(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년09월16일
(11) 등록번호 10-2444814
(24) 등록일자 2022년09월14일

(51) 국제특허분류(Int. Cl.)
G06F 16/332 (2019.01) *G06F 16/335* (2019.01)
G06Q 50/30 (2012.01) *H04L 51/00* (2022.01)
(52) CPC특허분류
G06F 16/3329 (2019.01)
G06F 16/335 (2019.01)
(21) 출원번호 10-2020-0163934
(22) 출원일자 2020년11월30일
심사청구일자 2020년11월30일
(65) 공개번호 10-2022-0075638
(43) 공개일자 2022년06월08일
(56) 선행기술조사문헌
KR1020190011570 A*
KR1020050012015 A
박남기, 인공지능과 윤리적 이슈, 언론정보연구 57권 3호, 2020.08. pp. 122-154. 1부.*
오신혁 외 6인, Dual WGAN 기반 페르소나 Multi-Turn 챗봇, 제31회 한글 및 한국어 정보처리 학술대회 논문집, 2019. pp. 49-53. 1부.*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김우주
서울특별시 서대문구 수색로 100, 301동 1704호(북가좌동, DMC 래미안 e편한세상)
(74) 대리인
특허법인우인

전체 청구항 수 : 총 3 항 심사관 : 최재귀

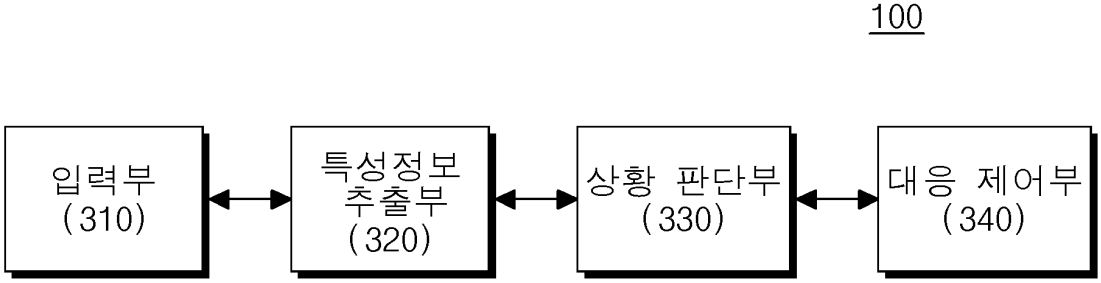
(54) 발명의 명칭 **비윤리 상황에서 사용자 특성정보 기반 챗봇 대응 제어 방법 및 그를 위한 장치**

(57) 요 약

비윤리 상황에서 사용자 특성정보 기반 챗봇 대응 제어 방법 및 그를 위한 장치를 개시한다.

본 발명의 실시예에 따른 챗봇 대응 제어 방법은, 사용자 또는 외부 장치로부터 입력 데이터를 수신하는 입력 단계; 상기 입력 데이터에서 사용자 특성정보와 대화 문맥 정보를 추출하는 특성정보 추출 단계; 상기 사용자 특성정보 및 상기 대화 문맥 정보로 최종 상황 판단 결과를 도출하고, 최종 상황 판단 결과를 기반으로 비윤리 대화 상황 여부를 판단하여 대화 상황 판단결과를 생성하는 상황 판단 단계; 및 상기 대화 상황 판단결과를 기반으로 비윤리 상황에 대한 대응 방안을 결정하고, 결정된 상기 대응 방안에 대한 대응 데이터가 출력되도록 제어하는 대응 제어 단계를 포함할 수 있다.

대 표 도 - 도3



(52) CPC특허분류
G06Q 50/30 (2015.01)
H04L 51/02 (2022.05)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711116331
과제번호 2016-0-00562-005
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송연구개발사업
연구과제명 상대방의 감성을 추론, 판단하여 그에 맞추어 대화하고 대응할 수 있는 감성 지능 연구개발
기 여 율 1/1
과제수행기관명 한국과학기술원
연구기간 2020.03.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

챗봇 대응 제어장치에서 챗봇 대응을 제어하는 방법에 있어서,

사용자 또는 외부 장치로부터 입력 데이터를 수신하는 입력 단계;

상기 입력 데이터에서 사용자 특성정보와 대화 문맥 정보를 추출하는 특성정보 추출 단계;

상기 사용자 특성정보 및 상기 대화 문맥 정보로 최종 상황 판단 결과를 도출하고, 최종 상황 판단 결과를 기반으로 비윤리 대화 상황 여부를 판단하여 대화 상황 판단결과를 생성하는 상황 판단 단계; 및

상기 대화 상황 판단결과를 기반으로 비윤리 상황에 대한 대응 방안을 결정하고, 결정된 상기 대응 방안에 대한 대응 데이터가 출력되도록 제어하는 대응 제어 단계를 포함하되,

상기 상황 판단 단계는, 상기 사용자 특성정보를 기반으로 제1 상황 판단을 수행하고, 상기 대화 문맥 정보를 기반으로 제2 상황 판단을 수행하며,

상기 제1 상황 판단 결과에 상기 사용자 특성정보에 기 설정된 우선순위에 대한 제1 가중치를 부여하고, 상기 제2 상황 판단 결과에 상기 문맥 정보에 기 설정된 우선순위에 대한 제2 가중치를 부여한 후 가중치가 적용된 제1 상황 판단 결과 및 제2 상황 판단 결과를 합산하여 산출된 판단 결과값을 기 설정된 판단 등급 기준과 비교하여 상기 최종 상황 판단 결과를 기반으로 상기 비윤리 대화 상황 여부를 판단하되,

상기 대응 제어 단계는, 기 학습된 모델을 불러와 사용자 특성정보와 비윤리 대화 상황 정보를 학습된 모델의 입력 값으로 삽입하여 모델 출력 값으로 도출된 사용자 별 대응 방안에 대한 상기 대응 데이터가 csv 파일 형태로 출력되도록 제어하는 것을 특징으로 하는 챗봇 대응 제어 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 대응 제어 단계는,

화제전환, 동조, 제지 및 침묵 중 적어도 하나의 방안에 대한 상기 대응 데이터가 출력되도록 제어하는 것을 특징으로 하는 챗봇 대응 제어 방법.

청구항 5

사용자 또는 외부 장치로부터 입력 데이터를 수신하는 입력부;

상기 입력 데이터에서 사용자 특성정보와 대화 문맥 정보를 추출하는 특성정보 추출부;

상기 사용자 특성정보 및 상기 대화 문맥 정보로 최종 상황 판단 결과를 도출하고, 최종 상황 판단 결과를 기반으로 비윤리 대화 상황 여부를 판단하여 대화 상황 판단결과를 생성하는 상황 판단부; 및

상기 대화 상황 판단결과를 기반으로 비윤리 상황에 대한 대응 방안을 결정하고, 결정된 상기 대응 방안에 대한 대응 데이터가 출력되도록 제어하는 대응 제어부를 포함하되,

상기 상황 판단부는, 상기 사용자 특성정보를 기반으로 제1 상황 판단을 수행하고, 상기 대화 문맥 정보를 기반으로 제2 상황 판단을 수행하며,

상기 제1 상황 판단 결과에 상기 사용자 특성정보에 기 설정된 우선순위에 대한 제1 가중치를 부여하고, 상기 제2 상황 판단 결과에 상기 문맥 정보에 기 설정된 우선순위에 대한 제2 가중치를 부여한 후 가중치가 적용된 제1 상황 판단 결과 및 제2 상황 판단 결과를 합산하여 산출된 판단 결과값을 기 설정된 판단 등급 기준과 비교하여 상기 최종 상황 판단 결과를 기반으로 상기 비윤리 대화 상황 여부를 판단하되,

상기 대응 제어부는, 기 학습된 모델을 불러와 사용자 특성정보와 비윤리 대화 상황 정보를 학습된 모델의 입력값으로 삽입하여 모델 출력 값으로 도출된 사용자 별 대응 방안에 대한 상기 대응 데이터가 csv 파일 형태로 출력되도록 제어하는 것을 특징으로 하는 챗봇 대응 제어장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 비윤리 상황이 발생한 경우 사용자 특성정보를 기반으로 챗봇 대응을 제어하는 방법 및 그를 위한 장치에 관한 것이다.

배 경 기 술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0003] 감성 ICT(Information & Communication Technology) 산업의 전세계 시장 규모는 2015년 185억 달러 규모에서, 2019년 270억 달러로 성장할 것으로 전망되며, 국내 감성 ICT 시장 규모는 2015년 6조원 규모에서 2019년에는 10조원 규모로 성장할 것으로 파악되고 있다.

[0004] 국내외적으로 대화 에이전트 및 감성 ICT 기술을 개발 및 실용화한 기관 및 기업이 증가하고 있으며 AI 스피커를 접목한 대화 에이전트에 대한 기술이 활발하게 진행되고 있다. 하지만, 기존까지의 대화 시스템은 사용자의 정보 및 상황, 대화의 맥락 등을 고려하지 않고 특정 목적에 제한적으로 적용이 가능한 경우가 많다. 또한, 사용자와의 대화에서 비윤리 상황에 대한 기준이 명확하지 않고, 비윤리 상황에서 챗봇의 대응을 제어하는 것을 어렵다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 상대방과의 대화시 비이성적인 욕설 등의 윤리적 판단이 필요한 상황에서 적절한 대화 및 대응하여 인간다운 감성 및 심리를 반영한 디지털 동반자 기술 및 응용서비스를 제공하는 비윤리 상황에서 사용자 특성정보 기반 챗봇 대응 제어 방법 및 그를 위한 장치를 제공하는 데 주된 목적이 있다.

과제의 해결 수단

[0006] 본 발명의 일 측면에 의하면, 상기 목적을 달성하기 위한 챗봇 대응 제어 방법은, 사용자 또는 외부 장치로부터 입력 데이터를 수신하는 입력 단계; 상기 입력 데이터에서 사용자 특성정보와 대화 문맥 정보를 추출하는 특성정보 추출 단계; 상기 사용자 특성정보 및 상기 대화 문맥 정보로 최종 상황 판단 결과를 도출하고, 최종 상황 판단 결과를 기반으로 비윤리 대화 상황 여부를 판단하여 대화 상황 판단결과를 생성하는 상황 판단 단계; 및 상기 대화 상황 판단결과를 기반으로 비윤리 상황에 대한 대응 방안을 결정하고, 결정된 상기 대응 방안에 대한 대응 데이터가 출력되도록 제어하는 대응 제어 단계를 포함할 수 있다.

[0007] 또한, 본 발명의 다른 측면에 의하면, 상기 목적을 달성하기 위한 챗봇 대응 제어장치는, 사용자 또는 외부 장치로부터 입력 데이터를 수신하는 입력부; 상기 입력 데이터에서 사용자 특성정보와 대화 문맥 정보를 추출하는 특성정보 추출부; 상기 사용자 특성정보 및 상기 대화 문맥 정보로 최종 상황 판단 결과를 도출하고, 최종 상황 판단 결과를 기반으로 비윤리 대화 상황 여부를 판단하여 대화 상황 판단결과를 생성하는 상황 판단부; 및 상기 대화 상황 판단결과를 기반으로 비윤리 상황에 대한 대응 방안을 결정하고, 결정된 상기 대응 방안에 대한 대응 데이터가 출력되도록 제어하는 대응 제어부를 포함할 수 있다.

발명의 효과

[0008] 이상에서 설명한 바와 같이, 본 발명은 사용자의 정보 및 상황, 대화의 맥락 등을 고려하여 대화 상황을 판단할 수 있고, 이에 대한 적절한 챗봇 대응을 수행할 수 있는 효과가 있다.

도면의 간단한 설명

[0009] 도 1은 본 발명의 실시예에 따른 챗봇 대응 제어장치의 구성을 개략적으로 나타낸 도면이다.

도 2는 본 발명의 실시예에 따른 챗봇 대응 제어장치의 동작을 설명하기 위한 예시도이다.

도 3은 본 발명의 실시예에 따른 챗봇 대응 제어장치를 개략적으로 나타낸 블록 구성도이다.

도 4는 본 발명의 실시예에 따른 챗봇 제어 방법을 설명하기 위한 예시도이다.

도 5는 본 발명의 실시예에 따른 입력 데이터를 나타낸 예시도이다.

도 6은 본 발명의 실시예에 따른 챗봇 제어 프로그램의 프로세스를 나타낸 예시도이다.

도 7a 및 도 7b는 본 발명의 실시예에 따른 입력 데이터의 변수를 나타낸 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0010] 이하, 본 발명의 바람직한 실시예를 첨부된 도면들을 참조하여 상세히 설명한다. 본 발명을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다. 또한, 이하에서 본 발명의 바람직한 실시예를 설명할 것이나, 본 발명의 기술적 사상은 이에 한정하거나 제한되지 않고 당업자에 의해 변형되어 다양하게 실시될 수 있음은 물론이다. 이하에서는 도면들을 참조하여 본 발명에서 제안하는 비윤리 상황에서 사용자 특성정보 기반 챗봇 대응 제어 방법 및 그를 위한 장치에 대해 자세하게 설명하기로 한다.

[0011] 본 발명에 따른 챗봇 대응 제어장치는 스마트 기기와 연계된 감성지능형 개인비서 시스템, 타겟 마케팅 및 컨텐츠 추천, 의료 산업, 감성교육 (social & emotional learning), 라이프 로깅 (life logging) 가능한 개인 일기 시스템, 엔터테인먼트 및 게임 등과 같이 다양한 분야에 적용 가능하다.

[0012] 도 1은 본 발명의 실시예에 따른 챗봇 대응 제어장치의 구성을 개략적으로 나타낸 도면이다.

[0013] 도 1에 도시된 챗봇 대응 제어장치(100)는 컴퓨팅 기기로 구현될 수 있으며, 적어도 하나의 프로세서(110), 컴퓨터 판독 가능한 저장매체(120) 및 통신 버스(160)를 포함한다.

[0014] 챗봇 대응 제어장치(100)의 입력부(110)는 입출력 인터페이스(140) 또는 통신 인터페이스(150)에 대응할 수 있고, 특성정보 추출부(320), 상황 판단부(330) 및 대응 제어부(340)는 프로세서(110)에 대응할 수 있다.

[0015] 프로세서(110)는 챗봇 대응 제어장치(100)의 동작을 제어할 수 있다. 예컨대, 프로세서(110)는 컴퓨터 판독 가능한 저장매체(120)에 저장된 하나 이상의 프로그램들을 실행할 수 있다. 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 컴퓨터 실행 가능 명령어는 프로세서(110)에 의해 실행되는 경우 챗봇 대응 제어장치(100)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0016] 컴퓨터 판독 가능한 저장매체(120)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능한 저장매체(120)에 저장된 프로그램(130)은 프로세서(110)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독한 가능 저장매체(120)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 챗봇 대응 제어장치(100)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장매체, 또는 이들의 적합한 조합일 수 있다.

[0017] 통신 버스(160)는 프로세서(110), 컴퓨터 판독 가능한 저장매체(120)를 포함하여 챗봇 대응 제어장치(100)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0018] 챗봇 대응 제어장치(100)는 또한 하나 이상의 입출력 장치를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(140) 및 하나 이상의 통신 인터페이스(150)를 포함할 수 있다. 입출력 인터페이스(140) 및 통신 인터페이스(150)는 통신 버스(160)에 연결된다. 입출력 장치는 입출력 인터페이스(140)를 통해 챗봇 대응 제어장

치(100)의 다른 컴포넌트들에 연결될 수 있다.

[0019] 도 2는 본 발명의 실시예에 따른 챗봇 대응 제어장치의 동작을 설명하기 위한 예시도이다.

[0020] 본 발명의 실시예에 따른 챗봇 대응 제어장치(100)는 상대방과의 대화시 비이성적인 욕설 등의 윤리적 판단이 필요한 상황에서 적절한 대화 및 대응하여 인간다운 감성 및 심리를 반영한 디지털 동반자 기술 및 응용서비스를 제공하는 것을 목표로 한다.

[0021] 구체적으로, 챗봇 대응 제어장치(100)는 스스로의 판단에 따라, 욕설 등 반사회적 비윤리적 대화를 하지 않고 사회적 윤리적 가치와 대화를 유지하는 윤리지능 보유하는 것을 목표로 한다. 이를 위해, 챗봇 대응 제어장치(100)는 인공신경망 모델을 적용하여 사용자 특성정보와 비윤리 상황 정보를 반영하여 적절한 대응방식(동조, 제지, 침묵, 화제전환 등)을 추론하는 동작을 수행한다.

[0022] 도 2는 챗봇 대응 제어장치(100)의 모델 아키텍처를 나타낸다.

[0023] 딥 러닝(deep learning)은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화(abstractions, 다량의 데이터나 복잡한 자료들 속에서 핵심적인 내용 또는 기능을 요약하는 작업)를 시도하는 기계 학습 알고리즘의 집합으로 정의되며, 큰 틀에서 사람의 사고방식을 컴퓨터에게 가르치는 기계학습의 한 분야이다. 어떠한 데이터가 있을 때 이를 컴퓨터가 알아 들을 수 있는 형태(예를 들어 이미지의 경우는 픽셀정보를 열벡터로 표현하는 등)로 표현(representation)하고 이를 학습에 적용하기 위해 많은 연구(어떻게 하면 더 좋은 표현기법을 만들고 또 어떻게 이것들을 학습할 모델을 만들지에 대한)가 진행되고 있으며, 이러한 노력의 결과로 deep neural networks, convolutional deep neural networks, deep belief networks와 같은 다양한 딥 러닝 기법들이 컴퓨터 비전, 음성인식, 자연어 처리, 음성/신호처리 등의 분야에 적용되어 최첨단의 결과들을 보여주고 있다.

[0024] 심층신경망(Deep Neural Network)은 딥러닝 알고리즘 중 하나로 입력층(input layer, 210)과 출력층(output layer, 230) 사이에 여러 개의 은닉층(hidden layer)들로 이루어진 인공신경망층(Artificial Neural Network layer, 220)이며 복잡한 비선형 관계(non-linear relationship)들을 모델링할 수 있다.

[0025] 인공 신경망은 노드들의 그룹으로 연결되어 있으며 이들은 뇌의 방대한 뉴런의 네트워크와 유사하다. 위 그림에서 각 원모양의 노드는 인공 뉴런을 나타내고 화살표는 하나의 뉴런의 출력에서 다른 하나의 뉴런으로의 입력을 나타낸다

[0026] 심층 신경망은 표준 오류역전파 알고리즘으로 학습될 수 있다. 이때, 가중치(weight, w)들은 아래의 등식을 이용한 확률적 경사 하강법(stochastic gradient descent)을 통하여 수학식 1과 같이 갱신될 수 있다.

수학식 1

[0027] $$\Delta w_{ij}(t+1) = \Delta w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}}$$

[0028] 여기서 η는 학습률(learning rate)를 의미하며, C는 비용함수(cost function)을 의미한다. 비용함수의 선택은 학습의 형태(지도학습, 비지도학습, 강화학습 등)와 활성화 함수(activation function)같은 요인들에 의해서 결정된다. 예를 들어서 다종 분류 문제(multiclass classification problem)에 지도 학습을 수행할 때, 일반적으로 활성화함수와 비용함수는 각각 softmax 함수와 교차 엔트로피 함수(cross entropy function)로 결정된다. softmax 함수는 $p_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)}$ 로 정의된다, 이때,p_j는 클래스 확률(class probability)을 나타내며 x_j와 x_k는 각각 유닛 j로의 전체 입력(total input)과 유닛 k로의 전체 입력을 나타낸다.

[0029] 교차 엔트로피는 $C = -\sum_j d_j \log(p_j)$ 로 정의된다. 이 때 d_j는 출력 유닛 j에 대한 목표 확률(target probability)을 나타내며, p_j는 해당 활성화함수를 적용한 이후의 j 에 대한 확률 출력(probability output)이다.

[0030] 도 2의 모델 입력층(210)은 도 5에 도시된 바와 같은 형태의 입력 데이터를 입력 받을 수 있다. 입력 데이터는 개개인의 사용자 특성정보, 비윤리 상황 정보 등을 포함할 수 있다. 또한, 사용자 특성정보는 도 7a 및 도 7b에

도시된 바와 같이, 다양한 변인을 포함할 수 있다.

[0031] 도 2의 신경망층(220)은 히든 레이어르 포함하며, 다양한 형태의 신경망으로 구성될 수 있다. 신경망층(220)은 deep neural networks, convolutional deep neural networks, deep belief networks 등과 같은 다양한 딥 러닝 기법으로 구현될 수 있다.

[0032] 도 2의 모델 출력층(230)은 비윤리 상황으로 판단된 경우에 대해 입력에 따라 각 개인이 적절하다고 판단된 대응방식들을 출력할 수 있다. 여기서, 모델 출력층(230)은 화제전환, 동조, 제지, 침묵 등의 대응 방식으로 출력될 수 있다.

[0033] 도 3은 본 발명의 실시예에 따른 챗봇 대응 제어장치를 개략적으로 나타낸 블록 구성도이다.

[0034] 본 실시예에 따른 챗봇 대응 제어장치(100)는 입력부(310), 특성정보 추출부(320), 상황 판단부(330) 및 대응 제어부(340)를 포함한다. 도 3의 챗봇 대응 제어장치(100)는 일 실시예에 따른 것으로서, 도 3에 도시된 모든 블록이 필수 구성요소는 아니며, 다른 실시예에서 챗봇 대응 제어장치(100)에 포함된 일부 블록이 추가, 변경 또는 삭제될 수 있다. 한편, 챗봇 대응 제어장치(100)는 컴퓨팅 디바이스로 구현될 수 있고, 챗봇 대응 제어장치(100)에 포함된 각 구성요소들은 각각 별도의 소프트웨어 프로그램으로 구현되거나, 소프트웨어가 결합된 별도의 하드웨어 장치로 구현될 수 있다.

[0035] 챗봇 대응 제어장치(100)는 윤리적 판단과 의사결정에 영향을 미치는 개인적 특성을 연구하여 개인 윤리 성향에 근거한 윤리적 대응 추론을 수행한다. 여기서, 챗봇 대응 제어장치(100)는 사용자 특성과 상황 문맥의 상호작용을 고려하여 적합한 대응 방식을 추론하였다. 이를 위해 챗봇 대응 제어장치(100)는 인공신경망 모델을 활용하여 개인 윤리 성향 특성(Personality) 판단을 위한 대화 에이전트 초기화 문항 구축 및 판단 모듈을 개발하였다. 이하, 챗봇 대응 제어장치(100)에 포함된 구성요소 각각에 대해 기재하도록 한다.

[0036] 입력부(310)는 사용자 또는 외부 장치로부터 입력 데이터를 수신한다. 여기서, 입력 데이터는 대화 데이터, 사용자 특성정보 등을 포함할 수 있다.

[0037] 특성정보 추출부(320)는 입력 데이터에서 사용자 특성정보와 대화 문맥 정보를 추출한다. 여기서, 사용자 특성정보는 인구 통계, 성격 특성, 심리 특성, 심리 특성, 인공지능 관련 특성, 기 설정된 시나리오 특성 등에 대한 변수를 포함할 수 있다.

[0038] 상황 판단부(330)는 사용자 특성정보를 기반으로 제1 상황 판단을 수행한다.

[0039] 이후, 상황 판단부(330)는 대화 문맥 정보를 기반으로 제2 상황 판단을 수행한다.

[0040] 상황 판단부(330)는 제1 상황 판단 결과와 제2 상황 판단 결과를 혼합하여 최종 상황 판단 결과를 도출한다. 상황 판단부(330)는 사용자 특성정보에 기 설정된 우선순위에 대한 제1 가중치와 문맥 정보에 기 설정된 우선순위에 대한 제2 가중치를 고려하여 최종 상황 판단 결과를 도출할 수 있다.

[0041] 상황 판단부(330)는 제1 상황 판단 결과에 제1 가중치를 부여하고, 제2 상황 판단 결과에 제2 가중치를 부여한 후 제1 상황 판단 결과와 제2 상황 판단 결과를 합산하여 판단 결과값을 산출하고, 판단 결과값을 기 설정된 판단 등급 기준과 비교하여 최종 상황 판단 결과를 도출할 수도 있다.

[0042] 상황 판단부(330)는 최종 상황 판단 결과를 기반으로 비윤리 대화 상황 여부를 판단한다.

[0043] 대응 제어부(340)는 대화 상황 판단결과를 기반으로 비윤리 상황에 대한 대응 방안을 결정하고, 결정된 대응 방안에 대한 대응 데이터가 출력되도록 제어한다. 여기서, 대응 방안은 화제전환, 동조, 제지, 침묵 등과 같은 방안을 포함할 수 있다.

[0044] 도 4는 본 발명의 실시예에 따른 챗봇 제어 방법을 설명하기 위한 예시도이다.

[0045] 챗봇 대응 제어장치(100)는 사용자 또는 외부 장치로부터 입력 데이터를 수신한다(S410). 여기서, 입력 데이터는 대화 데이터, 사용자 특성정보 등을 포함할 수 있다.

[0046] 챗봇 대응 제어장치(100)는 입력 데이터에서 사용자 특성정보와 대화 문맥 정보를 추출한다(S420). 여기서, 사용자 특성정보는 인구 통계, 성격 특성, 심리 특성, 심리 특성, 인공지능 관련 특성, 기 설정된 시나리오 특성 등에 대한 변수를 포함할 수 있다.

[0047] 챗봇 대응 제어장치(100)는 사용자 특성정보 및 대화 문맥 정보로 최종 상황 판단 결과를 도출하고, 최종 상황

판단 결과를 기반으로 비윤리 대화 상황 여부를 판단한다(S430).

[0048] 단계 S430의 판단 결과, 비윤리적 상황인 것으로 판단된 경우 챗봇 대응 제어장치(100)는 비윤리 상황에 대한 대응 방안을 결정하고, 결정된 대응 방안에 대한 대응 데이터가 출력되도록 제어한다(S450).

[0049] 한편, 단계 S430의 판단 결과, 비윤리적 상황이 아닌 것으로 판단된 경우 챗봇 대응 제어장치(100)는 정상 대화 프로세스를 수행하여 정상 대화 데이터가 출력되도록 제어한다(S442).

[0050] 도 4에서는 각 단계를 순차적으로 실행하는 것으로 기재하고 있으나, 반드시 이에 한정되는 것은 아니다. 다시 말해, 도 4에 기재된 단계를 변경하여 실행하거나 하나 이상의 단계를 병렬적으로 실행하는 것으로 적용 가능할 것이므로, 도 4는 시계열적인 순서로 한정되는 것은 아니다.

[0051] 도 4에 기재된 본 실시예에 따른 챗봇 제어 방법은 애플리케이션(또는 프로그램)으로 구현되고 단말장치(또는 컴퓨터)로 읽을 수 있는 기록매체에 기록될 수 있다. 본 실시예에 따른 챗봇 제어 방법을 구현하기 위한 애플리케이션(또는 프로그램)이 기록되고 단말장치(또는 컴퓨터)가 읽을 수 있는 기록매체는 컴퓨팅 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치 또는 매체를 포함한다.

[0052] 도 6은 본 발명의 실시예에 따른 챗봇 제어 프로그램의 프로세스를 나타낸 예시도이다.

[0053] 챗봇 대응 제어장치(100)는 챗봇 대화에서 도 6과 같은 프로세스로 사용자 특성 정보와 비윤리 상황 정보를 학습한다.

[0054] 챗봇 대응 제어장치(100)는 학습 결과에 따라 입력 파일에 따른 상황을 판단하고, 현재 상황에 정답인 것으로 판단되는 적절한 대응방식(화제전환, 동조, 제지, 침묵 등)을 도출하여 출력한다.

[0055] 각각의 프로세스 단계에 대한 파일 및 그에 대한 동작을 [표 1]과 같이 정의될 수 있다.

표 1

[0056]

파일명	설명
argparse_test.py	사용자에 대한 정보와 비윤리 상황 정보를 학습된 모델의 입력 값으로 삽입하여 적절한 대응방식을 csv 파일로 출력하는 파일
ethical_reaction_model.py	학습된 모델을 불러오는 파일
get_score.py	공인인증평가를 위한 정확도를 도출하는 파일
ethical_reaction/ethical_reaction_model.h5	학습된 모델 파일
ethical_reaction/scaler.save	학습된 모델 파일
test_x_ReactionModel.csv	모델의 입력 값으로 활용될 사용자에 대한 정보 및 비윤리 상황 정보가 담긴 csv 파일
test_onehot_label_y_ReactionModel.csv	사용자 별 적절한 대응방식에 대한 정답지(Label)
ethical_reaction/ethical_reaction_result.csv	모델 출력 값으로 도출된 사용자 별 적절한 대응방식(argparse_test.py 파일 실행으로 도출)

[0057] 이상의 설명은 본 발명의 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명의 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 발명의 실시예들은 본 발명의 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

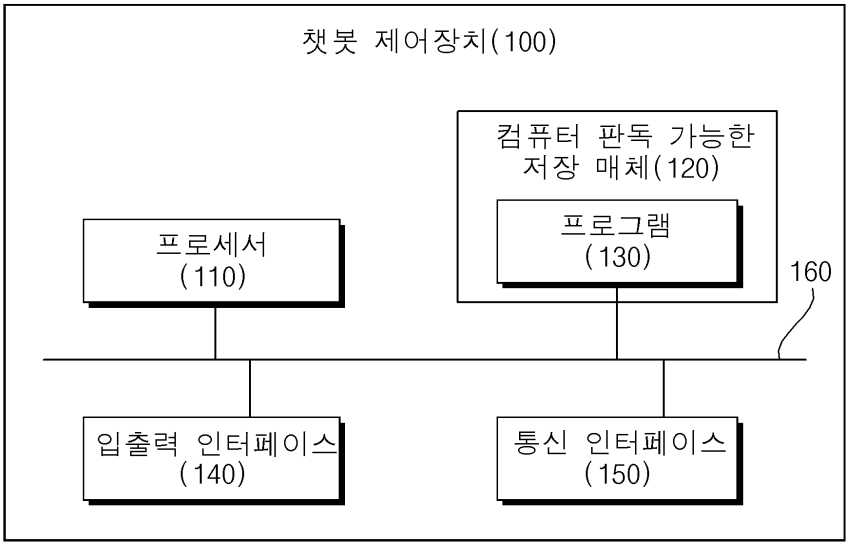
[0058] 100: 챗봇 대응 제어장치

310: 입력부 320: 특성정보 추출부

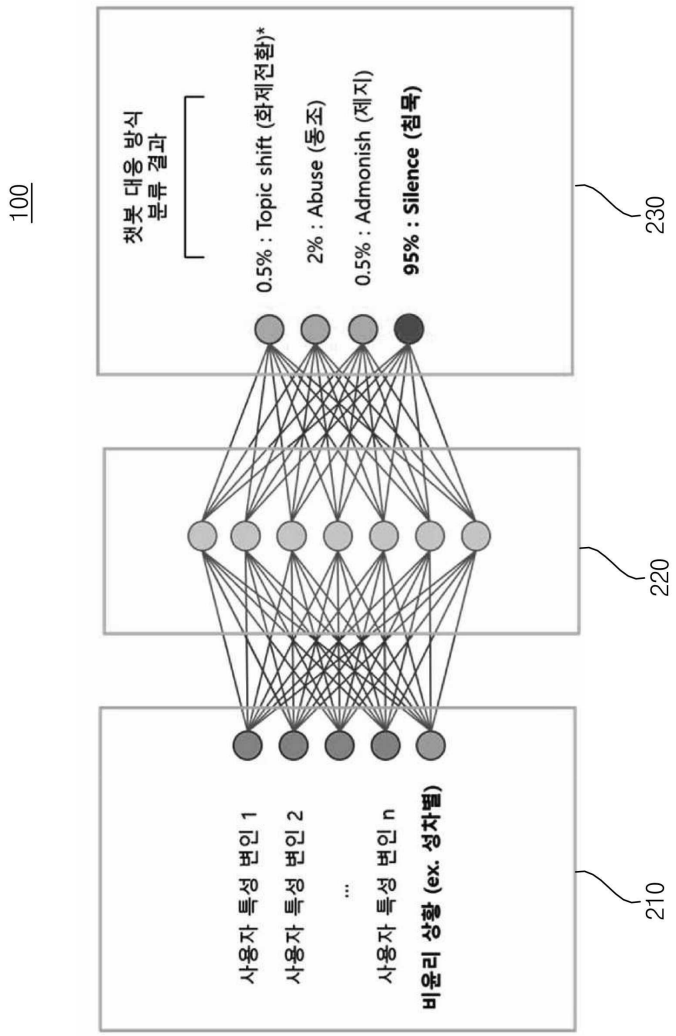
330: 상황 판단부 340: 대응 제어부

도면

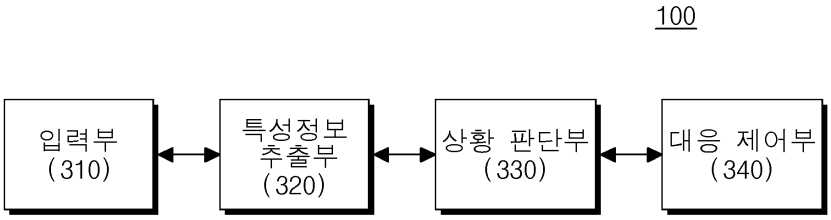
도면1



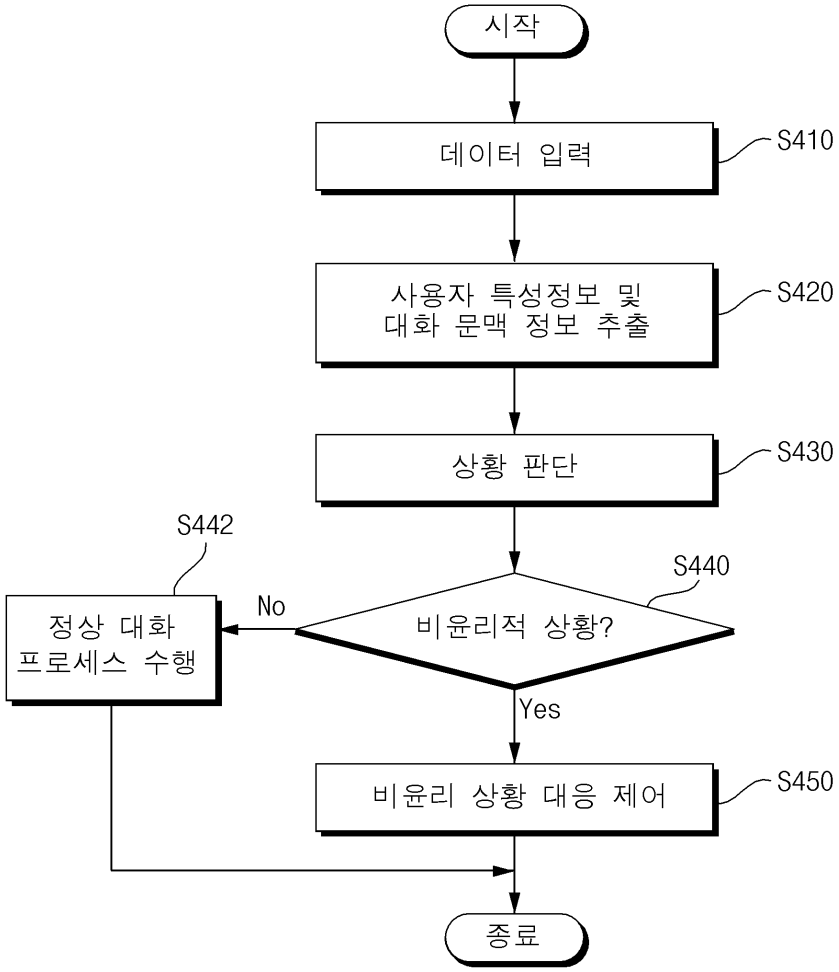
도면2


100
챗봇 대응 방식
분류 결과
0.5% : Topic shift (화제전환)*
2% : Abuse (동조)
0.5% : Admonish (제지)
95% : Silence (침묵)
230
220
사용자 특성 변인 1
사용자 특성 변인 2
...
사용자 특성 변인 n
비윤리 상황 (ex. 성차별)
210

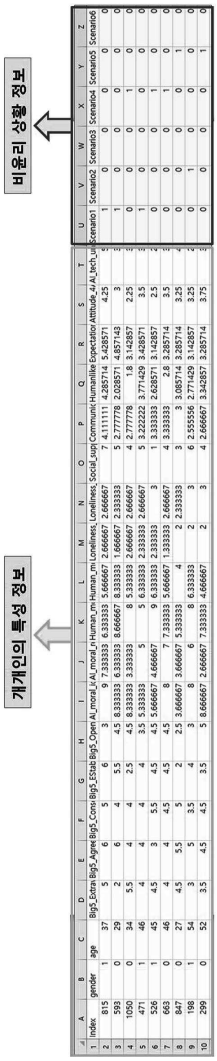
도면3


100
입력부
(310)
특성정보
추출부
(320)
상황 판단부
(330)
대응 제어부
(340)

도면4

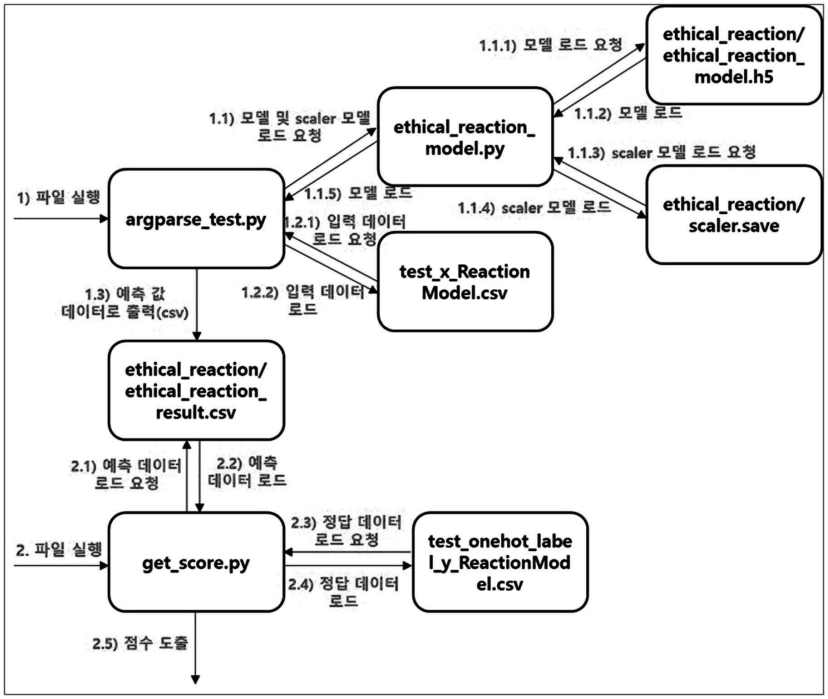

시작
데이터 입력
S410
사용자 특성정보 및
대화 문맥 정보 추출
S420
상황 판단
S430
S442
S440
정상 대화
프로세스 수행
No
비윤리적 상황?
Yes
비윤리 상황 대응 제어
S450
종료

도면5



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Index	gender	age	Big5_Extrai	Big5_Agree	Big5_Consi	Big5_Estab	Big5_Open	AI_moral	AI_moral_d	Human_mo	Human_mo	Loneliness	Loneliness_	Social_sup	Communic	Humanlike	Expectatio	Attitude_d	AI_tech_us	Scenario1	Scenario2	Scenario3	Scenario4	Scenario5	Scenario6
2	815	1	37	5	6	5	6	3	9	7.333333	6.333333	5.666667	2.666667	2.666667	7	4.111111	4.285714	5.428571	4.25	5	1	0	0	0	0	0
3	593	0	29	2	6	4	5.5	4.5	8.333333	6.333333	8.666667	8.333333	1.666667	2.333333	5	2.777778	2.028571	4.857143	3	5	1	0	0	0	0	0
4	1050	0	34	5.5	4	4	2.5	4.5	8.333333	3.333333	8	5.333333	2.666667	2.666667	4	2.777778	1.8	3.142857	2.25	3	0	0	0	1	0	0
5	471	1	46	4	4	4	4	3.5	5.333333	5	5	6.333333	2.333333	2.666667	5	3.222222	3.771429	3.428571	3.5	3	1	0	0	0	0	0
6	526	1	45	4.5	3	5.5	4.5	4.5	5.666667	4.666667	9	6.333333	2.333333	3	1	3.333333	2.628571	3.142857	2.5	2	0	0	0	1	0	0
7	663	0	46	4	4	4.5	4.5	4.5	8	7.333333	5.666667	1.333333	2.666667	4	4	3.333333	2.8	3.285714	3.5	2	0	0	0	1	0	0
8	847	0	27	4.5	5.5	5	2	3	3.666667	3.666667	7.333333	5.333333	4	2	2.333333	3	3.085714	2.885714	3.25	2	0	0	0	0	1	0
9	198	1	54	3	5	3.5	4	3	6	8	8	6.333333	2	2.333333	3	6	2.555556	2.771429	3.142857	3.25	2	0	1	0	0	1
10	299	0	52	3.5	4.5	4.5	3.5	5	8.666667	2.666667	7.333333	4.666667	2	3	4	2.666667	3.342857	3.285714	3.75	3	0	0	0	0	0	0

도면6



도면7a

개인특성	변수명	변인	설명
-	Index	Primary Key	사용자에 대한 개인 식별 번호. 모델 학습에는 활용되지 않음
인구통계	gender, age	(1) 연령, (2) 성별	
성격특성	Big5_Extraversion Big5_Agreeableness Big5_Conscientiousness Big5_EStability Big5_Openness	Personality	Big5 (TIPI)
심리특성	Human_moral_idealism	인간의 윤리성향 (절대주의)	• 인간은 고의적으로 다른 사람에게 조금이라도 해를 입히는 행동을 절대로 하지 말아야 한다. • 인간은 심리적으로든 물질적으로든 다른 사람에게 피해를 주면 절대 안 된다.
	Human_moral_relativism	인간의 윤리성향 (상대주의)	• 무엇이 윤리적인 것인지는 상황이나 사회에 따라 다르다. • 도덕적 기준은 개인적인 문제로 봐야 한다. 즉, 어떤 사람이 도덕적이라고 생각하는 것을 다른 사람은 비도덕적이라고 판단할 수도 있는 것이다.
	Loneliness_emotional Loneliness_social	외로움 (Loneliness)	• 귀하는 평소에 얼마나 외롭다고 느끼십니까? • 귀하는 가깝다고 느끼는 사람이 얼마나 있습니까? ®
	Social_support	사회적 지지 (Social support)	• 나의 기쁨과 슬픔을 함께할 수 있는 특별한 사람이 있다.
	Communicaion_competence	커뮤니케이션 유능감 (Perceived Social Support)	• 귀하는 평소에 사람들과 소통하는 데 어려움을 느끼십니까?

도면7b

개인특성	변수명	변인	설명
심리특성	AI_moral_idealism	기술에 대한 윤리성향 (절대주의)	• 인공지능은 사람에게 조금이라도 해를 입혀서는 안 된다. • 인공지능은 심리적으로든 물질적으로든 사람에게 피해를 주면 절대 안 된다. • 인공지능의 사용목적과 종류 그리고 기술개발의 허용범위를 정해놓고 모두가 이를 따르도록 해야 한다.
	AI_moral_relativism	기술에 대한 윤리성향 (상대주의)	• 인공지능의 윤리적인 개발과 사용에 대한 문제는 단정짓기 어렵다. 왜냐하면 무엇이 윤리적인지 비윤리적인지에 대한 기준은 개인마다 다르기 때문이다. • 인공지능을 사용할 때 윤리적 문제가 발생한다면 이에 대한 윤리적 가치 판단은 개개인마다 다르기 때문에, 한 사람에게 비윤리적으로 받아들여지는 것이 다른 사람에게는 윤리적으로 받아들여질 수도 있다고 생각한다. • 모든 인공지능에 적용할 수 있는 윤리적 규칙은 만들어 질 수 없다. 왜냐하면 여러 가지 상황에 따라서 규칙 또한 다르게 적용될 수 있기 때문이다.
인공지능 기술관련	Humanlikeness	Humanlikenss	• 귀하께서는 인공지능 기기가 사람과 얼마나 비슷한 존재라고 생각하십니까
	Expectation4AI_CASA	CASA vs. CAM	• AI에 대한 기대를 측정. 자체제작 7문항의 축약
	Attitude_4AI_general	일반적인 태도	• 일반적 태도를 측정. 자체제작 4문항의 축약
	AI_tech_understand_1	AI기술에 대한 이해도	• 나는 현재 인공지능 기술 수준에 대해 잘 알고 있다.
-	Scenario1	시나리오 상황 1	• 말기암 환자의 자살 기도
	Scenario2	시나리오 상황 2	• 가짜뉴스 유포 - 차별금지법
	Scenario3	시나리오 상황 3	• 가짜뉴스 유포 - 테러방지법
	Scenario4	시나리오 상황 4	• 운전 중 욕설
	Scenario5	시나리오 상황 5	• 전공 선택에 있어 성차별 - 여성
	Scenario6	시나리오 상황 6	• 전공 선택에 있어 성차별 - 남성