



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0094057
(43) 공개일자 2024년06월25일

(51) 국제특허분류(Int. Cl.)
G06F 40/53 (2020.01) G06N 20/00 (2019.01)
G06N 3/04 (2023.01)
(52) CPC특허분류
G06F 40/53 (2020.01)
G06N 20/00 (2021.08)
(21) 출원번호 10-2022-0155705
(22) 출원일자 2022년11월18일
심사청구일자 2022년11월18일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애기능생활관 311호 NLP&AI연구실
손수현
서울특별시 서대문구 성산로 573-54, 204호
(74) 대리인
김동용

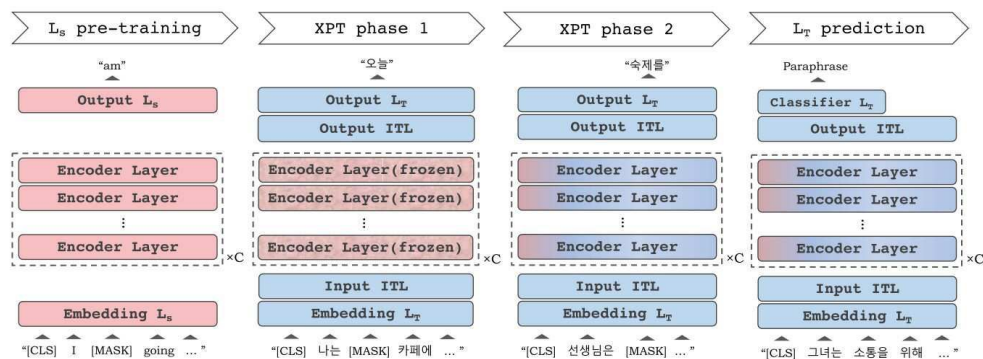
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 언어 간 전이학습을 통한 한국어 언어 모델 구축 방법 및 장치

(57) 요약

언어 간 전이학습을 통한 한국어 언어 모델 생성 방법 및 장치가 개시된다. 상기 한국어 언어 모델 생성 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 언어 모델 생성 방법으로서, 임베딩 레이어(embedding layer), 복수의 인코더 레이어들(encoder layers), 및 출력 레이어를 포함하는 사전학습 언어 모델의 구조를 변경하여 기본 모델을 생성하는 단계, 상기 기본 모델에 대하여 제1 단계의 학습을 수행하는 단계, 및 상기 제1 단계의 학습이 완료된 상기 기본 모델에 대하여 제2 단계의 학습을 수행하는 단계를 포함한다.

대표도 - 도1



(52) CPC특허분류

G06N 3/0454 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711159969
과제번호	2018-0-01405-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명세서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 언어 모델 생성 방법에 있어서,
임베딩 레이어(embedding layer), 복수의 인코더 레이어들(encoder layers), 및 출력 레이어를 포함하는 사전학습 언어 모델의 구조를 변경하여 기본 모델을 생성하는 단계;
상기 기본 모델에 대하여 제1 단계의 학습을 수행하는 단계; 및
상기 제1 단계의 학습이 완료된 상기 기본 모델에 대하여 제2 단계의 학습을 수행하는 단계를 포함하고,
언어 모델 생성 방법.

청구항 2

제1항에 있어서,
상기 사전학습 언어 모델은 제1 언어로 사전학습된 언어 모델이고,
상기 제1 단계의 학습과 상기 제2 단계의 학습은 상기 제1 언어와는 상이한 제2 언어로 된 학습 데이터를 이용하여 수행되는,
언어 모델 생성 방법.

청구항 3

제2항에 있어서,
상기 기본 모델을 생성하는 단계는,
상기 임베딩 레이어와, 상기 복수의 인코더 레이어들 중 첫번째 인코더 레이어 사이에 제1 ITL(Implicit Translation Layer)를 추가하고, 상기 복수의 인코더 레이어들 중 마지막 인코더 레이어와 상기 출력 레이어 사이에 제2 ITL을 추가하는,
언어 모델 생성 방법.

청구항 4

제3항에 있어서,
상기 제1 단계의 학습을 수행하는 단계는,
상기 복수의 인코더 레이어들 각각의 파라미터들을 고정시킨 상태에서 진행되는,
언어 모델 생성 방법.

청구항 5

제3항에 있어서,
상기 제1 단계의 학습을 수행하는 단계는,
상기 임베딩 레이어의 파라미터들과 상기 출력 레이어의 파라미터들을 임의의 값으로 초기화한 후 진행되는,
언어 모델 생성 방법.

청구항 6

제4항 또는 제5항에 있어서,

상기 제1 단계의 학습을 수행하는 단계는 상기 임베딩 레이어, 상기 제1 ITL, 상기 제2 ITL, 및 상기 출력 레이어의 파라미터들을 갱신하는,

언어 모델 생성 방법.

청구항 7

제6항에 있어서,

상기 제2 단계의 학습을 수행하는 단계는 상기 임베딩 레이어, 상기 제1 ITL, 상기 복수의 인코더 레이어들, 상기 제2 ITL, 및 상기 출력 레이어의 파라미터들을 갱신하는,

언어 모델 생성 방법.

발명의 설명

기술 분야

[0001] 본 발명은 사전학습된 언어 모델의 파라미터를 활용하여, 리소스가 적은 언어(Low Resource Language, LRL)에 대한 언어 모델을 구축하는 방법 및 장치에 관한 것이다.

배경 기술

[0002] 언어 모델(Language Model, LM)을 사전학습하여 구축하기 위해서는 대량의 단일 언어 코퍼스(monolingual corpus)가 요구된다. 로우-리소스(low-resource) 환경에서 언어 모델을 학습하기 위한 기존의 대표적인 연구는 1) 다국어 코퍼스로 사전학습을 진행하는 Multi-lingual pre-training 방법론과 2) 하이-리소스 언어(high resource language)에서 LRL(Low Resource Language)로의 언어 전이(language transfer)를 목적으로 하는 cross lingual transfer learning approaches의 두 가지가 존재한다. 본 발명은 이 중 후자에 속한다.

[0003] 기존의 PLM(Pre-trained Language Model) 연구에는 크게 3가지의 한계점이 존재한다. 첫째, 언어적 한계점이 존재한다. 현재 이루어지고 있는 대부분의 언어 모델 연구의 대다수는 영어를 기반으로 이루어지고 있다. 위키피디아에는 303종의 언어가 존재하지만, 대부분의 문서가 영어로 작성되어 있으며, 절반 이상에 해당하는 154종의 언어에는 1만건의 문서도 존재하지 않는다. 이에 따라 대부분의 연구는 영어를 중심으로 이루어지고 있으며, 영어 외 다른 언어에 대한 연구에는 데이터 양의 한계점으로 인해 활발한 연구가 이루어지고 있지 않은 실정이다. 학습 데이터의 양은 언어 모델의 성능과 직결되는 요소이기 때문에, 언어 리소스 불균형(language resource imbalance)은 성능 향상의 장애물이 될 수 있다.

[0004] 둘째, 모델적 한계점이다. Open AI의 GPT-3는 1750억개의 파라미터와 3000억개로 구성된 데이터셋으로 학습했으며 딥마인드도 2800억 파라미터의 고퍼(Gopher)를 개발하였다. 더 나아가 Google의 스위치 트랜스포머는 1조 6000억 개의 파라미터로 최초의 조 단위 파라미터 모델을 개발하였다. 위와 같은 연구들은 대부분 영어 기반이다. 이렇게 대용량의 파라미터와 데이터를 요구하는 최신 모델들을 LRL에 적용하는 것은 쉽지 않으며, 원본 논문에서 요구하는 양과 파라미터를 충족하면서 재구현하기란 불가능에 가깝다.

[0005] 셋째, 서비스적 한계점이다. 파라미터가 너무 큼으로 인하여 실생활에서 서비스를 진행하기란 불가능에 가깝다. 큰 스케일의 데이터를 처리하기 위해서는 충분한 컴퓨팅 파워(예컨대, GPU)를 갖춘 서비스 환경이 요구된다. 모델을 훈련할 때 많은 양의 파라미터들과 데이터를 요구하기 때문에 서버나 GPU 환경이 충분하지 않은 기업은 이러한 최신 모델을 이용한 서비스 환경 구성 및 성능 개선에서 많은 어려움이 발생한다.

[0006] 이에, 본 발명에서는 사전학습된 언어 모델에 대한 전이학습을 진행하여 LRL에 대한 높은 성능을 갖는 언어 모델을 생성하는 방법 및 장치를 제안하고자 한다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 대한민국 등록특허 제2368064호 (2022.02.25. 공고)

(특허문헌 0002) 대한민국 공개특허 제2020-0057291호 (2020.05.26. 공개)

(특허문헌 0003) 대한민국 공개특허 제2021-0121972호 (2021.10.08. 공개)

발명의 내용

해결하려는 과제

[0008] 본 발명이 이루고자 하는 기술적인 과제는 언어간 전이학습을 통해 리소스가 적은 언어에 대한 언어 모델을 생성하는 방법 및 장치를 제공하는 것이다.

과제의 해결 수단

[0009] 본 발명의 일 실시예에 따른 언어 모델 생성 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 언어 모델 생성 방법으로서, 임베딩 레이어(embedding layer), 복수의 인코더 레이어들(encoder layers), 및 출력 레이어를 포함하는 사전학습 언어 모델의 구조를 변경하여 기본 모델을 생성하는 단계, 상기 기본 모델에 대하여 제1 단계의 학습을 수행하는 단계, 및 상기 제1 단계의 학습이 완료된 상기 기본 모델에 대하여 제2 단계의 학습을 수행하는 단계를 포함한다.

발명의 효과

[0010] 본 발명의 실시예에 따른 언어 모델 생성 장치 및 방법에 의할 경우, 리소스가 풍부한 언어로 사전학습된 언어 모델에 사후학습을 진행하여 리소스가 적은 타겟 언어의 언어 모델을 생성할 수 있다.

[0011] 예컨대, 트랜스포머 기반 영어 사전학습 언어 모델에 사후학습을 사용하여 타겟 언어의 언어 모델로 변신시키는 기법인 XPT(Cross-lingual Post Training)를 활용하여 효율적으로 한국어 언어 모델을 구축할 수 있다.

[0012] 제안하는 기법은 별도의 사전학습을 진행하지 않아도 되며 대용량이 데이터와 파라미터를 요구하지 않기 때문에 LRL에도 적용하기 쉽다는 장점이 있다. XPT는 마이너 언어인 한국어를 대상으로 개발이 진행되었으며, 영어 사전학습 언어 모델을 활용하여 적은 양(4M)의 한국어 모노 코퍼스 데이터만으로도 기존의 한국어 언어 모델 등을 뛰어넘거나 비슷한 성능을 보여주었다.

도면의 간단한 설명

[0013] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명의 일 실시예에 따른 언어 모델 생성 방법을 설명하기 위한 개념도이다.

도 2는 본 발명의 일 실시예에 따른 언어 모델 생성 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0014] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0015] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0016] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0017] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어"

있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

- [0018] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0019] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.
- [0020] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.
- [0021] 도 1에 도시된 바와 같이, 언어 간 전이학습(Cross-lingual Post Training, XPT) 방법은 사전학습 언어 모델(Pre-trained Language Model)에 소량의 타겟 언어 코퍼스를 사용하여 언어 전이(language transfer)를 수행한다. 이를 위해, 두 가지 단계(phase)를 거쳐 제1 언어(예컨대, 영어)의 언어 모델(예컨대, RoBERTa)에 사후학습(post-training)을 진행한다.
- [0022] 본 발명에서는 사전학습 언어 모델(예컨대, RoBERTa)의 구조를 임베딩 레이어(embedding layer, 입력 레이어(input layer)로 명명될 수도 있음), 히든 레이어(hidden layer, 인코딩 레이어(encoding layer)나 인코더 레이어(encoder layer)로 명명될 수도 있음), 및 출력 레이어(output layer)로 분류하였다. 제1 단계의 학습에서는 인코더 레이어의 파라미터를 고정(freeze)한 후 임베딩 레이어와 출력 레이어의 파라미터만이 갱신되도록 학습을 진행할 수 있다. 실시예에 따라, 임베딩 레이어와 출력 레이어의 파라미터를 임의의 값으로 설정(즉, 초기화)한 후 학습을 진행할 수 있다. 또한, positional embeddings 및 language modeling head와 같이 언어에 독립적인 파라미터들(language-independent parameters)은 고정하지 않고 학습을 진행할 수 있다.
- [0023] 제1 단계의 학습에 후속되는 제2 단계의 학습은 임베딩 레이어와 출력 레이어의 파라미터뿐만 아니라 인코더 레이어의 파라미터를 모두 갱신하도록 하는 학습이 진행된다.
- [0024] 두 단계에 걸친 학습을 통해 언어 모델의 임베딩 레이어와 출력 레이어는 단어사전(Vocab)에 종속되게 학습이 되며, 인코더 레이어는 이미 공간에 투사된 단어 벡터들을 문맥을 바탕으로 맥락화(contextualize)하여 양방향 학습을 진행하게 된다. 즉, 인코더 레이어의 파라미터들은 언어간 공유가 가능하며 재활용이 가능하다. 이후, 타겟 언어(L_T)의 다운스트림 태스크(downstream task)에 대해 미세조정(fine tuning)하여 성능 평가를 진행할 수 있다.
- [0025] 제1 단계의 학습에서는 소스 언어 모델(source language model)의 인코더 레이어의 파라미터를 고정한 뒤, 임베딩 레이어, 인코더 레이어의 입력과 출력에 추가로 삽입된 ITL(Implicit Translation Layer), 및 출력 레이어만 학습을 진행한다. 따라서, 인코더 레이어의 파라미터를 고정하고, 임베딩 레이어, ITLs, 및 출력 레이어의 파라미터만 업데이트된다. 인코더 레이어를 고정하는 것은 기존 영어 모델로 잘 학습된 파라미터를 유지해 주는 효과를 갖는다.
- [0026] 제2 단계의 학습에서는 임베딩 레이어, 인코더 레이어, 출력 레이어, 및 ITLs 모두 학습한다. 이는 인코더 레이어에 대한 전이학습을 진행하여 타겟 언어와 관련없는 정보를 회색하는 역할을 하여, 제1 언어의 언어 모델을 완벽하게 목표 언어의 언어 모델로 변환하는 역할을 수행한다.
- [0027] 이러한, 사후학습 과정을 통해 XPT는 소스 언어 모델의 파라미터를 적절히 재사용하여 학습하기 때문에 대용량의 타겟 언어 코퍼스를 이용한 사전학습 과정이 필요하지 않다는 장점을 갖는다. 또한, 대용량의 언어 모델을 학습하는 기존의 모델들에 비해 연산 리소스(computational resource)가 저렴하다.

[0028] 이하에서는, 사후학습(즉, 타겟 언어의 언어 모델을 구축하는 방법)에 대해 보다 구체적으로 설명한다.

[0029] 언어 모델은 토큰 시퀀스(sequence of tokens) $\tau = [t_1, t_2, \dots, t_n]$ 를 입력받고 이에 대한 맥락화된 표현(contextualized representation)을 출력한다. 여기서, t_i 는 입력 문장의 i 번째 토큰을, n 은 토큰의 개수를 나타낸다. 이 경우, 언어 모델 LM의 출력은 수학적 1과 같이 표현될 수 있다.

[0030] [수학적 1]

$$\begin{aligned} E_{LS} &= [e_1^S; e_2^S; \dots; e_V^S], e_i^S \in \mathbb{R}^d \\ \mathcal{H}_0 &= \text{Embedding}(\tau, E) \\ &= [e_{t_1}^S, e_{t_2}^S, \dots, e_{t_n}^S] \\ \mathcal{H}_l &= \text{Encoder}_l(\mathcal{H}_{l-1}, \Theta_l) \\ &= [h_1^l, h_2^l, \dots, h_n^l] \\ LM(\tau, \Theta) &= \mathcal{H}_C \\ &= [h_1^C, h_2^C, \dots, h_n^C] \end{aligned}$$

[0031]

[0032] 상기 수학적식에서, E_{LS} 은 L_S 의 임베딩 행렬(embedding matrix)을, H_l 은 l 번째 레이어의 은닉 상태(hidden states)를, 그리고 C 는 LM의 인코더 레이어의 개수를 나타낸다. 또한, Θ 는 $[E_{LS}, \theta_1, \theta_2, \dots, \theta_C]$ 이다.

[0033] 그러나, 전이학습에서, 모든 파라미터들 Θ 를 학습하기 위해서는 매우 많은 연산이 필요하고, 이는 파라미터들이 타겟 언어에 유용함에도 불구하고 잘 학습된 가중치들을 다시 학습함으로써 잘 학습된 가중치들을 희석하는 문제를 초래한다. 이는 타겟 언어에 유용한 파라미터들과 불필요한 파라미터들이 모두 Θ 에 포함되어 있음을 의미한다.

[0034] 효율적인 전이학습을 위해, 타겟 언어에 필요한(유용한) 파라미터들만이 학습될 필요가 있다. 이를 위해, 본 발명은 두 단계에 걸친 학습을 제안한다. 제1 단계의 학습에서, 레이어들은 학습에 유용한 레이어와 불필요한 레이어로 분리된다.

[0035] 예시적으로, 사전학습 RoBERTa의 구조는 임베딩 레이어, 히든 레이어, 및 출력 레이어로 구성된다. 제1 단계의 학습에서, 인코더 레이어의 파라미터들은 고정하고, 임베딩 레이어와 출력 레이어의 파라미터들은 임의의 값으로 초기화된 후 학습이 진행될 수 있다. 여기서, positional embeddings 및 language modeling head와 같은 언어에 독립적인 파라미터들은 고정되지 않고 학습될 수 있다. 다음으로, 제2 단계의 학습에서, 모든 레이어들이 학습된다.

[0036] 인코더 레이어의 파라미터들을 고정하고 임베딩 레이어와 출력 레이어의 파라미터들을 초기화함으로써 두 가지의 주된 문제점이 발생한다. 첫번째 문제로, 워드 얼라인먼트(word alignment)이다. 두 언어들(소스 언어와 타겟 언어) 사이에서 일대일로 대응되는 단어가 존재하지 않을 수 있다. 두번째 문제는 언어적인 차이점(linguistic difference)이다. 두 언어들 사이의 문법적 차이에 따라, 워드 순서(word order)와 의미적 표현(semantic expression)이 상이할 수 있다. 그러나, 랜덤하게 워드 임베딩을 초기화하고 학습하는 방법만으로는, 이러한 차이점을 반영하기에 한계가 있고, 이러한 차이점을 학습할 수 있는 별도의 모듈이나 레이어가 필요하다.

[0037] 따라서, 본 발명에서는, 언어들 사이의 차이점을 학습하기 위하여, ITLs(Chanhee Lee, Kisu Yang, Taesun Whang, Chanjun Park, Andrew Matteson, and Heuiseok Lim. 2021. Exploring the data efficiency of cross-lingual post-training in pretrained language models. Applied Sciences, 11(5):1974.)를 도입한다. ITL은 인코더 레이어와 동일한 구조를 갖는다. ITL_{in}은 입력 레이어(임베딩 레이어)와 인코더 레이어 사이에 추가되고, ITL_{out}은 인코더 레이어와 출력 레이어 사이에 추가될 수 있다. 즉, 하나 이상의 ITL이 추가될 수 있다. ITLs를 갖는 XPT 모델의 출력은 수학적 2와 같이 표현될 수 있다.

[0038] [수학식 2]

$$E_{L_T} = [e_1^T; e_1^T; \dots; e_1^T], e_i^T \in \mathbb{R}^d$$

$$\mathcal{H}_{L_T} = \text{Embedding}(\tau, E_{L_T})$$

$$\mathcal{H}_0 = \text{ITL}_{in}(\mathcal{H}_{L_T}, \theta_{\text{ITL}_{in}})$$

$$LM(\tau, \Theta') = \text{ITL}_{out}(\mathcal{H}_C, \theta_{\text{ITL}_{out}})$$

[0039]

[0040]

위 수학식에서, E_{L_T} 는 타겟 언어 모델 L_T 의 임베딩 행렬을, τ 는 토큰 시퀀스를 나타내고, $\mathcal{H}_C$ 는 수학식 1에서와 같다. ITLs를 갖는 XPT는 이전 모델(사전학습 언어 모델)의 아키텍처와 구조적으로 일치하기 때문에, 계산적 및 공간적 복잡도(computational and spatial complexities)를 보장할 수 있다.

[0041]

도 2는 본 발명의 일 실시예에 따른 언어 모델 생성 방법을 설명하기 위한 흐름도이다. 언어 모델 생성 방법은 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 다시 말해, 언어 모델 생성 방법을 구성하는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수도 있다. 따라서, 컴퓨팅 장치는 언어 모델 생성 장치로 명명될 수도 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(server) 등을 포함한다.

[0042]

도 2를 참조하면, 기본 모델이 생성될 수 있다(S110). 기본 모델은 제1 언어(예컨대, 영어)로 사전학습된 사전학습 언어 모델(PLM)의 구조를 변경함으로써 생성될 수 있다. 사전학습 언어 모델은 임베딩 레이어(또는 입력 레이어), 복수의 인코더 레이어들, 및 출력 레이어로 구성된다. 기본 모델은 사전학습 언어 모델에 적어도 하나의 ITL을 추가(또는 삽입)함으로써 생성될 수 있다. ITL은 임베딩 레이어와 인코더 레이어 사이에, 및/또는 인코더 레이어와 출력 레이어 사이에 구비될 수 있다. 추가될 ITL의 파라미터들은 임의의 값을 가질 수 있다. ITL은 인코더 레이어와 동일한 구조를 갖는 어댑터 레이어(adapter layer)로써, 또다른 인코더 레이어로 이해될 수도 있다.

[0043]

다음으로, 기본 모델에 대한 제1 단계의 학습이 수행될 수 있다(S120). 제1 단계의 학습은 구조가 변경된 사전학습 언어 모델의 파라미터들 중 적어도 일부의 파라미터를 고정시킨 상태에서 수행될 수 있다. 예시적으로, 인코더 레이어의 파라미터들 중 적어도 일부만을 고정한 상태에서 학습이 진행될 수 있다. 또한, 학습을 위한 학습 데이터는 제2 언어(예컨대, 한국어)의 말뭉치(코퍼스)일 수 있다.

[0044]

실시예에 따라, 제1 단계의 학습은 일부 레이어의 파라미터들을 초기화한 후 진행될 수 있다. 예시적으로, 임베딩 레이어와 출력 레이어의 파라미터들이 임의의 값으로 초기화될 수 있다.

[0045]

실시예에 따라, 제1 단계의 학습에 이용되는 학습 데이터에 포함되는 문장들 각각은 분절화(토큰화)되어 있을 수도 있다. 이와는 달리, 학습 데이터에 포함되는 문장들에 대한 분절화 동작을 수행함으로써 기본 모델에 대한 입력 데이터를 생성할 수 있다.

[0046]

마지막으로, 제2 단계의 학습이 수행될 수 있다(S130). 즉, 제1 단계의 학습이 완료된 기본 모델에 대한 추가적인 제2 단계의 학습이 진행될 수 있다. 학습 데이터는 제2 언어의 말뭉치로써, 제1 단계에서 사용된 학습 데이터와 동일하거나 상이할 수 있다.

[0047]

제2 단계의 학습은 파라미터에 대한 고정 없이 모든 레이어의 파라미터를 갱신하도록 학습될 수 있다. 제2 단계의 학습이 완료됨에 따라, 제2 언어의 언어 모델(예컨대, 한국어 언어 모델)이 생성될 수 있다.

[0048]

실시예에 따라, 제2 단계의 학습에 이용되는 학습 데이터에 포함되는 문장들 각각은 분절화(토큰화)되어 있을 수도 있다. 이와는 달리, 학습 데이터에 포함되는 문장들에 대한 분절화 동작을 수행함으로써 기본 모델에 대한 입력 데이터를 생성할 수 있다.

[0049]

이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 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(Operation System, 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 사

용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 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.

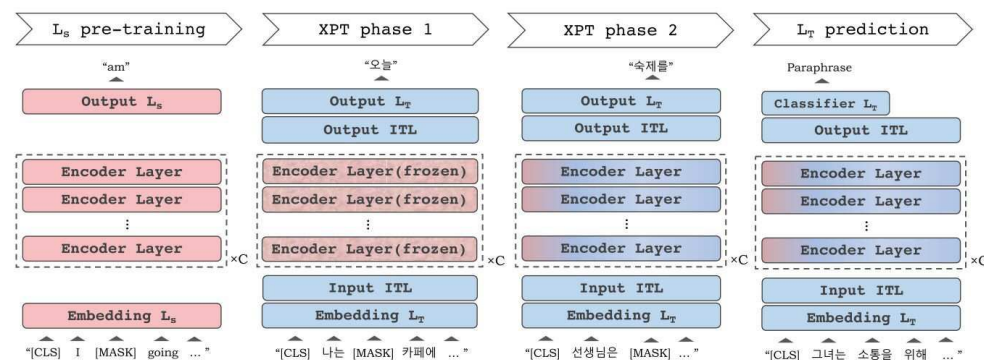
[0050] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 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.

[0051] 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), 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. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0052] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

도면1



도면2

