스에서 검색하여 형태소 단위인 "이루어지", "고", " ", "있", "다", "."로 구분할 수 있다.

[0050] 또한, 전처리부(100)는 형태소 데이터베이스에서 정의된 바와 같이, 구분한 형태소들 각각의 품사 정보를 결정할 수 있다. 예를 들면, 전처리부(100)는 형태소 "이루어지"에 대해 동사로, "고"에 대해 연결 어미로, " "에 대해 띄어쓰기로, "있"에 대해 보조 용언으로, "다"에 대해 종결 어미로, "."에 대해 마침표로 각각 품사 정보를 결정할 수 있다.

[0051] 전처리부(100)는 입력 문장을 복수의 형태소들로 구분하고 복수의 형태소들 각각의 품사 정보를 결정한 이후 형태소를 품사 정보와 매핑할 수 있다.

[0052] 전처리부(100)는 형태소들 및 품사 정보에 벡터화 알고리즘을 적용하여 형태소들 및 품사 정보에 대한 임베딩 벡터를 각각 생성한다.

[0053] 이 경우, 하나의 형태소가 복수의 음절로 구성된 경우, 전처리부(100)는 복수의 음절들 각각의 임베딩 벡터를 생성한다. 예를 들면, 전처리부(100)는 형태소 "이루어지"의 임베딩 벡터를 생성하는 경우, "이"의 임베딩 벡터, "루"의 임베딩 벡터, "어"의 임베딩 벡터 및 "지"의 임베딩 벡터를 생성할 수 있다.

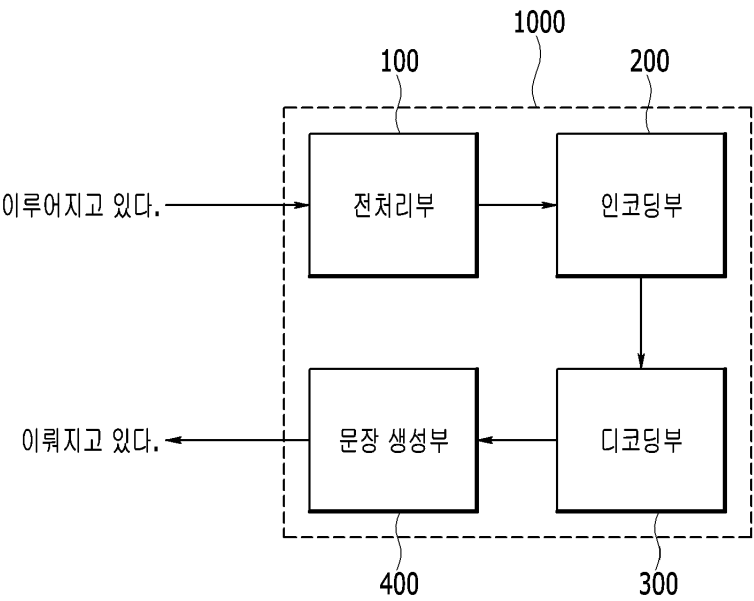
[0054] 도 3을 참고하면, 인코딩부(200)는 입력 문장에 위치한 순서대로 형태소 및 형태소에 대응하는 품사 정보의 임베딩 벡터를 제1 GRU에 입력한다.

- [0055] 예를 들면, 인코딩부(200)는 형태소 "이"의 임베딩 벡터를 제1 GRU에 입력하여 은닉 값을 결정할 수 있다. 이 경우 형태소 "이루어진"전체가 품사 정보 "동사"이므로, 인코딩부(200)는 형태소 "루"의 임베딩 벡터, "어"의 임베딩 벡터 및 "진"의 임베딩 벡터를 순차적으로 입력하여 은닉 값을 결정하고, 이후 품사 정보인 "동사"의 임베딩 벡터 및 이전 은닉 값을 제1 GRU에 입력하여 은닉 값을 순차적으로 결정할 수 있다. 인코딩부(200)는 상기 과정을 순차적으로 적용하여 입력 문장 "이루어지고 있다."의 임베딩 벡터인 컨텍스트 벡터를 생성할 수 있다.
- [0056] 디코딩부(300)는 인코딩부(200)로부터 수신한 컨텍스트 벡터 및 개시 명령어 "SOS"의 임베딩 벡터를 제2 GRU에 입력하여 은닉 값을 결정하고 음절 "이"를 출력할 수 있다. 이후, 디코딩부(300)는 음절 "이"의 임베딩 벡터 및 이전 은닉 값을 제2 GRU에 입력하여 은닉 값을 결정하고 음절 "뤄"를 출력할 수 있다.
- [0057] 상기 과정을 순차적으로 적용하여 문장의 끝을 알리는 종료 음절인 "EOS"가 출력되는 경우, 문장 생성부(400)는 "EOS"가 출력되기 이전의 음절들 "이", "뤄", "지", "고", " ", "있", "다", "."를 순차적으로 연결하여 형태소 "이루어지", "고", " ", "있", "다", "."가 합성된 문장인 "이뤄지고 있다."를 생성할 수 있다.
- [0058] 한편, 인코딩부(200) 및 디코딩부(300)는 시퀀스 투 시퀀스(Seq2Seq, Sequence to Sequence) 모델에 기반하며, 인코딩부(200) 및 디코딩부(300)를 형태소 합성 모델로 지칭할 수 있다. 도 1 내지 도 3에서 설명한 형태소 합성을 위해, 형태소 합성 모델은 학습 데이터를 통해 학습되어야 하며, 이하 형태소 합성 장치(1000)가 형태소 합성 모델을 학습하는 방법을 설명한다.
- [0059] 도 4는 형태소 합성 장치가 시퀀스 투 시퀀스 모델에 기반한 형태소 합성 모델을 학습시키는 방법을 설명하는 도면이다.
- [0060] 도 4를 참고하면, 형태소 합성 장치(1000)는 입력값 및 입력값에 대응하는 목표값으로 구성된 학습 데이터를 수신한다(S100).
- [0061] 구체적으로, 학습 데이터의 입력값은 복수의 형태소들 및 상기 복수의 형태소들 각각의 품사 정보로 구성되고, 목표값은 복수의 형태소들이 합성된 문장으로 구성된다.
- [0062] 예를 들면, 학습 데이터의 입력값은 "이", "루", "어", "지", "동사", "고", "연결어미", " ", "떨어쓰기", "있", "보조용언", "다", "종결어미", ".", "마침표"일 수 있고, 목표값은 "이", "뤄", "지", "고", " ", "있", "다", ".", "EOS"일 수 있다.
- [0063] 입력값은 형태소 및 형태소에 대응하는 품사 정보가 순차적으로 위치하며, 입력값에서 복수의 형태소들 및 각각의 품사 정보는 임베딩 벡터 형태로 구현될 수 있으며, 하나의 형태소가 복수의 음절들로 구성된 경우 복수의 음절들 각각이 임베딩 벡터 형태로 구현될 수 있다.
- [0064] 형태소 합성 장치(1000)는 입력값의 임베딩 벡터를 시퀀스 투 시퀀스 모델의 인코더에 입력한다(S110).
- [0065] 구체적으로, 형태소 합성 장치(1000)는 입력값에 포함된 복수의 형태소들 및 복수의 형태소들 각각의 품사 정보에 벡터화 알고리즘을 적용하여 복수의 형태소들의 임베딩 벡터 및 각각의 품사 정보의 임베딩 벡터를 생성한다.
- [0066] 또한, 형태소 합성 장치(1000)는 복수의 형태소들의 임베딩 벡터 및 복수의 형태소들 각각의 품사 정보의 임베딩 벡터를 교차하여 인코더에 입력한다.
- [0067] 예를 들면, 형태소 합성 장치(1000)는 "이"의 임베딩 벡터, "루"의 임베딩 벡터, "어"의 임베딩 벡터, "지"의 임베딩 벡터 형태소 및 형태소 "이루어지"의 품사 정보인 "동사"의 임베딩 벡터를 인코더에 순차적으로 입력할 수 있다.
- [0068] 형태소 합성 장치(1000)는 인코더의 출력이 시퀀스 투 시퀀스 모델의 디코더에 입력되면, 목표값이 상기 디코더의 출력으로 결정되도록 시퀀스 투 시퀀스 모델의 가중치를 결정한다.
- [0069] 구체적으로, 형태소 합성 장치(1000)는 인코더의 출력으로 입력값의 임베딩 벡터인 컨텍스트 벡터를 생성하고(S120), 컨텍스트 벡터를 디코더에 입력하여 목표값 "이", "뤄", "지", "고", "SPACE", "있", "다", ".", "EOS"가 출력으로 결정되도록 가중치를 결정한다(S130).
- [0070] 이 경우, 인코더 및 디코더는 GRU(Gated Recurrent Unit)로 구현될 수 있다.
- [0071] 도 5는 한 실시예에 따른 형태소 합성 장치가 입력 문장에 포함된 형태소들이 합성된 문장을 생성하는 방법을 설명하는 도면이다.

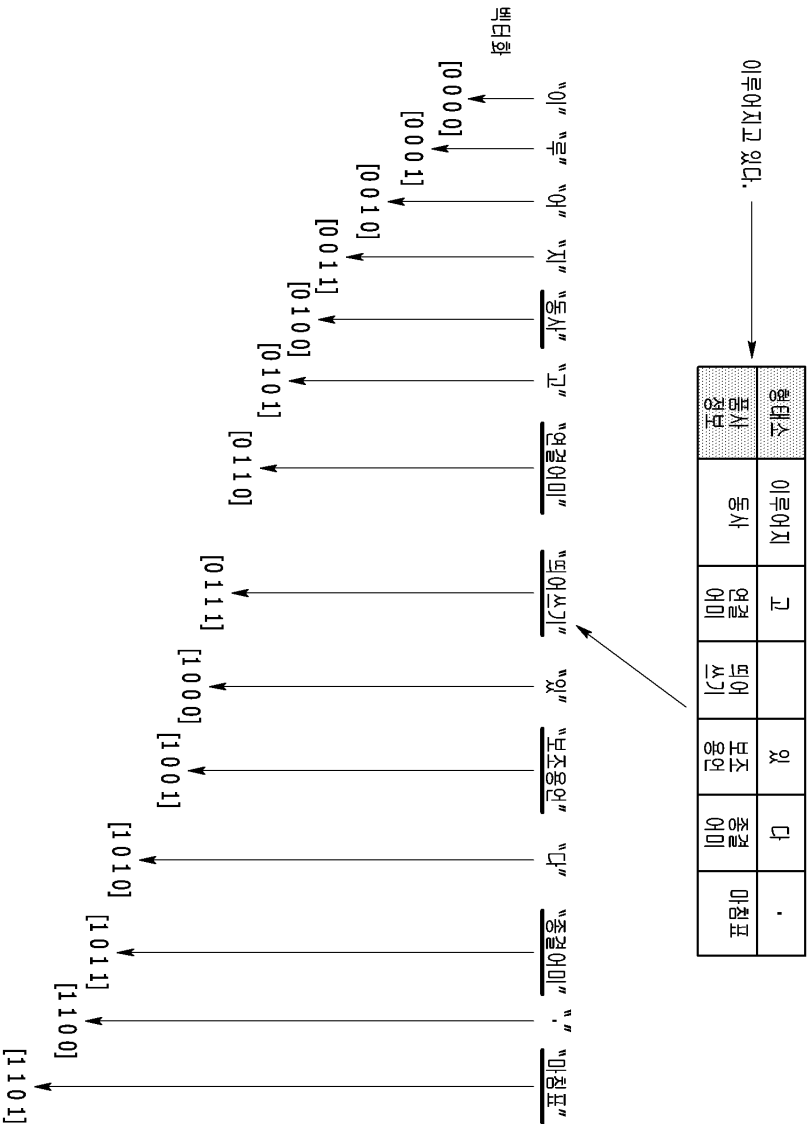
- [0072] 도 5에서, 도 1 내지 도 4와 동일한 내용은 자세한 설명을 생략한다.
- [0073] 도 5를 참고하면, 형태소 합성 장치(1000)는 입력 문장을 수신하고(S200), 입력 문장을 구성하는 제1 형태소 및 제2 형태소를 추출하고, 제1 형태소에 대한 제1 품사 정보 및 상기 제2 형태소에 대한 제2 품사 정보를 결정한다(S210).
- [0074] 형태소 합성 장치(1000)는 벡터화 알고리즘을 적용하여 제1 형태소 임베딩 벡터, 제2 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터 및 제2 품사 정보 임베딩 벡터를 각각 생성한다(S220).
- [0075] 이 경우, 제1 형태소는 입력 문장에서 제2 형태소보다 앞에 위치한다.
- [0076] 형태소 합성 장치(1000)는 제1 형태소 임베딩 벡터, 제1 품사 정보 임베딩 벡터, 제2 형태소 임베딩 벡터, 제2 품사 정보 임베딩 벡터 순으로 제1 GRU(Gated Recurrent Unit)에 입력하고, 생성되는 은닉 값들을 이용하여 입력 문장에 대한 컨텍스트 벡터를 생성한다(S230).
- [0077] 구체적으로, 형태소 합성 장치(1000)는 제1 형태소 임베딩 벡터를 제1 GRU에 입력하여 제1 은닉 값을 결정하고, 제1 은닉 값과 제1 품사 정보 임베딩 벡터를 제1 GRU에 입력하여 제2 은닉 값을 결정하고, 제2 은닉 값과 제2 형태소 임베딩 벡터를 제1 GRU에 입력하여 제3 은닉 값을 결정하고, 제3 은닉 값과 제2 품사 정보 임베딩 벡터를 제1 GRU에 입력하여 컨텍스트 벡터를 생성한다.
- [0078] 형태소 합성 장치(1000)는 컨텍스트 벡터 및 문장의 시작을 알리는 개시 임베딩 벡터를 제2 GRU에 입력하여 제1 음절을 출력하고, 제1 음절의 임베딩 벡터를 제2 GRU에 입력하여 제2 음절을 출력한다(S240).
- [0079] 구체적으로, 형태소 합성 장치(1000)는 컨텍스트 벡터 및 개시 임베딩 벡터를 제2 GRU에 입력하여 제4 은닉 값을 결정하고 제1 음절을 출력하며, 제4 은닉 값 및 제1 음절의 임베딩 벡터를 제2 GRU에 입력하여 제2 음절을 출력한다.
- [0080] 형태소 합성 장치(1000)는 제1 음절 및 제2 음절을 연결하여 제1 형태소 및 제2 형태소가 합성된 문장을 생성한다(S250).
- [0081] 구체적으로, 형태소 합성 장치(1000)는 문장의 끝을 알리는 종료 음절이 출력되는 경우, 종료 음절의 이전에 출력된 제1 음절 및 제2 음절을 연결하여 제1 형태소 및 제2 형태소가 합성된 문장을 생성한다.
- [0082] 본 발명에 따르면, 형태소 합성 규칙을 정의하지 않고 심층 신경망 알고리즘 기반의 형태소 합성 모델을 이용하여 한국어 형태소를 분석하고 합성하는바, 형태소 합성 규칙을 정의하는 번거로움이 없고, 규칙이 정의되지 않은 상황에서도 한국어 형태소 분석 및 합성이 가능하다.
- [0083] 이상에서 설명한 본 발명의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.
- [0084] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

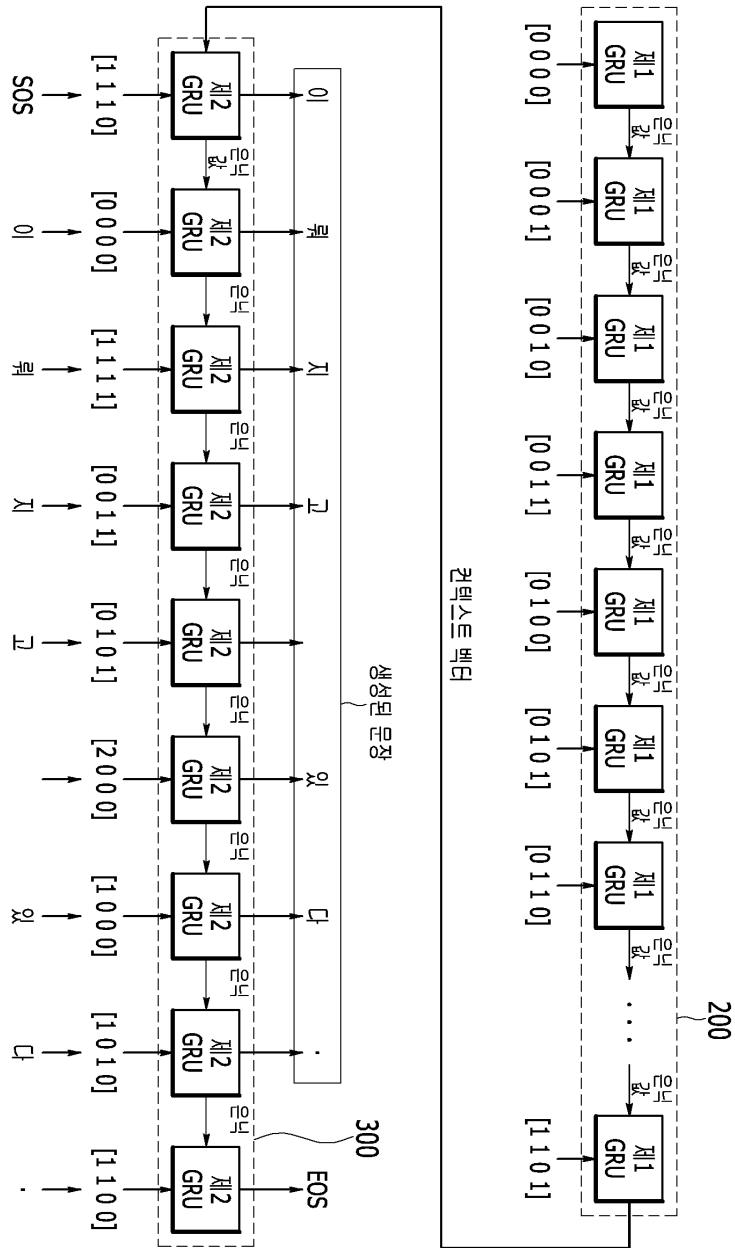
도면1



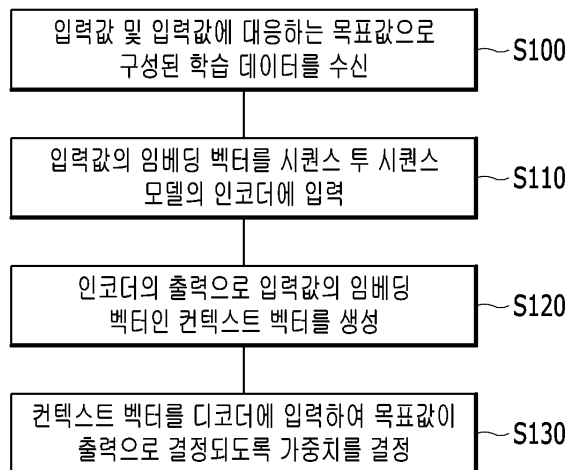
도면2



도면3



도면4



도면5

