(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0056633
(43) 공개일자 2022년05월06일

(51) 국제특허분류(Int. Cl.)
G06F 16/332 (2019.01) *G06F 16/33* (2019.01)
(52) CPC특허분류
G06F 16/3329 (2019.01)
G06F 16/3346 (2019.01)
(21) 출원번호 10-2020-0141388
(22) 출원일자 2020년10월28일
심사청구일자 2020년10월28일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
김광인
울산광역시 울주군 언양읍 유니스트길 50
강청웅
울산광역시 울주군 언양읍 유니스트길 50
(74) 대리인
특허법인 무한

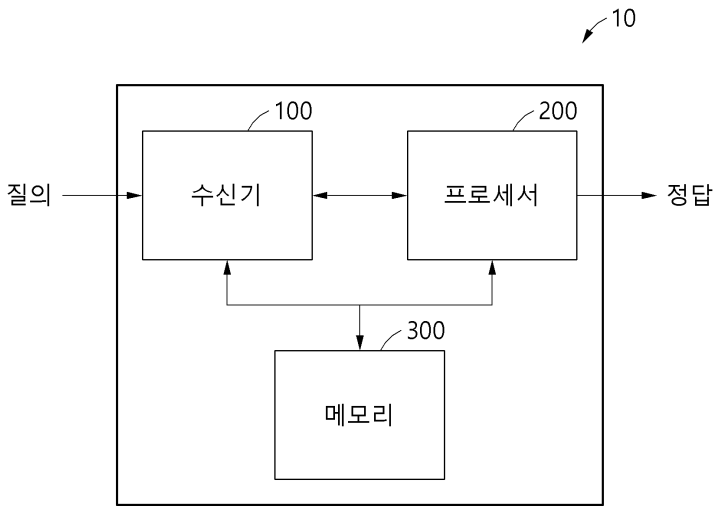
전체 청구항 수 : 총 17 항

(54) 발명의 명칭 **기계 독해 방법 및 장치**

(57) 요 약

기계 독해 방법 및 장치가 개시된다. 일 실시예에 따른 기계 독해 방법은, 자연어로 이루어진 컨텍스트 및 질문을 수신하는 단계와, 상기 컨텍스트 및 질문에 기초하여 임베딩 벡터를 생성하는 단계와, 상기 임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행하는 단계와, 인코딩된 임베딩 벡터에 기초하여 상기 질문에 대한 정답을 예측하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/3347 (2019.01)
G06N 20/00 (2021.08)
G06N 3/08 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711102896
과제번호	2017-0-00255-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	ICT융합산업원천기술개발
연구과제명	(지능정보-총괄/1세부) 자율지능 디지털 동반자 프레임워크 및 응용 연구개발
기 여 율	1/1
과제수행기관명	전자부품연구원
연구기간	2020.01.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

자연어로 이루어진 컨텍스트 및 질문을 수신하는 단계;

상기 컨텍스트 및 질문에 기초하여 임베딩 벡터를 생성하는 단계;

상기 임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행하는 단계; 및

인코딩된 임베딩 벡터에 기초하여 상기 질문에 대한 정답을 예측하는 단계

를 포함하는 기계 독해 방법.

청구항 2

제1항에 있어서,

상기 임베딩 벡터를 생성하는 단계는,

상기 컨텍스트에 기초하여 제1 임베딩 벡터를 생성하는 단계; 및

상기 질문에 기초하여 제2 임베딩 벡터를 생성하는 단계

를 포함하는 기계 독해 방법

청구항 3

제2항에 있어서,

상기 제1 임베딩 벡터 및 상기 제2 임베딩 벡터에 기초하여 어텐션을 수행함으로써 상기 컨텍스트와 상기 질문에 포함된 단어들을 매칭하는 단계

를 더 포함하는 기계 독해 방법.

청구항 4

제1항에 있어서,

상기 인코딩을 수행하는 단계는,

상기 임베딩 벡터에 포지셔널 인코딩(positional encoding)을 수행하는 단계;

포지셔널 인코딩이 수행된 임베딩 벡터에 컨볼루션 연산을 수행함으로써 지역 특징(local feature)를 추출하는 단계; 및

상기 지역 특징에 기초하여 셀프 어텐션을 수행함으로써 전역 특징(global feature)을 추출하는 단계

를 포함하는 기계 독해 방법.

청구항 5

제1항에 있어서,

상기 적어도 한 번의 인코딩을 수행하는 단계는,

잔차 블록(residual block)을 포함하는 뉴럴 네트워크에 의해 수행되는

기계 독해 방법.

청구항 6

제1항에 있어서,

상기 인코딩된 임베딩 벡터에 기초하여 인코딩을 수행함으로써 상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수를 계산하는 단계

를 더 포함하는 기계 독해 방법.

청구항 7

제1항에 있어서,

상기 정답을 예측하는 단계는,

상기 정답의 시작 단어의 확률을 계산하는 단계;

상기 정답의 끝 단어의 확률을 계산하는 단계; 및

상기 시작 단어의 확률 및 상기 끝 단어의 확률에 기초하여 상기 정답을 예측하는 단계

를 포함하는 기계 독해 방법.

청구항 8

제7항에 있어서,

상기 정답을 예측하는 단계는,

상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수에 기초하여 예측된 정답을 보정하는 단계

를 더 포함하는 기계 독해 방법.

청구항 9

하드웨어와 결합되어 제1항 내지 제8항 중 어느 하나의 항의 방법을 실행시키기 위하여 매체에 저장된 컴퓨터 프로그램.

청구항 10

자연어로 이루어진 컨텍스트 및 질문을 수신하는 수신기; 및

상기 컨텍스트 및 질문에 기초하여 임베딩 벡터를 생성하고, 상기 임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행하고, 인코딩된 임베딩 벡터에 기초하여 상기 질문에 대한 정답을 예측하는 프로세서

를 포함하는 기계 독해 장치.

청구항 11

제10항에 있어서,

상기 프로세서는,

상기 컨텍스트에 기초하여 제1 임베딩 벡터를 생성하고,

상기 질문에 기초하여 제2 임베딩 벡터를 생성하는

기계 독해 장치.

청구항 12

제11항에 있어서,

상기 프로세서는,

상기 제1 임베딩 벡터 및 상기 제2 임베딩 벡터에 기초하여 어텐션을 수행함으로써 상기 컨텍스트와 상기 질문에 포함된 단어들을 매칭하는

기계 독해 장치.

청구항 13

제10항에 있어서,

상기 프로세서는,

상기 임베딩 벡터에 포지셔널 인코딩(positional encoding)을 수행하고,

포지셔널 인코딩이 수행된 임베딩 벡터에 컨볼루션 연산을 수행함으로써 지역 특징(local feature)를 추출하고,

상기 지역 특징에 기초하여 셀프 어텐션을 수행함으로써 전역 특징(global feature)을 추출하는

기계 독해 장치.

청구항 14

제10항에 있어서,

상기 프로세서는,

잔차 블록(residual block)을 포함하는 뉴럴 네트워크를 이용하여 상기 적어도 한 번의 인코딩을 수행하는

기계 독해 장치.

청구항 15

제10항에 있어서,

상기 프로세서는,

상기 인코딩된 임베딩 벡터에 기초하여 인코딩을 수행함으로써 상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수를 계산하는

기계 독해 장치.

청구항 16

제10항에 있어서,

상기 프로세서는,

상기 정답의 시작 단어의 확률을 계산하고,

상기 정답의 끝 단어의 확률을 계산하고,

상기 시작 단어의 확률 및 상기 끝 단어의 확률에 기초하여 상기 정답을 예측하는

기계 독해 장치.

청구항 17

제10항에 있어서,

상기 프로세서는,

상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수에 기초하여 예측된 정답을 보정하는

기계 독해 장치.

발명의 설명

기 술 분 야

[0001] 실시예들은 기계 독해 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 기계 독해(machine reading comprehension)이란 사용자의 자연어 질의에 대한 정답을 문맥에서 찾아 반환하는 기술이다. 기계 독해 방식을 이용하면 사람이 텍스트를 읽고 질문 답변을 추론하듯이 기계 독해 알고리즘이 문장 속에서 의미를 찾고 답변할 수 있다.

[0003] 종래의 기계 독해 방식은 시작 단어와 끝 단어의 확률이 독립적으로 분석될 뿐만 아니라, 문장들 사이의 경계가 명시적이지 않아, 정답의 첫 단어와 마지막 단어가 같은 문장에서 추출되지 않아 모델의 성능이 저하되는 문제가 있다.

발명의 내용

해결하려는 과제

[0004] 실시예들은 기계 독해 기술을 제공할 수 있다. 다만, 기술적 과제는 상술한 기술적 과제들로 한정되는 것은 아니며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0005] 일 실시예에 따른 기계 독해 방법은, 자연어로 이루어진 컨텍스트 및 질문을 수신하는 단계와, 상기 컨텍스트 및 질문에 기초하여 임베딩 벡터를 생성하는 단계와, 상기 임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행하는 단계와, 인코딩된 임베딩 벡터에 기초하여 상기 질문에 대한 정답을 예측하는 단계를 포함한다.

[0006] 상기 임베딩 벡터를 생성하는 단계는, 상기 컨텍스트에 기초하여 제1 임베딩 벡터를 생성하는 단계와, 상기 질문에 기초하여 제2 임베딩 벡터를 생성하는 단계를 포함할 수 있다.

[0007] 상기 기계 독해 방법은, 상기 제1 임베딩 벡터 및 상기 제2 임베딩 벡터에 기초하여 어텐션을 수행함으로써 상기 컨텍스트와 상기 질문에 포함된 단어들을 매칭하는 단계를 더 포함할 수 있다.

[0008] 상기 인코딩을 수행하는 단계는, 상기 임베딩 벡터에 포지셔널 인코딩(positional encoding)을 수행하는 단계와, 포지셔널 인코딩이 수행된 임베딩 벡터에 컨볼루션 연산을 수행함으로써 지역 특징(local feature)를 추출하는 단계와, 상기 지역 특징에 기초하여 셀프 어텐션을 수행함으로써 전역 특징(global feature)을 추출하는 단계를 포함할 수 있다.

[0009] 상기 적어도 한 번의 인코딩을 수행하는 단계는, 잔차 블록(residual block)을 포함하는 뉴럴 네트워크에 의해

수행될 수 있다.

[0010] 상기 기계 독해 방법은, 상기 인코딩된 임베딩 벡터에 기초하여 인코딩을 수행함으로써 상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수를 계산하는 단계를 더 포함할 수 있다.

[0011] 상기 정답을 예측하는 단계는, 상기 정답의 시작 단어의 확률을 계산하는 단계와, 상기 정답의 끝 단어의 확률을 계산하는 단계와, 상기 시작 단어의 확률 및 상기 끝 단어의 확률에 기초하여 상기 정답을 예측하는 단계를 포함할 수 있다.

[0012] 상기 정답을 예측하는 단계는, 상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수에 기초하여 예측된 정답을 보정하는 단계를 더 포함할 수 있다.

[0013] 일 실시예에 따른 기계 독해 장치는, 자연어로 이루어진 컨텍스트 및 질문을 수신하는 수신기와, 상기 컨텍스트 및 질문에 기초하여 임베딩 벡터를 생성하고, 상기 임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행하고, 인코딩된 임베딩 벡터에 기초하여 상기 질문에 대한 정답을 예측하는 프로세서를 포함한다.

[0014] 상기 프로세서는, 상기 컨텍스트에 기초하여 제1 임베딩 벡터를 생성하고, 상기 질문에 기초하여 제2 임베딩 벡터를 생성할 수 있다.

[0015] 상기 프로세서는, 상기 제1 임베딩 벡터 및 상기 제2 임베딩 벡터에 기초하여 어텐션을 수행함으로써 상기 컨텍스트와 상기 질문에 포함된 단어들을 매칭할 수 있다.

[0016] 상기 프로세서는, 상기 임베딩 벡터에 포지셔널 인코딩(positional encoding)을 수행하고, 포지셔널 인코딩이 수행된 임베딩 벡터에 컨볼루션 연산을 수행함으로써 지역 특징(local feature)를 추출하고, 상기 지역 특징에 기초하여 셀프 어텐션을 수행함으로써 전역 특징(global feature)을 추출할 수 있다.

[0017] 상기 프로세서는, 잔차 블록(residual block)을 포함하는 뉴럴 네트워크를 이용하여 상기 적어도 한 번의 인코딩을 수행할 수 있다.

[0018] 상기 프로세서는, 상기 인코딩된 임베딩 벡터에 기초하여 인코딩을 수행함으로써 상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수를 계산할 수 있다.

[0019] 상기 프로세서는, 상기 정답의 시작 단어의 확률을 계산하고, 상기 정답의 끝 단어의 확률을 계산하고, 상기 시작 단어의 확률 및 상기 끝 단어의 확률에 기초하여 상기 정답을 예측할 수 있다.

[0020] 상기 프로세서는, 상기 컨텍스트 및 상기 질문에 포함된 문장에 대한 점수에 기초하여 예측된 정답을 보정할 수 있다.

도면의 간단한 설명

[0021] 도 1은 일 실시예에 따른 기계 독해 장치의 개략적인 블록도를 나타낸다.

도 2는 도 1에 도시된 기계 독해 장치의 동작을 나타낸다.

도 3은 도 2에 도시된 인코딩 블록의 구조를 나타낸다.

도 4는 셀프 어텐션의 예를 나타낸다.

도 5는 도 1에 도시된 기계 독해 장치의 동작의 순서도를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0022] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0023] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0024] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0025] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0026] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술분야에서 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.

[0027] 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.

[0029] 본 명세서에서의 모듈(module)은 본 명세서에서 설명되는 각 명칭에 따른 기능과 동작을 수행할 수 있는 하드웨어를 의미할 수도 있고, 특정 기능과 동작을 수행할 수 있는 컴퓨터 프로그램 코드를 의미할 수도 있고, 또는 특정 기능과 동작을 수행시킬 수 있는 컴퓨터 프로그램 코드가 탑재된 전자적 기록 매체, 예를 들어 프로세서 또는 마이크로 프로세서를 의미할 수 있다.

[0030] 다시 말해, 모듈이란 본 발명의 기술적 사상을 수행하기 위한 하드웨어 및/또는 상기 하드웨어를 구동하기 위한 소프트웨어의 기능적 및/또는 구조적 결합을 의미할 수 있다.

[0032] 도 1은 일 실시예에 따른 기계 독해 장치의 개략적인 블록도를 나타낸다.

[0033] 도 1을 참조하면, 기계 독해 장치(10)는 자연어로 이루어진 텍스트를 분석하여 주어진 질문 또는 질의에 대한 정답을 예측할 수 있다. 기계 독해 장치(10)는 뉴럴 네트워크(neural network)를 이용하여 자연어를 분석하고, 질문에 대한 정답을 예측할 수 있다.

[0034] 기계 독해 장치(10)는 사용자 질문에 대한 정답을 관련된 문서에서 찾아서 반환하거나 질문에 대한 적절한 해답이 없는 경우 이를 사용자에게 알릴 수 있다. 기계 독해 장치(10)는 적어도 하나의 인코더 블록을 이용하여 질문에 대한 정답을 예측할 수 있다.

[0035] 기계 독해 장치(10)는 기존의 정보검색, 질의-응답 기술과는 달리, 자연어로 이루어진 질의를 입력 받아 질의문에 형식에 관계없이 정답을 예측할 수 있다. 기계 독해 장치(10)는 질의 관련 문서를 제공하는 대신, 질의에 대한 정답을 직접적으로 사용자에게 반환하기 때문에 사용자에게 큰 편의성을 제공할 수 있다.

[0036] 뉴럴 네트워크(또는 인공 신경망)는 기계학습과 인지과학에서 생물학의 신경을 모방한 통계학적 학습 알고리즘을 포함할 수 있다. 뉴럴 네트워크는 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델 전반을 의미할 수 있다.

[0037] 뉴럴 네트워크는 심층 뉴럴 네트워크 (Deep Neural Network)를 포함할 수 있다. 뉴럴 네트워크는 CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), 퍼셉트론(perceptron), FF(Feed Forward), RBF(Radial Basis Network), DFF(Deep Feed Forward), LSTM(Long Short Term Memory), GRU(Gated Recurrent Unit), AE(Auto Encoder), VAE(Variational Auto Encoder), DAE(Denoising Auto Encoder), SAE(Sparse Auto Encoder), MC(Markov Chain), HN(Hopfield Network), BM(Boltzmann Machine), RBM(Restricted Boltzmann Machine), DBN(Depp Belief Network), DCN(Deep Convolutional Network), DN(Deconvolutional Network), DCIGN(Deep Convolutional Inverse Graphics Network), GAN(Generative Adversarial Network), LSM(Liquid State Machine), ELM(Extreme Learning Machine), ESN(Echo State Network), DRN(Deep Residual Network), DNC(Differentiable Neural Computer), NTM(Neural Turning Machine), CN(Capsule Network), KN(Kohonen Network) 및 AN(Attention Network)를 포함할 수 있다.

[0038] 기계 독해 장치(10)는 모바일 단말(mobile terminal)의 형태로 구현될 수 있다. 예를 들어, 기계 독해 장치(10)는 IoT 장치, Machine-type 통신 장치 또는 휴대용 전자 장치 등으로 구현될 수 있다.

[0039] 휴대용 전자 장치는 랩탑(laptop) 컴퓨터, 이동 전화기, 스마트 폰(smart phone), 태블릿(tablet) PC, 모바일 인터넷 디바이스(mobile internet device(MID)), PDA(personal digital assistant), EDA(enterprise digital assistant), 디지털 스틸 카메라(digital still camera), 디지털 비디오 카메라(digital video camera), PMP(portable multimedia player), PND(personal navigation device 또는 portable navigation device), 휴대용 게임 콘솔(handheld game console), e-북(e-book), 스마트 디바이스(smart device)로 구현될 수 있다. 예를 들어, 스마트 디바이스는 스마트 워치(smart watch) 또는 스마트 밴드(smart band)로 구현될 수 있다.

[0040] 기계 독해 장치(10)는 수신기(100), 프로세서(200)를 포함한다. 기계 독해 장치(10)는 메모리(300)를 더 포함할 수 있다.

[0041] 수신기(100)는 수신 인터페이스를 포함할 수 있다. 수신기(100)는 자연어로 이루어진 컨텍스트(context) 및 질문(question)(또는, 질의(query))을 수신할 수 있다. 컨텍스트 및 질문은 텍스트의 형태로 이루어질 수 있다. 수신기(100)는 수신한 컨텍스트 및 질문을 프로세서(200)로 출력할 수 있다.

[0042] 프로세서(200)는 메모리(300)에 저장된 데이터를 처리할 수 있다. 프로세서(200)는 메모리(300)에 저장된 컴퓨터로 읽을 수 있는 코드(예를 들어, 소프트웨어) 및 프로세서(200)에 의해 유발된 인스트럭션(instruction)들을 실행할 수 있다.

[0043] "프로세서(200)"는 목적하는 동작들(desired operations)을 실행시키기 위한 물리적인 구조를 갖는 회로를 가지는 하드웨어로 구현된 데이터 처리 장치일 수 있다. 예를 들어, 목적하는 동작들은 프로그램에 포함된 코드(code) 또는 인스트럭션들(instructions)을 포함할 수 있다.

[0044] 예를 들어, 하드웨어로 구현된 데이터 처리 장치는 마이크로프로세서(microprocessor), 중앙 처리 장치(central processing unit), 프로세서 코어(processor core), 멀티-코어 프로세서(multi-core processor), 멀티프로세서(multiprocessor), ASIC(Application-Specific Integrated Circuit), FPGA(Field Programmable Gate Array)를 포함할 수 있다.

[0045] 프로세서(200)는 컨텍스트 및 질문에 기초하여 임베딩(embedding) 벡터를 생성할 수 있다. 프로세서(200)는 컨텍스트에 기초하여 제1 임베딩 벡터를 생성할 수 있다. 프로세서(200)는 질문에 기초하여 제2 임베딩 벡터를 생성할 수 있다.

[0046] 프로세서(200)는 임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행할 수 있다. 프로세서(200)는 잔차 블록을 포함하는 뉴럴 네트워크를 이용하여 적어도 한 번의 인코딩을 수행할 수 있다.

[0047] 프로세서(200)는 임베딩 벡터에 포지셔널 인코딩(positional encoding)을 수행할 수 있다. 프로세서(200)는 포지셔널 인코딩이 수행된 임베딩 벡터에 컨볼루션 연산을 수행함으로써 지역 특징(local feature)를 추출할 수 있다. 프로세서(200)는 지역 특징에 기초하여 셀프 어텐션을 수행함으로써 전역 특징(global feature)을 추출할 수 있다.

[0048] 프로세서(200)는 생성한 임베딩 벡터에 기초하여 컨텍스트 및 질문에 포함된 단어에 대한 매칭을 수행할 수 있다. 예를 들어, 프로세서(200)는 제1 임베딩 벡터 및 제2 임베딩 벡터에 기초하여 어텐션을 수행함으로써 컨텍스트와 질문에 포함된 단어들을 매칭할 수 있다.

[0049] 프로세서(200)는 인코딩된 임베딩 벡터에 기초하여 인코딩을 수행함으로써 컨텍스트 및 질문에 포함된 문장에 대한 점수를 계산할 수 있다. 프로세서(200)는 인코딩된 임베딩 벡터에 기초하여 상기 질문에 대한 정답을 예측할 수 있다.

[0050] 프로세서(200)는 정답의 시작 단어의 확률을 계산하고, 정답의 끝 단어의 확률을 계산할 수 있다. 프로세서(200)는 시작 단어의 확률 및 끝 단어의 확률에 기초하여 정답을 예측할 수 있다.

[0051] 프로세서(200)는 컨텍스트 및 질문에 포함된 문장에 대한 점수에 기초하여 예측된 정답을 보정할 수 있다.

[0052] 메모리(300)는 프로세서(200)에 의해 실행가능한 인스트럭션들(또는 프로그램)을 저장할 수 있다. 예를 들어, 인스트럭션들은 프로세서의 동작 및/또는 프로세서의 각 구성의 동작을 실행하기 위한 인스트럭션들을 포함할 수 있다.

[0053] 메모리(300)는 휘발성 메모리 장치 또는 불휘발성 메모리 장치로 구현될 수 있다.

[0054] 휘발성 메모리 장치는 DRAM(dynamic random access memory), SRAM(static random access memory), T-

RAM(thyristor RAM), Z-RAM(zero capacitor RAM), 또는 TTRAM(Twin Transistor RAM)으로 구현될 수 있다.

[0055] 불휘발성 메모리 장치는 EEPROM(Electrically Erasable Programmable Read-Only Memory), 플래시(flash) 메모리, MRAM(Magnetic RAM), 스핀전달토크 MRAM(Spin-Transfer Torque(STT)-MRAM), Conductive Bridging RAM(CBRAM), FeRAM(Ferroelectric RAM), PRAM(Phase change RAM), 저항 메모리(Resistive RAM(RRAM)), 나노 튜브 RRAM(Nanotube RRAM), 폴리머 RAM(Polymer RAM(PoRAM)), 나노 부유 게이트 메모리(Nano Floating Gate Memory(NFGM)), 홀로그래픽 메모리(holographic memory), 분자 전자 메모리 소자(Molecular Eelectronic Memory Device), 또는 절연 저항 변화 메모리(Insulator Resistance Change Memory)로 구현될 수 있다.

[0057] 도 2는 도 1에 도시된 기계 독해 장치의 동작을 나타낸다.

[0058] 도 2를 참조하면, 프로세서(200)는 뉴럴 네트워크 또는, 뉴럴 네트워크 모델(210)(이하, 모델(210)이라 함)을 이용하여 질문에 대한 정답을 예측할 수 있다.

[0059] 프로세서(200)는 문장 단위로 정보를 인코딩하여 각 문장이 정답을 포함하고 있을 확률을 계산하는 모듈과 정답 구문의 위치를 찾는 두 개의 모듈을 이용하여 질문에 대한 정답을 예측할 수 있다.

[0060] 프로세서(200)는 최종적으로 각 구문이 정답이 될 확률에 정답 문장의 확률을 곱하는 방식으로 정답을 예측할 수 있다. 이를 통해, 프로세서(200)는 실제 정답 구문의 확률을 매우 정확하게 예측할 수 있다.

[0061] 또한, 상술한 두 개의 모듈은 일부 자유 변수(free-parameters)를 공유하여 최적화함으로써 뉴럴 네트워크에 대한 결합 학습을 수행할 수 있다. 프로세서(200)는 결합 학습을 통해 각 모듈의 일반화 (Generalization) 효과를 극대화할 수 있다.

[0062] 모델(210)은 베이스라인(211) 및 센텐스 스코어러(sentence scorer)(213)을 포함할 수 있다. 다시 말해, 상술한 두 개의 모듈은 베이스라인(211) 및 센텐스 스코어러(213)일 수 있다.

[0063] 베이스라인(211)은 정답 구문의 위치를 찾을 수 있다. 센텐스 스코어러(213)는 각 문장이 정답을 포함하고 있을 확률을 계산할 수 있다.

[0064] 센텐스 스코어러(213)는 임의의 문장이 정답을 포함하는 경우에는 1의 확률을 아닌 경우는 0의 확률을 출력할 수 있다. 프로세서(200)는 센텐스 스코어러(213)의 출력을 이용하여 예측된 정답을 보정할 수 있다.

[0065] 프로세서(200)는 컨텍스트 및 질문에 포함된 단어에 대하여 임베딩을 수행함으로써 임베딩 벡터를 생성할 수 있다. 프로세서(200)는 입력된 단어에 대한 임베딩을 수행함으로써 단어를 연속 벡터 공간(continuous vector space)에 매핑할 수 있다.

[0066] 프로세서(200)는 벡터와 1:1 매핑하는 단어 임베딩과 입력된 단어의 문자들을 CNN(convolutional neural network)로 인코딩한 문자 기반 단어 임베딩의 결합을 이용하여 컨텍스트 및 질문에 포함된 단어를 벡터화하여 임베딩 벡터를 생성할 수 있다. 이때, 컨텍스트 및 질문을 벡터화하는 신경망은 똑 같은 파라미터를 공유할 수 있다.

[0067] 임베딩 인코더 블록은 입력된 입베딩 벡터로 변환된 문서와 질문의 문맥을 인코딩할 수 있다. 임베딩 인코더 블록은 동일한 신경망을 사용하고, 파라미터를 공유할 수 있다. 즉, 컨텍스트에 대응하는 임베딩 인코더 블록 및 질문에 대응하는 임베딩 인코더 블록은 파라미터를 공유할 수 있다.

[0068] 프로세서(200)는 컨텍스트-질의 어텐션 구조를 사용하여 임베딩 인코딩된 문서와 질의에 대응하는 임베딩 벡터로부터 질의에 대한 컨텍스트의 단어와 컨텍스트에 대한 질문의 단어의 비중을 계산할 수 있다.

[0069] 프로세서(200)는 제1 임베딩 벡터 및 제2 임베딩 벡터에 기초하여 컨텍스트-질의 어텐션을 수행함으로써 컨텍스트와 질문에 포함된 단어들을 매칭할 수 있다. 다시 말해, 프로세서(200)는 컨텍스트의 단어들과 질문의 단어들이 서로 어떤 단어에 매칭될지를 결정할 수 있다.

[0070] 프로세서(200)는 컨텍스트의 히든 매트릭스(hidden matrix)를 문장별로 분리할 수 있다. 프로세서(200)는 맥스 폴링을 이용하여 문장에 속한 단어들로부터 특징을 추출하는 방식으로 문장에 대한 벡터 표현인 문장 임베딩을 구축할 수 있다.

[0071] 프로세서(200)는 최종적인 인코더 블록을 통해 문장들 사이의 상관관계를 분석하고, 각 문장이 정답이 될 확률

을 시그모이드 함수를 통해 0 내지 1 사이로 나타낼 수 있다.

[0072] 프로세서(200)는 계산한 비중 결과를 결합하여, 컨텍스트에 포함된 단어들과 질문에 포함된 단어들과 이루고 있는 구조를 추출할 수 있다. 프로세서(200)는 추출한 구조를 인코더 블록을 이용하여 다시 인코딩할 수 있다.

[0073] 프로세서(200)는 단어별 인코딩뿐만 아니라, 센텐스-와이즈 맥스-풀링을 수행하여 각 문장 벡터를 생성하고, 문장 별 적합도를 나타내는 센텐스 스코어를 계산할 수 있다.

[0074] 센텐스 스코어는 질문에 대한 문장의 관련도를 나타내고, 시그모이드(sigmoid) 함수를 이용하여 0 내지 1사이의 값을 나타낼 수 있다. 프로세서(200)는 계산된 센텐스 스코어에 기초하여 예측된 정답을 보정할 수 있다.

[0075] 프로세서(200)는 최종적으로 각 단어는 FFNN(Feed Forward Neural network)으로 각 단어의 정답 확률을 계산하고, 이전에 계산한 센텐스 스코어를 각 단어에 곱하여 최종 확률을 계산하여 정답을 예측할 수 있다.

[0076] 프로세서(200)는 정답 구문의 시작 단어의 확률과 끝 단어의 확률 분포를 계산하여 컨텍스트의 어떤 구문이 질문에 대한 정답이 될지를 예측할 수 있다. 예를 들어, 프로세서(200)는 각 문장에 정답이 존재할 확률과 각 문단이 정답이 될 확률을 곱하는 방식으로 질문에 대한 정답을 예측할 수 있다.

[0077] 인코더 블록(encoder block)은 패시지(passage) 인코더 블록 및 센텐스(sentence) 인코더 블록을 포함할 수 있다. 패시지 인코더 블록은 관련된 문단을 인코딩하여 정답을 찾을 수 있다. 센텐스 인코더 블록은 관련 문장을 인코딩하여 정답이 존재하는 문장의 위치를 독립적으로 분석할 수 있다. 인코더 블록은 도 3을 참조하여 자세하게 설명한다.

[0078] 프로세서(200)는 기계 독해를 위해 뉴럴 네트워크를 학습시킬 수 있다. 프로세서(200)는 정답의 구문 예측 결과와 정답 문장 예측 결과에 대한 손실 함수를 동시에 학습할 수 있다. 이를 통해, 프로세서(200)는 정답 구문의 문장의 경계를 정확하게 알 수 있다.

[0079] 프로세서(200)는 정답 후보 구문 추출과 정답 문장 예측을 위해 뉴럴 네트워크를 학습시키는 과정에서 자유 변수를 공유하고, 손실함수를 결합함으로써 기계 독해 성능을 향상시킬 수 있다.

[0080] 프로세서(200)는 다음 스크립트를 이용하여 기계 독해 또는 뉴럴 네트워크의 학습을 수행할 수 있다.

[0081] 1. Main.py: train, test 코드를 정의

[0082] 2. Config.py: 모델 구성에 필요한 하이퍼 파라미터를 정의

[0083] 3. Layer.py: 신경망 구성에 필요한 모듈들을 정의

[0084] 4. Model.py: MRC 신경망 모델을 정의

[0085] 5. Prepro.py: Tokenize, 임베딩과 같이 텍스트 분석 및 학습, 추론에 필요한 데이터 처리

[0086] 6. Util.py: Evaluation, batch와 같이 학습 추론에 필요한 기타 모듈들을 정의

[0087] 프로세서(200)는 학습을 위해, 우선 Prepro.py를 기반으로 전처리(preprocessing)를 진행하여, 학습 데이터를 학습 및 추론에 필요한 입출력 형태로 가공할 수 있다.

[0088] 프로세서(200)는 main.py의 내용을 기반으로 학습과 테스트를 수행할 수 있고, 뉴럴 네트워크 모델 및 구성은 model.py와 config.py을 통해 변경할 수 있다.

[0090] 도 3은 도 2에 도시된 인코딩 블록의 구조를 나타내고, 도 4는 셀프 어텐션의 예를 나타낸다.

[0091] 도 3 및 도 4를 참조하면, 모델(210)은 적어도 하나의 인코더 블록(215)을 포함할 수 있다. 도 3의 예시에 따르면, 모델(210)은 임베딩 인코더 블록, 패시지 인코더 블록 및 센텐스 인코더 블록을 포함할 수 있다.

[0092] 인코더 블록(215)는 입력에 대한 인코딩을 수행할 수 있다. 인코더 블록(215)은 포지셔널 인코딩(positional encoding) 블록, CNN, 셀프 어텐션(self-attention) 블록 및 FFNN를 포함할 수 있다.

[0093] 포지셔널 인코딩 블록은 입력된 임베딩 벡터의 절대 위치에 따라 신호를 부여할 수 있다. 포지셔널 인코딩 블록은 임베딩 벡터로 변환된 단어의 상대적 위치를 각 단어의 절대 위치로 변활할 수 있다.

[0094] 예를 들어, 포지셔널 인코딩 블록은 sin 또는 cos 함수를 이용하여 각 단어의 절대 위치를 정규화한 값을 나타

낼 수 있다. 포지셔널 인코딩 블록은 sin(x+y)를 sin(x)와 sin(y)의 선형 결합으로 변환하는 과정을 통해 포지셔널 인코딩을 수행할 수 있다.

[0095] 포지셔널 인코딩 블록은 각 단어의 절대 위치를 정규화한 값을 입력된 임베딩 벡터에 더하는 방식으로 각 단어의 의미와 위치를 표현할 수 있다.

[0096] CNN은 포지셔널 인코딩이 수행된 임베딩 벡터에 적어도 한 번의 컨볼루션 연산을 수행함으로써 임베딩 벡터에 대응되는 단어의 지역 특징(local feature)을 추출할 수 있다.

[0097] 도 4의 예시와 같이, 셀프 어텐션 블록은 문맥에서 각 단어들을 그래프 구조로 나타내고, 입력 임베딩 벡터로부터 전역 특징(global feature)을 추출할 수 있다.

[0098] FFNN은 CNN과 셀프 어텐션 블록으로부터 추출한 지역 특징 및 전역 특징에 비선형 변환을 수행할 수 있다.

[0099] 프로세서(200)는 CNN, 셀프 어텐션 블록 및 FFNN에서 출력을 입력과 더하는 잔차 블록(residual block)을 적용함으로써 심층 학습의 효율을 향상시킬 수 있다.

[0101] 도 5는 도 1에 도시된 기계 독해 장치의 동작의 순서도를 나타낸다.

[0102] 도 5를 참조하면, 수신기(100)는 자연어로 이루어진 컨텍스트 및 질문을 수신할 수 있다(510).

[0103] 프로세서(200)는 컨텍스트 및 질문에 기초하여 임베딩 벡터를 생성할 수 있다(530). 프로세서(200)는 컨텍스트에 기초하여 제1 임베딩 벡터를 생성할 수 있다. 프로세서(200)는 질문에 기초하여 제2 임베딩 벡터를 생성할 수 있다.

[0104] 프로세서(200)는 제1 임베딩 벡터 및 제2 임베딩 벡터에 기초하여 어텐션을 수행함으로써 컨텍스트와 질문에 포함된 단어들을 매칭할 수 있다.

[0105] 프로세서(200)는 임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행할 수 있다. 프로세서(200)는 잔차 블록을 포함하는 뉴럴 네트워크를 이용하여 적어도 한번의 인코딩을 수행할 수 있다.

[0106] 프로세서(200)는 임베딩 벡터에 포지셔널 인코딩을 수행할 수 있다. 프로세서(200)는 포지셔널 인코딩이 수행된 임베딩 벡터에 컨볼루션 연산을 수행함으로써 지역 특징을 추출할 수 있다. 프로세서(200)는 지역 특징에 기초하여 셀프 어텐션을 수행함으로써 전역 특징을 추출할 수 있다.

[0107] 프로세서(200)는 인코딩된 임베딩 벡터에 기초하여 질문에 대한 정답을 예측할 수 있다. 프로세서(200)는 정답의 시작 단어의 확률을 계산할 수 있다. 프로세서(200)는 정답의 끝 단어의 확률을 계산할 수 있다. 프로세서(200)는 시작 단어의 확률 및 끝 단어의 확률에 기초하여 정답을 예측할 수 있다.

[0108] 프로세서(200)는 인코딩된 임베딩 벡터에 기초하여 인코딩을 수행함으로써 컨텍스트 및 질문에 포함된 문장에 대한 점수를 계산할 수 있다. 프로세서(200)는 컨텍스트 및 질문에 포함된 문장에 대한 점수에 기초하여 예측된 정답을 보정할 수 있다.

[0110] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 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 설명된 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.

[0111] 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 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 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0112] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있으며 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.

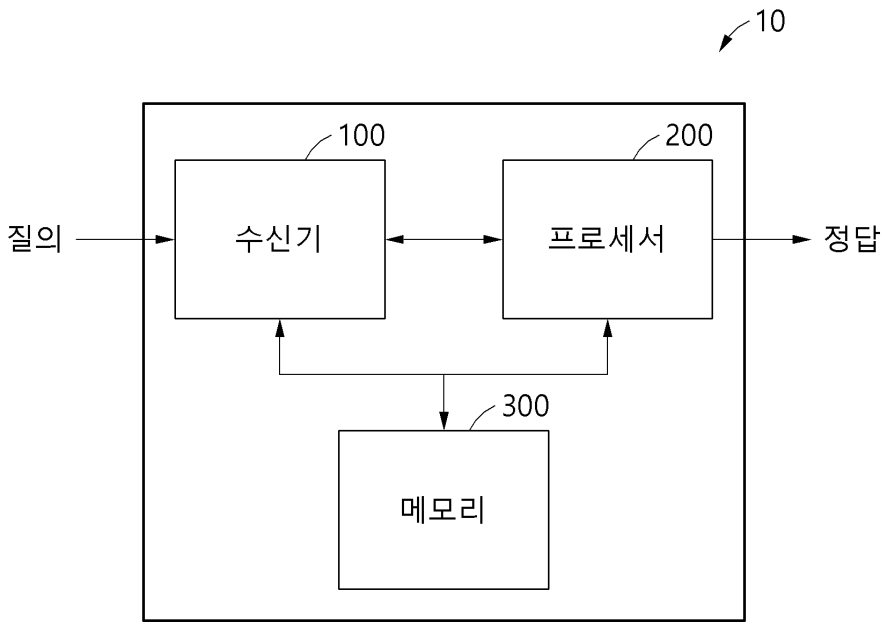
[0113] 위에서 설명한 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 또는 복수의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0114] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 이를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 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.

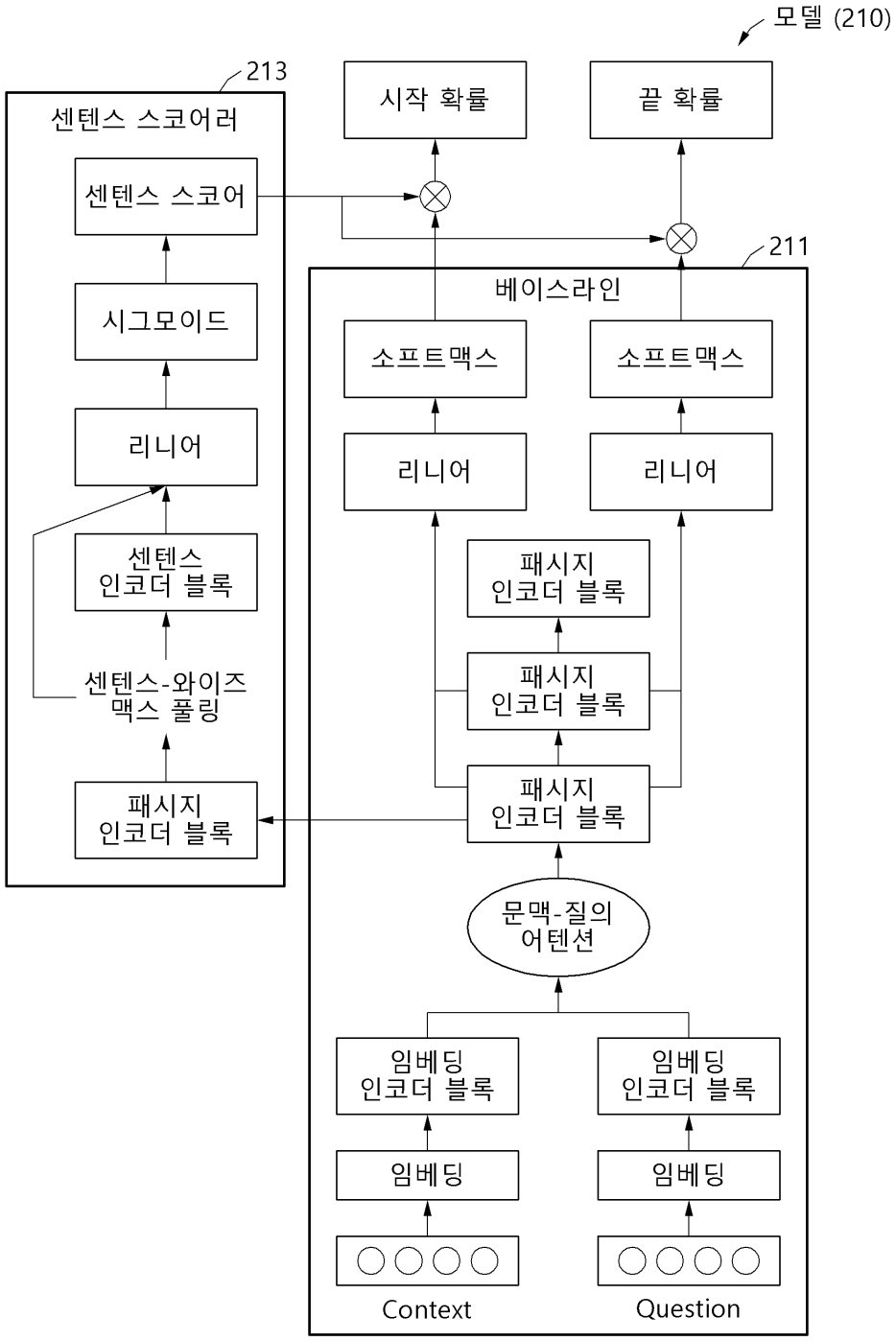
[0115] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

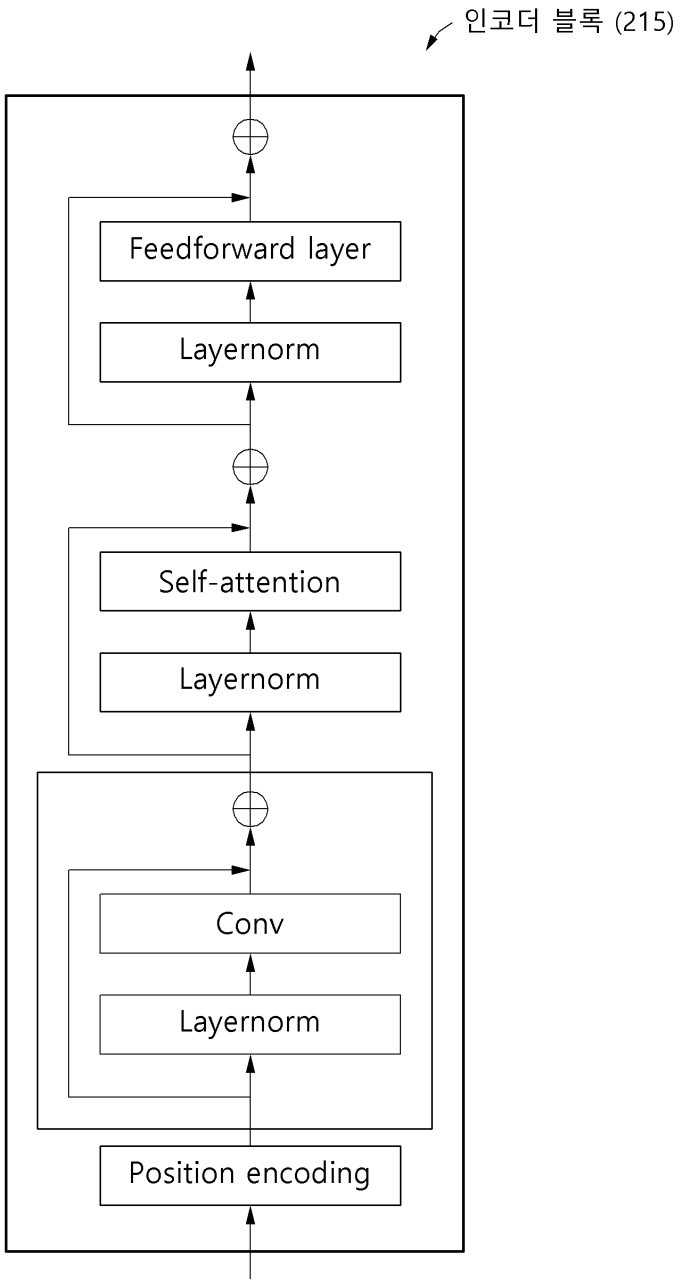
도면1



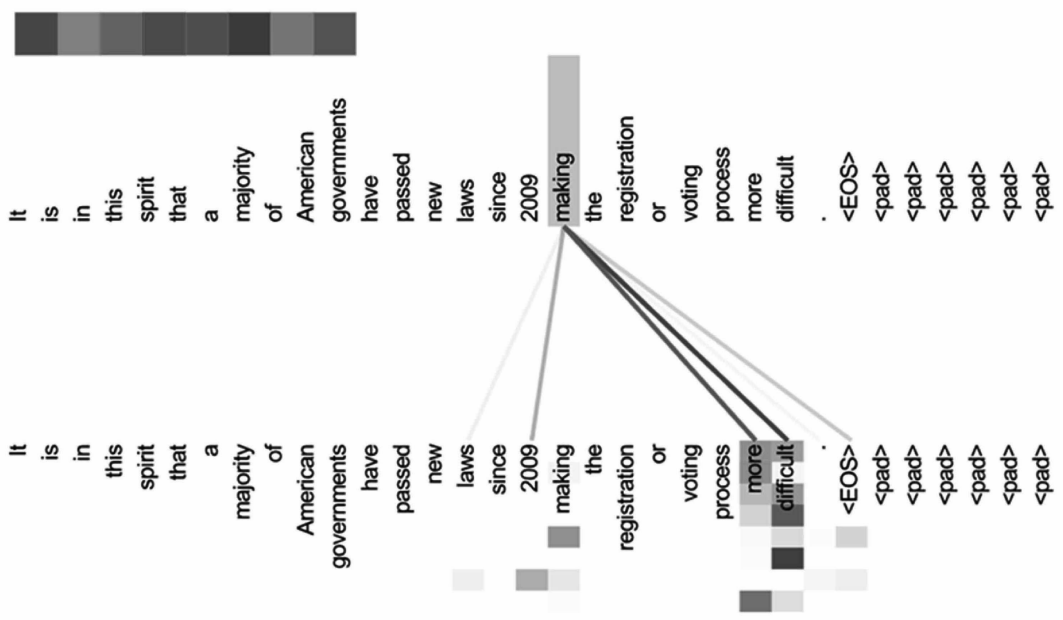
도면2


모델 (210)
213
센텐스 스코어러
센텐스 스코어
시그모이드
리니어
센텐스
인코더 블록
센텐스-와이즈
맥스 풀링
패시지
인코더 블록
시작 확률
끝 확률
211
베이스라인
소프트맥스
소프트맥스
리니어
리니어
패시지
인코더 블록
패시지
인코더 블록
패시지
인코더 블록
문맥-질의
어텐션
임베딩
인코더 블록
임베딩
인코더 블록
임베딩
임베딩
Context
Question

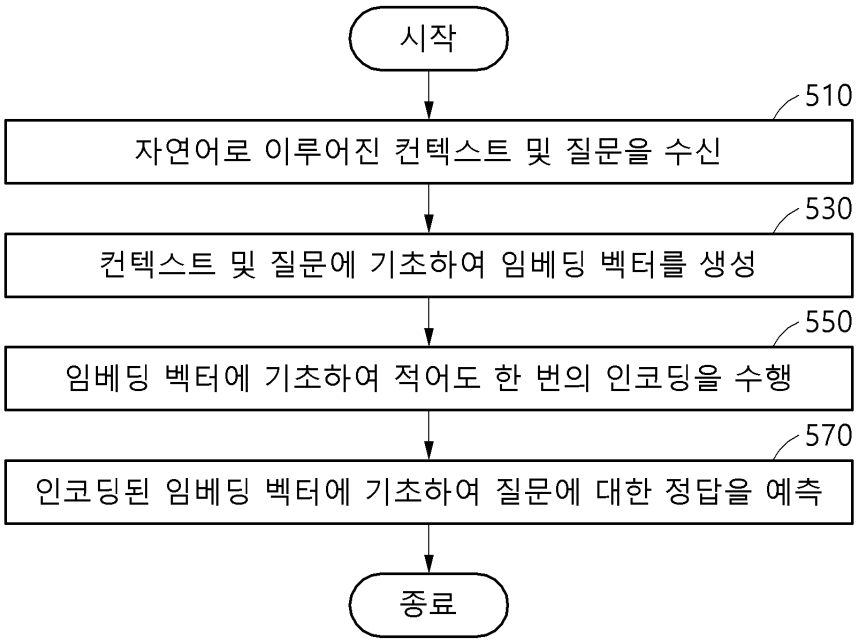
도면3


인코더 블록 (215)
Feedforward layer
Layernorm
Self-attention
Layernorm
Conv
Layernorm
Position encoding

도면4


It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>
It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult . <EOS> <pad> <pad> <pad> <pad> <pad> <pad>

도면5


시작
510
자연어로 이루어진 컨텍스트 및 질문을 수신
530
컨텍스트 및 질문에 기초하여 임베딩 벡터를 생성
550
임베딩 벡터에 기초하여 적어도 한 번의 인코딩을 수행
570
인코딩된 임베딩 벡터에 기초하여 질문에 대한 정답을 예측
종료