(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년09월04일
(11) 등록번호 10-2152339
(24) 등록일자 2020년08월31일

(51) 국제특허분류(Int. Cl.)
G10L 15/22 (2006.01) *G10L 15/04* (2006.01)
(52) CPC특허분류
G10L 15/22 (2013.01)
G10L 15/04 (2013.01)
(21) 출원번호 10-2018-0116575
(22) 출원일자 2018년09월28일
심사청구일자 2018년09월28일
(65) 공개번호 10-2019-0059201
(43) 공개일자 2019년05월30일
(30) 우선권주장
1020170156779 2017년11월22일 대한민국(KR)
(56) 선행기술조사문헌
JP2014224857 A*
KR1020070102267 A*
KR1020100016909 A*
KR1020110099434 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
서강대학교 산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
서정연
서울특별시 서초구 신반포로16길 15-20, 103동 2702호 (반포동, 반포힐스테이트)
구명완
서울특별시 양천구 목동동로 100, 1312동 1304호 (신정동, 목동13단지아파트)
허광호
서울특별시 강서구 공항대로48길 66, 301호
(74) 대리인
유미특허법인

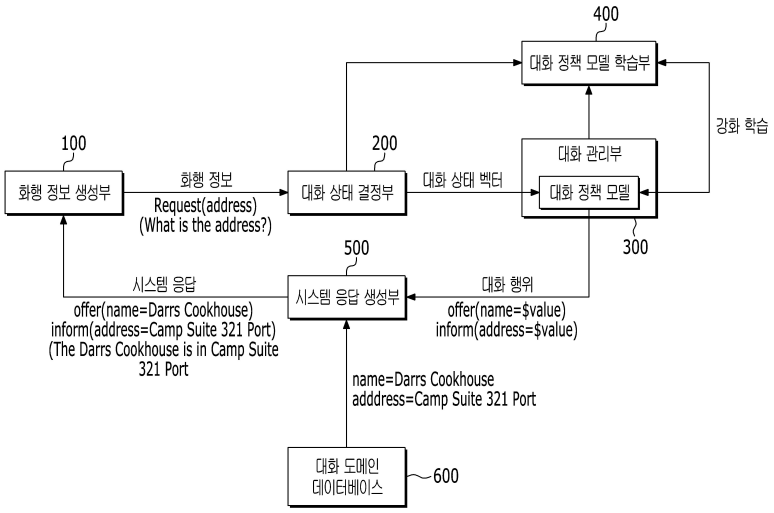
전체 청구항 수 : 총 6 항 심사관 : 김경완

(54) 발명의 명칭 **대화 정책 모델의 최적화 방법 및 이를 구현하는 대화 시스템**

(57) 요 약

대화 시스템으로서, 사용자 발화에 대한 화행(speech act) 정보를 수신하고, 상기 화행 정보를 이용하여 대화 상태를 결정하고, 상기 대화 상태를 벡터화하여 대화 상태 벡터를 생성하는 대화 상태 결정부, 상기 대화 상태 벡터를 수신하고, 상기 대화 상태 벡터를 대화 정책 모델에 입력하여 상기 대화 상태에서 가능한 대화 행위를 결정하는 대화 관리부, 그리고 대화 도메인 데이터베이스와 연동하여 상기 대화 행위의 슬롯에 삽입될 정보를 결정하고, 상기 정보를 상기 대화 행위의 슬롯에 삽입하여 시스템 응답을 생성하는 시스템 응답 생성부를 포함한다.

대 표 도



(52) CPC특허분류
G10L 2015/221 (2013.01)

명 세 서

청구범위

청구항 1

대화 시스템으로서,

사용자 발화에 대한 화행(speech act) 정보를 수신하고, 상기 화행 정보에 포함된 특정 슬롯의 값을 이용하여 상기 사용자 발화의 목표, 발화 방식 그리고 요청대상 중 적어도 하나를 포함하는 대화 상태를 결정하고, 상기 대화 상태에 대한 사용자 의도를 추정하고, 상기 대화 상태와 상기 사용자 의도가 포함된 대화 상태 벡터를 생성하는 대화 상태 결정부,

상기 대화 상태 벡터를 수신하고, 상기 대화 상태 벡터를 대화 정책 모델에 입력하여 상기 대화 상태에서 가능한 대화 행위를 결정하는 대화 관리부, 그리고

대화 도메인 데이터베이스를 검색하여 상기 대화 행위의 슬롯에 삽입될 정보를 결정하고, 상기 정보를 상기 슬롯에 삽입하여 시스템 응답을 생성하는 시스템 응답 생성부를 포함하고,

상기 대화 도메인 데이터베이스는 임의의 슬롯과 상기 임의의 슬롯에 매칭될 수 있는 복수의 제한 값들의 쌍을 포함하는, 대화 시스템.

청구항 2

삭제

청구항 3

제1항에서,

상기 대화 관리부는

상기 대화 상태 벡터를 상기 대화 정책 모델에 입력하여 상기 대화 상태에서 수행될 수 있는 후보 대화 행위들의 가치 값들을 출력하고, 가장 큰 가치 값을 가진 후보 대화 행위를 상기 대화 상태에서 가능한 대화 행위로 결정하는 대화 시스템.

청구항 4

제1항에서,

상기 대화 상태, 상기 대화 상태에서 가능한 대화 행위, 상기 대화 행위에 대한 보상 및 다음 대화 상태를 이용하여 상기 대화 정책 모델을 강화 학습시키는 대화 정책 모델 학습부

를 더 포함하는 대화 시스템.

청구항 5

대화 시스템이 대화 상태에서 가능한 대화 행위를 결정하는 대화 정책 모델을 강화 학습시키는 방법으로서,

사용자 발화에 대한 화행(speech act) 정보를 수신하는 단계,

상기 화행 정보에 포함된 특정 슬롯의 값을 이용하여 상기 사용자 발화의 목표, 발화 방식 그리고 요청대상을 추출하고, 상기 목표, 상기 발화 방식 그리고 상기 요청대상을 바탕으로 상기 화행 정보가 수신된 시점에서의 대화 상태를 결정하는 단계,

상기 대화 상태에 대한 사용자 의도를 추정하고, 상기 대화 상태와 상기 사용자 의도가 포함된 대화 상태 벡터를 생성하고, 상기 대화 상태 벡터를 상기 대화 정책 모델에 입력하여 상기 대화 상태에서 가능한 대화 행위를 결정하는 단계,

상기 대화 행위에 대응하는 보상 및 상기 대화 행위의 슬롯에 정보를 삽입하여 시스템 응답으로 출력한 이후의

대화 상태를 결정하는 단계,

상기 대화 상태, 상기 대화 행위, 상기 보상 및 상기 이후의 대화 상태를 이용하여 경험 데이터를 생성하는 단계, 그리고

상기 경험 데이터를 이용하여 상기 대화 정책 모델에 대한 강화 학습을 수행하는 단계를 포함하고,

상기 정보는 임의의 슬롯에 매칭될 수 있는 복수의 제한 값들 중에서 선택된 것인, 대화 정책 모델 강화 학습 방법.

청구항 6

삭제

청구항 7

삭제

청구항 8

제5항에서,

상기 대화 행위를 결정하는 단계는

상기 대화 상태에서 수행될 수 있는 후보 대화 행위들의 가치 값들을 출력하는 단계, 그리고

가장 큰 가치 값을 가진 후보 대화 행위를 상기 대화 상태에서 가능한 대화 행위로 결정하는 단계

를 포함하는 대화 정책 모델 강화 학습 방법.

청구항 9

제5항에서,

상기 대화 정책 모델에 대한 강화 학습을 수행하는 단계는

상기 화행 정보를 수신하는 단계, 상기 대화 상태를 결정하는 단계, 상기 대화 행위를 결정하는 단계, 상기 보상 및 이후 대화 상태를 결정하는 단계 및 상기 경험 데이터를 생성하는 단계를 복수회 수행하여 경험 데이터 집합을 생성하는 단계,

상기 경험 데이터 집합에서 임의의 경험 데이터를 추출하는 단계, 그리고

상기 임의의 경험 데이터를 이용하여 상기 대화 정책 모델의 가중치를 업데이트하는 단계

를 포함하는 정책 모델 강화 학습 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 대화 정책 모델의 최적화 방법 및 이를 구현하는 대화 시스템에 관한 것이다.

배 경 기 술

[0002] 대화 시스템은 사용자와 자연어로 대화하는 방식을 통해 사용자의 목표를 달성해주는 시스템으로서, 사용자의 발화에서 사용자의 의도를 이해하고 의도에 대응하는 작업을 처리한 후 처리결과를 사용자에게 자연어로 응답한다.

[0003] 한편, 대화 시스템의 컴포넌트 중 하나인 대화 관리부는 다른 컴포넌트들의 행동을 조율하고, 사용자와의 대화 흐름을 제어하며 외부 프로그램과의 연동을 관리하는 등 중심적인 역할을 담당한다. 대화 관리부의 행동 방식을 결정하는 대화 정책 모델(Dialog Policy Model)은 종래에는 규칙을 정의하는 방식으로 설계되었으며, 대표적으로 Call-flow 기반 대화 관리 방법 및 Form-filling 기반 대화 관리 방법이 있다.

[0004] Call-flow 기반 대화 관리 방법은 대화 관리부의 행동 방식을 Call-flow 그래프로 정의하며, Call-flow 그래프

는 대화 시스템의 대화 상태를 나타내는 노드(Node)와 사용자의 질의를 나타내는 호(Arc)로 구성된다. Call-flow 그래프 기반의 대화 관리부는 비교적 간단하고 대화 시스템의 행위가 예측 가능하다는 장점이 있다.

[0005] Form-filling 기반 대화 관리 방법은 사용자의 의도를 분석하는 부분과 사용자와의 대화 흐름을 제어하는 부분을 분리하여 관리하는 것이 특징이며, 사용자의 의도는 개념을 나타내는 슬롯(Slot)과 해당 개념에 대한 정보를 나타내는 필러(Filler)로 표현된다. Form-filling 기반의 대화 관리부는 Call-flow 그래프 기반의 대화 관리부보다 유연성이 있고, 사용자 주도적인 대화가 가능하다는 장점이 있다.

[0006] 그러나, 위와 같은 종래의 대화 관리 방법은 다음과 같은 한계를 가진다. 구체적으로, Call-flow 기반 대화 관리 방법은 대화를 진행하는데 있어 가능한 대화 상태를 Call-flow 그래프로 표현하기 때문에 도메인에 종속적이며, Call-flow 그래프에 의하여 대화 방식이 결정되기 때문에 대화 흐름이 시스템 주도적으로 진행되고 사용자의 발화가 제한적이다. 또한, Form-filling 기반 대화 관리 방법의 경우 대화를 진행하기 위해 가능한 대화 상태를 슬롯으로 표현하여 도메인에 종속적이고 확장이 어렵다.

[0007] 이에 따라, 규칙 정의 방식을 통해 대화 정책 모델을 설계하는 방식이 가지는 한계를 극복하기 위해, 대화를 장기적 결정과정으로 간주하고 대화 정책 모델을 학습시키는 강화 학습 방법이 연구되고 있으며, 대표적인 온라인 학습 알고리즘으로 Q-learning이 있다.

[0008] Q-learning 알고리즘은 환경으로부터 받은 보상을 이용하여 주어진 상태에서 수행할 수 있는 행위에 대한 가치인 Q 값(Q-value)을 계산하여 Q 값이 최대치를 가지는 행위를 결정한다. 그러나, Q-learning 알고리즘을 이용하여 정책을 최적화하는 것은 상당한 계산복잡도가 따르기 때문에 상태 공간이 상대적으로 작아야 한다는 조건을 전제로 한다. 따라서, 대화 시스템처럼 가능한 상태가 무한대에 가까운 경우 상태 공간이 매우 방대하기 때문에 통상적인 Q-learning 알고리즘을 이용하여 대화 정책 모델을 최적화하는 것은 매우 어려운 문제가 있다.

발명의 내용

해결하려는 과제

[0009] 본 발명이 해결하고자 하는 과제는 화행 정보들에 의해 결정된 대화 상태를 벡터화하여 대화 상태 벡터를 생성하고, 대화 상태 벡터를 대화 정책 모델에 입력하여 대화 행위를 결정하는 기술을 제공하는 것이다.

[0010] 또한, 본 발명이 해결하고자 하는 과제는 시간 단계마다 수신한 화행 정보를 이용하여 경험 데이터 집합을 생성하고, 경험 데이터 집합을 이용하여 대화 정책 모델을 강화 학습시키는 기술을 제공하는 것이다.

과제의 해결 수단

[0011] 본 발명의 일 실시예에 따른 대화 시스템은 사용자 발화에 대한 화행(speech act) 정보를 수신하고, 상기 화행 정보를 이용하여 대화 상태를 결정하고, 상기 대화 상태를 벡터화하여 대화 상태 벡터를 생성하는 대화 상태 결정부, 상기 대화 상태 벡터를 수신하고, 상기 대화 상태 벡터를 대화 정책 모델에 입력하여 상기 대화 상태에서 가능한 대화 행위를 결정하는 대화 관리부, 그리고 대화 도메인 데이터베이스와 연동하여 상기 대화 행위의 슬롯에 삽입될 정보를 결정하고, 상기 정보를 상기 대화 행위의 슬롯에 삽입하여 시스템 응답을 생성하는 시스템 응답 생성부를 포함한다.

[0012] 상기 대화 상태 벡터는 사용자 목표, 발화 방식 또는 사용자 요청 중 적어도 하나를 포함한다.

[0013] 상기 대화 관리부는 상기 대화 상태 벡터를 상기 대화 정책 모델에 입력하여 상기 대화 상태에서 수행될 수 있는 후보 대화 행위들의 가치 값들을 출력하고, 가장 큰 가치 값을 가진 후보 대화 행위를 상기 대화 상태에서 가능한 대화 행위로 결정한다.

[0014] 상기 대화 시스템은 상기 대화 상태, 상기 대화 상태에서 가능한 대화 행위, 상기 대화 행위에 대한 보상 및 다음 대화 상태를 이용하여 상기 대화 정책 모델을 강화 학습시키는 대화 정책 모델 학습부를 더 포함한다.

[0015] 본 발명의 일 실시예에 따른 대화 시스템이 대화 상태에서 가능한 대화 행위를 결정하는 대화 정책 모델을 강화 학습시키는 방법은 사용자 발화에 대한 화행(speech act) 정보를 수신하는 단계, 상기 화행 정보가 수신된 시점에서의 대화 상태를 결정하는 단계, 상기 대화 상태에서 가능한 대화 행위를 결정하는 단계, 상기 대화 행위에 대응하는 보상 및 이후 대화 상태를 결정하는 단계, 상기 대화 상태, 상기 대화 행위, 상기 보상 및 상기 이후 대화 상태를 이용하여 경험 데이터를 생성하는 단계, 그리고 상기 경험 데이터를 이용하여 대화 정책 모델에 대한 강화 학습을 수행하는 단계를 포함한다.

[0016] 상기 대화 상태에서 가능한 대화 행위를 결정하는 단계는 상기 대화 상태를 벡터화하여 대화 상태 벡터를 생성하는 단계, 그리고 상기 대화 상태 벡터를 상기 대화 정책 모델에 입력하여 상기 대화 상태에서 가능한 대화 행위를 결정하는 단계를 포함한다.

[0017] 상기 대화 상태 벡터는 사용자 목표, 발화 방식 또는 사용자 요청 중 적어도 하나를 포함한다.

[0018] 상기 대화 상태 벡터를 상기 대화 정책 모델에 입력하여 상기 대화 상태에서 가능한 대화 행위를 결정하는 단계는 상기 대화 상태 벡터를 상기 대화 정책 모델에 입력하여 상기 대화 상태에서 수행될 수 있는 후보 대화 행위들의 가치 값들을 출력하는 단계, 그리고 가장 큰 가치 값을 가진 후보 대화 행위를 상기 대화 상태에서 가능한 대화 행위로 결정하는 단계를 포함한다.

[0019] 상기 대화 정책 모델에 대한 강화 학습을 수행하는 단계는 상기 화행 정보를 수신하는 단계, 상기 대화 상태를 결정하는 단계, 상기 대화 행위를 결정하는 단계, 상기 보상 및 이후 대화 상태를 결정하는 단계 및 상기 경험 데이터를 생성하는 단계를 복수회 수행하여 경험 데이터 집합을 생성하는 단계, 상기 경험 데이터 집합에서 임의의 경험 데이터를 추출하는 단계, 그리고 상기 임의의 경험 데이터를 이용하여 상기 대화 정책 모델의 가중치를 업데이트하는 단계를 포함한다.

발명의 효과

[0020] 본 발명에 따르면, 대화 상태를 사용자 발화의도의 신뢰점수를 포함한 연속적 벡터로 표현하는바, 사용자 발화에 대한 정보를 효율적으로 대화 상태에 반영할 수 있다.

[0021] 또한, 본 발명에 따르면, 시간 단계마다 수신한 데이터를 이용하는 경험재현 기법을 사용하는바, 경험 데이터를 재사용하고 샘플간의 상관관계를 감소시켜 데이터 효율성을 높일 수 있다.

도면의 간단한 설명

[0022] 도 1은 한 실시예에 따른 대화 시스템을 설명하는 도면이다.

도 2는 대화 관리부가 대화 상태 벡터와 대화 정책 모델을 이용하여 대화 상태에서 가능한 대화 행위를 결정하는 방법을 설명하는 도면이다.

도 3은 대화 정책 모델 학습부가 대화 정책 모델을 강화 학습시키는 알고리즘을 도시한 도면이다.

도 4는 대화 시스템이 사용자 발화에 대한 대화 행위를 결정하는 대화 정책 모델을 최적화하는 방법을 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0023] 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.

[0024] 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.

[0025] 도 1은 한 실시예에 따른 대화 시스템을 설명하는 도면이고, 도 2는 대화 관리부가 대화 상태 벡터와 대화 정책 모델을 이용하여 대화 상태에서 가능한 대화 행위를 결정하는 방법을 설명하는 도면이고, 도 3은 대화 정책 모델 학습부가 대화 정책 모델을 강화 학습시키는 알고리즘을 도시한 도면이다.

[0026] 도 1을 참고하면, 대화 시스템(1000)은 화행 정보 생성부(100), 대화 상태 결정부(200), 대화 관리부(300), 대화 정책 모델 학습부(400), 시스템 응답 생성부(500) 및 대화 도메인 데이터베이스(600)를 포함한다.

[0027] 화행 정보 생성부(100)는 사용자 발화를 이용하여 화행 정보(speech act)를 생성하고, 생성한 화행 정보를 대화 상태 결정부(200)로 전송한다.

[0028] 예를 들면, 화행 정보 생성부(100)는 사용자 발화 "What is the addrees?"를 이용하여 화행 정보 "request(address)"를 생성하고 대화 상태 결정부(200)로 전송할 수 있다.

[0029] 또한, 화행 정보 생성부(100)는 대화 시작 시점에서 초기 사용자 목표를 설정하고, 초기 사용자 목표에 대한 화행 정보를 생성하여 대화 상태 결정부(200)로 전송할 수 있다. 이후, 화행 정보 생성부(100)는 초기 화행 정보에 대응하여 결정된 시스템 응답에 따라 초기 사용자 목표를 수정하고, 수정된 사용자 목표에 대한 화행 정보를 생성하여 대화 상태 결정부(200)로 전송함으로써 지속적인 대화를 진행할 수 있다.

[0030] 한편, 본 발명에서는 화행 정보 생성부(100)를 통해 사용자 발화를 이용하여 화행 정보를 생성하나, 다른 실시예에서는 사용자로부터 사용자 발화를 수신하여 텍스트로 변환해주는 음성 인식기(미도시) 및 변환된 텍스트에서 사용자의 의도를 분석하여 화행 정보를 생성하는 의미 분석기(미도시)에 의해 화행 정보를 생성할 수도 있다.

[0031] 대화 상태 결정부(200)는 사용자 발화에 대한 화행 정보를 수신하고, 화행 정보를 이용하여 대화 상태를 결정하고, 대화 상태를 벡터화하여 대화 상태 벡터를 생성한다.

[0032] 구체적으로, 대화 상태 결정부(200)는 수신된 화행 정보들을 이용하여 대화 상태를 결정한다.

[0033] 예를 들면, 수신된 화행 정보들이 표 1과 같은 경우, 대화 상태 결정부(200)는 수신된 화행 정보들을 이용하여 표 2와 같은 대화 상태를 결정할 수 있다.

표 1

[0034]

턴	발화자	발화 예시	화행 정보
1	시스템	안녕하세요. 무엇을 도와드릴까요?	welcomemsg()
	사용자	호주 음식을 먹고 싶어.	inform(food=austrian)
2	시스템	남쪽 지역에 darrs cookhouse 레스토랑이 있습니다.	offer(name=darrs cookhouse), inform(area=south)
3	사용자	특색메뉴는?	request(signature)
	시스템	Darrs cookhouse의 특색메뉴는 cake이고 전화번호는 123입니다.	offer(name=darrs cookhouse), inform(signature=cake), inform(phone=123)

표 2

[0035]

대화 상태		
사용자 목표	발화 방식	사용자 요청
food=austian	by name	signature

[0036] 표 2에서, 대화 상태는 사용자 목표, 발화 방식 또는 사용자 요청 중 적어도 하나를 포함한다. 사용자 목표는 수신된 화행 정보들 중에서 사용자가 제시한 inform 슬롯의 값으로 구성된다. 발화 방식은 수신된 화행 정보들 중에서 최근 화행 정보에 나타난 사용자 발화 방식에 대한 정보를 나타내며 "by name", "by constraints", "by alternatives", "finished" 또는 "none"중 어느 하나를 갖는다. 사용자 요청은 수신된 화행 정보들 중에서 사용자가 제시한 request 슬롯의 값으로 구성된다.대화 상태 결정부(200)는 사용자 목표, 발화 방식 또는 사용자 요청 중 적어도 하나를 벡터화 하여 대화 상태 벡터를 생성한다. 예를 들면, 대화 상태 결정부(200)는 사용자 목표에 대한 정보 및 발화 방식에 대한 정보를 5차원 벡터로 생성하고, 사용자 요청에 대한 정보를 9차원 벡터로 생성하여 대화 상태 벡터를 생성할 수 있다.

[0037] 한편, 대화 상태 결정부(200)는 화행 정보들을 이용하여 미리 설정된 수만큼의 사용자 의도를 추정하고, 사용자가 제시한 제약조건을 이용하여 연동된 데이터베이스에서 검색을 수행하여 검색된 결과를 결정하고, 대화 상태, 사용자 의도 및 검색 결과를 벡터화하여 대화 상태 벡터를 생성할 수 있다.

[0038] 예를 들면, 대화 상태 결정부(200)는 사용자 의도를 대화 상태에 반영하기 위해 사용자 발화에 대한 상위 3개의 인식 결과를 대화 상태 벡터에 추가하며 하나의 인식 결과 당 78차원 벡터로 표현할 수 있다.

[0039] 또한, 대화 상태 결정부(200)는 사용자가 제시한 제약조건으로 검색된 데이터베이스 결과 수를 대화 상태 벡터에 추가하며 한 결과당 1차원 벡터로 표현할 수 있다. 제약 조건은 사용자 목표에 대한 추정 결과를 이용하며, 데이터베이스에서 검색된 결과 수를 대화상태에 포함하는 것은 이후 대화 정책 모델이 검색된 결과에 따라 "추천"(offer) 또는 "도울 수 없음"(canthelp)의 화행을 결정할 수 있도록 하기 위함이다.

[0040] 표 3은 예시적인 대화 상태 벡터를 나타낸 표이다.

표 3

[0041]

대화 상태 벡터 (254차원)						
대화 상태			사용자 의도			검색 결과
사용자 목표	발화 방식	사용자 요청	제1 결과	제2 결과	제3 결과	검색된 결과 수
5차원	5차원	9차원	78차원	78차원	78차원	1차원

[0042] 대화 관리부(300)는 대화 상태 결정부(200)로부터 대화 상태 벡터를 수신하고, 대화 상태 벡터를 대화 정책 모델에 입력하여 대화 상태에서 가능한 대화 행위를 결정한다.예를 들면, 도 2를 참고하면, 대화 정책 모델은 254개의 노드를 가진 입력층, 120개의 노드를 가진 2개의 은닉층 및 51개의 노드를 가진 출력층으로 구성될 수 있으며, 대화 상태 벡터를 입력으로 받아 대화 상태에서 수행될 수 있는 후보 대화 행위들에 대한 가치 값들을 출력한다.

[0043] 만일 대화 정책 모델이 Q-learning 알고리즘으로 구현된 경우, 대화 정책 모델은 후보 대화 행위들에 대한 Q 값(Q-value)을 출력할 수 있다.

[0044] 대화 관리부(300)는 후보 대화 행위들에 대한 가치 값들을 비교하고, 가장 큰 가치 값을 가진 후보 대화 행위를 대화 상태에서 가능한 대화 행위로 결정한다.

[0045] 예를 들면, 대화 관리부(300)는 대화 상태에서 가능한 대화 행위로서 "offer(name=$value) inform(address=$value)"을 출력할 수 있다. 대화 행위에서 괄호는 대화 행위의 슬롯을 의미한다.

[0046] 한편, 대화 정책 모델은 화행 정보를 이용하여 결정한 대화 상태, 대화 상태에 대응하는 대화 행위, 대화 행위에 대한 보상 및 이후 대화 상태를 이용하여 강화 학습되었으며, 이에 대한 내용은 대화 정책 모델 학습부(400)를 통해 자세히 설명한다.

[0047] 대화 정책 모델 학습부(400)는 화행 정보를 이용하여 결정한 대화 상태, 대화 상태에 대응하는 대화 행위, 대화 행위에 대한 보상 및 다음 대화 상태를 이용하여 대화 정책 모델을 강화 학습시킨다.

[0048] 구체적으로, 강화 학습은 학습 에이전트(Agent)가 환경과의 일련의 상호작용을 통하여 학습하는 머신러닝 방법론으로, 특정 시점에서 강화학습 에이전트는 상태 s∈S에서 행위 a∈A를 취하고 상태전이함수 T(s,a,s')에 의하여 다음 상태 s'로 전이되며 환경의 보상함수 R(s,a,s')에 의해 보상 r을 받는다. 에이전트의 행위는 대화 정책 모델 π:S->A에 의해 결정되며 에이전트의 목표는 장기적으로 얻는 보상의 누적 합을 최대화하는 것이다.

[0049] 한편, 강화 학습 알고리즘 중 Q-learning 알고리즘은 모든 가능한 상태와 행위의 조합인 Q-value 함수 Q(s,a)를 추정한다. 다만, 대화 시스템과 같이 가능한 상태의 수가 무한대에 가까운 경우, Q-learning 알고리즘으로 Q-value 함수를 추정하는 것은 비현실적이므로, 함수 근사자 Θ를 사용하여 Q-value 함수 Q(s,a)를 $Q(s,a;\Theta)$로 근사화하여 근사화된 Q-value 함수 $Q(s,a;\Theta)$를 추정할 수 있다.

[0050] Q-value 함수를 추정하는 함수 근사자 Θ로 선형함수를 사용할 수 있지만, 신경망과 같은 비선형 함수를 사용할 수 있으며 가중치 Θ를 가진 신경망 비선형 함수를 Q-network이라고 한다. Q-network은 수학식 1과 같이 반복(iteration) i마다 변하는 손실 함수 $L_i(\Theta_i)$를 최소화하는 것으로 학습시킬 수 있다.

수학식 1

[0051]
$$L_i(\theta_i)=E(r+\gamma \max_{a'} Q(s',a';\theta_{i-1})-Q(s,a;\theta_i))^2]$$

[0052] 수학식 1에서, $L_i(\Theta_i)$는 손실 함수, r은 보상, s'는 다음 상태, a'는 다음 행위, s는 현재 상태, a는 현재 행위, Θ_{i-1}는 기존 가중치, Θ_i는 현재 가중치, γ는 할인 인자이다.

[0053] 반복 i에서 임의의 경험(s,a,r,s')이 주어졌을 때 Q-network 가중치 Θ_i에 대한 학습은 다음과 같은 방식으로 진행된다. 우선 Q-network의 기존 가중치 Θ_{i-1}를 이용하여 다음 상태 s'에서 선택할 수 있는 모든 행위 a'중 최

대 Q-value를 구하고 벨만 방정식에 의해 할인 인자 γ를 곱하고 보상 r을 더하여 목표 값(target)으로 설정한다. 다음 목표 값과 $Q(s,a;\Theta)$의 제곱오차가 최소가 되도록 Q-network의 가중치 Θ_i를 업데이트한다. 계산적 효율성을 위하여 상기 수학식 1에 대한 최적화 과정은 경사 하강법(SGD, Stochastic Gradient Descent) 알고리즘으로 진행할 수 있다.

[0054] 이하, 대화 정책 모델 학습부(400)가 Q-learning 알고리즘을 사용하여 Q-network 즉, 대화 정책 모델을 강화 학습시키는 방법을 설명한다.

[0055] 대화 정책 모델 학습부(400)는 대화 상태 결정부(200)가 결정한 대화 상태들, 대화 관리부(300)가 결정한 대화 행위들 및 대화 행위들 각각에 대한 보상 및 다음 대화 상태를 이용하여 경험 데이터 집합을 생성하고, 경험 데이터 집합에 포함된 경험 데이터를 이용하여 대화 정책 모델의 가중치를 업데이트 한다.

[0056] 구체적으로, 대화 정책 모델 학습부(400)는 시간 단계마다 발생하는 경험 데이터 $e_t=(s_t,a_t,r_t,s_{t+1})$를 경험 데이터 집합 $D=e_1, e_2, \ldots, e_t$에 저장한다. 경험 데이터 e_t에서, s_t는 대화 상태, a_t는 대화 상태에서 가능한 대화 행위, r_t는 대화 행위에 대한 보상, s_{t+1}는 대화 행위에 대한 다음 대화 상태를 의미한다.

[0057] 대화 정책 모델 학습부(400)는 경험 데이터 집합 D에서 임의의 경험 데이터를 추출하여 대화 정책 모델의 가중치 업데이트 과정에 사용한다.

[0058] 해당 방법을 통해, 대화 시스템(1000)은 경험 데이터를 재사용하고, 경험 데이터들 간의 상관관계를 감소시켜 데이터 효율성을 높일 수 있다. 도 3은 대화 정책 모델 학습부(400)가 대화 정책 모델을 강화 학습시키는 알고리즘을 도시한 도면이다.

[0059] 시스템 응답 생성부(500)는 대화 관리부(300)로부터 대화 행위를 수신하고, 대화 도메인 데이터베이스(600)와 연동하여 대화 행위의 슬롯에 삽입될 정보를 결정하고, 정보를 대화 행위의 슬롯에 삽입하여 시스템 응답을 생성한다.

[0060] 구체적으로, 대화 도메인 데이터베이스(600)는 대화 도메인 온톨로지를 이용하여 생성된 복수의 장소 정보들을 저장한다.

[0061] 복수의 장소 정보들은 복수의 슬롯들로 구성되며, 각 슬롯은 복수의 제한 값들을 포함한다.

[0062] 예를 들면, 대화 도메인 데이터베이스(600)에 포함된 장소 정보들은 지역(area), 음식의 종류(food), 장소명(name), 가격(pricerange), 주소(addr), 전화번호(phone), 우편번호(postcode)와 시그네이처(signature)의 8개의 슬롯으로 구성될 수 있고, 지역 슬롯은 5개의 제한 값, 음식의 종류는 91개의 제한 값, 장소명은 113개의 제한 값, 가격은 3개의 제한 값들을 가질 수 있다. 이 경우, 제한 값은 해당 슬롯의 구체적인 정보를 지칭한다.

[0063] 시스템 응답 생성부(500)는 대화 행위의 슬롯을 대화 도메인 데이터베이스(600)에서 검색하며, 검색 결과는 대화 행위의 슬롯과 매칭되는 슬롯과 제한 값의 쌍으로 구성된다. 예를 들면, 대화 행위의 슬롯이 food인 경우 검색 결과는 (food=Austrian)일 수 있고, 대화 행위의 슬롯이 pricerange인 경우 검색 결과는 (pricerange=cheap)일 수 있다.

[0064] 시스템 응답 생성부(500)는 검색 결과에서 제한 값을 대화 행위의 슬롯에 삽입될 정보로 결정하고, 제한 값을 대화 행위의 슬롯에 삽입하여 시스템 응답을 생성한다.

[0065] 도 4는 대화 시스템이 사용자 발화에 대한 대화 행위를 결정하는 대화 정책 모델을 최적화하는 방법을 설명하는 도면이다.

[0066] 도 4에서, 도 1 내지 도 3과 동일한 내용은 그 설명을 생략한다.

[0067] 도 4를 참고하면, 대화 시스템(1000)은 사용자 발화에 대한 화행 정보를 수신한다(S100).

[0068] 대화 시스템(1000)은 화행 정보가 수신된 시점에서의 대화 상태를 결정한다(S110).

[0069] 대화 시스템(1000)은 대화 상태에서 가능한 대화 행위, 대화 행위에 대응하는 보상 및 이후 대화 상태를 결정한다(S120).

[0070] 구체적으로, 대화 시스템(1000)은 대화 상태를 벡터화하여 대화 상태 벡터를 생성하고, 대화 상태 벡터를 대화 정책 모델에 입력하여 대화 상태에서 가능한 대화 행위, 해당 대화 행위에 대한 보상 및 이후 대화 상태를 결정

한다.

[0071] 대화 시스템(1000)은 대화 상태, 대화 행위, 보상 및 이후 대화 상태를 이용하여 경험 데이터를 생성하고(S130), 수신되는 화행 정보마다 단계 S110 내지 S130을 반복하여 경험 데이터 집합을 생성한다(S140).

[0072] 대화 시스템(1000)은 경험 데이터를 이용하여 대화 정책 모델에 대한 강화 학습을 수행한다.

[0073] 구체적으로, 대화 시스템(1000)은 경험 데이터 집합에서 임의의 경험 데이터를 추출한다(S150). 이 경우, 대화 시스템(1000)은 메모리에 저장된 경험 데이터 집합에서 일정한 확률로 경험 데이터를 추출할 수 있다.

[0074] 대화 시스템(1000)은 추출한 경험 데이터를 이용하여 대화 정책 모델의 가중치를 업데이트한다(S160). 만일 대화 정책 모델이 Q-learning 알고리즘으로 구현된 경우, 대화 시스템(1000)은 대화 정책 모델에서 Q 값을 결정하기 위한 가중치를 추출한 경험 데이터를 이용하여 업데이트시킬 수 있다.

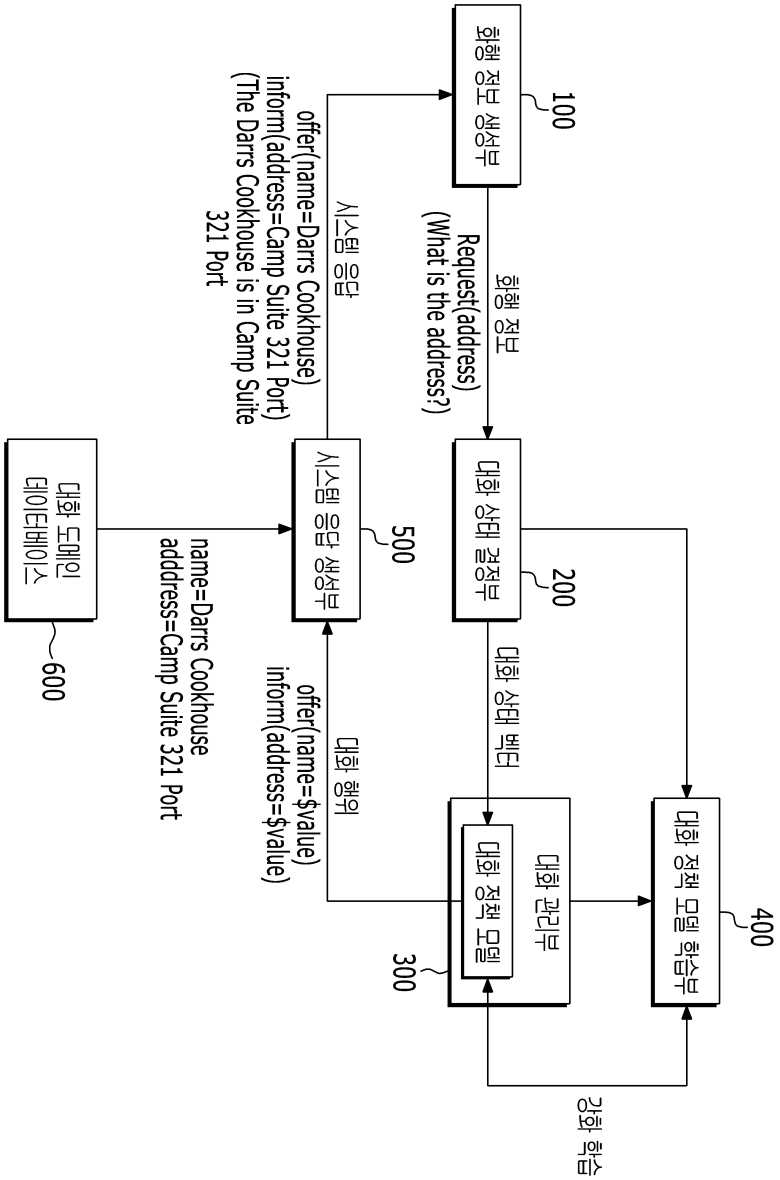
[0075] 본 발명에 따르면, 대화 상태를 사용자 발화의도의 신뢰점수를 포함한 연속적 벡터로 표현하는바, 사용자 발화에 대한 정보를 효율적으로 대화 상태에 반영할 수 있다.

[0076] 또한, 본 발명에 따르면, 시간 단계마다 결정한 데이터를 이용하는 경험재현 기법을 사용하는바, 경험 데이터를 재사용하고 샘플간의 상관관계를 감소시켜 데이터 효율성을 높일 수 있다.

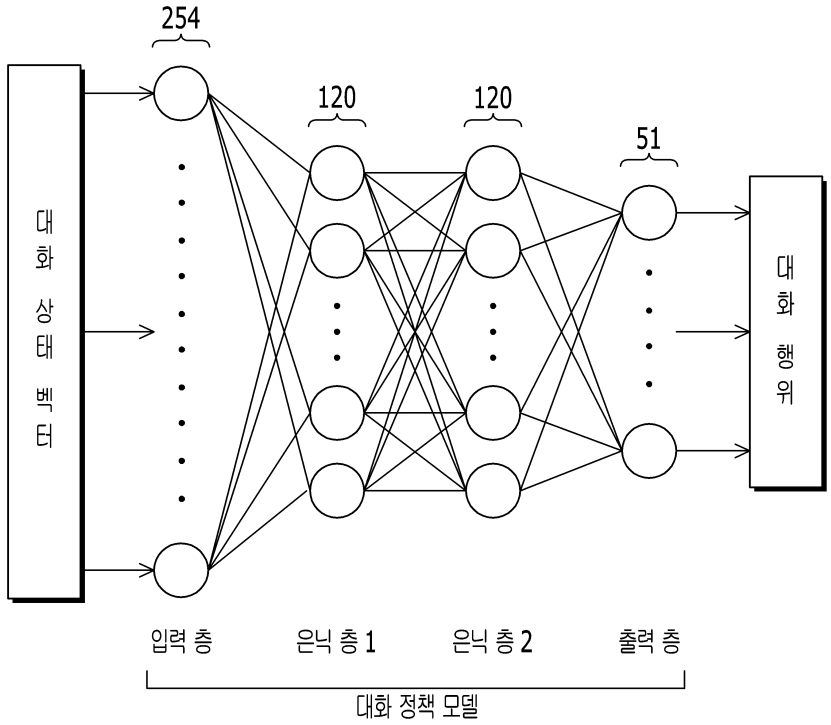
[0077] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

도면1


화행 정보 생성부
100
Request(address)
(What is the address?)
화행 정보
대화 상태 결정부
200
대화 상태 벡터
대화 관리부
대화 정책 모델
300
대화 정책 모델 학습부
400
강화 학습
대화 행위
offer(name=$value)
inform(address=$value)
시스템 응답 생성부
500
시스템 응답
offer(name=Darrs Cookhouse)
inform(address=Camp Suite 321 Port)
(The Darrs Cookhouse is in Camp Suite
321 Port
name=Darrs Cookhouse
address=Camp Suite 321 Port
대화 도메인
데이터베이스
600

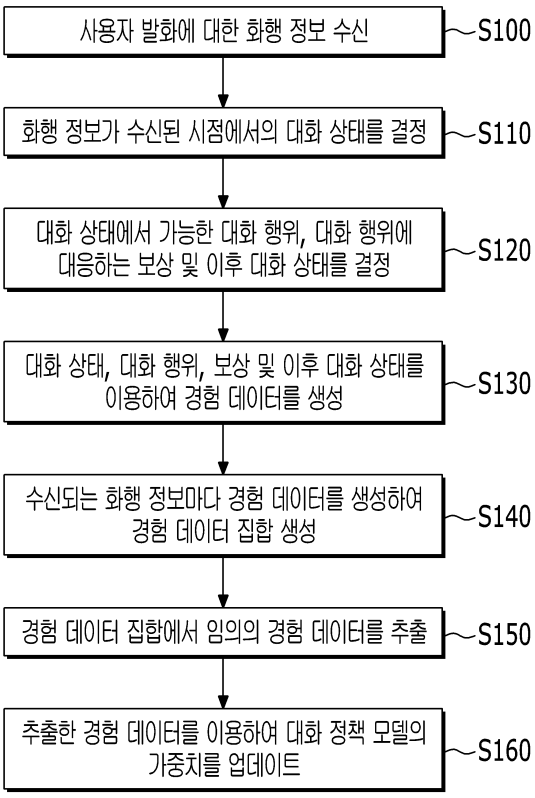
도면2



도면3

1: Initialize replay memory D to capacity N
2: Initialize action-value function Q with random weights
3: for episode = 1 to M do
4: Initialized state s_1
5: for t=1, T do
6: With probability $\in$ select a random action a_t
7: otherwise select $a_t = \max_a Q^*(s_t, a;\theta)$
8: Execute action a_t and observe reward r_t and next state s_{t+1}
9: Store experience (s_t, a_t, r_t, s_{t+1}) in D
10: Sample random mini-batch of experiences (s_j, a_j, r_j, s_{j+1}) from D
11: Set $y_j = \begin{cases} r_j & \text{for terminal state } s_{j+1} \\ r_j + \gamma \max_{a'} Q(s_{j+1}, a';\theta) & \text{for non terminal } s_{j+1} \end{cases}$
12: Perform a gradient descent step on $(y_j - Q(s_j, a_j;\theta))^2$

도면4


사용자 발화에 대한 화행 정보 수신
S100
화행 정보가 수신된 시점에서의 대화 상태를 결정
S110
대화 상태에서 가능한 대화 행위, 대화 행위에 대응하는 보상 및 이후 대화 상태를 결정
S120
대화 상태, 대화 행위, 보상 및 이후 대화 상태를 이용하여 경험 데이터를 생성
S130
수신되는 화행 정보마다 경험 데이터를 생성하여 경험 데이터 집합 생성
S140
경험 데이터 집합에서 임의의 경험 데이터를 추출
S150
추출한 경험 데이터를 이용하여 대화 정책 모델의 가중치를 업데이트
S160