(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0000902
(43) 공개일자 2020년01월06일

(51) 국제특허분류(Int. Cl.)
G06Q 50/30 (2012.01) *G06N 3/08* (2006.01)
G06N 99/00 (2019.01) *G10L 15/16* (2006.01)
(52) CPC특허분류
G06Q 50/30 (2013.01)
G06N 20/00 (2019.01)
(21) 출원번호 10-2018-0073153
(22) 출원일자 2018년06월26일
심사청구일자 없음

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
조성배
서울특별시 강남구 선릉로76길 12, 101동 201호(대치동, 대치한신휴플러스)
부석준
경기도 고양시 일산동구 중앙로 1347(장항동)
(뒷면에 계속)
(74) 대리인
민영준

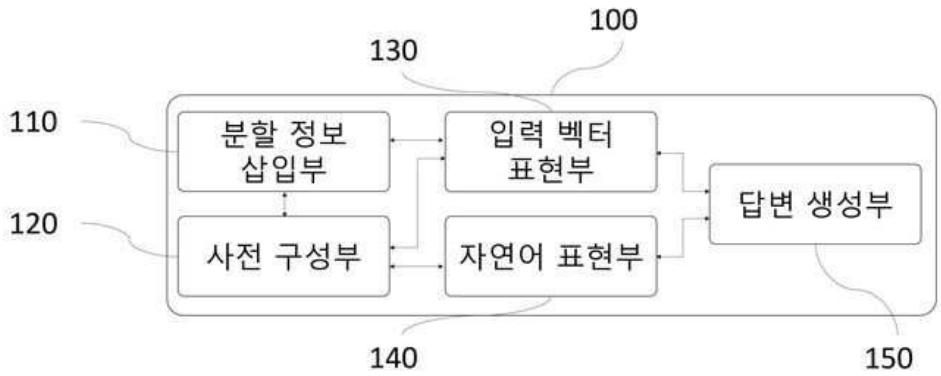
전체 청구항 수 : 총 7 항

(54) 발명의 명칭 **형태소 정보를 추가한 딥러닝 기반 자동 대화생성 방법**

(57) 요 약

형태소 정보를 추가한 딥러닝 기반 자동 대화생성 방법 및 시스템이 개시된다. 개시된 시스템은, ,딥러닝 모델을 사용한 데이터 기반 학습을 사용하는 챗봇에 대해서 문장 시퀀스를 형태소에 따라 분할하고, 분할된 요소에 형태소 정보를 추가하기 위한 분할 정보 삽입부: 분할된 시퀀스를 사전으로 구성하는 사전 구성부:입력 사전을 통해 입력 시퀀스를 벡터로 변환하는 벡터 표현부: 및 출력 사전을 통해 출력 시퀀스를 자연어로 변환하는 자연어 표현부를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2013.01)
G10L 15/16 (2013.01)
(72) 발명자
서수인
인천광역시 계양구 장기서로16번길 13(장기동, 신동아타운)
김진영
경기도 하남시 미사강변대로 165, 110동 1303호(망월동, 미사강변 푸르지오)

이 발명을 지원한 국가연구개발사업
과제고유번호 2016-0-00562
부처명 과학기술정보통신부
연구관리전문기관 정보통신기술진흥센터(NIPA산하)
연구사업명 정보통신방송연구개발사업
연구과제명 [이지바로][주관/한국과학기술원]상대방의 감성을 추론, 판단하여 그에 맞추어 대화하고 대응할 수 있는 감성지능 기술 연구개발(2/5)
기 여 율 1/1
주관기관 한국과학기술원
연구기간 2017.09.01 ~ 2018.06.30

명 세 서

청구범위

청구항 1

딥러닝 모델을 사용한 데이터 기반 학습을 사용하는 챗봇에 대해서

문장 시퀀스를 형태소에 따라 분할하고, 분할된 요소에 형태소 정보를 추가하기 위한 분할 정보 삽입부:

분할된 시퀀스를 사전으로 구성하는 사전 구성부:

입력 사전을 통해 입력 시퀀스를 벡터로 변환하는 벡터 표현부: 및

출력 사전을 통해 출력 시퀀스를 자연어로 변환하는 자연어 표현부를 포함한 것을 특징으로 하는 챗봇 시스템

청구항 2

제1항에 있어서, 사전 구성부는 학습 데이터의 입력 시퀀스로부터 사전을 구성하기 위한 입력 사전 구성부: 및 학습 데이터의 출력 시퀀스로부터 사전을 구성하기 위한 출력 사전 구성부를 포함하는 것을 특징으로 하는 챗봇 시스템

청구항 3

제1항에 있어서, 벡터 표현부는 입력 사전 구성부로부터 생성된 입력 시퀀스 사전을 기반으로 입력 시퀀스를 벡터로 변환하는 것을 특징으로 하는 챗봇 시스템

청구항 4

제1항에 있어서, 자연어 표현부는 출력 사전 구성부로부터 생성된 출력 시퀀스 사전을 기반으로 출력 시퀀스 벡터를 자연어로 변환하는 것을 특징으로 하는 챗봇 시스템

청구항 5

형태소 단위로 분할된 입력 시퀀스에 각 분할 요소의 정보를 포함한 입력 시퀀스를 사용하여 답변을 생성하는 챗봇에 있어서,

(a) 입력 문장을 형태소 분석하고 분할 지점을 결정하는 단계:

(b) 분할 지점을 기준으로 시퀀스를 분할하고, 각 요소의 형태소 정보를 입력 시퀀스에 추가하는 단계:

(c) 분할된 입력 시퀀스를 기반으로 분할 정보 단위 사전을 구성하고 이를 활용하여 입력 시퀀스를 입력 벡터로 변환하는 단계: 및

(d) 분할된 출력 시퀀스를 기반으로 입력 사전과 별개의 출력 사전을 구성하고 이를 활용하여 출력 벡터를 출력 시퀀스로 변환하는 단계를 포함하는 것을 특징으로 하는 챗봇 시스템 구성 방법

청구항 6

제5항에 있어서,

(c2) 입력 시퀀스 변환 과정에서 one-hot representation 방식으로 시퀀스를 표현하는 것을 특징으로 하는 챗봇

시스템 구성 방법.

청구항 7

제6항에 기재된 챗봇 시스템 구성 방법을 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체

발명의 설명

기 술 분 야

[0001] 본 발명은 자동 대화 생성 방법에 관한 것으로서, 더욱 상세하게는 딥러닝 기반의 자동 대화 생성 방법에 관한 것이다.

배 경 기 술

[0003] 소셜 미디어 서비스의 사용량이 증가하면서 대화 데이터의 수집이 손쉬워지고, 동시에 대용량의 데이터를 처리할 수 있는 컴퓨터 기반이 갖춰지면서 데이터 기반의 대화 에이전트 연구가 활발해지고 있다. 특히 2015년 Google에서 개발된 Neural Conversational Model은 Encoder-Decoder 기반의 딥러닝 기술로서 학습 데이터만을 활용하여 자연스러운 대화를 생성하는 챗봇을 생성할 수 있게 되었다.

[0004] 이렇게 형성된 데이터 기반 챗봇은 문법 등의 언어상의 정해진 규칙이 제공되지 않은 주어진 대화 스크립트만으로 입력 문장에 대한 답변을 학습하는 것이 가능하다. 이와 같은 데이터 기반 모델은 데이터의 양과 종류가 결과에 큰 영향을 미친다.

발명의 내용

해결하려는 과제

[0006] 기존에 연구된 데이터 기반 챗봇 모델은 단어의 시퀀스인 문장만을 입력하기 때문에, 같은 형태이지만 다른 의미를 띄는 동음이의어나 다의어의 경우 의미의 차이를 구별하지 못하며, 의미적 분할이 단어 수준에서 이루어지지 않는 언어에 적용하기 어렵다는 단점이 있다.

[0007] 예를 들면 한국어의 경우, 어간과 어미가 함께 사용되면서 하나의 어절을 형성하기 때문에 하나의 어간이 주어진 경우, 어미에 따라 다양한 조합이 가능하게 되어 문장을 공백을 기준으로 하여 나누게 되면 한 어간에 대해서 수많은 어휘가 존재하게 되어 문장 스크립트의 전체 어휘의 개수가 매우 많아지게 된다.

과제의 해결 수단

[0009] 상기한 기술적 과제를 해결하기 위한 딥러닝 기반 챗봇은 단어 시퀀스의 형태소 분석을 통해서 문장 성분을 분할하고 형태소 정보를 추가하기 위한 분할 정보 삽입부: 각 분할된 시퀀스와 이에 대한 형태소 정보를 포함하는 요소를 사전으로 구성하는 사전 구성부: 사전을 통해 입력 시퀀스를 벡터로 변환하는 벡터 표현부: 시퀀스 벡터를 입력받은 후 딥러닝 모델을 거쳐 출력 시퀀스 벡터를 출력하는 답변 생성부: 및 사전을 통해 출력 시퀀스 벡터를 자연어로 변환하는 자연어 표현부를 포함하는 것을 특징으로 한다.

[0010] 여기서 상기 분할 정보 삽입부는 입력된 문장의 형태소를 분석하고 이를 기준으로 문장의 분할 지점을 결정하는 분할 지점 분석부: 분할 지점 분석부의 동작에 따라 시퀀스에 분할된 형태소 정보를 추가하는 시퀀스 분할부를 포함할 수 있다.

[0011] 또한, 상기 사전 구성부는 입력 시퀀스를 통하여 사전을 구성하기 위한 입력 사전 구성부와 출력 시퀀스로부터 사전을 구성하기 위한 출력 사전 구성부를 포함할 수 있다.

발명의 효과

[0013] 이상과 같이, 본 발명은 입력 시퀀스의 형태소 단위로 분할된 각 시퀀스 요소에 형태소 정보를 포함한 사전을 구성하고 이를 활용한 딥러닝 기반 챗봇을 구성하는 방법으로, 동음이의어 및 다의어의 단어 의미 명확화(Word Sense Disambiguation)나 언어의 의미적 분할(Semantic Segmentation)을 해결함으로써 데이터 기반 챗봇의 질이 향상되는 효과를 얻을 수 있다.

도면의 간단한 설명

[0015] 도 1은 발명의 실시예에 따른 전체 챗봇(100) 구조의 블록도이다.

도 2는 분할 지점 분석부와 시퀀스 분할부를 포함한 분할 정보 삽입부(110)의 블록도이다.

도 3은 사전 구성부(120)의 구체적인 실시예를 나타내는 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0016] 이하에서는 도면을 참고하여 본 발명의 바람직한 실시예들을 상세히 설명한다. 이하 설명 및 첨부된 도면들의 구성 요소들이 같은 개체를 나타낼 경우, 통일된 부호를 사용하였으며, 중복 설명을 생략하기로 한다. 또한, 본 발명을 설명하는 과정에서 관련된 공지기능 및 구성에 대한 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그에 대한 상세한 내용은 생략하도록 한다.

[0017] 도 1은 본 발명의 실시예에 따른 데이터 기반 챗봇을 구성하는 블록도이다. 본 실시예에 따른 챗봇(100)은 자연어 문장을 입력 및 출력으로 하며, 모든 소프트웨어에 탑재될 수 있다. 인공비서, 질의응답 시스템 등에 탑재된 챗봇이 이에 대한 예시이다.

[0018] 본 실시예에 따른 챗봇(100)은 사용자가 자연어 문장을 입력하면 이를 벡터의 시퀀스로 변환하고 딥러닝 모델을 통해 답변 벡터 시퀀스를 출력한다. 답변 벡터 시퀀스는 다시 자연어 문장으로 변환하여 사용자에게 자연어 형식으로 보여주는 형식으로 동작한다. 이때 벡터-자연어의 변환은 학습 데이터를 통해 구성된 사전을 활용한다.

[0019] 도 1을 참조하면, 본 실시예에 따른 챗봇(100)은 문장 시퀀스의 형태소 정보를 탐색하고 이를 삽입하기 위한 분할 정보 삽입부(110), 학습 데이터를 활용하여 사전을 구성하는 사전 구성부(120), 입력 분할 조각을 벡터로 변환하는 입력 벡터 표현부(130), 답변 벡터를 자연어로 변환하는 자연어 표현부(140), 입력 벡터로부터 답변 벡터를 생성해내는 답변 생성부(150) 등을 포함하여 이루어진다.

[0020] 입력 벡터 표현부(130)는 형태소로 분할된 자연어 조각을 답변 생성부(150)에서 활용할 수 있도록 벡터 형태로 변환하는 역할을 한다. 이 과정을 위해 입력 시퀀스 사전(122)을 사용한다. 변환되는 벡터는 입력 시퀀스 사전(122)에 포함된 시퀀스 차원의 수의 크기를 가지며, 전체 어휘의 인덱스를 활용해 자연어를 벡터로 변환한다.

[0021] 자연어 표현부(140)는 답변 생성부(150)가 생성한 답변 벡터를 다시 자연어로 복구하는 역할을 하며, 이를 위해 출력 시퀀스 사전(124)을 사용한다. 자연어 표현부(140)에서 출력 시퀀스 사전(124) 내에 포함된 어휘를 벡터로 표현하고, 이 벡터들과 답변 벡터 사이의 거리를 계산한다. 그리고 각 자연어 벡터 중 답변 벡터와의 가장 거리가 가까운 벡터가 의미하는 어휘를 출력 단어로 선택한다. 여기서 거리는 Euclidean distance, Cosine similarity 등이 사용될 수 있다.

[0022] 답변 생성부(150)는 딥러닝 모델로 이루어져 있으며, 여기서는 그 중 하나로서 Seq2Seq 구조를 사용한 실시예로, 순환신경망(Recurrent Neural Network)를 활용한 Encoder-Decoder 구조를 사용하였지만 다른 모델을 사용하여도 무방하다. 답변 생성부(150)는 입력 벡터 표현부(130)가 생성한 벡터를 입력으로 사용하므로 입력 벡터의 차원은 입력 시퀀스 사전(122)이 표현하는 벡터의 차원과 같다. 출력은 자연어 표현부(140)로 전달되어 자연어 단어를 생성하므로 출력 벡터의 차원은 출력 시퀀스 사전(124)이 포함하는 벡터의 차원과 같다.

[0023] 분할 정보 삽입부(110)는 입력 시퀀스를 형태소 분석하여 이를 바탕으로 시퀀스를 각 시퀀스 요소로 분할할 것인지를 것인지 결정하고, 분할된 시퀀스 요소들에 각 요소의 형태소 정보를 삽입하는 역할을 한다.

[0024] 도 2는 분할 정보 삽입부(110)의 구체적인 실시예를 나타내는 블록도이다. 도 2를 참조하면 분할 정보 삽입부

(110)는 분할 지점 분석부(111)와 시퀀스 분할부(112)로 이루어진다.

[0025] 분할 지점 분석부(111)는 입력 시퀀스를 형태소 분석기를 사용하여 입력 시퀀스를 구성하는 형태소를 탐색하고 각 형태소의 경계 지점을 분할 지점으로 지정하여 시퀀스 분할부(112)로 정보를 전달한다.

[0026] 시퀀스 분할부(112)는 주어진 시퀀스를 분할하고, 분할된 요소의 형태소 정보를 시퀀스 요소 내에 추가한다. 시퀀스 분할부(112)의 분할 기준은 분할 지점 분석부(111)의 기준을 따른다.

[0027] 사전 구성부(120)는 대화 에이전트의 입력 벡터 표현부(130)와 자연어 표현부(140)에 사용하기 위한 사전을 구성한다. 사전은 데이터 기반 대화 에이전트가 동작하기 전에 학습 데이터로부터 미리 구성된다. 도 3은 사전 구성부(120)의 구체적인 실시예를 나타내는 블록도이다. 도 3을 참조하면 사전 구성부(120)는 입력 사전 구성부(121)와 입력 시퀀스 사전(122), 출력 사전 구성부(123), 출력 시퀀스 사전(124)으로 이루어진다.

[0028] 입력 사전 구성부(121)는 입력 시퀀스로부터 입력 시퀀스 사전(122)을 구성한다. 입력 시퀀스가 주어지면, 이를 분할 정보 삽입부(110)에 입력하여 시퀀스 요소로 분할한다. 그리고 이 시퀀스 요소가 입력 시퀀스 사전(122)에 존재하지 않는 경우 새로 추가하는 과정을 통해 사전을 구성한다.

[0029] 입력 시퀀스 사전(122)은 입력 시퀀스 요소를 인덱스를 통해 관리한다. 사전 내에 학습 데이터를 통해 형성한 어휘뿐만 아니라 <UNK> (Unknown Word), <EOS> (End of Speech) 등의 특수한 상황을 표현하는 요소를 포함하기도 한다. 사전 내의 시퀀스 요소를 탐색할 경우, 시퀀스 사전은 해당 시퀀스 요소에 대한 인덱스 값을 반환한다.

[0030] 출력 사전 구성부(123)는 학습 데이터의 출력 시퀀스를 통해 출력 시퀀스 사전(124)을 구성한다.

[0031] 출력 시퀀스 사전(124)은 출력 시퀀스의 요소를 인덱스를 통해 관리한다. 입력 시퀀스 사전(122)과 마찬가지로 특수한 상황을 표현하는 요소를 함께 저장하며, 시퀀스 요소의 탐색 요청이 입력된 경우, 해당 요소의 사전 내 인덱스를 반환한다.

[0032] 본 발명의 챗봇 시스템(100)은 입력 사전과 출력 사전을 별개로 분리하고 사전의 어휘 수를 감소시킴으로써 답변 생성부(150)가 보다 어휘의 탐색 공간이 작아지게 된다. 또한, 시퀀스의 형태소 정보가 입력 시퀀스 사전(122)에 포함되어, 더 작은 크기의 사전으로 같은 양의 정보를 함유하도록 할 수 있으며, 이는 결과적으로 입력 벡터 표현부(130)가 입력 시퀀스를 벡터로 변환 시 Unknown Word가 발생할 비율을 감소시킨다. 이를 통해 챗봇(100)의 답변 탐색 능력을 높이게 되어 챗봇은 더 자연스러운 자연어를 출력할 수 있게 된다.

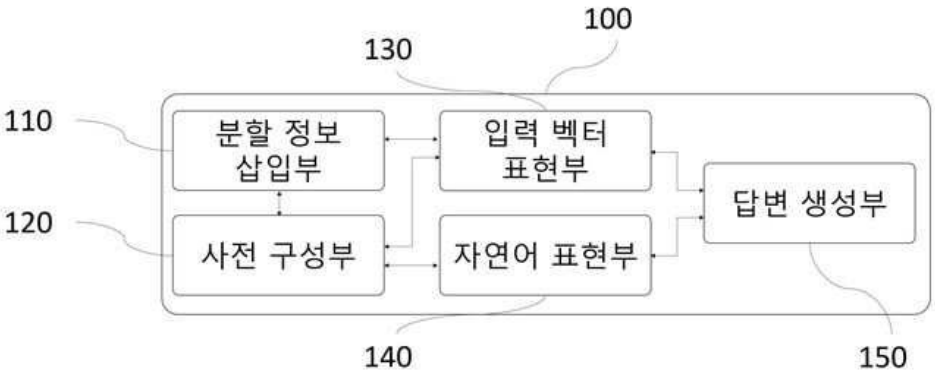
[0033] 한편, 본 발명은 컴퓨터 혹은 유사한 기기에서 실행될 수 있는 프로그램으로 작성 가능하며, 컴퓨터 혹은 유사한 기기로 읽을 수 있는 기록 매체를 이용하여 상기 프로그램을 동작시키는 범용 디지털 컴퓨터에서 구현될 수 있다. 상기 컴퓨터로 읽을 수 있는 기록 매체는 마그네틱 저장 매체(롬, 플로피 디스크, 하드 디스크 등), 광학적 판독 매체(예를 들면, 시디롬, 디브이디 등) 및 캐리어 웨이브(예를 들면, 인터넷을 통한 전송)와 같은 저장 매체를 포함한다.

[0034] 지금까지 본 발명에 대하여 그 바람직한 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 이를 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다.

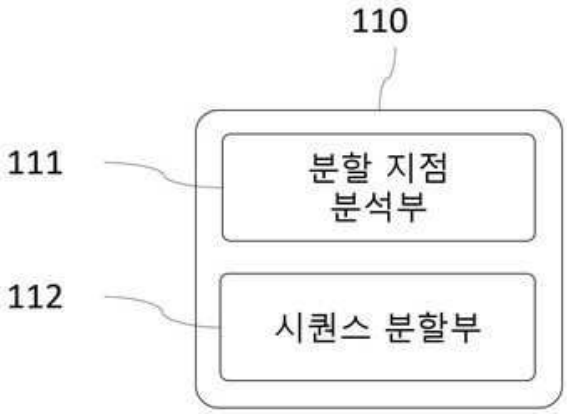
[0035] 본 발명의 범위는 전술한 설명이 아니라 특허 청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면

도면1



도면2



도면3