(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0089615
(43) 공개일자 2024년06월20일

(51) 국제특허분류(Int. Cl.)
G06N 3/096 (2023.01) *G06F 40/20* (2020.01)
G06F 40/30 (2020.01) *G06N 3/045* (2023.01)
(52) CPC특허분류
G06N 3/096 (2023.01)
G06F 40/20 (2022.01)
(21) 출원번호 10-2024-7015690
(22) 출원일자(국제) 2022년08월17일
심사청구일자 없음
(85) 번역문제출일자 2024년05월10일
(86) 국제출원번호 PCT/US2022/040530
(87) 국제공개번호 WO 2023/064033
국제공개일자 2023년04월20일
(30) 우선권주장
63/254,740 2021년10월12일 미국(US)
17/735,651 2022년05월03일 미국(US)

(71) 출원인
오라클 인터내셔날 코포레이션
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500
(72) 발명자
부, 탄 티엔
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500
팜, 투옌 팡
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500
(뒷면에 계속)
(74) 대리인
특허법인에이아이피

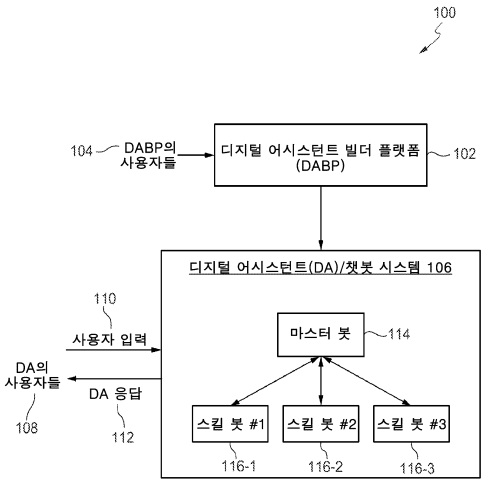
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **사전-트레이닝된 언어 모델의 단일 트랜스포머 계층으로부터의 다중-헤드 네트워크의 미세-튜닝**

(57) 요 약

다수의 계층들을 포함하며 오디오 또는 텍스트 언어 입력을 프로세싱하도록 구성된 기계-학습 모델의 사전-트레이닝된 버전을 맞춤화하거나 또는 미세-튜닝하기 위한 기술들이 제공된다. 다수의 계층들 각각은 복수의 파라미터들에 대응하는 복수의 계층-특정 사전-트레이닝된 파라미터 값들로 구성되고, 다수의 계층들 각각은 다중-헤드 어텐션 기술을 구현하도록 구성된다. 대응하는 계층-특정 사전-트레이닝된 파라미터 값들이 클라이언트 데이터 세트를 사용하여 미세-튜닝될 다수의 계층들의 불완전한 서브세트가 식별된다. 기계-학습 모델은 기계-학습 모델의 업데이트된 버전을 생성하기 위해 클라이언트 데이터 세트를 사용하여 미세-튜닝되며, 여기서 불완전한 서브세트에 포함되지 않은 다수의 계층들 중 하나 이상의 계층의 각각의 계층에 대해 구성된 계층-특정 사전-트레이닝된 파라미터 값들은 미세-튜닝 동안 프리즈된다. 기계-학습 모델의 업데이트된 버전의 사용이 가능하게 된다.

대 표 도 - 도1



(52) CPC특허분류
G06F 40/30 (2020.01)
G06N 3/045 (2023.01)
(72) 발명자
네자미, 오미드 모하마드
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500
존슨, 마크 에드워드
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500
동, 탄 룽
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500
호앙, 콩 주이 부
미국, 캘리포니아 94065, 레드우드 쇼어스 엠에스 5오피7, 오라클 파크웨이 500

명 세 서

청구범위

청구항 1

방법으로서,

다수의 계층들을 포함하며 오디오 또는 텍스트 언어 입력을 프로세싱하도록 구성된 기계-학습 모델의 사전-트레이닝된 버전에 액세스하는 단계로서, 상기 다수의 계층들 각각은 복수의 파라미터들에 대응하는 복수의 계층-특정 사전-트레이닝된 파라미터 값들로 구성되고, 상기 다수의 계층들 각각은 다중-헤드 어텐션(attention) 기술을 구현하도록 구성되는, 단계;

클라이언트 데이터 세트에 액세스하는 단계;

대응하는 계층-특정 사전-트레이닝된 파라미터 값들이 상기 클라이언트 데이터 세트를 사용하여 미세-튜닝될 상기 다수의 계층들의 불완전한 서브세트를 식별하는 단계;

상기 기계-학습 모델의 업데이트된 버전을 생성하기 위해 상기 클라이언트 데이터 세트를 사용하여 상기 기계-학습 모델을 미세-튜닝하는 단계로서, 상기 불완전한 서브세트에 포함되지 않은 상기 다수의 계층들 중 하나 이상의 계층의 각각의 계층에 대해 구성된 상기 계층-특정 사전-트레이닝된 파라미터 값들은 상기 미세-튜닝 동안 프리즈(freeze)되는, 단계; 및

상기 기계-학습 모델의 상기 업데이트된 버전의 사용을 가능하게 하는 단계를 포함하는, 방법.

청구항 2

청구항 1에 있어서,

상기 다중-헤드 어텐션 기술은 셀프-어텐션 기술을 포함하는, 방법.

청구항 3

청구항 1에 있어서,

상기 다수의 계층들의 상기 불완전한 서브세트는 2개의 계층들 또는 하나의 계층을 포함하며, 상기 사전-트레이닝된 기계-학습 모델은 적어도 5개의 계층들을 포함하는, 방법.

청구항 4

청구항 1에 있어서,

상기 복수의 파라미터들은 적어도 500,000개의 파라미터들을 포함하며, 상기 복수의 파라미터 값들을 업데이트하는 단계는 하나 이상의 중앙 프로세싱 유닛들을 사용하여 수행되는, 방법.

청구항 5

청구항 1에 있어서,

상기 파라미터 값들을 업데이트하는 단계는,

상기 불완전한 서브세트에 포함되지 않은 상기 다수의 계층들 중 상기 하나 이상의 계층에 대응하는 캐싱된 값들의 세트에 액세스하는 단계를 포함하는, 방법.

청구항 6

청구항 1에 있어서,

상기 다수의 계층들의 상기 불완전한 서브세트에 대해 구성된 상기 복수의 파라미터 값들의 수량은 상기 기계-학습 모델에 대해 구성된 파라미터들의 수량과 비교하여 30% 이하인, 방법.

청구항 7

청구항 1에 있어서,

상기 업데이트된 버전의 사용을 가능하게 하는 단계는 언어 입력을 명령(command) 또는 쿼리로 번역하는 단계를 포함하는, 방법.

청구항 8

청구항 1에 있어서,

상기 다수의 계층들은 적어도 4개의 계층들을 포함하며, 상기 다수의 계층들의 상기 불완전한 서브세트의 각각은 상기 다수의 계층들에서 계층 2 이상에 위치되는, 방법.

청구항 9

시스템으로서,

하나 이상의 프로세서들; 및

상기 하나 이상의 프로세서들에 결합되는 메모리로서, 상기 메모리는 상기 하나 이상의 프로세서들에 의해 실행 가능한 복수의 명령어들을 저장하며, 상기 복수의 명령어들은 상기 하나 이상의 프로세서들에 의해 실행될 때 상기 하나 이상의 프로세서들이 액션들의 세트를 수행하게끔 하며, 상기 액션들의 세트는,

다수의 계층들을 포함하며 오디오 또는 텍스트 언어 입력을 프로세싱하도록 구성된 기계-학습 모델의 사전-트레이닝된 버전에 액세스하는 액션으로서, 상기 다수의 계층들 각각은 복수의 파라미터들에 대응하는 복수의 계층-특정 사전-트레이닝된 파라미터 값들로 구성되고, 상기 다수의 계층들 각각은 다중-헤드 어텐션(attention) 기술을 구현하도록 구성되는, 액션;

클라이언트 데이터 세트에 액세스하는 액션;

대응하는 계층-특정 사전-트레이닝된 파라미터 값들이 상기 클라이언트 데이터 세트를 사용하여 미세-튜닝될 상기 다수의 계층들의 불완전한 서브세트를 식별하는 액션;

상기 기계-학습 모델의 업데이트된 버전을 생성하기 위해 상기 클라이언트 데이터 세트를 사용하여 상기 기계-학습 모델을 미세-튜닝하는 액션으로서, 상기 불완전한 서브세트에 포함되지 않은 상기 다수의 계층들 중 하나 이상의 계층의 각각의 계층에 대해 구성된 상기 계층-특정 사전-트레이닝된 파라미터 값들은 상기 미세-튜닝 동안 프리즈(freeze)되는, 액션; 및

상기 기계-학습 모델의 상기 업데이트된 버전의 사용을 가능하게 하는 액션을 포함하는, 시스템.

청구항 10

청구항 9에 있어서,

상기 다중-헤드 어텐션 기술은 셀프-어텐션 기술을 포함하는, 시스템.

청구항 11

청구항 9에 있어서,

상기 다수의 계층들의 상기 불완전한 서브세트는 2개의 계층들 또는 하나의 계층을 포함하며, 상기 사전-트레이닝된 기계-학습 모델은 적어도 5개의 계층들을 포함하는, 시스템.

청구항 12

청구항 9에 있어서,

상기 복수의 파라미터들은 적어도 500,000개의 파라미터들을 포함하며, 상기 복수의 파라미터 값들을 업데이트하는 액션은 하나 이상의 중앙 프로세싱 유닛들을 사용하여 수행되는, 시스템.

청구항 13

청구항 9에 있어서,

상기 파라미터 값들을 업데이트하는 액션은,

상기 불완전한 서브세트에 포함되지 않은 상기 다수의 계층들 중 상기 하나 이상의 계층에 대응하는 캐싱된 값들의 세트에 액세스하는 액션을 포함하는, 시스템.

청구항 14

청구항 9에 있어서,

상기 다수의 계층들의 상기 불완전한 서브세트에 대해 구성된 상기 복수의 파라미터 값들의 수량은 상기 기계-학습 모델에 대해 구성된 파라미터들의 수량과 비교하여 30% 이하인, 시스템.

청구항 15

청구항 9에 있어서,

상기 업데이트된 버전의 사용을 가능하게 하는 액션은 언어 입력을 명령 또는 쿼리로 번역하는 액션을 포함하는, 시스템.

청구항 16

청구항 9에 있어서,

상기 다수의 계층들은 적어도 4개의 계층들을 포함하며, 상기 다수의 계층들의 상기 불완전한 서브세트의 각각은 상기 다수의 계층들에서 계층 2 이상에 위치되는, 시스템.

청구항 17

하나 이상의 프로세서들에 의해 실행가능한 복수의 명령어들을 저장하는 비-일시적 컴퓨터-판독가능 메모리로서, 상기 복수의 명령어들은 상기 하나 이상의 프로세서들에 의해 실행될 때 상기 하나 이상의 프로세서들이 액션들의 세트를 수행하게끔 하며, 상기 액션들의 세트는,

다수의 계층들을 포함하며 오디오 또는 텍스트 언어 입력을 프로세싱하도록 구성된 기계-학습 모델의 사전-트레이닝된 버전에 액세스하는 액션으로서, 상기 다수의 계층들 각각은 복수의 파라미터들에 대응하는 복수의 계층-특정 사전-트레이닝된 파라미터 값들로 구성되고, 상기 다수의 계층들 각각은 다중-헤드 어텐션(attention) 기술을 구현하도록 구성되는, 액션;

클라이언트 데이터 세트에 액세스하는 액션;

대응하는 계층-특정 사전-트레이닝된 파라미터 값들이 상기 클라이언트 데이터 세트를 사용하여 미세-튜닝될 상기 다수의 계층들의 불완전한 서브세트를 식별하는 액션;

상기 기계-학습 모델의 업데이트된 버전을 생성하기 위해 상기 클라이언트 데이터 세트를 사용하여 상기 기계-학습 모델을 미세-튜닝하는 액션으로서, 상기 불완전한 서브세트에 포함되지 않은 상기 다수의 계층들 중 하나 이상의 계층의 각각의 계층에 대해 구성된 상기 계층-특정 사전-트레이닝된 파라미터 값들은 상기 미세-튜닝 동안 프리즈(freeze)되는, 액션; 및

상기 기계-학습 모델의 상기 업데이트된 버전의 사용을 가능하게 하는 액션을 포함하는, 비-일시적 컴퓨터-판독가능 메모리.

청구항 18

청구항 17에 있어서,

상기 다중-헤드 어텐션 기술은 셀프-어텐션 기술을 포함하는, 비-일시적 컴퓨터-판독가능 메모리.

청구항 19

청구항 17에 있어서,

상기 다수의 계층들의 상기 불완전한 서브세트는 2개의 계층들 또는 하나의 계층을 포함하며, 상기 사전-트레이닝된 기계-학습 모델은 적어도 5개의 계층들을 포함하는, 비-일시적 컴퓨터-판독가능 메모리.

청구항 20

청구항 17에 있어서,

상기 복수의 파라미터들은 적어도 500,000개의 파라미터들을 포함하며, 상기 복수의 파라미터 값들을 업데이트하는 액션은 하나 이상의 중앙 프로세싱 유닛들을 사용하여 수행되는, 비-일시적 컴퓨터-판독가능 메모리.

발명의 설명

기 술 분 야

[0001] 관련 출원에 대한 상호 참조

[0002] 본 출원은 2021년 10월 12일자로 출원된 미국 가특허 출원 제63/254,740호에 대한 이익 및 우선권을 주장하는 2022년 05월 03일자로 출원된 미국 정규 출원 제17/735,651호에 대한 우선권을 주장한다. 이로써 이러한 출원들의 각각은 모든 목적들을 위해 그 전체가 참조로서 통합된다.

[0003] 기술분야

[0004] 본 개시내용은 전반적으로 사전-트레이닝된 언어 모델의 선택 계층들을 선택적으로 미세-튜닝하는 것에 관한 것이다. 보다 구체적으로, 선택적 미세-튜닝은, 트레이닝 동안 사전-트레이닝된 언어 모델의 하나 이상의 다른 계층들 내의 파라미터들을 프리즈(freeze)하면서 사전-트레이닝된 언어 모델의 세트의 불완전한 서브세트 내의 파라미터들이 업데이트되도록 업데이트하는 단계를 포함하는 트레이닝을 수행하는 단계를 포함한다.

배 경 기 술

[0005] 전세계의 다수의 사용자들은 즉각적인 반응을 얻기 위해 인스턴트 메시징 또는 채팅 플랫폼을 사용하고 있다. 조직들은 종종 이러한 인스턴트 메시징 또는 채팅 플랫폼들을 사용하여 고객들(또는 최종 사용자들)과의 실시간 대화에 참여한다. 그러나, 조직들이 고객들 또는 최종 사용자들과 실시간 커뮤니케이션에 참여하여 위해 서비스 인력을 고용하는 것은 매우 많은 비용이 들어갈 수 있다. 특히 인터넷을 통해 최종 사용자들과의 대화들을 시뮬레이션하기 위해 챗봇(Chatbot)들 또는 봇(bot)들이 개발되기 시작했다. 최종 사용자들은, 최종 사용자들이 이미 설치하여 사용하고 있는 메시징 앱들을 통해 봇들과 소통할 수 있다. 일반적으로 인공 지능(artificial intelligence; AI)을 기반으로 하는 지능형 봇은 실시간 대화에서 지능적이고 컨텍스트적으로 소통할 수 있으며, 따라서 개선된 대화 경험을 위해 봇과 최종 사용자들 사이의 더 자연스러운 대화를 가능하게 할 수 있다. 최종 사용자가 봇이 어떻게 응답해야 하는지를 아는 키워드들 또는 명령들의 고정된 세트를 학습하는 대신에, 지능형 봇은 자연어로 된 사용자의 발화(utterance)들에 기초하여 최종 사용자의 의도를 이해하고 그에 따라 응답할 수 있다.

[0006] 그러나, 이러한 자동화된 해법들이 특정 분야에 대한 특정 지식 및 전문 개발자들의 능력들 내에만 있을 수 있는 특정 기술들의 적용을 필요로 하기 때문에, 챗봇들을 구축하는 것은 어렵다. 이러한 챗봇들을 구축하는 것의 부분으로서, 개발자는 먼저 기업들과 최종 사용자들의 요구들을 이해할 수 있다. 그런 다음, 개발자는, 예를 들어, 분석을 위해 사용될 데이터 세트들을 선택하는 것, 분석을 위한 입력 데이터 세트들을 준비하는 것(예를 들어, 데이터를 클렌징(cleanse)하는 것, 분석 이전에 데이터를 추출하거나, 포맷팅하거나, 및/또는 변환하는 것, 데이터 특징 엔지니어링을 수행하는 것, 등), 분석을 수행하기 위한 적절한 기계 학습(machine learning; ML) 기술(들) 또는 모델(들)을 식별하는 것, 및 피드백에 기초하여 결과(result)들/성과(outcome)들을 개선하기 위해 기술 또는 모델을 개선하는 것과 관련된 결정들을 분석하고 결정할 수 있다. 적절한 모델을 식별하는 태스크(task)는, 다수의 모델들을 개발하는 것, 사용을 위한 특정 모델(또는 모델들)을 식별하기 이전에 아마도 병렬로, 이러한 모델들로 반복적으로 테스트하고 실험하는 것을 포함할 수 있다. 또한, 지도 학습-기반 해법들은 전형적으로, 그 뒤에 적용(즉, 추론) 단계가 이어지는 트레이닝 단계, 및 트레이닝 단계와 적용 단계 사이의 반복 루프들을 수반한다. 개발자는 최적 해법들을 달성하기 위해 이러한 단계들을 주의 깊게 구현하고 모니터링하는 것을 담당할 수 있다.

[0007] 따라서, 이에 의해 특정 사용 케이스에 대한 모델이 더 효율적으로 개발될 수 있는 기술을 식별하는 것이 유리할 것이다.

발명의 내용

[0008] 본 명세서에서 개시되는 기술들은 전반적으로 챗봇들에 관한 것이다. 더 구체적으로 그리고 비제한적으로, 본 명세서에서 개시되는 기술들은 사전-트레이닝된 모델에 액세스하고 모델에서 다수의 계층들의 불완전한 서브세트 내의 파라미터들을 미세-튜닝함으로써 챗봇들을 트레이닝시키는 것에 관한 것이다.

[0009] 다양한 실시예들에서, 컴퓨터-구현형 방법이 제공되며, 컴퓨터 구현형 방법은: 다수의 계층들을 포함하며 오디오 또는 텍스트 언어 입력을 프로세싱하도록 구성된 기계-학습 모델의 사전-트레이닝된 버전에 액세스하는 단계로서, 다수의 계층들 각각은 복수의 파라미터들에 대응하는 복수의 계층-특정 사전-트레이닝된 파라미터 값들로 구성되고, 다수의 계층들 각각은 다중-헤드 어텐션(attention) 기술을 구현하도록 구성되는, 단계; 클라이언트 데이터 세트에 액세스하는 단계; 대응하는 계층-특정 사전-트레이닝된 파라미터 값들이 클라이언트 데이터 세트를 사용하여 미세-튜닝될 다수의 계층들의 불완전한 서브세트를 식별하는 단계; 기계-학습 모델의 업데이트된 버전을 생성하기 위해 클라이언트 데이터 세트를 사용하여 기계-학습 모델을 미세-튜닝하는 단계로서, 불완전한 서브세트에 포함되지 않은 다수의 계층들 중 하나 이상의 계층의 각각의 계층에 대해 구성된 계층-특정 사전-트레이닝된 파라미터 값들은 미세-튜닝 동안 프리즈(freeze)되는, 단계; 및 기계-학습 모델의 업데이트된 버전의 사용을 가능하게 하는 단계를 포함한다.

[0010] 일부 실시예들에서, 사전-트레이닝된 기계-학습 모델은 셀프-어텐션 모델을 포함하며, 여기서 다수의 계층들의 불완전한 세트는 셀프-어텐션 모델 내의 적어도 하나의 신경망을 포함한다.

[0011] 일부 실시예들에서, 다수의 계층들의 불완전한 서브세트는 2개의 계층들 또는 하나의 계층을 포함하며, 여기서 사전-트레이닝된 기계-학습 모델은 적어도 5개의 계층들을 포함한다.

[0012] 일부 실시예들에서, 복수의 파라미터들은 적어도 500,000개의 파라미터들을 포함하며, 여기서 복수의 파라미터 값들을 업데이트하는 단계는 하나 이상의 중앙 프로세싱 유닛들을 사용하여 수행된다.

[0013] 일부 실시예들에서, 파라미터 값들을 업데이트하는 단계는: 불완전한 서브세트에 포함되지 않은 다수의 계층들

중 하나 이상의 계층에 대응하는 캐싱된 값들의 세트에 액세스하는 단계를 포함한다.

[0014] 일부 실시예들에서, 다수의 계층들의 불완전한 서브세트에 대해 구성된 복수의 파라미터 값들의 수량은 기계-학습 모델에 대해 구성된 파라미터들의 수량과 비교하여 30% 이하이다.

[0015] 일부 실시예들에서, 업데이트된 버전의 사용을 가능하게 하는(facilitate) 단계는 언어 입력을 명령 또는 쿼리로 번역하는 단계를 포함한다.

[0016] 일부 실시예들에서, 다수의 계층들은 적어도 4개의 계층들을 포함하며, 여기서 다수의 계층들의 불완전한 서브세트의 각각은 다수의 계층들에서 계층 2 이상에 위치된다.

[0017] 다양한 실시예들에서, 하나 이상의 데이터 프로세서들 및 하나 이상의 데이터 프로세서들 상에서 실행될 때 하나 이상의 데이터 프로세서들이 본 명세서에 개시된 하나 이상의 방법들 중 부분 또는 전부를 수행하게끔 하는 명령어들을 포함하는 비-일시적 컴퓨터 판독가능 저장 매체를 포함하는 시스템이 제공된다.

[0018] 다양한 실시예들에서, 비-일시적 기계-판독가능 저장 매체에 유형적으로 구현되며, 하나 이상의 데이터 프로세서들이 본 명세서에서 개시된 하나 이상의 방법들의 부분 또는 전부를 수행하게끔 하는 명령어들을 포함하는 컴퓨터-프로그램 제품이 제공된다.

[0019] 이상에서 그리고 이하에서 설명된 기술들은 다수의 방식들로 그리고 다수의 컨텍스트들에서 구현될 수 있다. 몇몇 예시적인 구현예들 및 컨텍스트들은 이하에서 더 상세하게 설명되는 바와 같이 다음의 도면들을 참조하여 제공된다. 그러나, 다음의 구현예들 및 컨텍스트들은 많은 것들 중 일부에 불과하다.

도면의 간단한 설명

[0020] 도 1은 예시적인 실시예를 통합하는 분산형 환경의 간략화된 블록도이다.

도 2는 특정 실시예들에 따른 마스터 봇을 구현하는 컴퓨팅 시스템의 간략화된 블록도이다.

도 3은 특정 실시예들에 따른 스킬 봇(skill bot)을 구현하는 컴퓨팅 시스템의 간략화된 블록도이다.

도 4는 다양한 실시예들에 따른 챗봇 트레이닝 및 배포 시스템의 간략화된 블록도이다.

도 5는 다양한 실시예들에 따른 트레이닝 데이터 세트를 키워드들로 증강하기 위한 프로세스 흐름을 예시한다.

도 6은 다양한 실시예들을 구현하기 위한 분산형 시스템의 간략화된 도면을 도시한다.

도 7은, 다양한 실시예들에 따른, 이에 의해 일 실시예의 시스템의 하나 이상의 구성요소들에 의해 제공되는 서비스들이 클라우드 서비스들로서 제공될 수 있는, 시스템 환경의 하나 이상의 구성요소들의 간략화된 블록도이다.

도 8은 다양한 실시예들을 구현하기 위해 사용될 수 있는 예시적인 컴퓨터 시스템을 예시한다.

도 9a 내지 도 9c는 다양한 실시예들에 따라 미세튜닝될 수 있는 예시적인 모델의 아키텍처들을 예시한다.

도 10은, 모델의 미세-튜닝 트레이닝을 위한 다양한 접근방식들에 대응하는 데이터를 나타내는 표를 표시한다.

도 11은 본 발명의 선택적인 실시예들에 따라 사전-트레이닝되고 미세-튜닝될 수 있는 모델 내의 계층들을 예시한다.

발명을 실시하기 위한 구체적인 내용

[0021] 다음의 설명에서, 설명의 목적들을 위해, 특정 세부사항들이 특정 실시예들의 완전한 이해를 제공하기 위해 기술된다. 그러나 다양한 실시예들은 이러한 특정 세부사항들 없이 실시될 수 있음이 명백할 것이다. 도면들 및 설명은 제한적인 것으로 의도되지 않는다. 단어 "예시적인"은 "예, 예증, 또는 예시로서 기능함"을 의미하도록 본 명세서에서 사용된다. "예시적인" 것으로서 본 명세서에서 설명된 임의의 실시예 또는 설계는 다른 실시예들 또는 설계들에 비해 반드시 바람직하거나 또는 유리한 것으로서 해석되지는 않아야 한다.

[0022] **서론**

[0023] 디지털 어시스턴트(assistant)는, 사용자들이 자연어 대화들에서 다양한 태스크들을 달성하는 것을 돕는 인공지능 구동형 인터페이스이다. 각각의 디지털 어시스턴트에 대해, 고객은 하나 이상의 스킬들을 조립할 수 있다. 스킬들(본 명세서에서 챗봇들, 봇들, 또는 스킬 봇들로도 설명됨)은, 재고 추적, 타임 카드 제출, 및 비용 보고

서 생성과 같은 특정 유형들의 태스크들에 초점이 맞춰진 개별적인 봇들이다. 최종 사용자들이 디지털 어시스턴트에 참여할 때, 디지털 어시스턴트는 최종 사용자 입력을 평가하고 대화를 적절한 챗봇으로 그리고 적절한 챗봇으로부터 라우팅한다. 디지털 어시스턴트는, FACEBOOK® 메신저, SKYPE MOBILE® 메신저, 또는 단문 메시지 서비스(Short Message Service; SMS)와 같은 다양한 채널들을 통해 최종 사용자들에게 이용가능해 질 수 있다. 채널들은 다양한 메시지 플랫폼들 상에서 최종 사용자들로부터 디지털 어시스턴트 및 그 다양한 챗봇들로 채팅을 주고받는다. 채널들은 또한 사용자 에이전트 에스컬레이션(escalation), 이벤트-개시형 대화들, 및 테스트를 지원할 수 있다.

[0024] 의도(intent)들은 챗봇이, 사용자가 챗봇이 무엇을 수행하는 것을 원하는지 이해하는 것을 가능하게 한다. 의도들은 전형적인 사용자 요청들 및 진술(statement)들의 순열(permutation)들로 구성되며, 이들은 또한 발화들(예를 들어, 계정 잔액 가져오기, 구매하기, 등)로도 지칭된다. 본 명세서에서 사용되는 바와 같이, 발화 또는 메시지는 챗봇과의 대화 동안 주고받는 단어들의 세트(예를 들어, 하나 이상의 문장들)을 의미할 수 있다. 의도들은, 어떤 사용자 액션(예를 들어, 피자 주문)을 예시하는 명칭을 제공하고, 일반적으로 액션을 트리거하는 것과 연관된 실제 사용자 진술들, 또는 발화들의 세트를 컴파일링함으로써 생성될 수 있다. 챗봇의 인지가 이러한 의도들로부터 도출되기 때문에, 각각의 의도는 강력하고 다양한 데이터 세트(1 내지 2 더즌(dozen)의 발화들)로부터 생성될 수 있으며, 그 결과 챗봇은 모호한 사용자 입력을 해석할 수 있다. 발화들의 풍부한 세트는, 챗봇이 동일한 것을 의미하지만 상이하게 표현되는 "이 주문 취소!(Forget this order!)" 또는 "배달 취소!(Cancel delivery!)"와 같은 메시지들을 수신할 때 챗봇이 사용자가 원하는 것을 이해하는 것을 가능하게 한다. 집합적으로, 의도들, 및 이들에 속하는 발화들은 챗봇에 대한 트레이닝 코퍼스(corpus)를 구성한다. 코퍼스로 모델을 트레이닝시킴으로써, 고객은 본질적으로 해당 모델을, 최종 사용자 입력을 단일 의도로 리졸빙(resolve)하기 위한 참조 툴로 전환할 수 있다. 고객은, 의도 테스트 및 의도 트레이닝의 라운드들을 통해 챗봇의 인지의 예리함을 개선할 수 있다.

[0025] 그러나, 사용자 발화들에 기초하여 최종 사용자의 의도들을 결정할 수 있는 챗봇을 구축하는 것은, 자연어들의 미묘함과 모호함, 및 입력 공간의 차원(예를 들어, 가능한 사용자 발화들) 및 출력 공간(예를 들어, 발화들의 수)의 크기에 부분적으로 기인하여 어려운 태스크이다. 이러한 어려움의 예시적인 예는, 의도를 표현하기 위한 완곡어법들, 동의어들 또는 비문법적 스피치를 이용하는 것과 같은 자연어의 특성들로부터 발생한다. 예를 들어, 발화는 피자, 주문, 또는 배달을 명시적으로 언급하지 않고 피자를 주문하기 위한 의도를 표현할 수 있다. 예를 들어, 특정 지역적 방언(vernacular)들에서, "피자"는 "파이"로 지칭된다. 부정확성 또는 가변성과 같은 자연어에서의 이러한 경향들은, 예를 들어, 키워드들의 포함을 통한 의도의 명시적 표시와는 대조적으로, 불확실성을 야기하고 의도의 예측을 위한 파라미터로서 신뢰도를 도입한다. 이와 같이, 챗봇은 챗봇의 성능 및 챗봇과의 사용자 경험을 개선하기 위해 트레이닝되고, 모니터링되며, 디버깅되고, 재트레이닝되어야 할 수 있다. 통상적인 시스템들에서, 음성 언어 이해(spoken language understanding; SLU) 및 자연어 프로세싱(natural language processing; NLP)에서 디지털 어시스턴트 또는 챗봇의 기계-학습 모델들을 트레이닝시키고 재-트레이닝시키기 위한 트레이닝 시스템들이 제공된다. 통상적으로, 챗봇 시스템들에 대해 사용되는 모델들은 임의의 의도에 대한 "제조된" 발화들로 NLP에서 트레이닝된다. 예를 들어, 발화 "가격을 변경하나요?(Do you do price changes?)"는 챗봇 시스템의 분류기 모델을 트레이닝시켜서 이러한 유형의 발화를 의도 - "가격 매칭을 제공하나요(Do you offer a price match)"로서 분류하기 위해 사용될 수 있다. 제조된 발화들을 이용하는 모델들의 트레이닝은 서비스들을 제공하기 위해 챗봇 시스템을 초기에 트레이닝시키는 것을 도우며, 그런 다음 챗봇 시스템은, 일단 챗봇 시스템이 배포되고 사용자들로부터 실제 발화들을 획득하기 시작하면 재-트레이닝될 수 있다.

[0026] 발화에 대한 NLP 프로세싱의 부분으로서, 디지털 어시스턴트는 발화의 의미를 이해하도록 트레이닝되며, 이는 발화에 대응하는 하나 이상의 의도들 및 하나 이상의 개체(entity)들을 식별하는 것을 수반한다. 디지털 어시스턴트에서의 개체 추출은 2개의 단계들을 갖는다: 개체명 인식 및 개체명 인식기를 통한 개체 레졸루션. 본 명세서에서 다뤄지는 특정 트레이닝 문제는 개체명 인식기(named entity recognizer; NER)에 관한 것이다. 개체들은 명사들로서 이해될 수 있으며, 종종 슬롯들로 지칭된다. 개체들은 전형적으로, 날짜, 시간, 도시들, 명칭들, 브랜드들, 등과 같은 것들이다(본 명세서에서 예들로서 사용되는 몇 가지 일반적인 개체들은 도메인 독립적이며 시스템 개체들이라고 한다: 사람(PERSON), 숫자(NUMBER), 통화(CURRENCY), 날짜_시간(DATE_TIME); 그러나, 본 개시내용이 이러한 시스템 개체들에 한정되지 않으며, 복합 개체들, 개체 역할들, 개체 리스트들, 및 유사한 것과 같은 임의의 개체 유형에 적용가능하다는 것이 이해되어야 한다). 예를 들어, 여행 봇의 경우, 출발 도시, 목적지, 여행 모드, 가격, 날짜 및 시간을 캡처하는 것은 인터페이스의 기초에서 이루어진다. 그러나, 사용자들이 다양한 컨텍스트(context)들에서 다양한 언어로 특정한 순서 없이 랜덤하게 데이터를 입력하기 때문에, 개체

들을 캡처하는 것은 디지털 어시스턴트에게 어려운 태스크이다. 다음의 표 1은 개체 추출 문제 입력(발화들) 및 출력들(개체들)의 몇 가지 예들을 보여준다:

[0027] 표 1:

[0028]

입력 발화	개체 추출 출력
12/2 택시 요금 $60	통화: $60 시간_날짜: 12/2
로랜드(rowland)의 지시를 보여줘	사람: 로랜드(rowland)

[0029] 그러나, 이러한 개체들을 캡처하는 것은 디지털 어시스턴트가 사용자의 의도에 기초하여 액션을 취하는 데 중요하다.

[0030] 개체명 인식을 위한 모델들의 통상적인 트레이닝은 사전-라벨링된 데이터로 시작한다. 그러나, 챗봇 시스템이 수신할 수 있는 고유 토큰들(예를 들어, 단어들, 심볼들, 날짜 표현들, 시간 표현들, 등)의 수가 엄청나게 많을 수 있으며, 토큰들의 다양한 순서화된 조합들의 양이 몇 자릿수 더 많을 수 있다. 크고 복잡한 입력 공간을 고려하면, 개체명 인식기들은 관심이 있는 개체들을 정확하게 식별하고 분류할 수 있도록 하기 위해 매우 많은 수의 파라미터들(예를 들어, 수억 개의 파라미터들)을 포함한다. 상이한 토큰들 및 토큰 조합들을 신뢰할 수 있게 검출하고 토큰들 및 토큰 조합들을 대응하는 개체들에 정확하게 매핑하도록 개체명 인식기를 트레이닝시키는 것은 매우 큰 트레이닝 세트 및 상당한 양의 트레이닝 시간을 사용해야 한다.

[0031] 트레이닝 데이터 세트의 크기 및 트레이닝 시간을 감소시키기 위한 하나의 접근방식은 전이 학습(transfer learning)을 사용하는 것이다. 즉, 모델은 제1 컨텍스트와 연관되어 학습된 파라미터 값들로 초기화될 수 있으며, 그런 다음 값들은 제2 컨텍스트와 연관된 트레이닝 데이터를 사용하여 미세-튜닝될 수 있다. 전이 학습이 엄청난 효율성 이점들을 제공할 수 있지만, 미세-튜닝에 투입되는 시간 및 자원들은 여전히 값들이 미세-튜닝되는 파라미터들의 수에 의존한다. 따라서, 챗봇-시스템 컨텍스트에서, 기계-학습 모델이 전이 학습을 사용하여 트레이닝되는 경우에도, 트레이닝은 여전히 임계 성능 메트릭들을 달성하기 위해 트레이닝에 대해 큰 크기의 트레이닝 데이터 및 시간 및 컴퓨팅 자원들의 광범위한 투입을 필요로 할 수 있다.

[0032] 따라서, 이러한 문제들을 해결하기 위해 상이한 접근방식이 필요하다. 본 명세서에서 설명되는 일부 실시예들은 클라이언트-특정 데이터를 사용하여 사전-트레이닝된 모델의 일 부분을 미세-튜닝함으로써 주어진 챗봇-시스템 사용 케이스에 대한 모델을 정의하는 것에 관한 것이다. 클라이언트-특정 데이터는, 개별적인 클라이언트에 의해 제공된 데이터, 주어진 플랫폼과 연관된 데이터, 주어진 사용과 연관된 데이터, 주어진 컨텍스트와 연관된 데이터, 및/또는 주어진 시간 기간과 연관된 데이터를 포함할 수 있다.

[0033] 일부 실시예들에서, 기계-학습 모델은 NER이며, 트레이닝 데이터 세트를 사용하여 트레이닝된다. 기계-학습 모델은, 적어도 2개의, 적어도 3개의, 적어도 5개의, 적어도 7개의, 적어도 10개의, 또는 적어도 15개의 계층들을 포함할 수 있다. 기계-학습 모델은, 다중-헤드 어텐션(attention) 기술(예를 들어, 다중-헤드 어텐션 네트워크를 포함함)을 사용하는 적어도 2개의 계층들, 다중-헤드 어텐션 기술을 사용하는 적어도 3개의 계층들, 다중-헤드 어텐션 기술을 사용하는 적어도 5개의 계층들, 다중-헤드 어텐션 기술을 사용하는 적어도 7개의 계층들, 다중-헤드 어텐션 기술을 사용하는 적어도 10개의 계층들, 또는 다중-헤드 어텐션 기술을 사용하는 적어도 15개의 계층들을 포함할 수 있다. 다중-헤드 어텐션 기술은 셀프-어텐션 기술을 포함할 수 있다. 다중-헤드 어텐션 기술을 사용하는 것은 (예를 들어) 셀프-어텐션 모델을 사용하는 것, 다중-헤드 모델을 사용하는 것 또는 트랜스포머(transformer) 모델을 사용하는 것을 포함할 수 있다. 기계-학습 모델은 BERT(Bidirectional Encoder Representations from Transformers) 모델을 포함할 수 있다. 트레이닝된 기계-학습 모델은 다국어 BERT 모델을 포함할 수 있다. 트레이닝된 기계-학습 모델은 하나 이상의 트랜스포머 계층들(예를 들어, 적어도 2개의 트랜스포머 계층들, 적어도 4개의 트랜스포머 계층들, 적어도 6개의 트랜스포머 계층들, 또는 적어도 8개의 트랜스포머 계층들) 및 임베딩(embedding) 계층을 포함할 수 있다.

[0034] 기계-학습 모델을 트레이닝시키는 것은 기계-학습 모델의 파라미터들의 세트의 각각에 대한 값을 학습하는 것을 포함할 수 있다. 파라미터들의 세트는 (예를 들어) 적어도 100,000개의 파라미터들, 적어도 200,000개의 파라미터들, 적어도 500,000개의 파라미터들, 적어도 1,000,000개의 파라미터들, 적어도 5,000,000개의 파라미터들,

적어도 10,000,000개의 파라미터들, 적어도 50,000,000개의 파라미터들, 적어도 100,000,000개의 파라미터들, 적어도 200,000,000개의 파라미터들, 또는 적어도 500,000,000개의 파라미터들을 포함할 수 있다.

[0035] 트레이닝 데이터는, 주어진 사용 케이스, 주어진 사용 케이스들의 조합, 주어진 애플리케이션, 주어진 애플리케이션들의 조합, 주어진 클라이언트, 및/또는 주어진 클라이언트들의 조합과 연관될 수 있다(예를 들어, 이들에 의해 제공되거나, 이들에 대해 액세스되거나, 이들에 의해 선택되거나, 또는 이들에 대해 자동으로 식별될 수 있다). 그 후에, 트레이닝된 모델은, 트레이닝 데이터 세트에 대한 상이한 사용 케이스, 상이한 사용 케이스들의 조합, 상이한 애플리케이션, 상이한 애플리케이션들의 조합, 상이한 클라이언트, 및/또는 상이한 클라이언트들의 조합에 대응하는 클라이언트 데이터 세트를 사용하여 미세-튜닝되는 모델의 초기화를 위해 사용될 수 있다.

[0036] 초기화 이후에, 적어도 하나의 계층, 적어도 2개의 계층들, 적어도 3개의 계층들, 적어도 5개의 계층들, 적어도 8개의 계층들과 연관된 파라미터들의 값들은 적어도 일시적으로 파라미터 값들의 수정을 방지하기 위해 적어도 일시적으로 프리즈될 수 있다. 파라미터들의 값들이 프리즈된 계층들 중 하나, 그 이상, 또는 전부 각각은, 파라미터들의 값들이 프리즈되지 않은 각각의 계층보다 더 낮은 레벨에 있을 수 있다. 예를 들어, 파라미터 값들이 프리즈될 계층들을 결정하는 것은, 임계 계층 아래의 계층들의 파라미터들의 값들이 프리즈되며 반면 임계 계층 및 임계 계층 위의 임의의 계층의 파라미터들의 값들은 프리즈되지 않도록(그리고 클라이언트 데이터 세트를 사용하여 미세-튜닝될 수 있도록) 임계 계층을 선택하는 것을 포함한다.

[0037] 일부 대안적인 또는 추가적인 실시예들에서, 모델들의 계층들 중 하나 이상의 계층의 각각의 계층에 대해(예를 들어, 파라미터 값들을 프리즈하기 위해 선택되지 않은 각각의 계층에 대해), 계층의 파라미터들의 제1 서브세트는 파라미터 값들을 프리즈하기 위해 식별되며 계층의 파라미터들의 제2 서브세트는 (값들이 클라이언트 데이터 세트를 사용하여 미세-튜닝될 수 있도록) 파라미터 값들을 프리즈하는 것을 억제하기 위해 식별된다.

[0038] 제1 서브세트는 계층 내의 하나 이상의 네트워크들 및/또는 서브-모델들의 파라미터들을 포함할 수 있으며, 제2 서브세트는 계층 내의 하나 이상의 다른 네트워크들 및/또는 서브-모델들의 파라미터들을 포함할 수 있다. 예를 들어, 제1 서브세트는 다중-헤드 어텐션 네트워크 또는 셀프-어텐션 네트워크와 연관될 수 있으며, 제2 서브세트는 피드 포워드 네트워크를 포함할 수 있다.

[0039] 이러한 접근방식의 하나의 이점은, 기계-학습 모델이, 기계-학습 모델에서 파라미터들 모두를 미세-튜닝하기 위해 필요한 것보다 더 적은 데이터를 사용하여 미세-튜닝될 수 있다는 점이다. 추가로, 파라미터들의 불완전한 서브세트의 값들의 미세-튜닝을 수행하는 것은 파라미터들 모두의 값들을 미세-튜닝하는 것에 비해 더 빠르게 수행될 수 있다.

[0040] 임계 계층 아래의 하나 이상의 계층들의 파라미터들의 값들이 프리즈되는 경우, 프리즈된 계층들로부터의 출력들에 대응하는 중간 값들은 하나의 미세-튜닝 반복 동안 계산되고 캐싱되며, 미세-튜닝에서의 다른 반복들에 걸쳐 사용될 수 있다. 따라서, 모델은 트레이닝 데이터 내의 각각의 데이터 세트에 대해 중간 값들을 결정해야 할 필요가 없다. 이러한 캐싱은 미세-튜닝의 속도를 감소시킬 수 있으며, 미세-튜닝 동안 사용되는 계산 자원들을 감소시킬 수 있다.

[0041] 따라서, 모델의 파라미터들의 불완전한 서브세트만의 값들이 조정되는 방식으로 기계-학습 모델을 미세-튜닝하는 것은 더 많은 또는 모든 파라미터들의 값들을 조정하는 것에 의해 모델을 미세-튜닝하는 것보다 이점들을 갖는다. 예를 들어, 미세-튜닝은, 모델의 파라미터들의 불완전한 서브세트만이 조정될 때 더 작은 데이터 세트를 사용하여, 더 빠르게, 및/또는 더 적은 계산 자원들을 사용하여 수행될 수 있다. 중간 값들의 캐싱을 가능하게 함으로써, 이러한 이점들은 더 높은 정도로 달성된다. 그럼에도 불구하고, 본 발명자들은, 파라미터들의 불완전한 서브세트만의 값들이 미세-튜닝 동안 업데이트된 기계-학습 모델에 의해 생성된 예측들의 정확도가 모든 파라미터들의 값들이 미세-튜닝 동안 업데이트된 비교할 만한 모델의 정확도에 필적한다는 것을 발견하였다.

[0042] **봇 및 분석 시스템들**

[0043] 봇(스킬, 챗봇, 채터봇(chatterbot), 또는 토크봇(talkbot)으로도 지칭됨)은 최종 사용자들과 대화를 수행할 수 있는 컴퓨터 프로그램이다. 봇은 일반적으로 자연어 메시지들을 사용하는 메시징 애플리케이션을 통해 자연어 메시지들(예를 들어, 질문들 또는 코멘트들)에 응답할 수 있다. 기업들은 메시징 애플리케이션을 통해 최종 사용자들과 소통하기 위해 하나 이상의 봇 시스템들을 사용할 수 있다. 채널로서 지칭될 수 있는 메시징 애플리케이션은, 최종 사용자가 이미 설치했고 익숙한 최종 사용자 선호 메시징 애플리케이션일 수 있다. 따라서, 최종 사용자는 봇 시스템과 채팅하기 위해 새로운 애플리케이션들을 다운로드하여 설치해야 할 필요가 없다. 메시징

애플리케이션은, 예를 들어, 오버-더-톱(over-the-top; OTT) 메시징 채널들(예컨대, 페이스북 메신저(Facebook Messenger), 페이스북 왓츠앱(Facebook WhatsApp), 위챗(WeChat), 라인(Line), 킥(Kik), 텔레그램(Telegram), 토크(Talk), 스카이프(Skype), 슬랙(Slack), 또는 SMS), 가상 개인 어시스턴트들(예컨대 아마존 닷(Amazon Dot), 에코(Echo), 또는 쇼(Show), 구글 홈(Google Home), 애플 홈팟(Apple HomePod), 등), 채팅 능력들, 또는 음성 기반 입력(예컨대, 시리(Siri), 코타나(Cortana), 구글 보이스(Google Voice), 또는 상호작용을 위한 다른 스피치(speech) 입력)으로 기본 또는 하이브리드/반응형 모바일 앱들 또는 웹 애플리케이션들을 확장하는 모바일 및 웹 앱 확장들을 포함할 수 있다.

[0044] 일부 예들에서, 봇 시스템은 통합 자원 식별자(Uniform Resource Identifier; URI)와 연관될 수 있다. URI는 문자열을 사용하여 봇 시스템을 식별할 수 있다. URI는 하나 이상의 메시징 애플리케이션 시스템들에 대한 웹훅(webhook)으로서 사용될 수 있다. URI는, 예를 들어, 통합 자원 로케이터(Uniform Resource Locator; URL) 또는 통합 자원 명칭(Uniform Resource Name; URN)을 포함할 수 있다. 봇 시스템은 메시징 애플리케이션 시스템으로부터 메시지(예를 들어, 하이퍼텍스트 전송 프로토콜(hypertext transfer protocol; HTTP) 포스트 콜(post call) 메시지)를 수신하도록 설계될 수 있다. HTTP 포스트 콜 메시지는 메시징 애플리케이션 시스템으로부터 URI로 보내질 수 있다. 일부 실시예들에서, 메시지는 HTTP 포스트 콜 메시지와는 상이할 수 있다. 예를 들어, 봇 시스템은 단문 메시지 서비스(Short Message Service; SMS)로부터 메시지를 수신할 수 있다. 본 명세서의 논의가 봇 시스템이 메시지로서 수신하는 통신들을 참조할 수 있지만, 메시지는 HTTP 포스트 콜 메시지, SMS 메시지, 또는 2개의 시스템들 사이의 임의의 다른 유형의 통신일 수 있다는 것이 이해되어야 한다.

[0045] 최종 사용자들은 사람들 사이의 상호작용들과 마찬가지로 대화형 상호작용(때때로 대화형 사용자 인터페이스(user interface; UI)로 지칭됨)을 통해 봇 시스템과 상호작용할 수 있다. 일부 경우들에서, 상호작용은, 사용자가 "안녕하세요(Hello)"라고 봇에게 말하고 봇이 "안녕(Hi)"으로 응답하고 최종 사용자에게 어떻게 도움을 줄 수 있는지 묻는 것을 포함할 수 있다. 일부 경우들에서, 상호작용은 또한, 예를 들어, 하나의 계좌로부터 다른 계좌로 돈을 이체하는 것과 같은 뱅킹 봇과의 거래 상호작용; 예를 들어, 남은 휴가를 체크하기 위한 HR 봇과의 정보 상호작용; 또는, 예를 들어, 구매한 상품의 반품을 논의하거나 또는 기술 지원을 요청하는 것과 같은 리테일 봇(retail bot)과의 상호작용일 수 있다.

[0046] 일부 실시예들에서, 봇 시스템은 봇 시스템의 관리자 또는 개발자와의 상호작용 없이 최종 사용자 상호작용들을 지능적으로 핸들링할 수 있다. 예를 들어, 최종 사용자는 희망되는 목표를 달성하기 위해 하나 이상의 메시지들을 봇 시스템으로 전송할 수 있다. 메시지는, 특정 컨텐츠, 예컨대 텍스트, 이모티콘(emojis), 오디오, 이미지, 비디오, 또는 메시지를 운반하는 다른 방법을 포함할 수 있다. 일부 실시예들에서, 봇 시스템은 컨텐츠를 표준화된 형태(예를 들어, 적절한 파라미터들을 갖는 기업 서비스들에 대한 표현 상태 전환(representational state transfer; REST) 콜)로 변환하고 자연어 응답을 생성할 수 있다. 봇 시스템은 또한 추가적인 입력 파라미터들에 대해 최종 사용자에게 프롬프트(prompt)하거나 또는 다른 추가적인 정보를 요청할 수 있다. 일부 실시예들에서, 봇 시스템은 또한 최종 사용자 발화들에 수동적으로 응답하는 것이 아니라 최종 사용자와의 소통을 개시할 수도 있다. 봇 시스템의 명시적 호출(invocation)을 식별하고 호출되는 봇 시스템에 대한 입력을 결정하기 위한 다양한 기술들이 본 명세서에 개시된다. 특정 실시예들에서, 명시적 호출 분석은 발화에서 호출 명칭을 검출하는 것에 기초하여 마스터 봇에 의해 수행된다. 호출 명칭의 검출에 응답하여, 발화는 호출 명칭과 연관된 스킬 봇에 대한 입력을 위해 정제될 수 있다.

[0047] 봇과의 대화는 다수의 상태들을 포함하는 특정 대화 흐름을 따를 수 있다. 흐름은 입력에 기초하여 다음에 어떤 것이 일어날지를 정의할 수 있다. 일부 실시예들에서, 사용자 정의형 상태들(예를 들어, 최종 사용자 의도들) 및 상태들에서 또는 상태별로 취할 액션들을 포함하는 상태 머신이 봇 시스템을 구현하기 위해 사용될 수 있다. 대화는 최종 사용자 입력에 기초하여 상이한 경로들을 취할 수 있으며, 이는 흐름에 대해 봇이 결정하는 결정에 영향을 줄 수 있다. 예를 들어, 각각의 상태에서, 최종 사용자 입력 또는 발화들에 기초하여, 봇은 취해야 할 적절한 다음 액션을 결정하기 위해 최종 사용자의 의도를 결정할 수 있다. 본 명세서에서 그리고 발화의 맥락에서 사용되는 바와 같이, 용어 "의도"는 발화를 제공한 사용자의 의도를 의미한다. 예를 들어, 사용자는, 사용자의 의도가 발화 "피자 주문(Order pizza)"을 통해 표현될 수 있도록 피자를 주문하기 위한 대화에 봇을 참여시킬 것을 의도할 수 있다. 사용자 의도는, 사용자가 챗봇이 사용자를 대신해 수행하기를 원하는 특정 태스크로 전달될 수 있다. 따라서, 발화들은 사용자의 의도를 반영하는 질문들, 명령들, 요청들, 및 유사한 것으로서 표현(phrase)될 수 있다. 의도는 최종 사용자가 달성하기를 원하는 목표를 포함할 수 있다.

[0048] 챗봇의 구성의 맥락에서, 용어 "의도"는 본 명세서에서, 사용자의 발화를 챗봇이 수행할 수 있는 특정 태스크/액션 또는 태스크/액션의 카테고리에 매핑하기 위한 구성 정보를 지칭하기 위해 사용된다. 발화의 의도(즉, 사

용자 의도)와 챗봇의 의도를 구별하기 위해, 챗봇의 의도는 때때로 "봇 의도"로서 본 명세서에서 지칭된다. 봇 의도는 의도와 연관된 하나 이상의 발화들의 세트를 포함할 수 있다. 예를 들어, 피자를 주문하기 위한 의도는, 피자를 주문하려는 욕구를 표현하는 발화들의 다양한 순열들을 가질 수 있다. 이러한 연관된 발화들은, 의도 분류기가 그 후에 사용자로부터의 입력 발화가 피자 주문 의도와 매칭되는지 여부를 결정하는 것을 가능하게 하기 위해 챗봇의 의도 분류기를 트레이닝시키기 위해 사용될 수 있다. 봇 의도는, 사용자와의 대화를 시작하기 위한 그리고 특정 상태에서의 하나 이상의 다이얼로그(dialog) 흐름들과 연관될 수 있다. 예를 들어, 피자 주문 의도에 대한 첫 번째 메시지는 질문 "어떤 종류의 피자를 원하십니까?(What kind of pizza would you like?)"일 수 있다. 연관된 발화들에 추가하여, 봇 의도는 의도와 관련된 개체명들을 더 포함할 수 있다. 예를 들어, 피자 주문 의도는, 피자를 주문하는 태스크를 수행하기 위해 사용되는 변수들 또는 파라미터들, 예를 들어, 토핑 1, 토핑 2, 피자 유형, 피자 크기, 피자 수량, 및 유사한 것을 포함할 수 있다. 개체의 값은 전형적으로 사용자와의 대화를 통해 획득된다.

[0049] 도 1은 특정 실시예들에 따른 챗봇 시스템을 통합하는 환경(100)의 간략화된 블록도이다. 환경(100)은, 디지털 어시스턴트 빌더 플랫폼(digital assistant builder platform; DABP)(102)의 사용자들이 디지털 어시스턴트들 또는 챗봇 시스템들을 생성하고 배포하는 것을 가능하게 하는 DABP(102)를 포함한다. DABP(102)는 하나 이상의 디지털 어시스턴트(digital assistant; DA)들 또는 챗봇 시스템들을 생성하기 위해 사용될 수 있다. 예를 들어, 도 1에 도시된 바와 같이, 특정 기업을 대표하는 사용자(104)는 특정 기업의 사용자들에 대한 디지털 어시스턴트(106)를 생성하고 배포하기 위해 DABP(102)를 사용할 수 있다. 예를 들어, DABP(102)는 은행의 고객들이 사용하기 위한 하나 이상의 디지털 어시스턴트들을 생성하기 위해 은행에 의해 사용될 수 있다. 동일한 DABP(102) 플랫폼은 디지털 어시스턴트들을 생성하기 위해 다수의 기업들에 의해 사용될 수 있다. 다른 예로서, 레스토랑(예를 들어, 피자 가게)의 소유자는, 레스토랑의 고객들이 음식을 주문(예를 들어, 피자를 주문)하는 것을 가능하게 하는 디지털 어시스턴트들을 생성하고 배포하기 위해 DABP(102)를 사용할 수 있다.

[0050] 본 개시내용의 목적들을 위해, "디지털 어시스턴트"는, 디지털 어시스턴트의 사용자들이 자연어 대화들을 통해 다양한 태스크들을 달성하는 것을 돕는 개체이다. 디지털 어시스턴트는 소프트웨어만을 사용하여(예를 들어, 디지털 어시스턴트는 하나 이상의 프로세서들에 의해 실행가능한 프로그램들, 코드, 또는 명령어들을 사용하여 구현되는 디지털 개체임), 하드웨어를 사용하여, 또는 하드웨어와 소프트웨어의 조합을 사용하여 구현될 수 있다. 디지털 어시스턴트는 다양한 물리적 시스템들 또는 디바이스들에, 예컨대 컴퓨터, 모바일 폰, 시계, 가전기기, 차량, 및 유사한 것에 실현되거나 또는 구현될 수 있다. 디지털 어시스턴트는 때때로 챗봇 시스템으로도 지칭된다. 따라서, 본 개시내용의 목적들을 위해, 용어들 디지털 어시스턴트 및 챗봇 시스템은 상호교환가능하다.

[0051] DABP(102)를 사용하여 구축된 디지털 어시스턴트(106)와 같은 디지털 어시스턴트는 디지털 어시스턴트와 그 사용자들(108) 사이의 자연어-기반 대화들을 통해 다양한 태스크들을 수행하기 위해 사용될 수 있다. 대화의 부분으로서, 사용자는 하나 이상의 사용자 입력들(110)을 디지털 어시스턴트(106)에 제공하고 디지털 어시스턴트(106)로부터 다시 응답들(112)을 얻을 수 있다. 대화는 입력들(110) 및 응답들(112) 중 하나 이상을 포함할 수 있다. 이러한 대화들을 통해, 사용자는 디지털 어시스턴트에 의해 수행될 하나 이상의 태스크들을 요청할 수 있으며, 이에 응답하여, 디지털 어시스턴트는 사용자-요청 태스크들을 수행하고 적절한 응답들로 사용자에게 응답하도록 구성된다.

[0052] 사용자 입력들(110)은 일반적으로 자연어 형태이며 발화들로 지칭된다. 사용자 발화(110)는, 사용자가 문장, 질문, 텍스트 단편, 또는 심지어 단일 단어로 타이핑하고 이를 디지털 어시스턴트(106)에 대한 입력으로 제공할 때와 같이, 텍스트 형태일 수 있다. 일부 실시예들에서, 사용자 발화(110)는, 사용자가 디지털 어시스턴트(106)에 대한 입력으로서 제공되는 어떤 것을 말하거나 또는 이야기할 때와 같이 오디오 입력 또는 스피치 형태일 수 있다. 발화들은 전형적으로 사용자(108)가 말하는 언어이다. 예를 들어, 발화들은 영어 또는 어떤 다른 언어일 수 있다. 발화가 스피치 형태일 때, 스피치 입력은 해당 언어의 텍스트 형태 발화들로 변환되며, 그런 다음 텍스트 발화들이 디지털 어시스턴트(106)에 의해 프로세싱된다. 다양한 스피치-대-텍스트 프로세싱 기술들은 스피치 또는 오디오 입력을 텍스트 발화로 변환하기 위해 사용될 수 있으며, 텍스트 발화는 그런 다음 디지털 어시스턴트(106)에 의해 프로세싱된다. 일부 실시예들에서, 스피치-대-텍스트 변환은 디지털 어시스턴트(106) 자체에 의해 이루어진다.

[0053] 텍스트 발화 또는 스피치 발화일 수 있는 발화는, 단편, 다수의 문장들, 하나 이상의 단어들, 하나 이상의 질문들, 전술한 유형들의 조합들, 및 유사한 것일 수 있다. 디지털 어시스턴트(106)는, 사용자 입력의 의미를 이해하기 위해 자연어 이해(natural language understanding; NLU) 기술들을 발화에 적용하도록 구성된다. 발화에 대한 NLU 프로세싱의 부분으로서, 디지털 어시스턴트(106)는 발화의 의미를 이해하기 위한 프로세싱을 수행하도

록 구성되며, 이는 발화에 대응하는 하나 이상의 의도들 및 하나 이상의 개체들을 식별하는 것을 수반한다. 발화의 의미를 이해할 때, 디지털 어시스턴트(106)는 이해된 의미 또는 의도들에 응답하여 하나 이상의 액션들 또는 동작들을 수행할 수 있다. 본 개시내용의 목적들을 위해, 발화들이 디지털 어시스턴트(106)의 사용자(108)에 의해 직접적으로 제공된 텍스트 발화들이거나 또는 텍스트 형태로의 입력 스피치 발화들의 변환의 결과들인 텍스트 발화들인 것으로 가정된다. 그러나, 이는 어떠한 방식으로도 제한적이거나 또는 제한하려고 의도되지 않는다.

[0054] 예를 들어, 사용자(108)는 "나는 피자를 주문하고 싶어(I want to order a pizza)"와 같은 발화를 제공함으로써 피자가 주문되도록 요청할 수 있다. 이러한 발화의 수신 시에, 디지털 어시스턴트(106)는 발화의 의미를 이해하고 적절한 액션들을 취하도록 구성된다. 적절한 액션들은, 예를 들어, 사용자가 주문하고 싶은 피자의 유형, 피자의 크기, 피자에 대한 임의의 토핑들, 및 유사한 것에 대한 사용자 입력을 요청하는 질문들로 사용자에게 응답하는 것을 수반할 수 있다. 디지털 어시스턴트(106)에 의해 제공되는 응답들도 또한 자연어 형태이며 전형적으로 입력 발화와 동일한 언어이다. 이러한 응답들을 생성하는 것을 부분으로서, 디지털 어시스턴트(106)는 자연어 생성(natural language generation; NLG)을 수행할 수 있다. 사용자와 디지털 어시스턴트(106) 사이의 대화를 통해 피자를 주문하는 사용자에 대해, 디지털 어시스턴트는 피자 주문을 위한 모든 필수 정보를 제공하도록 사용자를 가이드하고, 그런 다음 대화의 끝에서 피자가 주문되게 할 수 있다. 디지털 어시스턴트(106)는 피자가 주문되었다는 것을 나타내는 정보를 사용자에게 출력함으로써 대화를 종료할 수 있다.

[0055] 개념적 레벨에서, 디지털 어시스턴트(106)는 사용자로부터 수신된 발화에 응답하여 다양한 프로세싱을 수행한다. 일부 실시예들에서, 이러한 프로세싱은, 예를 들어, 입력 발화의 의미를 이해하는 것(때때로 자연어 이해(Natural Language Understanding; NLU)로 지칭됨), 발화에 응답하여 수행될 액션을 결정하는 것, 적절한 경우 액션이 수행되게 하는 것, 사용자 발화에 응답하여 사용자에게 출력될 응답을 생성하는 것, 사용자에게 응답을 출력하는 것, 및 유사한 것을 포함하여, 일련의 프로세싱 단계들 또는 프로세싱 단계들의 파이프라인을 수반한다. NLU 프로세싱은, 발화의 구조 및 의미를 이해하기 위해 수신된 입력 발화를 파싱(parse)하는 것, 발화에 대해 더 잘 이해할 수 있는 형태(예를 들어, 논리적 형태) 또는 구조를 개발하기 위해 발화를 정제하고 리포밍(reform)하는 것을 포함할 수 있다. 응답을 생성하는 것은 NLG 기술들을 사용하는 것을 포함할 수 있다.

[0056] 디지털 어시스턴트(106)와 같은 디지털 어시스턴트에 의해 수행되는 NLU 프로세싱은 문장 파싱(예를 들어, 토큰화, 레마타이징(lemmatizing), 문장에 대한 품사 태그들을 식별하는 것, 문장 내의 개체명들을 식별하는 것, 문장 구조를 표현하기 위한 종속성 트리들을 생성하는 것, 문장을 절들로 분할하는 것, 개별적인 절들을 분석하는 것, 대용어(anaphora)들을 리졸빙하는 것, 청킹(chunking)을 수행하는 것, 및 유사한 것)과 같은 다양한 NLP 관련 프로세싱을 포함할 수 있다. 특정 실시예들에서, NLU 프로세싱 또는 이의 부분들은 디지털 어시스턴트(106) 자체에 의해 수행된다. 일부 다른 실시예들에서, 디지털 어시스턴트(106)는 NLU 프로세싱의 부분들을 수행하기 위해 다른 자원들을 사용할 수 있다. 예를 들어, 입력 발화 문장의 신택스(syntax) 및 구조는 파서, 품사 태거(tagger), 및/또는 개체명 인식기를 사용하여 문장을 프로세싱함으로써 식별될 수 있다. 일 구현예에서, 영어에 대해, 스탠포드 자연어 프로세싱(Natural Language Processing; NLP) 그룹에 의해 제공되는 것들과 같은 파서, 품사 태거, 및 개체명 인식기가 문장 구조 및 신택스를 분석하기 위해 사용된다. 이들은 스탠포드 CoreNLP 툴키트(toolkit)의 부분으로서 제공된다.

[0057] 본 개시내용에서 제공되는 다양한 예들이 발화들을 영어로 보여주지만, 이는 단지 일 예를 의미한다. 특정 실시예들에서, 디지털 어시스턴트(106)는 또한 영어 이외의 언어들로 된 발화들도 핸들링할 수 있다. 디지털 어시스턴트(106)는, 상이한 언어들에 대해 프로세싱을 수행하도록 구성된 서브시스템들(예를 들어, NLU 기능성을 구현하는 구성요소들)을 제공할 수 있다. 이러한 서브시스템들은, NLU 코어 서버로부터의 서비스 콜들을 사용하여 호출될 수 있는 플러그가능 유닛들로서 구현될 수 있다. 이는, 프로세싱의 상이한 순서들을 허용하는 것을 포함하여, NLU 프로세싱을 각각의 언어에 대해 유연하고 확장가능하게 만든다. 언어 팩이 개별적인 언어들에 대해 제공될 수 있으며, 여기서 언어 팩은 NLU 코어 서버로부터 서비스될 수 있는 서브시스템들의 리스트를 등록할 수 있다.

[0058] 도 1에 도시된 디지털 어시스턴트(106)와 같은 디지털 어시스턴트는 다양하고 상이한 채널들을 통해, 예컨대 비제한적으로, 특정 애플리케이션들을 통해, 소셜 미디어 플랫폼들을 통해, 다양한 메시징 서비스들 및 애플리케이션들을 통해, 그리고 다른 애플리케이션들 또는 채널들을 통해 그 사용자들(108)에게 이용가능해지거나 또는 액세스가능해질 수 있다. 단일 디지털 어시스턴트는, 상이한 서비스들 상에서 동시에 실행되고 액세스될 수 있도록 단일 디지털 어시스턴트에 대해 구성된 몇몇 채널들을 가질 수 있다.

[0059] 디지털 어시스턴트 또는 챗봇 시스템은 일반적으로 하나 이상의 스킬들을 포함하거나 또는 이와 연관된다. 특정 실시예들에서, 이러한 스킬들은, 사용자들과 상호작용하고 재고 추적, 타임카드 제출, 비용 보고서 생성, 음식 주문, 은행 계좌 체크, 예약, 위젯 구매, 및 유사한 것과 같은 특정 유형들의 태스크들을 이행하도록 구성된 개별적인 챗봇들(스킬 봇들로 지칭됨)이다. 예를 들어, 도 1에 도시된 실시예에 대해, 디지털 어시스턴트 또는 챗봇 시스템(106)은 스킬들(116-1, 116-2, 등)을 포함한다. 본 개시내용의 목적들을 위해, 용어들 "스킬" 및 "스킬들"은 각각 용어들 "스킬 봇" 및 "스킬 봇들"과 동의적으로 사용된다.

[0060] 디지털 어시스턴트와 연관된 각각의 스킬은 디지털 어시스턴트의 사용자가 사용자와의 대화를 통해 태스크를 완료하는 것을 도우며, 여기서 대화는 사용자에 의해 제공되는 텍스트 또는 오디오 입력들의 조합 및 스킬 봇들에 의해 제공되는 응답들을 포함할 수 있다. 이러한 응답들은 사용자에 대한 텍스트 또는 오디오 메시지들의 형태일 수 있거나 및/또는 사용자가 선택할 수 있도록 사용자에게 표시되는 간단한 사용자 인터페이스 요소들(예를 들어, 선택 리스트들)을 사용할 수 있다.

[0061] 스킬 또는 스킬 봇이 디지털 어시스턴트와 연관되거나 또는 이에 추가될 수 있는 다양한 방식들이 있다. 일부 경우들에서, 스킬 봇은 DABP(102)를 사용하여 기업에 의해 개발되고 그런 다음 디지털 어시스턴트에 추가될 수 있다. 다른 경우들에서, 스킬 봇은 DABP(102)를 사용하여 개발되며 생성되고 그런 다음 DABP(102)를 사용하여 생성된 디지털 어시스턴트에 추가될 수 있다. 또 다른 경우들에서, DABP(102)는 광범위한 태스크들에 대한 다수의 스킬들을 제공하는 온라인 디지털 스토어("스킬 스토어"로 지칭됨)를 제공한다. 스킬 스토어를 통해 제공되는 스킬들은 또한 다양한 클라우드 서비스들을 노출할 수 있다. 스킬을 DABP(102)를 사용하여 생성되는 디지털 어시스턴트에 추가하기 위해, DABP(102)의 사용자는 DABP(102)를 통해 스킬 스토어에 액세스하고, 희망되는 스킬을 선택하며, 선택된 스킬이 DABP(102)을 사용하여 생성된 디지털 어시스턴트에 추가될 것임을 표시할 수 있다. 스킬 스토어로부터의 스킬은 있는 그대로 또는 수정된 형태로 디지털 어시스턴트에 추가될 수 있다(예를 들어, DABP(102)의 사용자는 스킬 스토어에 의해 제공되는 특정 스킬 봇을 선택하고 복제하며, 선택된 스킬 봇에 대한 맞춤화 또는 수정을 수행하고, 그런 다음 수정된 스킬 봇을 DABP(102)를 사용하여 생성된 디지털 어시스턴트에 추가할 수 있다).

[0062] 다양한고 상이한 아키텍처들이 디지털 어시스턴트 또는 챗봇 시스템을 구현하기 위해 사용될 수 있다. 예를 들어, 특정 실시예들에서, DABP(102)를 사용하여 생성되고 배포된 디지털 어시스턴트들은 마스터 봇/차일드(또는 서브) 봇 패러다임 또는 아키텍처를 사용하여 구현될 수 있다. 이러한 패러다임에 따르면, 디지털 어시스턴트는 스킬 봇들인 하나 이상의 차일드 봇들과 상호작용하는 마스터 봇으로서 구현된다. 예를 들어, 도 1에 도시된 실시예에서, 디지털 어시스턴트(106)는 마스터 봇(114) 및 마스터 봇(114)의 차일드 봇들인 스킬 봇들(116-1, 116-2, 등)을 포함한다. 특정 실시예들에서, 디지털 어시스턴트(106)는 그 자체가 마스터 봇으로서 역할하는 것으로 간주된다.

[0063] 마스터-차일드 봇 아키텍처에 따라 구현된 디지털 어시스턴트는 디지털 어시스턴트의 사용자들이 통합된 사용자 인터페이스를 통해, 즉 마스터 봇을 통해 다수의 스킬들과 상호작용하는 것을 가능하게 한다. 사용자가 디지털 어시스턴트에 참여할 때, 사용자 입력은 마스터 봇에 의해 수신된다. 그런 다음, 마스터 봇은 사용자 입력 발화의 의미를 결정하기 위한 프로세싱을 수행한다. 그런 다음, 마스터 봇은 발화에서 사용자에 의해 요청된 태스크가 마스터 봇 자체에 의해 핸들링될 수 있는지 여부를 결정하며, 그렇지 않은 경우 마스터 봇은 사용자 요청을 핸들링하기 위한 적절한 스킬 봇을 선택하고 대화를 선택된 스킬 봇으로 라우팅한다. 이는 사용자가 공통 단일 인터페이스를 통해 디지털 어시스턴트와 대화하는 것을 가능하게 하며, 특정 태스크들을 수행하도록 구성된 몇몇 스킬 봇들을 사용할 수 있는 능력을 계속해서 제공한다. 예를 들어, 기업을 위해 개발된 디지털 어시스턴트에 대해, 디지털 어시스턴트의 마스터 봇은 특정 기능성들을 갖는 스킬 봇들, 예컨대 고객 관계 관리(customer relationship management; CRM)와 관련된 기능들을 수행하기 위한 CRM 봇, 전사적 자원 관리(enterprise resource planning; ERP)와 관련된 기능들을 수행하기 위한 ERP 봇, 인적 자본 관리(human capital management; HCM)와 관련된 기능들을 수행하기 위한 HCM 봇, 등과 인터페이스할 수 있다. 이러한 방식으로, 디지털 어시스턴트의 최종 사용자 또는 고객은 공통 마스터 봇 인터페이스를 통해 디지털 어시스턴트에 액세스하기 위한 방법만 알면 되며, 사용자 요청을 핸들링하기 위해 배후에서 다수의 스킬 봇들이 제공된다.

[0064] 특정 실시예들에서, 마스터 봇/차일드 봇 인프라스트럭처에서, 마스터 봇은 이용가능한 스킬 봇들의 리스트를 인식하도록 구성된다. 마스터 봇은 다양한, 이용가능한 스킬 봇들, 및 각각의 스킬 봇에 대한, 스킬 봇에 의해 수행될 수 있는 태스크들을 포함하는 스킬 봇의 능력들을 식별하는 메타데이터에 대한 액세스를 가질 수 있다. 발화 형태로 사용자 요청을 수신할 때, 마스터 봇은, 다수의 이용가능한 스킬 봇들로부터, 사용자 요청을 최상으로 서비스하거나 또는 핸들링할 수 있는 특정 스킬 봇을 식별하거나 또는 예측하도록 구성된다. 그런 다음,

마스터 봇은 추가적인 핸들링을 위해 발화(또는 발화의 일 부분)을 해당 특정 스킬 봇으로 라우팅한다. 따라서, 제어(control)는 마스터 봇으로부터 스킬 봇들로 흐른다. 마스터 봇은 다수의 입력 및 출력 채널들을 지원할 수 있다. 특정 실시예들에서, 라우팅하는 것은 하나 이상의 이용가능한 스킬 봇들에 의해 수행되는 프로세싱의 도움으로 수행될 수 있다. 예를 들어, 이하에서 논의되는 바와 같이, 스킬 봇은 발화에 대한 의도를 추론하고 추론된 의도가 스킬 봇이 구성된 의도와 매칭되는지 여부를 결정하도록 트레이닝될 수 있다. 따라서, 마스터 봇에 의해 수행되는 라우팅은, 스킬 봇이 발화를 핸들링하기에 적절한 의도로 구성되었는지 여부의 표시를 스킬 봇이 마스터 봇으로 통신하는 것을 수반할 수 있다.

[0065] 도 1의 실시예가 마스터 봇(114) 및 스킬 봇들(116-1, 116-2, 및 116-3)을 포함하는 디지털 어시스턴트(106)를 도시하지만, 이는 제한적으로 의도되지 않는다. 디지털 어시스턴트는, 디지털 어시스턴트의 기능성들을 제공하는 다양한 다른 구성요소들(예를 들어, 다른 시스템들 및 서브시스템들)을 포함할 수 있다. 이러한 시스템들 및 서브시스템들은 소프트웨어(예를 들어, 컴퓨터-판독가능 매체에 저장되고 하나 이상의 프로세서들에 의해 실행가능한 코드, 명령어들)로만, 하드웨어로만, 또는 소프트웨어와 하드웨어의 조합을 사용하는 구현들로 구현될 수 있다.

[0066] DABP(102)는, DABP(102)의 사용자가 디지털 어시스턴트와 연관된 하나 이상의 스킬 봇들을 포함하는 디지털 어시스턴트를 생성하는 것을 가능하게 하는 인프라스트럭처 및 다양한 서비스들과 특징들을 제공한다. 일부 경우들에서, 스킬 봇은, 기존 스킬 봇을 복제함으로써, 예를 들어, 스킬 스토어에 의해 제공된 스킬 봇을 복제함으로써 생성될 수 있다. 이상에서 표시된 바와 같이, DABP(102)는, 다양한 태스크들을 수행하기 위한 다수의 스킬 봇들을 제공하는 스킬 스토어 또는 스킬 카탈로그를 제공한다. DABP(102)의 사용자는 스킬 스토어로부터 스킬 봇을 복제할 수 있다. 필요에 따라, 수정들 또는 맞춤화들이 복제된 스킬 봇에 대해 이루어질 수 있다. 일부 다른 경우들에서, DABP(102)의 사용자는 DABP(102)에 의해 제공되는 툴들 및 서비스들을 사용하여 스킬 봇을 처음부터 생성한다. 이상에서 표시된 바와 같이, DABP(102)에 의해 제공되는 스킬 스토어 또는 스킬 카탈로그는 다양한 태스크들을 수행하기 위한 다수의 스킬 봇들을 제공할 수 있다.

[0067] 특정 실시예들에서, 하이 레벨에서, 스킬 봇을 생성하는 것 또는 맞춤화하는 것은 다음의 단계들을 수반한다:

[0068] (1) 새로운 스킬 봇에 대한 세팅들을 구성하는 단계

[0069] (2) 스킬 봇에 대한 하나 이상의 의도들을 구성하는 단계

[0070] (3) 하나 이상의 의도들에 대한 하나 이상의 개체들을 구성하는 단계

[0071] (4) 스킬 봇을 트레이닝시키는 단계

[0072] (5) 스킬 봇에 대한 다이얼로그 흐름을 생성하는 단계

[0073] (6) 필요에 따라 스킬 봇에 맞춤형 구성요소들을 추가하는 단계

[0074] (7) 스킬 봇을 테스트하고 배포하는 단계

[0075] 이상의 단계들 각각이 이하에서 간략하게 설명된다.

[0076] (1) 새로운 스킬 봇에 대한 세팅들을 구성하는 단계 - 다양한 세팅들이 스킬 봇에 대해 구성될 수 있다. 예를 들어, 스킬 봇 설계자는 생성되는 스킬 봇에 대한 하나 이상의 호출 명칭들을 지정할 수 있다. 그런 다음, 이러한 호출 명칭들은 스킬 봇을 명시적으로 호출하기 위해 디지털 어시스턴트의 사용자들에 의해 사용될 수 있다. 예를 들어, 사용자는 대응하는 스킬 봇을 명시적으로 호출하기 위해 사용자의 발화에서 호출 명칭을 입력할 수 있다.

[0077] (2) 스킬 봇에 대한 하나 이상의 의도들 및 연관된 예시적인 발화들을 구성하는 단계 - 스킬 봇 설계자는 생성되는 스킬 봇에 대해 하나 이상의 의도들(봇 의도들로도 지칭됨)을 지정한다. 그런 다음, 스킬 봇은 이러한 지정된 의도들에 기초하여 트레이닝된다. 이러한 의도들은, 스킬 봇이 입력 발화들에 대해 추론하도록 트레이닝되는 카테고리들 또는 클래스들을 나타낸다. 발화의 수신 시에, 트레이닝된 스킬 봇은 발화에 대한 의도를 추론하며, 여기서 추론된 의도는 스킬 봇을 트레이닝시키기 위해 사용된 미리 정의된 의도들의 세트로부터 선택된다. 그런 다음, 스킬 봇은 해당 발화에 대해 추론된 의도에 기초하여 발화에 응답하여 적절한 액션을 취한다. 일부 경우들에서, 스킬 봇에 대한 의도들은, 스킬 봇이 디지털 어시스턴트의 사용자들에 대해 수행할 수 있는 태스크들을 나타낸다. 각각의 의도에는 의도 식별자 또는 의도 명칭이 주어진다. 예를 들어, 은행에 대해 트레이닝된 스킬 봇에 대해, 해당 스킬 봇에 대해 지정된 의도들은 "잔액체크(CheckBalance)", "송금(TransferMoney)",

"예금체크(DepositCheck)" 및 유사한 것을 포함할 수 있다.

[0078] 스킬 봇에 대해 정의된 각각의 의도에 대해, 스킬 봇 설계자는 또한, 의도를 대표하고 예시하는 하나 이상의 예시적인 발화들을 제공할 수 있다. 이러한 예시적인 발화들은, 사용자가 해당 의도에 대해 스킬 봇에 입력할 수 있는 발화들을 나타내도록 의도된다. 예를 들어, 잔액체크 의도에 대해, 예시적인 발화들은, "내 저축 계좌 잔액이 얼마예요?(What's my savings account balance?)", "내 당좌예금 계좌에 얼마가 있지?(How much is in my checking account?)", "내 계좌에 돈이 얼마나 있지(How much money do I have in my account)", 및 유사한 것을 포함할 수 있다. 따라서, 전형적인 사용자 발화들의 다양한 순열들이 의도에 대한 예시적인 발화로서 지정될 수 있다.

[0079] 의도들 및 이들의 연관된 예시적인 발화들은 스킬 봇을 트레이닝시키기 위한 트레이닝 데이터로서 사용된다. 다양하고 상이한 트레이닝 기술들이 사용될 수 있다. 이러한 트레이닝의 결과로서, 입력으로서 발화를 취하고 예측 모델에 의해 발화에 대해 추론된 의도를 출력하도록 구성된 예측 모델이 생성된다. 일부 경우들에서, 입력 발화들은, 입력 발화에 대한 의도를 예측하거나 또는 추론하기 위해 트레이닝된 모델을 사용하도록 구성된 의도 분석 엔진에 제공된다. 그런 다음, 스킬 봇은 추론된 의도에 기초하여 하나 이상의 액션들을 취할 수 있다.

[0080] (3) 스킬 봇의 하나 이상의 의도들에 대한 개체들을 구성하는 단계 - 일부 경우들에서, 스킬 봇이 사용자 발화에 적절하게 응답하는 것을 가능하게 하기 위해 추가적인 컨텍스트가 필요할 수 있다. 예를 들어, 사용자 입력 발화가 스킬 봇에서 동일한 의도로 리졸빙되는 상황들이 있을 수 있다. 예를 들어, 이상의 예에서, 발화들 "내 저축 계좌 잔액이 얼마예요?(What's my savings account balance?)" 및 "내 당좌예금 계좌에 얼마가 있지?(How much is in my checking account?)" 둘 모두는 동일한 잔액체크 의도로 리졸빙되지만, 이러한 발화들은 상이한 것들을 요구하는 상이한 요청들이다. 이러한 요청들을 명확히 하기 위해, 하나 이상의 개체들이 의도에 추가된다. 뱅킹 스킬 봇 예를 사용하면, "당좌예금(checking)" 및 "저축(saving)"으로 지칭되는 값들을 정의하는 계정유형(AccountType)으로 지칭되는 개체는 스킬 봇이 사용자 요청을 파싱하고 적절하게 응답하는 것을 가능하게 할 수 있다. 이상의 예에서, 발화들이 동일한 의도로 리졸빙되지만, 계정유형 개체와 연관된 값은 2개의 발화들에 대해 상이하다. 이는 스킬 봇이, 2개의 의도들을 동일한 의도로 리졸빙함에도 불구하고 아마도 이러한 2개의 발화들에 대해 상이한 액션들을 수행하는 것을 가능하게 한다. 하나 이상의 개체들은 스킬 봇에 대해 구성된 특정 의도들에 대해 지정될 수 있다. 따라서, 개체들은 의도 자체에 컨텍스트를 추가하기 위해 사용된다. 개체들은 의도를 더 완전히 설명하고 스킬 봇이 사용자 요청을 완료하는 것을 가능하게 하는 것을 돕는다.

[0081] 특정 실시예들에서, 2가지 유형들의 개체들이 있다: (a) DABP(102)에 의해 제공되는 내장 개체들, 및 (2) 스킬 봇 설계자에 의해 지정될 수 있는 맞춤형 개체들. 내장 개체들은, 다양한 봇들과 함께 사용될 수 있는 일반 개체들이다. 내장 개체들의 예들은, 비제한적으로, 시간, 날짜, 주소들, 숫자들, 이메일 주소들, 지속기간, 반복(recurring) 시간 기간들, 통화들, 전화 번호들, URL들, 및 유사한 것과 관련된 개체들을 포함한다. 맞춤형 개체들은 더 맞춤화된 애플리케이션들에 대해 사용된다. 예를 들어, 뱅킹 스킬에 대해, 당좌예금, 저축, 및 신용카드들, 등과 같은 키워드들에 대해 사용자 입력을 체크함으로써 다양한 뱅킹 거래들을 가능하게 하는 계정유형(AccountType) 개체는 스킬 봇 설계자에 의해 정의될 수 있다.

[0082] (4) 스킬 봇을 트레이닝시키는 단계 - 스킬 봇은, 발화 형태의 사용자 입력을 수신하고, 수신된 입력을 파싱하거나 또는 달리 프로세싱하며, 수신된 사용자 입력과 관련된 의도를 식별하거나 또는 선택하도록 구성된다. 이상에서 표시된 바와 같이, 스킬 봇은 이를 위해 트레이닝되어야 한다. 특정 실시예들에서, 스킬 봇은, 스킬 봇이 사용자 입력 발화들을 그 구성된 의도들 중 하나로 리졸빙할 수 있도록 스킬 봇에 대해 구성된 의도들 및 의도들과 연관된 예시적인 발화들(집합적으로, 트레이닝 세트)에 기초하여 트레이닝된다. 특정 실시예들에서, 스킬 봇은, 트레이닝 데이터를 사용하여 트레이닝되며 스킬 봇이 사용자가 말하는 것(또는 일부 경우들에서, 사용자가 말하려고 하는 것)을 식별하는 것을 가능하게 하는 예측 모델을 사용한다. DABP(102)는, 다양한 기계-학습 기반 트레이닝 기술들, 규칙-기반 트레이닝 기술들, 및/또는 이들의 조합들을 포함하여, 스킬 봇을 트레이닝시키기 위해 스킬 봇 설계자에 의해 사용될 수 있는 다양하고 상이한 트레이닝 기술들을 제공한다. 특정 실시예들에서, 트레이닝 데이터의 일 부분(예를 들어, 80%)은 스킬 봇 모델을 트레이닝시키기 위해 사용되며, 다른 부분(예를 들어, 나머지 20%)은 모델을 테스트하거나 또는 검증하기 위해 사용된다. 일단 트레이닝되면, 트레이닝된 모델(때때로 트레이닝된 스킬 봇으로 지칭됨)은 그런 다음 사용자 발화들을 핸들링하고 이에 응답하기 위해 사용될 수 있다. 특정 경우들에서, 사용자의 발화는, 단일 답변만을 요구하며 추가적인 대화를 요구하지 않는 질문일 수 있다. 이러한 상황들을 핸들링하기 위해, Q&A(질문 및 답변) 의도가 스킬 봇에 대해 정의될 수 있다. 이는 스킬 봇이 다이얼로그 정의를 업데이트할 필요 없이 사용자 요청들에 대한 리플라이(reply)를 출력하는 것을 가능하게 한다. Q&A 의도들은 일반적인 의도들과 유사한 방식으로 생성된다. Q&A 의도들에 대한 다이얼로그

흐름은 일반적인 의도들에 대한 것과는 상이할 수 있다.

[0083] (5) 스킬 봇에 대한 다이얼로그 흐름을 생성하는 단계 -- 스킬 봇에 대해 지정된 다이얼로그 흐름은, 스킬 봇에 대한 상이한 의도들이 수신된 사용자 의도에 응답하여 리졸빙될 때 스킬 봇이 어떻게 반응하는지를 설명한다. 다이얼로그 흐름은, 스킬 봇이 취할 동작들 또는 액션들, 예를 들어, 스킬 봇이 사용자 발화들에 대해 어떻게 반응할지, 스킬 봇이 입력을 위해 어떻게 사용자들에게 프롬프트할지, 스킬 봇이 어떻게 데이터를 반환할지를 정의한다. 다이얼로그 흐름은, 스킬 봇이 따르는 흐름도와 유사하다. 스킬 봇 설계자는 마크다운(markdown) 언어와 같은 언어를 사용하여 다이얼로그 흐름을 지정한다. 특정 실시예에서, OBotML로 지칭되는 YAML의 버전이 스킬 봇에 대한 다이얼로그 흐름을 지정하기 위해 사용될 수 있다. 스킬 봇에 대한 다이얼로그 흐름 정의는, 스킬 봇 설계자가 스킬 봇이 서비스하는 사용자들과 스킬 봇 사이의 상호작용들을 편성(choreograph)하도록 하는 대화 자체에 대한 모델로서 역할한다.

[0084] 특정 실시예들에서, 스킬 봇에 대한 다이얼로그 흐름 정의는 3개의 섹션들을 포함한다:

[0085] (a) 컨텍스트 섹션

[0086] (b) 디폴트 전환 섹션

[0087] (c) 상태 섹션

[0088] 컨텍스트 섹션 -- 스킬 봇 설계자는, 컨텍스트 섹션에서 대화 흐름에서 사용되는 변수들을 정의할 수 있다. 컨텍스트 섹션에서 명명될 수 있는 다른 변수들은, 비제한적으로, 오류를 핸들링하기 위한 변수들, 내장 또는 맞춤형 개체들에 대한 변수들, 스킬 봇이 사용자 선호사항들을 인식하고 유지하는 것을 가능하게 하는 사용자 변수들, 및 유사한 것을 포함한다.

[0089] 디폴트 전환 섹션 -- 스킬 봇에 대한 전환들은 다이얼로그 흐름 상태 섹션에 또는 디폴트 전환 섹션에 정의된다. 디폴트 전환 섹션에 정의된 전환들은 폴백(fallback)으로서 역할하며, 소정의 상태 내에 정의된 적용 가능한 전환들이 없을 때 또는 상태 전환을 트리거하기 위해 필요한 조건들이 충족될 수 없을 때 트리거된다. 디폴트 전환 섹션은, 스킬 봇이 예상치 못한 사용자 액션들을 원활하게 핸들링하는 것을 가능하게 하는 라우팅을 정의하기 위해 사용될 수 있다.

[0090] 상태 섹션 - 다이얼로그 흐름 및 그 관련된 동작들은, 다이얼로그 흐름 내의 로직을 관리하는 일시적 상태들의 시퀀스로서 정의된다. 다이얼로그 흐름 정의 내의 각각의 상태 노드는, 다이얼로그의 해당 지점에서 필요한 기능성을 제공하는 구성요소를 명명한다. 따라서, 상태들은 구성요소들을 중심으로 구축된다. 상태는 구성요소-특정 속성들을 포함하며, 구성요소가 실행된 이후에 트리거된 다른 상태들로의 전환들을 정의한다.

[0091] 특수 케이스 시나리오들이 상태 섹션들을 사용하여 핸들링될 수 있다. 예를 들어, 디지털 어시스턴트 내의 제2 스킬에서 어떤 것을 하기 위해 사용자들이 참여한 제1 스킬을 일시적으로 떠날 수 있는 옵션을 사용자들에게 제공하기를 원하는 경우들이 있을 수 있다. 예를 들어, 사용자가 쇼핑 스킬과의 대화에 참여한 경우(예를 들어, 사용자가 구매를 위해 어떤 선택을 했음), 사용자는 뱅킹 스킬로 점프하고(예를 들어, 사용자는, 자신이 구매를 위해 충분한 돈을 가지고 있는지 확인하기를 원할 수 있음) 그런 다음 사용자의 주문을 완료하기 위해 쇼핑 스킬로 복귀하기를 원할 수 있다. 이를 해결하기 위해, 제1 스킬에서의 액션은 동일한 디지털 어시스턴트에서 제2 상이한 스킬과의 상호작용을 개시하고 그런 다음 원래의 흐름으로 복귀하도록 구성될 수 있다.

[0092] (6) 스킬 봇에 맞춤형 구성요소들을 추가하는 단계 - 이상에서 설명된 바와 같이, 스킬 봇에 대한 다이얼로그 흐름에서 지정된 상태들은 상태들에 대응하는 필요한 기능성을 제공하는 구성요소들을 명명한다. 구성요소들은 스킬 봇이 기능들을 수행하는 것을 가능하게 한다. 특정 실시예들에서, DABP(102)는 광범위한 기능들을 수행하기 위한 미리 구성된 구성요소들의 세트를 제공한다. 스킬 봇 설계자는 이러한 미리 구성된 구성요소들 중 하나 이상을 선택하고 이들을 스킬 봇에 대한 다이얼로그 흐름 내의 상태들과 연관시킬 수 있다. 스킬 봇 설계자는 또한 DABP(102)에 의해 제공된 툴들을 사용하여 맞춤형 또는 새로운 구성요소들을 생성하고 맞춤형 구성요소들을 스킬 봇에 대한 다이얼로그 흐름 내의 하나 이상의 상태들과 연관시킬 수 있다.

[0093] (7) 스킬 봇을 테스트하고 배포하는 단계 - DABP(102)는, 스킬 봇 설계자가 개발되고 있는 스킬 봇을 테스트하는 것을 가능하게 하는 몇몇 특징들을 제공한다. 그런 다음, 스킬 봇이 배포되고 디지털 어시스턴트에 포함될 수 있다.

[0094] 이상의 설명이 스킬 봇을 생성하는 방법을 설명하지만, 유사한 기술들은 또한 디지털 어시스턴트(또는 마스터 봇)를 생성하기 위해 사용될 수 있다. 마스터 봇 또는 디지털 어시스턴트 레벨에서, 내장 시스템 의도들은 디지

털 어시스턴트에 대해 구성될 수 있다. 이러한 내장 시스템 의도들은, 디지털 어시스턴트 자체(즉, 마스터 봇)가 디지털 어시스턴트와 연관된 스킬 봇을 호출하지 않고 핸들링할 수 있는 일반적인 태스크들을 식별하기 위해 사용된다. 마스터 봇에 대해 정의된 시스템 의도들의 예들은 다음을 포함한다: (1) 종료(Exit): 사용자가 디지털 어시스턴트에서 현재 대화 또는 컨텍스트를 종료하기 위한 희망을 시그널링할 때 적용된다; (2) 도움(Help): 사용자가 도움 또는 안내를 요청할 때 적용된다; 및 (3) 리졸빙되지않은의도(UnresolvedIntent): 종료 및 도움 의도들과 잘 매칭되지 않는 사용자 입력에 대해 적용된다. 디지털 어시스턴트는 또한 디지털 어시스턴트와 연관된 하나 이상의 스킬 봇들에 대한 정보를 저장한다. 이러한 정보는 마스터 봇이 발화를 핸들링하기 위한 특정 스킬 봇을 선택하는 것을 가능하게 한다.

[0095] 마스터 봇 또는 디지털 어시스턴트 레벨에서, 사용자가 디지털 어시스턴트에 구문 또는 발화를 입력할 때, 디지털 어시스턴트는 발화 및 관련된 대화를 라우팅할 방법을 결정하기 위한 프로세싱을 수행하도록 구성된다. 디지털 어시스턴트는 라우팅 모델을 사용하여 이를 결정하며, 라우팅 모델은 규칙-기반, AI-기반, 또는 이들의 조합일 수 있다. 디지털 어시스턴트는, 사용자 입력 발화에 대응하는 발화가 핸들링을 위해 특정 스킬로 라우팅될지, 내장 시스템 의도에 따라 디지털 어시스턴트 또는 마스터 봇 자체에 의해 핸들링될지, 또는 현재 대화 흐름 내의 상이한 상태로서 핸들링될지 여부를 결정하기 위해 라우팅 모델을 사용한다.

[0096] 특정 실시예들에서, 이러한 프로세싱의 부분으로서, 디지털 어시스턴트는, 사용자 입력 발화가 그 호출 명칭을 사용하여 스킬 봇을 명식적으로 식별하는지 여부를 결정한다. 호출 명칭이 사용자 입력에 존재하는 경우, 이는 호출 명칭에 대응하는 스킬 봇의 명시적 호출로서 처리된다. 이러한 시나리오에서, 디지털 어시스턴트는 사용자 입력을 추가적인 핸들링을 위해 명시적으로 호출된 스킬 봇으로 라우팅할 수 있다. 특정 또는 명시적 호출이 없는 경우, 특정 실시예들에서, 디지털 어시스턴트는 수신된 사용자 입력 발화를 평가하고, 디지털 어시스턴트와 연관된 스킬 봇들 및 시스템 의도들에 대한 신뢰도 스코어들을 계산한다. 스킬 봇 또는 시스템 의도에 대해 계산된 스코어는, 사용자 입력이 스킬 봇이 수행하도록 구성된 태스크를 나타내거나 또는 시스템 의도를 나타낼 가능성을 나타낸다. 임계 값(예를 들어, 신뢰도 임계 라우팅 파라미터)을 초과하는 연관된 계산된 신뢰도 스코어를 갖는 임의의 시스템 의도 또는 스킬 봇은 추가적인 평가를 위한 후보로서 선택된다. 그런 다음, 디지털 어시스턴트는, 식별된 후보들로부터, 사용자 입력 발화의 추가적인 핸들링을 위한 특정 시스템 의도 또는 스킬 봇을 선택한다. 특정 실시예들에서, 하나 이상의 스킬 봇들이 후보들로서 식별된 이후에, 이러한 후보 스킬들과 연관된 의도들은 (각각의 스킬에 대한 의도 모델에 따라) 평가되고, 신뢰도 스코어들이 각각의 의도에 대해 결정된다. 일반적으로, 임계 값(예를 들어, 70%)을 초과하는 신뢰도 스코어를 갖는 임의의 의도는 후보 의도로서 처리된다. 특정 스킬 봇이 선택되는 경우, 사용자 의도는 추가적인 프로세싱을 위해 해당 스킬 봇으로 라우팅된다. 시스템 의도가 선택되는 경우, 하나 이상의 액션들은 선택된 시스템 의도에 따라 마스터 봇 자체에 의해 수행된다.

[0097] 도 2는 특정 실시예들에 따른 마스터 봇(master bot; MB) 시스템(200)의 간략화된 블록도이다. MB 시스템(200)은 소프트웨어로만, 하드웨어로만, 또는 하드웨어와 소프트웨어의 조합으로 구현될 수 있다. MB 시스템(200)은 사전-프로세싱 서브시스템(210), 다중 의도 서브시스템(multiple intent subsystem; MIS)(220), 명시적 호출 서브시스템(explicit invocation subsystem; EIS)(230), 스킬 봇 호출기(240), 및 데이터 스토어(250)를 포함한다. 도 2에 도시된 MB 시스템(200)은 단지 마스터 봇 내의 구성요소들의 배열의 일 예이다. 당업자는 다수의 가능한 변형예들, 대안예들, 및 수정예들을 인식할 것이다. 예를 들어, 일부 구현예들에서, MB 시스템(200)은 도 2에 도시된 것보다 더 많거나 또는 더 적은 시스템들 또는 구성요소들을 가질 수 있거나, 2개 이상의 서브시스템들을 결합할 수 있거나, 또는 서브시스템들의 상이한 구성 또는 배열을 가질 수 있다.

[0098] 사전-프로세싱 서브시스템(210)은 사용자로부터 발화 "A"(202)를 수신하며, 언어 검출기(212) 및 언어 파서(214)를 통해 발화를 프로세싱한다. 이상에서 표시된 바와 같이, 발화는 오디오 또는 텍스트를 포함하여 다양한 방식들로 제공될 수 있다. 발화(202)는 문장 단편, 완전한 문장, 다수의 문장들, 및 유사한 것일 수 있다. 발화(202)는 구두점을 포함할 수 있다. 예를 들어, 발화(202)가 오디오로서 제공되는 경우, 사전-프로세싱 서브시스템(210)은, 구두점 마크들, 예를 들어, 콤마들, 세미콜론들, 마침표들을 결과적인 텍스트에 삽입하는 스피치-대-텍스트 변환기(미도시)를 사용하여 오디오를 텍스트로 변환한다.

[0099] 언어 검출기(212)는 발화(202)의 텍스트에 기초하여 발화(202)의 언어를 검출한다. 발화(202)가 핸들링되는 방식은, 각각의 언어가 그 자체의 문법 및 의미를 갖기 때문에 언어에 의존한다. 언어들 사이의 차이점들은, 발화의 신택스 및 구조를 분석할 때 고려된다.

[0100] 언어 파서(214)는 발화(202)에서 개별적인 언어학적 단위들(예를 들어, 단어들)에 대해 품사(part of speech;

POS) 태그들을 추출하기 위해 발화(202)를 파싱한다. POS 태그들은, 예를 들어, 명사(noun; NN), 대명사(pronoun; PN), 동사(verb; VB), 및 유사한 것을 포함한다. 언어 파서(214)는 또한 (예를 들어, 각각의 단어를 별개의 토큰으로 변환하기 위해) 발화(202)의 언어학적 단위들을 토큰화하고 단어들을 레마타이징(lemmatize)할 수 있다. 레마(lemma)는 사전에 표시되는 바와 같은 단어들의 세트의 기본 형태이다(예를 들어, "run"은 run, runs, ran, running, 등에 대한 레마이다). 언어 파서(214)가 수행할 수 있는 다른 유형들의 사전-프로세싱은, 복합 표현들의 청킹, 예를 들어, "신용(credit)" 및 "카드(card)"를 단일 표현 "신용_카드(credit_card)"로 결합하는 것을 포함한다. 언어 파서(214)는 또한 발화(202) 내의 단어들 사이의 관계들을 식별할 수 있다. 예를 들어, 일부 실시예들에서, 언어 파서(214)는, 발화의 어떤 부분(예를 들어, 특정 명사)이 직접 목적어인지, 발화의 어떤 부분이 전치사인지, 등을 나타내는 종속성 트리를 생성한다. 언어 파서(214)에 의해 수행된 프로세싱의 결과들은 추출된 정보(205)를 형성하고, 발화(202) 자체와 함께 MIS(220)에 대한 입력으로서 제공된다.

[0101] 이상에서 표시된 바와 같이, 발화(202)는 2개 이상의 문장을 포함할 수 있다. 다수의 의도들 및 명시적 호출을 검출하는 목적들을 위해, 발화(202)는, 발화가 다수의 문장들을 포함하는 경우 단일 유닛으로서 처리될 수 있다. 그러나, 특정 실시예들에서, 사전-프로세싱은, 예를 들어, 다중 의도 분석 및 명시적 호출 분석을 위해 다수의 문장들 중에서 단일 문장을 식별하기 위해 사전-프로세싱 서브시스템(210)에 의해 수행될 수 있다. 일반적으로, MIS(220) 및 EIS(230)에 의해 생성된 결과들은, 발화(202)가 개별적인 문장의 레벨로 프로세싱되거나 또는 다수의 문장들을 포함하는 단일 유닛으로서 프로세싱되는지 여부와 무관하게 실질적으로 동일하다.

[0102] MIS(220)는, 발화(202)가 다수의 의도들을 나타내는지 여부를 결정한다. MIS(220)가 발화(202)에서 다수의 의도들의 존재를 검출할 수 있지만, MIS(220)에 의해 수행되는 프로세싱은, 발화(202)의 의도들이 봇에 대해 구성된 임의의 의도들에 매칭되는지 여부를 결정하는 것을 수반하지 않는다. 그 대신, 발화(202)의 의도가 봇 의도와 매칭되는지 여부를 결정하기 위한 프로세싱은, MIB 시스템(200)의 의도 분류기(242)에 의해 또는 (예를 들어, 도 3의 실시예에 도시된 바와 같은) 스킬 봇의 의도 분류기에 의해 수행될 수 있다. MIS(220)에 의해 수행되는 프로세싱은, 발화(202)를 핸들링할 수 있는 봇(예를 들어, 특정 스킬 봇 또는 마스터 봇 자체)가 존재한다고 가정한다. 따라서, MIS(220)에 의해 수행되는 프로세싱은, 챗봇 시스템에 어떤 봇들이 있는지에 대한 지식(예를 들어, 마스터 봇에 등록된 스킬 봇들을 신원(identity)들) 또는 특정 봇에 대해 어떤 의도들이 구성되었는지에 대한 지식을 필요로 하지 않는다.

[0103] 발화(202)가 다수의 의도들을 포함한다는 것을 결정하기 위해, MIS(220)는 데이터 스토어(250) 내의 규칙들의 세트(252)로부터 하나 이상의 규칙들을 적용한다. 발화(202)에 적용되는 규칙들은 발화(202)의 언어에 의존하며, 다수의 의도들의 존재를 나타내는 문장 패턴들을 포함할 수 있다. 예를 들어, 문장 패턴은 문장의 2개의 부분들을 연결하는 등위 접속사(예를 들어, 접속사)를 포함할 수 있으며, 여기서 2개의 부분들 모두는 별도의 의도에 대응한다. 발화(202)가 문장 패턴과 매칭되는 경우, 발화(202)가 다수의 의도들을 나타낸다는 것이 추론될 수 있다. 다수의 의도들을 갖는 발화가 반드시 상이한 의도들(예를 들어, 상이한 봇들을 향한 의도들 또는 동일한 봇 내의 상이한 의도들을 향한 의도들)을 가져야 하는 것은 아님을 유의해야 한다. 그 대신, 발화는 동일한 의도의 별도의 인스턴스들, 예를 들어, "결제 계좌 X를 사용하여 피자를 주문하고, 그런 다음 결제 계좌 Y를 사용하여 피자를 주문해(Place a pizza order using payment account X, then place a pizza order using payment account Y)"를 가질 수 있다.

[0104] 발화(202)가 다수의 의도들을 나타낸다고 결정하는 것의 부분으로서, MIS(220)는 또한 발화(202)의 어떤 부분이 각각의 의도와 연관되는지를 결정한다. MIS(220)는, 다수의 의도들을 포함하는 발화에서 표현된 각각의 의도에 대해, 도 2에 도시된 바와 같이, 원래의 발화 대신에 별도의 프로세싱을 위한 새로운 발화, 예를 들어, 발화 "B"(206) 및 발화 "C"(208)를 구성한다. 따라서, 원래의 발화(202)는 한번에 하나씩 핸들링되는 2개 이상의 별도의 발화들로 분할될 수 있다. MIS(220)는, 추출된 정보(205)를 사용하여 및/또는 발화(202) 자체의 분석으로부터, 2개 이상의 발화들 중 어떤 것이 먼저 핸들링되어야 하는지를 결정한다. 예를 들어, MIS(220)는, 발화(202)가 특정 의도가 먼저 핸들링되어야 한다는 것을 나타내는 마커 단어를 포함한다는 것을 결정할 수 있다. 이러한 특정 의도에 대응하는 새롭게 형성된 발화(예를 들어, 발화(206) 또는 발화(208) 중 하나)는 EIS(230)에 의한 추가적인 프로세싱을 위해 가장 먼저 전송될 것이다. 제1 발화에 의해 트리거된 대화가 종료된 이후에(또는 일시적으로 중단된 이후에), 그런 다음 다음으로 높은 우선순위의 발화(예를 들어, 발화(206) 또는 발화(208) 중 다른 하나)가 프로세싱을 위해 EIS(230)로 전송될 수 있다.

[0105] EIS(230)는, 이것이 수신한 발화(예를 들어, 발화(206) 또는 발화(208))가 스킬 봇의 호출 명칭을 포함하는지 여부를 결정한다. 특정 실시예들에서, 챗봇 시스템 내의 각각의 스킬 봇에는, 스킬 봇을 챗봇 시스템 내의 다른 스킬 봇들과 구별하는 고유 호출 명칭이 할당된다. 호출 명칭들의 리스트는 데이터 스토어(250) 내의 스킬 봇

정보(254)의 부분으로서 유지될 수 있다. 발화가 호출 명칭과 매칭되는 단어를 포함할 때 발화는 명시적 호출인 것으로 여겨진다. 봇이 명시적으로 호출되지 않는 경우, EIS(230)에 의해 수신된 발화는 비-명시적 호출 발화(234)로서 여겨지며, 발화를 핸들링하기 위해 사용할 봇을 결정하기 위해 마스터 봇의 의도 분류기(예를 들어, 의도 분류기(242))로 입력된다. 일부 경우들에서, 의도 분류기(242)는, 마스터 봇이 비-명시적 호출 발화를 핸들링해야 한다는 것을 결정할 것이다. 다른 경우들에서, 의도 분류기(242)는 핸들링을 위해 발화를 라우팅할 스킬 봇을 결정할 것이다.

[0106] EIS(230)에 의해 제공되는 명시적 호출 기능성은 몇몇 이점들을 갖는다. 이는, 마스터 봇이 수행해야 할 프로세싱의 양을 감소시킬 수 있다. 예를 들어, 명시적 호출이 있을 때, 마스터 봇은 (예를 들어, 의도 분류기(242)를 사용하는) 임의의 의도 분류 분석을 수행할 필요가 없을 수 있거나, 또는 스킬 봇을 선택하기 위한 감소된 의도 분류 분석을 가져야 할 수 있다. 따라서, 명시적 호출 분석은 의도 분류 분석에 의존하지 않고 특정 스킬 봇의 선택을 가능하게 할 수 있다.

[0107] 또한, 다수의 스킬 봇들 사이의 기능성들의 중첩이 존재하는 상황들이 있을 수 있다. 이는, 예를 들어, 2개의 스킬 봇들에 의해 핸들링되는 의도들이 중첩되거나 또는 서로 매우 가까운 경우에 발생할 수 있다. 이러한 상황에서, 마스터 봇이 의도 분류 분석만에 기초하여 다수의 스킬 봇들 중 어떤 것을 선택할지 식별하기는 어려울 수 있다. 이러한 시나리오에서, 명시적 호출은 사용될 특정 스킬 봇을 명확하게 한다.

[0108] 발화가 명시적 호출이라고 결정하는 것에 추가하여, EIS(230)는, 발화의 임의의 부분이 명시적으로 호출된 스킬 봇에 대한 입력으로서 사용되어야 하는지 여부를 결정하는 것을 담당한다. 특히, EIS(230)는, 발화의 부분이 호출과 연관되지 않는지 여부를 결정할 수 있다. EIS(230)는 발화의 분석 및/또는 추출된 정보(205)의 분석을 통해 이러한 결정을 수행할 수 있다. EIS(230)는, EIS(230)에 의해 수신된 발화 전체를 전송하는 대신에 호출과 연관되지 않은 발화의 부분을 호출된 스킬 봇으로 전송할 수 있다. 일부 경우들에서, 호출된 스킬 봇에 대한 입력은 단순히 호출과 연관된 발화의 임의의 부분을 제거함으로써 형성된다. 예를 들어, "나는 피자 봇을 사용하여 피자를 주문하고 싶어(I want to order pizza using Pizza Bot)"는, "피자 봇을 사용하여(using Pizza Bot)"가 피자 봇의 호출과 관련되지만 피자 봇에 의해 수행될 임의의 프로세싱과는 무관하기 때문에 "나는 피자를 주문하고 싶어(I want to order pizza)"로 단축될 수 있다. 일부 경우들에서, EIS(230)는, 예를 들어, 완전한 문장을 형성하기 위해 호출된 봇으로 전송될 부분을 재포맷팅(reformat)할 수 있다. 따라서, EIS(230)는, 명시적 호출이 있다는 것뿐만 아니라, 명시적 호출이 있을 때 스킬 봇으로 어떤 것을 전송할지 결정한다. 일부 경우들에서, 호출되는 봇에 입력할 임의의 텍스트가 없을 수 있다. 예를 들어, 발화가 "피자 봇(Pizza Bot)"이었던 경우, EIS(230)는, 피자 봇이 호출되지만 피자 봇에 의해 프로세싱될 텍스트가 없다는 것을 결정할 수 있다. 이러한 시나리오들에서, EIS(230)는, 전송할 것이 없다는 것을 스킬 봇 호출기(240)에 표시할 수 있다.

[0109] 스킬 봇 호출기(240)는 다양한 방식들로 스킬 봇을 호출한다. 예를 들어, 스킬 봇 호출기(240)는, 특정 스킬 봇이 명시적 호출의 결과로서 선택되었다는 표시(235)를 수신하는 것에 응답하여 봇을 호출할 수 있다. 표시(235)는 명시적으로 호출된 스킬 봇에 대한 입력과 함께 EIS(230)에 의해 전송될 수 있다. 이러한 시나리오에서, 스킬 봇 호출기(240)는 대화의 제어를 명시적으로 호출된 스킬 봇으로 넘길 것이다. 명시적으로 호출된 스킬 봇은, 입력을 독립형 발화로서 처리함으로써 EIS(230)로부터의 입력에 대한 적절한 응답을 결정할 것이다. 예를 들어, 응답은 특정 상태에서 새로운 대화를 시작하거나 또는 특정 액션을 수행하는 것일 수 있으며, 여기서 새로운 대화의 초기 상태는 EIS(230)로부터 전송된 입력에 의존한다.

[0110] 스킬 봇 호출기(240)가 스킬 봇을 호출할 수 있는 다른 방식은 의도 분류기(242)를 사용하는 암시적 호출을 통하는 것이다. 의도 분류기(242)는, 발화가 특정 스킬 봇이 수행하도록 구성된 태스크를 나타낼 가능성을 결정하도록, 기계-학습 및/또는 규칙-기반 트레이닝 기술들을 사용하여, 트레이닝될 수 있다. 의도 분류기(242)는 각각의 스킬 봇에 대해 하나의 클래스씩 상이한 클래스들에 대해 트레이닝된다. 예를 들어, 새로운 스킬 봇이 마스터 봇에 등록될 때마다, 새로운 스킬 봇과 연관된 예시적인 발화들의 리스트는, 특정 발화가 새로운 스킬 봇이 수행할 수 있는 태스크를 나타낼 가능성을 결정하도록 의도 분류기(242)를 트레이닝시키기 위해 사용될 수 있다. 이러한 트레이닝의 결과로서 생성된 파라미터들(예를 들어, 기계-학습 모델의 파라미터들에 대한 값들의 세트)는 스킬 봇 정보(254)의 부분으로서 저장될 수 있다.

[0111] 특정 실시예들에서, 의도 분류기(242)는, 본 명세에서 추가로 상세하게 설명되는 바와 같이, 기계-학습 모델을 사용하여 구현된다. 기계-학습 모델의 트레이닝은, 기계-학습 모델의 출력으로서, 어떤 봇이 임의의 특정 트레이닝 발화를 핸들링하기 위한 정확한 봇인지에 대한 추론을 생성하기 위해 다양한 스킬 봇들과 연관된 예시적인 발화들로부터의 발화들의 적어도 서브세트를 입력하는 것을 수반할 수 있다. 각각의 트레이닝 발화에 대해, 트

레이닝 발화에 대해 사용할 정확한 봇의 표시는 그라운드 트루스(ground truth) 정보로서 제공될 수 있다. 그런 다음, 기계-학습 모델의 거동은 생성된 추론들과 그라운드 트루스 정보 사이의 차이를 최소화하기 위해 (예를 들어, 역-전파를 통해) 적응될 수 있다.

[0112] 특정 실시예들에서, 의도 분류기(242)는, 마스터 봇에 등록된 각각의 스킬 봇에 대해, 스킬 봇이 발화(예를 들어, EIS(230)로부터 수신된 비-명시적 호출 발화(234))를 핸들링할 수 있는 가능성을 나타내는 신뢰도 스코어를 결정한다. 의도 분류기(242)는 또한 구성된 각각의 시스템 레벨 의도(예를 들어, 도움, 종료)에 대한 신뢰도 스코어를 결정할 수 있다. 특정 신뢰도 스코어가 하나 이상의 기준들을 충족시키는 경우, 스킬 봇 호출기(240)는 특정 신뢰도 스코어와 연관된 봇을 호출할 것이다. 예를 들어, 임계 신뢰도 스코어 값이 충족되어야 할 수 있다. 따라서, 의도 분류기(242)의 출력(245)은 특정 스킬 봇의 식별(identification) 또는 시스템 의도의 식별이다. 일부 실시예들에서, 임계 신뢰도 스코어 값을 충족시키는 것에 더하여, 신뢰도 스코어는 특정 승리 마진만큼 다음으로 가장 높은 신뢰도 스코어를 초과해야 한다. 이러한 조건을 부과하는 것은, 다수의 스킬 봇들의 신뢰도 스코어들이 각각 임계 신뢰도 스코어 값을 초과할 때 특정 스킬 봇으로의 라우팅을 가능하게 할 것이다.

[0113] 신뢰도 스코어들의 평가에 기초하여 봇을 식별한 이후에, 스킬 봇 호출기(240)는 프로세싱을 식별된 봇으로 핸드 오버(hand over)한다. 시스템 의도의 경우에, 식별된 봇은 마스터 봇이다. 그렇지 않으면, 식별된 봇은 스킬 봇이다. 또한, 스킬 봇 호출기(240)는 식별된 봇에 대한 입력(247)으로서 어떤 것을 제공할지 결정할 것이다. 이상에서 표시된 바와 같이, 명시적 호출의 경우에, 입력(247)은 호출과 연관되지 않은 발화의 일 부분에 기초할 수 있거나, 또는 입력(247)은 아무 것도 아닌 것(nothing)(예를 들어, 빈 문자열)일 수 있다. 암시적 호출의 경우에, 입력(247)은 전체 발화일 수 있다.

[0114] 데이터 스토어(250)는 마스터 봇 시스템(200)의 다양한 서브시스템들에 의해 사용되는 데이터를 저장하는 하나 이상의 컴퓨팅 디바이스들을 포함한다. 이상에서 설명된 바와 같이, 데이터 스토어(250)는 규칙들(252) 및 스킬 봇 정보(254)를 포함한다. 규칙들(252)은, 예를 들어, MIS(220)에 의해, 발화가 다수의 의도들을 나타낼 때 및 다수의 의도들을 나타내는 발화를 분할하기 위한 방법을 결정하기 위한 규칙들을 포함한다. 규칙들(252)은, EIS(230)에 의해, 스킬 봇을 명시적으로 호출하는 발화의 어떤 부분들을 스킬 봇으로 전송할지를 결정하기 위한 규칙들을 더 포함한다. 스킬 봇 정보(254)는 챗봇 시스템 내의 스킬 봇들의 호출 명칭들, 예를 들어, 특정 마스터 봇에 등록된 모든 스킬 봇들의 호출 명칭들의 리스트를 포함한다. 스킬 봇 정보(254)는 또한, 챗봇 시스템 내의 각각의 스킬 봇에 대한 신뢰도 스코어, 예를 들어, 기계-학습 모델의 파라미터들을 결정하기 위해 의도 분류기(242)에 의해 사용되는 정보를 포함할 수 있다.

[0115] 도 3은 특정 실시예들에 따른 스킬 봇 시스템(300)의 간략화된 블록도이다. 스킬 봇 시스템(300)은, 소프트웨어로만, 하드웨어로만, 또는 하드웨어와 소프트웨어의 조합으로 구현될 수 있는 컴퓨팅 시스템이다. 도 1에 도시된 실시예와 같은 특정 실시예들에서, 스킬 봇 시스템(300)은 디지털 어시스턴트 내에 하나 이상의 스킬 봇들을 구현하기 위해 사용될 수 있다.

[0116] 스킬 봇 시스템(300)은 MIS(310), 의도 분류기(320), 및 대화 관리자(330)를 포함한다. MIS(310)는 도 2의 MIS(220)와 유사하며, 데이터 스토어(350)의 규칙들(352)을 사용하여: (1) 발화가 다수의 의도들을 나타내는지 여부 및, 그런 경우 (2) 발화를 복수의 의도들의 각각의 의도에 대한 별도의 발화로 분할하는 방법을 결정하도록 동작하는 것을 포함하여, 유사한 기능성을 제공한다. 특정 실시예들에서, 다수의 의도들을 검출하기 위해 또는 발화를 분할하기 위해 MIS(310)에 의해 적용되는 규칙들은 MIS(220)에 의해 적용되는 규칙들과 동일하다. MIS(310)는 발화(302) 및 추출된 정보(304)를 수신한다. 추출된 정보(304)는 도 1의 추출된 정보(205)와 유사하며, 언어 파서(214) 또는 스킬 봇 시스템(300)에 로컬인 언어 파서를 사용하여 생성될 수 있다.

[0117] 의도 분류기(320)는 도 2의 실시예와 관련하여 이상에서 논의되고 본 명세서에서 추가로 상세하게 설명되는 바와 같은 의도 분류기(242)와 유사한 방식으로 트레이닝될 수 있다. 예를 들어, 특정 실시예들에서, 의도 분류기(320)는 기계-학습 모델을 사용하여 구현된다. 의도 분류기(320)의 기계-학습 모델은, 트레이닝 발화들로서 특정 스킬 봇과 연관된 예시적인 발화들의 적어도 서브세트를 사용하여 특정 스킬 봇에 대해 트레이닝된다. 각각의 트레이닝 발화에 대한 그라운드 트루스는 트레이닝 발화와 연관된 특정 봇 의도일 것이다.

[0118] 발화(302)는 사용자로부터 직접적으로 수신되거나 또는 마스터 봇을 통해 공급될 수 있다. 발화(302)가, 예를 들어, 도 2에 도시된 실시예에서 MIS(220) 및 EIS(230)의 프로세싱의 결과로서 마스터 봇을 통해 공급될 때, MIS(310)는 MIS(220)에 의해 이미 수행된 프로세싱을 반복하는 것을 피하기 위해 바이패스될 수 있다. 그러나, 발화(302)가, 예를 들어, 스킬 봇으로의 라우팅 이후에 발생하는 대화를 통해 사용자로부터 직접적으로 수신되는 경우, MIS(310)는 발화(302)가 다수의 의도들을 나타내는지 여부를 결정하기 위해 발화(302)를 프로세싱할

수 있다. 그런 경우, MIS(310)는 발화(302)를 각각의 의도에 대한 별도의 발화, 예를 들어, 발화 "D"(306) 및 발화 "E"(308)로 분할하기 위해 하나 이상의 규칙들을 적용한다. 발화(302)가 다수의 의도들을 나타내지 않는 경우, MIS(310)는 발화(302)를 분할하지 않고 발화(302)를 의도 분류를 위해 의도 분류기(320)로 포워딩한다.

[0119] 의도 분류기(320)는 수신된 발화(예를 들어, 발화(306 또는 308))를 스킬 봇 시스템(300)과 연관된 의도에 매칭시키도록 구성된다. 이상에서 설명된 바와 같이, 스킬 봇은 하나 이상의 의도들로 구성될 수 있으며, 각각의 의도는 의도와 연관되며 분류기를 트레이닝시키기 위해 사용되는 적어도 하나의 예시적인 발화를 포함한다. 도 2의 실시예에서, 마스터 봇 시스템(200)의 의도 분류기(242)는 개별적인 스킬 봇에 대한 신뢰도 스코어들 및 시스템 의도들에 대한 신뢰도 스코어를 결정하도록 트레이닝된다. 유사하게, 의도 분류기(320)는 스킬 봇 시스템(300)과 연관된 각각의 의도에 대한 신뢰도 스코어를 결정하도록 트레이닝될 수 있다. 의도 분류기(242)에 의해 수행되는 분류는 봇 레벨인 반면, 의도 분류기(320)에 의해 수행되는 분류는 의도 레벨이고 따라서 더 미세하게 세분화된다. 의도 분류기(320)는 의도 정보(354)에 대한 액세스를 갖는다. 의도 정보(354)는, 스킬 봇 시스템(300)과 연관된 각각의 의도에 대해, 의도의 의미를 나타내고 예시하며 전형적으로 해당 의도에 의해 수행가능한 태스크와 연관되는 발화들의 리스트를 포함한다. 의도 정보(354)는 이러한 의도들의 리스트에 대한 트레이닝의 결과로서 생성된 파라미터들을 더 포함할 수 있다.

[0120] 대화 관리자(330)는, 의도 분류기(320)의 출력으로서, 의도 분류기(320)에 입력된 의도와 최상으로 매칭되는 것으로서 의도 분류기(320)에 의해 식별된 특정 의도의 표시(322)를 수신한다. 일부 경우들에서, 의도 분류기(320)는 어떠한 매칭되 결정할 수 없다. 예를 들어, 의도 분류기(320)에 의해 계산된 신뢰도 스코어들은, 의도가 시스템 의도 또는 상이한 스킬 봇의 의도에 관한 것인 경우 임계 신뢰도 스코어 아래로 떨어질 수 있다. 이러한 상황이 발생할 때, 스킬 봇 시스템(300)은, 예를 들어, 상이한 스킬 봇으로 라우팅하기 위해, 핸들링을 위해 발화를 마스터 봇에 위탁(refer)할 수 있다. 그러나, 의도 분류기(320)가 스킬 봇 내에서 의도를 식별하는 데 성공한 경우, 대화 관리자(330)는 사용자와의 대화를 개시할 것이다.

[0121] 대화 관리자(330)에 의해 개시된 대화는 의도 분류기(320)에 의해 식별된 의도에 특정한 대화이다. 예를 들어, 대화 관리자(330)는 식별된 의도에 대한 다이얼로그 흐름을 실행하도록 구성된 상태 머신을 사용하여 구현될 수 있다. 상태 머신은 (예를 들어, 의도가 임의의 추가적인 입력 없이 호출되는 경우에 대한) 디폴트 시작 상태 및 하나 이상의 추가적인 상태들을 포함할 수 있으며, 여기서 각각의 상태는 스킬 봇에 의해 수행될 액션들(예를 들어, 구매 트랜잭션을 실행하는 것) 및/또는 사용자에게 표시될 다이얼로그(예를 들어, 질문들, 응답들)과 연관되었다. 따라서, 대화 관리자(330)는 의도를 식별하는 표시(322)의 수신 시에 액션/다이얼로그(335)를 결정할 수 있으며, 대화 동안 수신되는 후속 발화들에 응답하여 추가적인 액션들 또는 다이얼로그를 결정할 수 있다.

[0122] 데이터 스토어(350)는 스킬 봇 시스템(300)의 다양한 서브시스템들에 의해 사용되는 데이터를 저장하는 하나 이상의 컴퓨팅 디바이스들을 포함한다. 도 3에 도시된 바와 같이, 데이터 스토어(350)는 규칙들(352) 및 의도 정보(354)를 포함한다. 특정 실시예들에서, 데이터 스토어(350)는 마스터 봇 또는 디지털 어시스턴트의 데이터 스토어(250), 예를 들어, 도 2의 데이터 스토어에 통합될 수 있다.

[0123] **미세-튜닝**

[0124] 다중-헤드 어텐션 기술을 구현하는 계층들의 불완전한 서브세트를 트레이닝시킴으로써 새로운 컨텍스트에 대해 기계-학습 모델(예를 들어, NER)을 미세-튜닝하는 것은, 모든 파라미터들이 미세-튜닝 동안 구성가능했던 비교할 만한 트레이닝의 효율과 비교하여 (예를 들어, 트레이닝 시간, 트레이닝을 위한 계산 자원들의 사용, 미세-튜닝에서 사용하기 위한 데이터 세트를 수집하는 시간, 및/또는 미세-튜닝에서 사용하기 위한 계산 자원들의 사용과 관련하여) 트레이닝의 효율을 극적으로 개선할 수 있다. 또한, 그럼에도 불구하고, 미세-튜닝된 모델의 정확도는 새로운 컨텍스트에 대응하는 클라이언트 데이터에 기초하여 모든 파라미터들을 트레이닝한 비교할 만한 모델과 긍정적으로 비교할 만할 수 있다(예를 들어, 비교할 만한 모델과 매칭될 수 있거나, 이의 적어도 90%일 수 있거나, 이의 적어도 80%일 수 있거나, 이의 적어도 70%일 수 있거나, 이의 적어도 60%일 수 있거나, 또는 이의 적어도 50%일 수 있다). 미세-튜닝될 기계-학습 모델 및/또는 미세-튜닝된 기계-학습 모델은 도 1, 도 2 및 도 3을 참조하여 설명된 바와 같은 채팅 시스템에 구현될 수 있다.

[0125] 도 4는, 텍스트 데이터에 기초하여, 예를 들어, 개체명 인식기로서 구현된, 하나 이상의 모델들을 트레이닝시키도록 구성된 챗봇 시스템(400)의 측면들을 예시하는 블록도이다. 도 4에 도시된 바와 같이, 이러한 예에서 챗봇 시스템(400)에 의해 수행되는 개체 인식은 다양한 스테이지들을 포함한다: (예측 모델 내의 각각의 파라미터가 학습될 수 있도록) 예측 모델을 구축하고 트레이닝시키는 사전-트레이닝 스테이지(402), (예측 모델의 파라미터들 중 적어도 일부의 각각이 수정될 수 있도록) 예측 모델을 미세-튜닝하는 미세-튜닝 스테이지(404) 및 하나

이상의 챗봇들을 구현하기 위한 챗봇 구현 스테이지(406). 따라서, 사전-트레이닝 스테이지(402)에서, 사전-트레이닝된 기계-학습 모델(408)이 생성되며; 미세-튜닝 스테이지(404)에서, 사전-트레이닝된 기계-학습 모델(408)이 미세-튜닝되어 미세-튜닝된 기계-학습 모델(410)을 생성하고; 그리고 챗봇 구현 스테이지(406)에서, 미세-튜닝된 기계-학습 모델(410)은 수신된 텍스트 데이터(412)를 프로세싱하는 하나 이상의 챗봇들을 지원하기 위해 사용된다.

[0126] 사전-트레이닝된 기계-학습 모델(408) 및 미세-튜닝된 기계-학습 모델(410) 각각은 (예를 들어) 발화에서 하나 이상의 개체들을 인식하도록 구성되고 사용되는 개체명 인식기 모델일 수 있다. 챗봇 구현 스테이지에서, 미세-튜닝된 기계-학습 모델(410)은, 하나 이상의 다른 모델들, 예컨대 발화가 특정 스킬 봇이 수행하도록 구성된 태스크를 나타낼 가능성을 결정하기 위한 다른 모델, 제1 유형의 스킬 봇에 대한 발화로부터의 의도를 예측하기 위한 다른 모델, 및/또는 제2 유형의 스킬 봇에 대한 발화로부터의 의도를 예측하기 위한 다른 모델과 조합되어 사용될 수 있다.

[0127] 사전-트레이닝된 기계-학습 모델(408) 및 미세-튜닝된 기계-학습 모델(410) 각각은 다음일 수 있거나 및/또는 다음을 포함할 수 있다: 트랜스포머, 다중-헤드 어텐션 기술을 사용하는 모델, 컨볼루션 신경망(convolutional neural network; "CNN"), 예를 들어, 인셉션(inception) 신경망, 잔차 신경망(residual neural network; "Resnet"), 순환 신경망, 예를 들어, 장단기 기억(long short-term memory; "LSTM") 모델들 또는 게이트 순환 유닛(gated recurrent unit; "GRU")들 모델들, 및/또는 심층 신경망(Deep Neural Network; "DNN")들의 다른 변형들(예를 들어, 단일 의도 분류를 위한 다중-라벨 n-바이너리 DNN 분류기 또는 다중-클래스 DNN 분류기).

[0128] 사전-트레이닝된 기계-학습 모델(408) 및 미세-튜닝된 기계-학습 모델(410) 각각은, 나이브 베이즈 분류기, 선형 분류기, 지원 벡터 머신, 배깅 모델들 예컨대 랜덤 포레스트 모델, 부스팅 모델들, 얕은 신경망(Shallow Neural Network)들, 또는, 이러한 기술들 중 하나 이상의 조합들, 예를 들어, CNN-HMM 또는 MCNN(Multi-Scale Convolutional Neural Network)과 같은 자연어 프로세싱을 위해 트레이닝된 임의의 다른 적절한 기계-학습 모델일 수 있거나 또는 이를 포함할 수 있다.

[0129] 사전-트레이닝된 기계-학습 모델(408) 및 미세-튜닝된 기계-학습 모델(410) 각각은, 다수의 계층들, 다중-헤드 어텐션 기술을 사용하는 다수의 계층들, 트랜스포머 모델을 포함하는 다수의 계층들, 다중-헤드 셀프-어텐션 기술을 사용하는 다수의 계층들을 포함할 수 있다. 다중-헤드 셀프-어텐션 기술에 대해, 다수의 헤드들의 각각은 상이한 쿼리, 키, 및 값 선형 투영들을 사용하여 셀프-어텐션 기술을 수행한다.

[0130] 사전-트레이닝된 기계-학습 모델(408) 및 미세-튜닝된 기계-학습 모델(410) 각각은 셀프-어텐션 기술을 사용하는 다수의 계층들을 포함할 수 있다. 셀프-어텐션 기술은 주어진 입력을 쿼리 표현(쿼리 선형 투영을 사용하여 생성됨), 키 표현(키 선형 투영에 의해 생성됨), 및 값 표현(값 선형 투영에 의해 생성됨) 각각으로 변환할 수 있다. 다른 위치에서의 값을 평가할 때 주어진 위치에서의 값에 대해 얼마나 많은 어텐션이 기울어져야 하는지를 나타내는 어텐션 스코어들을 생성하기 위해 쿼리 표현과 키 표현을 곱한다(그리고 잠재적으로 소프트맥스 함수와 같은 활성화 함수를 적용한다). 그런 다음, 어텐션 스코어들은 출력을 생성하기 위해 값 표현과 곱해질 수 있다.

[0131] 일부 경우들에서, 사전-트레이닝된 학습 모델(408)의 아키텍처 및/또는 하이퍼파라미터들은 미세-튜닝된 기계-학습 모델(410)의 아키텍처 및/또는 하이퍼파라미터들과 동일하다.

[0132] 챗봇 시스템(400)은 발화에서 하나 이상의 개체들을 인식하기 위해, 발화가 특정 스킬 봇이 수행하도록 구성된 태스크를 나타낼 가능성을 결정하고, 제1 유형의 스킬 봇에 대한 발화로부터의 의도를 예측하며, 그리고 제2 유형의 스킬 봇에 대한 발화로부터의 의도를 예측하는, 동일한 유형의 예측 모델 또는 상이한 유형들의 예측 모델들을 이용할 수 있다, 또 다른 유형들의 예측 모델들은 본 개시내용에 따라 다른 예들에서 구현될 수 있다.

[0133] 사전-트레이닝 스테이지(402)에서, 시스템이 구축되고 있는 모델을 트레이닝시키고 테스트할 수 있도록 사전-트레이닝 데이터 자산(asset)들(414)이 로딩되어 트레이닝 세트와 검증 세트로 분할된다. 사전-트레이닝 자산들(414)을 트레이닝 세트와 검증 세트로 분할하는 것은 랜덤하게(예를 들어, 90/10% 또는 70/30%)하게 수행될 수 있거나, 또는 분할하는 것은 샘플링 편향 및 오버피팅(overfitting)을 최소화하기 위해 K-폴드 교차-검증(K-Fold Cross-Validation), 리브-원-아웃 교차-검증(Leave-one-out Cross-Validation), 리브-원-그룹-컷 교차-검증(Leave-one-group-out Cross-Validation), 네스티드 교차-검증(Nested Cross-Validation) 또는 유사한 것과 같은 더 복잡한 검증 기술들에 따라 수행될 수 있다.

[0134] 사전-트레이닝 데이터 자산들(414)은 적어도 하나 이상의 스킬 봇들과 연관된 예시적인 발화들로부터의 발화들

의 서브세트를 포함할 수 있다. 이상에서 표시된 바와 같이, 발화는 오디오 또는 텍스트를 포함하여 다양한 방식들로 제공될 수 있다. 발화는 문장 단편, 완전한 문장, 다수의 문장들, 및 유사한 것일 수 있다. 일부 경우들에서, 예시적인 발화들은 이전 또는 기존 클라이언트 또는 고객에 의해 제공된다. 다른 경우들에서, 예시적인 발화들은 발화들의 이전 라이브러리들(예를 들어, 챗봇이 학습하도록 설계된 스킬에 특정한 라이브러리로부터의 발화들을 식별함)로부터 자동으로 생성된다. 사전-트레이닝 데이터 자산들(414)은 입력 테스트 또는 오디오(또는 텍스트 또는 오디오 프레임들의 입력 특징들)를 포함할 수 있다.

[0135] 사전-트레이닝 스테이지들에서, 사전-트레이닝 라벨들(416)이 또한 획득될 수 있으며, 여기서 사전-트레이닝 라벨 각각은 개별적인 사전-트레이닝 데이터 자산에 대응한다. 사전-트레이닝 라벨들은 값들의 매트릭스 또는 테이블에 저장되거나 또는 이로서 저장될 수 있다. 각각의 트레이닝 발화에 대해, 정확한 개체들의 표시 및 이에 의해 추론될 이의 분류는 사전-트레이닝 라벨들(416)에 대한 그라운드 트루스 정보로서 제공될 수 있다. 그런 다음, 사전-트레이닝되고 있는 모델의 거동은 다양한 개체들에 대해 생성된 추론들과 그라운드 트루스 정보 사이의 차이를 최소화하기 위해 (예를 들어, 역-전파를 통해) 적응될 수 있다.

[0136] 사전-트레이닝은, 사전-트레이닝된 기계-학습 모델(408)에 대한 하이퍼파라미터들을 선택하는 것 및, 사전-트레이닝된 기계-학습 모델(408)에 대한, 목적 함수를 최대화하거나 또는 최소화하는, 예를 들어, 손실 함수를 최소화하는 모델 파라미터들(예를 들어, 가중치들 및/또는 편향(bias)들)의 세트를 찾기 위해 사전-트레이닝 데이터 자산들(414) 중 적어도 일부로부터의 발화들을 사전-트레이닝된 기계-학습 모델(408)에 입력하는 반복적 동작들을 수행하는 것을 포함할 수 있다. 하이퍼파라미터들은, 사전-트레이닝된 기계-학습 모델(408)의 거동을 제어하기 위해 튜닝되거나 또는 최적화될 수 있는 세팅들이다. 대부분의 모델들은 메모리 또는 실행 비용과 같은 모델들의 상이한 측면들을 제어하는 하이퍼파라미터들을 명시적으로 정의한다. 그러나, 추가적인 하이퍼파라미터들은 모델을 특정 시나리오에 맞춰 적응시키도록 정의될 수 있다. 예를 들어, 하이퍼파라미터들은 모델의 은닉 유닛들 또는 계층들의 수, 모델의 학습률, 컨볼루션 커널 폭, 또는 모델에 대한 파라미터들의 수를 포함할 수 있다. 트레이닝의 각각의 반복은, 모델 파라미터들의 세트를 사용하는 목적 함수의 값이 이전 반복에서 모델 파라미터들의 상이한 세트를 사용하는 목적 함수의 값보다 더 작도록 (하이퍼파라미터들의 정의된 세트로 구성된) 모델(408)에 대한 모델 파라미터들의 세트를 찾는 것을 수반할 수 있다. 목적 함수는, 라벨들(416)을 사용하여 증강된 사전-트레이닝 데이터 자산들(414)의 서브세트에 주석으로 달린 그라운드 트루스와 모델(408)을 사용하여 추론된 출력들 사이의 차이를 측정하도록 구성될 수 있다.

[0137] 사전-트레이닝은, 예측을 생성하기 위해 모델의 현재 버전을 사용하여 하나 이상의 사전-트레이닝 데이터 자산들 각각을 프로세싱하는 것, 데이터 자산에 대응하는 라벨 및 예측에 기초하여 손실을 계산하는 것, 및 손실에 기초하여 모델의 하나 이상의 파라미터들을 업데이트하는 것을 포함할 수 있다. 예측을 생성하기 위해, 다수의 중간 값들이 각각의 사전-트레이닝 데이터 자산에 대해 생성될 수 있다. 예를 들어, 중간 값은 모델 내의 계층으로부터의 출력을 포함할 수 있다. 각각의 계층은 다수의 출력들(예를 들어, 100개 초과, 1,000개 초과, 10,000개 초과, 또는 100,000개 초과 출력들)을 생성할 수 있다. 일부 경우들에서, 모델은 전체적으로, 입력이 동일하고, 모델의 하이퍼파라미터들의 동일하며, 모델의 파라미터들이 동일한 한 일관된 출력을 생성하도록 구성된다.

[0138] 모델을 미세-튜닝하는 것은, 모델의 전체 파라미터들 중 일부를 미세-튜닝하면서 모델의 다른 파라미터들은 트레이닝 동안 식별된 값들에 고정된 상태로 남겨두는 것을 포함할 수 있다. 예를 들어, 미세-튜닝은, 하나 이상의 특정 계층과 연관된 파라미터들을 조정하면서 하나 이상의 하위 계층들에 대한 파라미터 값들을 고정된 상태로 유지하는 것을 포함할 수 있다. 이러한 경우에, 모델이 학습 반복들에 걸쳐 상이한 하이퍼파라미터들(예를 들어, 난수 시드들)을 사용하지 않는 한, 하나 이상의 하위 계층들로부터의 출력은 학습 반복들에 걸쳐 주어진 입력 데이터 자산에 대해 동일해야 한다. 따라서, 일단 모델이 사전-트레이닝되면, 미세-튜닝될 각각의 계층 아래의 계층에 의한 출력들에 대응하는 하나 이상의 중간 값들은 (데이터 자산의 식별자 및/또는 대응하는 라벨과 연관되어) 각각의 사전-트레이닝 데이터 자산들에 대해 캐시 데이터 스토어(418)에 저장될 수 있다. 일부 경우들에서, 미세-튜닝될 적어도 하나의 계층의 아키텍처는 고정될 적어도 하나의 계층의 아키텍처와 동일하다.

[0139] 중간 값들을 캐싱하는 것은 다수의 컨텍스트들 또는 다수의 환경들 각각에 대한 모델의 미세-튜닝을 용이하게 하는 이점을 갖는다. 예를 들어, 모델은 주어진 언어(예를 들어, 영어)의 요소들을 학습하도록 사전-트레이닝될 수 있으며, 주어진 웹 사이트 또는 주어진 판매자(merchant)와 연관되어 수신된 텍스트 또는 오디오에 어떻게 응답할지를 학습하기 위해 여러 번 미세-튜닝될 수 있다.

[0140] 미세-튜닝 스테이지(404)에서, 캐싱된 중간 값들 및 미세-튜닝 라벨들(420)은 사전-트레이닝 스테이지(402) 동

안 초기에 학습된 파라미터들의 서브세트를 미세-튜닝하기 위해 사용된다. 예를 들어, 사전-트레이닝 데이터 자산은 텍스트: "나는 배송을 받지 못했습니다(I have not received my shipment)"를 포함할 수 있다. 사전-트레이닝된 모델에 의한 출력은 배송 상태에 대한 업데이트 요청에 대응할 수 있다. 모델 내의 하위 계층으로부터의 출력은 단어들 "제품(product)" "수신되지 않음(not received)"을 나타낼 수 있다. 미세-튜닝된 모델에 의한 출력은 가능한 한 빨리 배송을 완료하라는 요청에 대응할 수 있다.

[0141] 일부 경우들에서, 미세-튜닝은, 캐싱된 데이터가 입력 데이터 세트들 대신에 사용가능할 수 있다는 것을 고려하면, 기본 데이터 자산들(414)을 사용하지 않고 수행될 수 있다. 모델을 미세-튜닝하는 것은, 셀프-어텐션 기술 및/또는 다중-헤드 어텐션 기술을 사용하여 하나 이상의 계층들에 대한 파라미터들을 학습시키는 것을 포함할 수 있다.

[0142] 미세-튜닝된 모델(420)은 사전-트레이닝 스테이지(402) 동안 학습된 적어도 일부 파라미터들 및 미세-튜닝 스테이지(404) 동안 학습된 적어도 일부 파라미터들로 구성되도록 정의될 수 있다.

[0143] 미세-튜닝된 모델(420)은 개체 인식 모델일 수 있거나 또는 이의 부분일 수 있다. 미세-튜닝된 모델(420)은 새로운 텍스트 데이터(424)를 프로세싱하기 위해 구현 스테이지(406)에서 사용될 수 있다.

[0144] 구현 스테이지(406)에서, 미세-튜닝된 모델(420)이 배포되고 하나 이상의 챗봇들을 구현하는 트레이닝된 개체 인식 모델로서 또는 이의 부분으로서 사용될 수 있다. 예를 들어, 하나 이상의 챗봇들은, 1명 이상의 사용자들로부터 텍스트 데이터(424)를 수신하며 하나 이상의 챗봇들에 의해 수신된 다양한 발화들로부터 개체들을 인식하고 추출하기 위해 미세-튜닝된 모델(420)로 구성될 수 있다. 텍스트 데이터(424)는, 이에 대해 미세-튜닝 기계-학습 모델(422)이 미세-튜닝된 것에 대응하는 환경 또는 컨텍스트에서 수신된 텍스트 데이터를 포함할 수 있다. 개체들은 텍스트 데이터(460)로부터 획득된 추출된 정보(예를 들어, 도 2 및 도 3에서 각각 설명된 추출된 정보(205; 304))의 부분일 수 있으며, 의도 분류와 같은 하류 프로세싱에서 사용될 수 있다. 그런 다음, 추출된 개체들에 기초하여 생성된 출력은 발화의 소스에 대응하는 디바이스로 송신되거나 및/또는 그 디바이스에 표시될 수 있다.

[0145] **주어진 환경에 대해 미세-튜닝된 모델을 생성하고 사용하기 위한 기술들**

[0146] 도 5는 특정 실시예들에 따른 주어진 환경에 대한 미세-튜닝된 모델을 생성하고 사용하기 위한 프로세스(500)를 예시하는 흐름도이다. 도 5에 도시된 프로세싱은 개별적인 시스템들, 하드웨어, 또는 이들의 조합의 하나 이상의 프로세싱 유닛들(예를 들어, 프로세서들, 코어들)에 의해 실행되는 소프트웨어(예를 들어, 코드, 명령어들, 프로그램)로 구현될 수 있다. 소프트웨어는 비-일시적 저장 매체 상에(예를 들어, 메모리 디바이스 상에) 저장될 수 있다. 도 5에 표현되고 이하에서 설명되는 방법은 예시적으로 그리고 비-제한적으로 의도된다. 도 5가 특성 시퀀스 또는 순서로 발생하는 다양한 프로세싱 단계들을 도시하지만, 이는 제한적으로 의도되지 않는다. 특정한 대안적인 실시예들에서, 단계들은 어떤 상이한 순서로 수행될 수 있으며, 일부 단계들이 또한 병렬로 수행될 수 있다. 도 1 내지 도 4에 도시된 실시예에서와 같이, 특정 실시예들에서, 도 5에 도시된 프로세싱은 하나 이상의 예측 모델들(예를 들어, 의도 분류기(242 또는 320) 또는 예측 모델들(425))에 의해 트레이닝을 위한 키워드 증강형 데이터 세트들을 생성하기 위해 사전-프로세싱 서브시스템(예를 들어, 사전-프로세싱 서브시스템(210) 또는 예측 모델 트레이닝 스테이지(410))에 의해 수행될 수 있다.

[0147] 블록(505)에서, 기계-학습 모델의 사전-트레이닝된 버전이 액세스된다. 기계-학습 모델은 셀프-어텐션 기술을 구현하도록 구성된 다수의 계층들(예를 들어, 다중-헤드 어텐션 기술을 구현하도록 구성된 다수의 계층들)을 포함한다.

[0148] 블록(510)에서, 클라이언트 데이터 세트가 액세스된다. 클라이언트 데이터 세트는 라벨들의 세트를 포함할 수 있으며, 여기서 각각의 라벨은 기계-학습 모델을 트레이닝시키기 위해 사용되는 입력 데이터에 대응한다. 추가적으로 또는 대안적으로, 클라이언트 데이터 세트는 입력 데이트 세트들 및 라벨들의 쌍들을 포함하며, 여기서 입력 데이터 세트들 및 라벨들 각각은 클라이언트에 의해 제공되거나, 선택되거나 및/또는 식별된다.

[0149] 블록(515)에서, 기계-학습 모델 내의 계층들의 불완전한 서브세트는 클라이언트 데이터 세트를 사용하여 모델을 미세-튜닝하기 위해 식별된다. 계층들의 불완전한 서브세트는, 미세-튜닝 스테이지 동안 어떤 계층(들)이 미세-튜닝될 것인지를 나타내는 저장된 정보에 기초하여 식별될 수 있다. 일부 경우들에서, 계층들의 불완전한 서브세트는 클라이언트 데이터 세트의 크기, 클라이언트 디바이스로부터의 선택, 기계-학습 모델 내의 계층들의 수, 및/또는 모델을 사전-트레이닝시키기 위해 사용된 데이터 및/또는 라벨들에 대응하는 분포와 비교한 클라이언트 데이터 세트 내의 데이터 및/또는 라벨들에 대응하는 분포의 예비 비교에 기초하여 식별된다. 계층들의 불완전

한 서브세트는, 서브세트에 있지 않고 각각의 다른 계층 위에 있는 하나 이상의 계층들을 포함할 수 있다. 예를 들어, 계층들의 불완전한 서브세트는 모델 내의 계층 8 및 그 이상을 포함할 수 있으며, 반면 계층들 1-7은 서브세트에 있지 않다(여기서 모델은, 계층 7로부터의 출력이 계층 8로 공급되도록 구성된다).

[0150] 블록(520)에서, 기계-학습 모델은 모델의 업데이트되고 미세-튜닝된 버전을 생성하기 위해 클라이언트 데이터세트를 사용하여 미세-튜닝된다. 미세-튜닝 동안, 계층들의 불완전한 서브세트 내에 있지 않은 계층들과 연관된 파라미터 값들은 프리즈될 수 있으며(예를 들어, 그리고 초기 모델 구성에 대해 캐시로부터 검색될 수 있으며), 반면 계층들의 서브세트에 대한 파라미터 값들은 학습될 수 있다.

[0151] 블록(525)에서, 기계-학습 모델의 업데이트된 미세-튜닝된 버전의 사용이 가능해진다. 예를 들어, 기계-학습 모델의 업데이트된 버전은 발화를 의도 표현으로 변환하기 위해 및/또는 발화에 대한 응답을 식별하기 위해 사용될 수 있다.

[0152] **예시적인 시스템들**

[0153] 도 6은 분산형 시스템(600)의 간략화된 도면을 도시한다. 예시된 실시예에서, 분산형 시스템(600)은 하나 이상의 통신 네트워크들(610)을 통해 서버(612)에 결합된 하나 이상의 클라이언트 컴퓨팅 디바이스들(602, 604, 606, 및 608)을 포함한다. 클라이언트 컴퓨팅 디바이스들(602, 604, 606, 및 608)은 하나 이상의 애플리케이션들을 실행하도록 구성될 수 있다.

[0154] 다양한 예들에서, 서버(612)는 본 개시내용에서 설명된 하나 이상의 실시예들을 가능하게 하는 하나 이상의 서비스들 또는 소프트웨어 애플리케이션들을 실행하도록 적응될 수 있다. 특정 예들에서, 서버(612)는 또한 비-가상 및 가상 환경들을 포함할 수 있는 다른 서비스들 또는 소프트웨어 애플리케이션들을 제공할 수 있다. 일부 예들에서, 이러한 서비스들은 웹-기반으로 또는 클라우드 서비스들로서, 예컨대 서비스형 소프트웨어(Software as a Service; SaaS) 모델 하에서 클라이언트 컴퓨팅 디바이스들(602, 604, 606, 및/또는 608)의 사용자들에게 제공될 수 있다. 클라이언트 컴퓨팅 디바이스들(602, 604, 606, 및/또는 608)을 조작하는 사용자들은 결과적으로 이러한 구성요소들에 의해 제공되는 서비스들을 사용하기 위하여 서버(612)와 상호작용하기 위해 하나 이상의 클라이언트 애플리케이션들을 사용할 수 있다.

[0155] 도 6에 도시된 구성에서, 서버(612)는, 서버(612)에 의해 수행되는 기능들을 구현하는 하나 이상의 구성요소들(618, 620 및 622)을 포함할 수 있다. 이러한 구성요소들은, 하나 이상의 프로세서들, 하드웨어 구성요소들, 또는 이들의 조합들에 의해 실행될 수 있는 소프트웨어 구성요소들을 포함할 수 있다. 분산형 시스템(600)과 상이할 수 있는 다양한 상이한 시스템 구성들이 가능하다는 것이 이해되어야 한다. 따라서, 도 6에 도시된 예는 예시적인 시스템을 구현하기 위한 분산형 시스템의 일 예이며, 제한하는 것으로 의도되지 않는다.

[0156] 사용자들은, 그 후에 본 개시내용의 교시들에 따라 구현되거나 또는 서비스될 수 있는 하나 이상의 이벤트들 또는 모델들을 생성할 수 있는 하나 이상의 애플리케이션들, 모델들 또는 챗봇들을 실행하기 위해 컴퓨팅 디바이스들(602, 604, 606, 및/또는 608)을 사용할 수 있다. 클라이언트 디바이스는, 클라이언트 디바이스의 사용자가 클라이언트 디바이스와 상호작용하는 것을 가능하게 하는 인터페이스를 제공할 수 있다. 클라이언트 디바이스는 또한 이러한 인터페이스를 통해 사용자에게 정보를 출력할 수 있다. 도 6이 4개의 클라이언트 컴퓨팅 디바이스들만을 도시하지만, 임의의 수의 클라이언트 컴퓨팅 디바이스들이 지원될 수 있다.

[0157] 클라이언트 디바이스들은 다양한 유형들의 컴퓨팅 시스템들 예컨대 휴대용 핸드헬드 디바이스들, 범용 컴퓨터들 예컨대 개인용 컴퓨터들 및 랩탑들, 워크스테이션 컴퓨터들, 착용형 디바이스들, 게이밍 시스템들, 씬(thin) 클라이언트들, 다양한 메시징 디바이스들, 센서들 또는 다른 센싱 디바이스들, 및 유사한 것을 포함할 수 있다. 이러한 컴퓨팅 디바이스들은, 다양한 모바일 운영 시스템들(예를 들어, Microsoft Windows Mobile®, iOS®, Windows Phone®, Android™, BlackBerry®, Palm OS®)을 포함하여, 다양한 유형들 및 버전들의 애플리케이션들 및 운영 시스템들(예를 들어, Microsoft Windows®, Apple Macintosh®, UNIX® 또는 UNIX-유사 운영 시스템들, Linux 또는 Linux-유사 운영 시스템들 예컨대 Google Chrome™ OS)을 실행할 수 있다. 휴대용 핸드헬드 디바이스들은 무선 전화기들, 스마트폰들(예를 들어, iPhone®), 태블릿들(예를 들어, iPad®), 개인용 디지털 보조기기(personal digital assistant; PDA), 및 유사한 것을 포함할 수 있다. 웨어러블(wearable) 디바이스들은 Google Glass® 머리 착용형 디스플레이, 및 다른 디바이스들을 포함할 수 있다. 게이밍 시스템들은 다양한 핸드헬드 게이밍 디바이스들, 인터넷-가능형 게이밍 디바이스들(예를 들어, Kinect® 제스처 입력 디바이스를 갖거나 또는 갖지 않는 Microsoft Xbox® 게이밍 콘솔, Sony PlayStation® 시스템, Nintendo®에 의해 제공되는 다양한 게이밍 시스템들, 및 다른 것들), 및 유사한 것을 포함할 수 있다. 클라이언트 디바이스들은, 다양한

인터넷-관련 앱들, 통신 애플리케이션들(예를 들어, 이-메일 애플리케이션들, 단문 메시지 서비스(short message service; SMS) 애플리케이션들)과 같은 다양하고 상이한 애플리케이션들을 실행할 수 있으며, 다양한 통신 프로토콜들을 사용할 수 있다.

[0158] 네트워크(들)(610)는, 비제한적으로, TCP/IP(transmission control protocol/Internet protocol), SNA(systems network architecture), IPX(Internet packet exchange), AppleTalk®, 및 유사한 것을 포함하는 다양한 이용가능 프로토콜들 중 임의의 것을 사용하여 데이터 통신을 지원할 수 있는 당업자들에게 익숙한 임의의 유형의 네트워크일 수 있다. 단지 예로서, 네트워크(들)(610)는, 근거리 네트워크(local area network; LAN), 이더넷 기반 네트워크들, 토큰-링(Token-Ring), 광역 네트워크(wide-area network; WAN), 인터넷, 가상 네트워크, 가상 사설 네트워크(virtual private network; VPN), 인트라넷, 엑스트라넷, 공중 교환 전화 네트워크(public switched telephone network; PSTN), 적외선 네트워크, 무선 네트워크(예를 들어, 전기 전자 기술자 협회(Institute of Electrical and Electronics; IEEE) 1002.11 프로토콜들의 묶음, Bluetooth®, 및/또는 임의의 다른 무선 프로토콜 중 임의의 것 하에서 동작하는 네트워크), 및/또는 이들 및/또는 다른 네트워크들의 임의의 조합일 수 있다.

[0159] 서버(612)는, 하나 이상의 범용 컴퓨터들, 특수 서버 컴퓨터들(예로서, PC(personal computer) 서버들, UNIX® 서버들, 중급 서버(mid-range server)들, 메인프레임 컴퓨터들, 랙-장착형 서버들, 등을 포함함), 서버 팜(server farm)들, 서버 클러스터(server cluster)들, 또는 임의의 다른 적절한 배열 및/또는 조합으로 구성될 수 있다. 서버들(612)은, 서버에 대한 가상 저장 디바이스를 유지하기 위해 가상화될 수 있는 논리적인 저장 디바이스들의 하나 이상의 유연한 풀(pool)들과 같은 가상화를 수반하는 다른 컴퓨팅 아키텍처들 또는 가상 운영 시스템들을 실행하는 하나 이상의 가상 머신들을 포함할 수 있다. 다양한 예들에서, 서버(612)는 이상의 개시내용에서 설명된 기능을 제공하는 하나 이상의 서비스들 또는 소프트웨어 애플리케이션들을 실행하도록 적응될 수 있다.

[0160] 서버(612) 내의 컴퓨팅 시스템들은, 임의의 상용 이용가능 서버 운영 시스템뿐만 아니라 이상에서 논의된 것들 중 임의의 것을 포함하는 하나 이상의 운영 시스템들을 실행할 수 있다. 서버(612)는 또한, HTTP(hypertext transport protocol) 서버들, FTP(file transfer protocol) 서버들, CGI(common gateway interface) 서버들, JAVA® 서버들, 데이터베이스 서버들, 및 유사한 것을 포함하는 다양한 추가적인 서버 애플리케이션들 및/또는 중간-계층(mid-tier) 애플리케이션들 중 임의의 것을 실행할 수 있다. 예시적인 데이터베이스 서버들은 비제한적으로, Oracle®, Microsoft®, Sybase®, IBM(International Business Machines)®, 및 유사한 것으로부터 상용적으로 이용가능한 것들을 포함한다.

[0161] 일부 구현예들에서, 서버(612)는, 클라이언트 컴퓨팅 디바이스들(602, 604, 606, 및 608)의 사용자들로부터 수신되는 데이터 피드(data feed)들 및/또는 이벤트 업데이트들을 분석하고 통합하기 위한 하나 이상의 애플리케이션들을 포함할 수 있다. 일 예로서, 데이터 피드들 및/또는 이벤트 업데이트들은 비제한적으로, Twitter® 피드들, Facebook® 업데이트들 또는 하나 이상의 제3자 정보 소스들 및 연속적인 데이터 스트림들로부터 수신되는 실-시간 업데이트들을 포함할 수 있으며, 이들은 센서 데이터 애플리케이션들, 금융 시세표시기들, 네트워크 성능 측정 툴들(예를 들어, 네트워크 모니터링 및 트래픽 관리 애플리케이션들), 클릭스트림(clickstream) 분석 툴들, 자동차 트래픽 모니터링, 및 유사한 것과 연관된 실-시간 이벤트들을 포함할 수 있다. 서버(612)는 또한 클라이언트 컴퓨팅 디바이스들(602, 604, 606, 및 608)의 하나 이상의 디스플레이 디바이스들을 통해 데이터 피드들 및/또는 실-시간 이벤트들을 디스플레이하기 위한 하나 이상의 애플리케이션을 포함할 수 있다.

[0162] 분산형 시스템(600)은 또한 하나 이상의 데이터 저장소들(614, 616)을 포함할 수 있다. 이러한 데이터 저장소들은 특정 예들에서 데이터 및 다른 정보를 저장하기 위해 사용될 수 있다. 예를 들어, 데이터 저장소들(614, 616) 중 하나 이상은, 다양한 실시예들에 따라 다양한 기능들을 수행할 때, 서버(612)에 의해 사용되는 챗봇들이 사용하기 위한 생성된 모델들 또는 챗봇 성능과 관련된 정보와 같은 정보를 저장하기 위해 사용될 수 있다. 데이터 저장소들(614 및 616)은 다양한 위치들에 존재할 수 있다. 예를 들어, 서버(612)에 의해 사용되는 데이터 저장소는 서버(612)에 대해 로컬에 존재할 수 있거나 또는 서버(612)로부터 원격에 존재하고 네트워크-기반 또는 전용 연결을 통해 서버(612)와 통신할 수 있다. 데이터 저장소들(614, 616)은 상이한 유형들일 수 있다. 특정 예들에서, 서버(612)에 의해 사용되는 데이터 저장소는 데이터베이스, 예를 들어, 관계형 데이터베이스, 예컨대 Oracle Corporation® 및 다른 판매사에 의해 제공되는 데이터베이스들일 수 있다. 이러한 데이터베이스들 중 하나 이상이 SQL-포맷형 명령들에 응답하여 데이터베이스로 그리고 데이터베이스로부터 데이터의 저장, 업데이트, 및 검색을 가능하게 하도록 적응될 수 있다.

[0163] 특정 예들에서, 데이터 저장소들(614, 616) 중 하나 이상이 또한 애플리케이션 데이터를 저장하기 위해 애플리케이션들에 의해 사용될 수 있다. 애플리케이션들에 의해 사용되는 데이터 저장소들은, 예를 들어, 키-값 저장 저장소, 객체 저장 저장소, 또는 파일 시스템에 의해 지원되는 범용 저장 저장소와 같은 상이한 유형들일 수 있다.

[0164] 특정 예들에 서, 본 개시내용에서 설명된 기능성들은 클라우드 환경을 통한 서비스들로서 제공될 수 있다. 도 7은, 특정 예들에 따른 그 내부에서 서비스들이 클라우드 서비스들로서 제공될 수 있는 클라우드-기반 시스템 환경의 간략화된 블록도이다. 도 7에 도시된 예에서, 클라우드 인프라스트럭처 시스템(702)은, 하나 이상의 클라이언트 컴퓨팅 디바이스들(704, 706, 및 708)을 사용하여 사용자들에 의해 요청될 수 있는 하나 이상의 클라우드 서비스들을 제공할 수 있다. 클라우드 인프라스트럭처 시스템(702)은 서버(612)에 대하여 이상에서 설명된 것들을 포함할 수 있는 하나 이상의 서버들 및/또는 컴퓨터들을 포함할 수 있다. 클라우드 인프라스트럭처 시스템(702) 내의 컴퓨터들은, 범용 컴퓨터들, 특수 서버 컴퓨터들, 서버 팜들, 서버 클러스터들, 또는 임의의 다른 적절한 배열 및/또는 조합으로서 조직될 수 있다.

[0165] 네트워크(들)(710)는 클라이언트들(704, 706, 및 708)과 클라우드 인프라스트럭처 시스템(702) 사이의 데이터의 교환 및 통신을 가능하게 할 수 있다. 네트워크(들)(710)는 하나 이상의 네트워크들을 포함할 수 있다. 네트워크들은 동일하거나 또는 상이한 유형들일 수 있다. 네트워크(들)(710)는, 통신들을 가능하게 하기 위하여, 무선 및/또는 유선 프로토콜들을 포함하는 하나 이상의 통신 프로토콜들을 지원할 수 있다.

[0166] 도 7에 도시된 예는 단지 클라우드 인프라스트럭처 시스템의 일 예이며, 제한하는 것으로 의도되지 않는다. 일부 다른 예들에서, 클라우드 인프라스트럭처 시스템(702)은 도 7에 도시된 것보다 더 많거나 또는 더 적은 구성요소들을 가질 수 있거나, 2개 이상의 구성요소들을 결합할 수 있거나, 또는 구성요소들의 상이한 구성 또는 배열을 가질 수 있다. 예를 들어, 도 7이 3개의 클라이언트 컴퓨팅 디바이스들을 도시하지만, 대안적인 예들에서 임의의 수의 클라이언트 컴퓨팅 디바이스들이 지원될 수 있다.

[0167] 용어 클라우드 서비스는 일반적으로, 주문형으로 그리고 서비스 제공자의 시스템들(예를 들어, 클라우드 인프라스트럭처 시스템(702))에 의해 인터넷과 같은 통신 네트워크를 통해 사용자들이 이용할 수 있게 되는 서비스를 지칭하기 위해 사용된다. 전형적으로, 공중 클라우드 환경에서, 클라우드 서비스 제공자의 시스템을 구성하는 서버들 및 시스템들은 고객의 자체적인 사내(on-premise) 서버들 및 시스템들과는 상이하다. 클라우드 서비스 제공자의 시스템들은 클라우드 서비스 제공자에 의해 관리된다. 따라서, 고객들은, 서비스들에 대한 별개의 라이센스들, 지원들, 또는 하드웨어 및 소프트웨어 자원들을 구매할 필요 없이 클라우드 서비스 제공자에 의해 제공되는 클라우드 서비스들 자체를 이용할 수 있다. 예를 들어, 클라우드 서비스 제공자의 시스템은 애플리케이션을 호스팅할 수 있으며, 사용자는, 사용자가 애플리케이션을 실행하기 위하여 인프라스트럭처 자원들을 구매할 필요 없이 인터넷을 통해 애플리케이션을 주문형으로, 주문 및 사용할 수 있다. 클라우드 서비스들은 애플리케이션들, 자원들 및 서비스들에 대한 용이한 스케일러블(scalable) 액세스를 제공하도록 설계된다. 몇몇 제공자들이 클라우드 서비스들을 제공한다. 예를 들어, 미들웨어 서비스들, 데이터베이스 서비스들, 자바 클라우드 서비스들 및 다른 것들과 같은 몇몇 클라우드 서비스들은, 캘리포니아, 레드우드 쇼어 소재의 Oracle Corporation®에 의해 제공된다.

[0168] 특정 예들에서, 클라우드 인프라스트럭처 시스템(702)은, 예컨대, 하이브리드 서비스 모델들을 포함하여, 서비스형 소프트웨어(Software as a Service; SaaS) 모델, 서비스형 플랫폼(Platform as a Service; PaaS) 모델, 서비스형 인프라스트럭처(Infrastructure as a Service; IaaS) 모델, 및 다른 것들 하에서 상이한 모델들을 사용하여 하나 이상의 클라우드 서비스들을 제공할 수 있다. 클라우드 인프라스트럭처 시스템(702)은, 다양한 클라우드 서비스들의 프로비저닝(provision)을 가능하게 하는 애플리케이션들, 미들웨어, 데이터베이스들, 및 다른 자원들의 묶음을 포함할 수 있다.

[0169] SaaS 모델은, 고객이 기초(underlying) 애플리케이션에 대한 하드웨어 또는 소프트웨어를 구매할 필요 없이, 애플리케이션 또는 소프트웨어가 서비스로서 인터넷과 같은 통신 네트워크를 통해 고객에게 전달되는 것을 가능하게 한다. 예를 들어, SaaS 모델은, 고객에게 클라우드 인프라스트럭처 시스템(702)에 의해 호스팅되는 주문형 애플리케이션들에 대한 액세스를 제공하기 위해 사용될 수 있다. Oracle Corporation®에 의해 제공되는 SaaS 서비스들의 예들은, 비제한적으로, 인사/자본 관리, 고객 관계 관리(customer relationship management; CRM), 전사적 자원 관리(enterprise resource planning; ERP), 공급망 관리(supply chain management; SCM), 전사적 성과 관리(enterprise performance management; EPM), 분석 서비스, 소셜 애플리케이션들, 및 다른 것들에 대한 다양한 서비스들을 포함한다.

[0170] IaaS 모델은 일반적으로, 탄력적인 컴퓨팅 및 자원 성능들을 제공하기 위한 클라우드 서비스로서 고객에게 인프라스트럭처 자원들(예를 들어, 서버들, 저장부들, 하드웨어 및 네트워킹 자원들)을 제공하기 위해 사용된다. Oracle Corporation®에 의해 다양한 IaaS 서비스들에 제공된다.

[0171] PaaS 모델은 일반적으로, 고객이 이러한 자원들을 입수하거나, 구축하거나, 또는 유지할 필요 없이, 고객들이 애플리케이션들 및 서비스들을 개발하고, 실행하며, 관리하는 것을 가능하게 하는 플랫폼 및 환경을 서비스로서 제공하기 위해 사용된다. Oracle Corporation®에 의해 제공되는 PaaS 서비스들의 예들은, 비제한적으로, 오라클 자바 클라우드 서비스(Java Cloud Service; JCS), 오라클 데이터베이스 클라우드 서비스(Database Cloud Service; DBCS), 데이터 관리 클라우드 서비스, 다양한 애플리케이션 개발 솔루션 서비스들, 및 다른 것들을 포함한다.

[0172] 클라우드 서비스들은 일반적으로, 주문형 셀프-서비스 기반의, 가입(subscription)-기반의, 탄력적으로 스케일러블하고, 신뢰할 수 있으며, 고도로 이용성이 높고 안전한 방식으로 제공된다. 예를 들어, 가입 주문을 통해, 고객은 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 하나 이상의 서비스들을 주문할 수 있다. 그러면, 클라우드 인프라스트럭처 시스템(702)은 고객의 가입 주문에서 요청되는 서비스들을 제공하기 위하여 프로세싱을 수행한다. 예를 들어, 사용자는, 이상에서 설명된 바와 같이, 클라우드 인프라스트럭처 시스템이 특정 액션(예를 들어, 의도)을 취하거나 및/또는 본 명세서에서 설명된 바와 같이 챗봇 시스템에 대해 서비스들을 제공하도록 요청하기 위해 발화들을 사용할 수 있다. 클라우드 인프라스트럭처 시스템(702)은 하나 또는 심지어 다수의 클라우드 서비스들을 제공하도록 구성될 수 있다.

[0173] 클라우드 인프라스트럭처 시스템(702)은 상이한 배포 모델들을 통해 클라우드 서비스들을 제공할 수 있다. 공개 클라우드 모델에서, 클라우드 인프라스트럭처 시스템(702)은 제3자 클라우드 서비스 제공자에 의해 소유될 수 있으며, 클라우드 서비스들은 임의의 일반적인 공개 고객에게 제공될 수 있고, 여기에서 고객은 개인 또는 기업일 수 있다. 특정한 다른 예들에서, 사설 클라우드 모델 하에서, 클라우드 인프라스트럭처 시스템(702)은 조직 내에서(예를 들어, 기업 조직 내에서) 운영될 수 있으며, 서비스들은 조직 내의 고객들에게 제공될 수 있다. 예를 들어, 고객들은 인사 부서, 경리 부서, 등과 같은 기업의 다양한 부서들 또는 심지어 기업 내의 개인들일 수 있다. 특정한 다른 예들에서, 커뮤니티(community) 클라우드 모델 하에서, 클라우드 인프라스트럭처 시스템(702) 및 제공되는 서비스들은 연관된 커뮤니티 내의 몇몇 조직들에 의해 공유될 수 있다. 이상에서 언급된 모델들의 하이브리드와 같은 다양한 다른 모델들이 또한 사용될 수 있다.

[0174] 클라이언트 컴퓨팅 디바이스들(704, 706, 및 708)은 (도 6에 도시된 클라이언트 컴퓨팅 디바이스들(602, 604, 606, 및 608)과 같이) 상이한 유형들일 수 있으며, 하나 이상의 클라이언트 애플리케이션들을 동작시킬 수 있다. 사용자는, 클라우드 인프라스트럭처 시스템(702)과 상호작용하기 위하여, 예컨대 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 서비스를 요청하기 위하여 클라이언트 디바이스를 사용할 수 있다. 예를 들어, 사용자는 본 개시내용에서 설명된 바와 같은 챗봇으로부터의 정보 또는 액션을 요청하기 위해 클라이언트 디바이스를 사용할 수 있다.

[0175] 일부 예들에서, 서비스들을 제공하기 위해 클라우드 인프라스트럭처 시스템(702)에 의해 수행되는 프로세싱은 모델 트레이닝 및 배포를 수반할 수 있다. 이러한 분석은 하나 이상의 모델들을 트레이닝시키고 배포하기 위해 데이터 세트들을 사용하는 것, 분석하는 것 및 조작하는 것을 수반할 수 있다. 이러한 분석은, 아마도 데이터를 병렬로 프로세싱하며, 데이터를 사용하여 시뮬레이션들을 수행하고, 유사한 것을 수행하는 하나 이상의 프로세서들에 의해 수행될 수 있다. 예를 들어, 빅 데이터 분석은, 챗봇 시스템에 대한 하나 이상의 모델들을 생성하고 트레이닝시키기 위해 클라우드 인프라스트럭처 시스템(702)에 의해 수행될 수 있다. 이러한 분석을 위해 사용되는 데이터는, 구조화된 데이터(예를 들어, 데이터베이스에 저장되거나 또는 구조화 모델에 따라 구조화된 데이터) 및/또는 구조화되지 않은 데이터(예를 들어, 데이터 블랍(data blob)들(이진 대형 객체들))를 포함할 수 있다.

[0176] 도 7의 예에 도시된 바와 같이, 클라우드 인프라스트럭처 시스템(702)은, 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 다양한 클라우드 서비스들의 프로비저닝을 가능하게 하기 위하여 사용되는 인프라스트럭처 자원들(730)을 포함할 수 있다. 인프라스트럭처 자원들(730)은, 예를 들어, 프로세싱 자원들, 저장 또는 메모리 자원들, 네트워킹 자원들, 및 유사한 것을 포함할 수 있다. 특정 예들에서, 애플리케이션들로부터 요청된 저장을 서비스하기 위해 이용가능한 저장 가상 머신들은 클라우드 인프라스트럭처 시스템(702)의 부분일 수 있다. 다른 예들에서, 저장 가상 머신들은 상이한 시스템들의 부분일 수 있다.

[0177] 특정 예들에서, 상이한 고객들에 대해 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 다양한 클라우드

서비스들을 지원하기 위한 이러한 자원들의 효율적인 프로비저닝을 가능하게 하기 위해, 자원들은 자원들의 세트들 또는 자원 모듈들("포드(pod)들"로서도 지칭됨)로 번들링(bundle)될 수 있다. 각각의 자원 모듈 또는 포드는 하나 이상의 유형들의 자원들의 미리-통합되고 최적화된 조합을 포함할 수 있다. 특정 예들에서, 상이한 포드들이 상이한 유형들의 클라우드 서비스들에 대하여 사전-프로비저닝될 수 있다. 예를 들어, 포드들의 제 1 세트는 데이터베이스 서비스를 위해 프로비저닝될 수 있으며, 포드들의 제 1 세트 내의 포드와는 상이한 자원들의 조합을 포함할 수 있는 포드들의 제 2 세트는 자바 서비스를 위해 프로비저닝될 수 있는 등이다. 일부 서비스들에 대하여, 서비스들을 프로비저닝하기 위해 할당된 자원들이 서비스들 사이에서 공유될 수 있다.

[0178] 클라우드 인프라스트럭처 시스템(702)은 자체적으로, 클라우드 인프라스트럭처 시스템(702)의 상이한 구성요소들에 의해 공유되며 클라우드 인프라스트럭처 시스템(702)에 의한 서비스들의 프로비저닝을 가능하게 하는 서비스들(732)을 내부적으로 사용할 수 있다. 이러한 내부 공유형 서비스들은, 비제한적으로, 보안 및 신원(identity) 서비스, 통합 서비스, 기업 저장소 서비스, 기업 관리자 서비스, 바이러스 스캐닝 및 화이트 리스트(white list) 서비스, 고 이용가능성, 백업 및 복원 서비스, 클라우드 지원을 가능하게 하기 위한 서비스, 이메일 서비스, 통지 서비스, 파일 전송 서비스, 및 유사한 것을 포함할 수 있다.

[0179] 클라우드 인프라스트럭처 시스템(702)은 다수의 서브시스템들을 포함할 수 있다. 이러한 서브시스템들은 소프트웨어, 또는 하드웨어, 또는 이들의 조합들로 구현될 수 있다. 도 7에 도시된 바와 같이, 서브시스템들은, 클라우드 인프라스트럭처 시스템(702)의 고객들 또는 사용자들이 클라우드 인프라스트럭처 시스템(702)와 상호작용하는 것을 가능하게 하는 사용자 인터페이스 서브시스템(712)을 포함할 수 있다. 사용자 인터페이스 서브시스템(712)은 웹 인터페이스(714), 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 클라우드 서비스들이 광고되고 고객에 의해 구매될 수 있는 온라인 스토어 인터페이스(716), 및 다른 인터페이스들(718)과 같은 다양하고 상이한 인터페이스들을 포함할 수 있다. 예를 들어, 고객은 클라이언트 디바이스를 사용하여, 인터페이스들(714, 716, 및 718) 중 하나 이상을 사용하여 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 하나 이상의 서비스들을 요청(서비스 요청(734))할 수 있다. 예를 들어, 고객은 온라인 스토어에 액세스하고, 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 클라우드 서비스들을 브라우징하며, 고객이 가입하기를 원하는 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 하나 이상의 서비스들에 대한 가입 주문을 할 수 있다. 서비스 요청은, 고객을 및 고객이 가입하기를 원하는 하나 이상의 서비스들을 식별하는 정보를 포함할 수 있다. 예를 들어, 고객은 클라우드 인프라스트럭처 시스템(702)에 의해 제공되는 서비스에 대해 가입 주문을 할 수 있다. 주문의 부분으로서, 고객은 서비스가 제공될 챗봇 시스템을 식별하는 정보, 및 선택적으로 챗봇 시스템에 대한 하나 이상의 자격증명서들을 제공할 수 있다.

[0180] 도 7에 도시된 예와 같은, 특정 예들에서, 클라우드 인프라스트럭처 시스템(702)은, 신규 주문을 프로세싱하도록 구성된 주문 관리 서브시스템(order management subsystem; OMS)(720)을 포함할 수 있다. 이러한 프로세싱의 부분으로서, OMS(720)은: 이전에 수행되지 않은 경우 고객에 대한 계정을 생성하고; 요청된 서비스를 고객에게 제공하기 위하여 고객에게 요금을 청구하기 위해 사용될 고객으로부터의 요금청구 및/또는 회계 정보를 수신하며; 고객 정보를 검증하고; 검증 시에, 고객에 대한 주문을 예약(book)하며; 및 프로비저닝하기 위해 주문을 준비하기 위한 다양한 작업흐름들을 편성(orchestrate)하도록 구성될 수 있다.

[0181] 일단 적절하게 검증되면, OMS(720)은, 프로세싱, 메모리, 및 네트워킹 자원들을 포함하는 주문에 대한 자원들을 프로비저닝하도록 구성된 주문 프로비저닝 서브시스템(order provisioning subsystem; OPS)(724)을 호출할 수 있다. 프로비저닝은, 주문에 대하여 자원들을 할당하는 것 및 고객 주문에 의해 요청된 서비스를 가능하게 하기 위하여 자원들을 구성하는 것을 포함할 수 있다. 주문에 대하여 자원들이 프로비저닝되는 방식 및 프로비저닝되는 자원들의 유형은, 고객에 의해 주문된 클라우드 서비스의 유형에 의존할 수 있다. 예를 들어, 하나의 작업흐름에 따르면, OPS(724)는 요청되는 특정 클라우드 서비스를 결정하고, 해당 특정 클라우드 서비스에 대하여 미리-구성되었을 수 있는 다수의 포드들을 식별하도록 구성될 수 있다. 주문에 대하여 할당되는 포드들의 수는 요청된 서비스의 크기/양/레벨/범위에 의존할 수 있다. 예를 들어, 할당될 포드들의 수는 서비스에 의해 지원될 사용자들의 수, 서비스가 요청되는 시간의 지속 기간, 및 유사한 것에 의존하여 결정될 수 있다. 그런 다음, 할당된 포드들은 요청된 서비스를 제공하기 위하여 특정한 요청 고객에 대해 맞춤화될 수 있다.

[0182] 특정 예들에서, 이상에서 설명된 바와 같은, 셋업 단계 프로세싱은 프로비저닝 프로세스의 부분으로서 클라우드 인프라스트럭처 시스템(702)에 의해 수행될 수 있다. 클라우드 인프라스트럭처(702)는 애플리케이션 ID를 생성하고, 클라우드 인프라스트럭처 시스템(702) 자체에 의해 제공된 저장 가상 머신들 중에서 또는 클라우드 인프라스트럭처 시스템(702) 이외의 다른 시스템들에 의해 제공된 저장 가상 머신들 중에서 애플리케이션에 대한 저

장 가상 머신을 선택할 수 있다.

[0183] 클라우드 인프라스트럭처 시스템(702)은, 요청된 서비스가 이제 사용할 준비가 된 때를 나타내기 위해 요청 고객에게 응답 또는 통지(744)를 전송할 수 있다. 일부 경우들에서, 고객이 요청된 서비스들의 사용을 시작하는 것 및 요청된 서비스들의 장점들을 이용하는 것을 가능하게 하는 정보(예를 들어, 링크)가 고객에게 전송될 수 있다. 특정 예들에서, 서비스를 요청하는 고객에 대해, 응답은, 클라우드 인프라스트럭처 시스템(702)에 의해 생성된 챗봇 시스템 ID, 및 챗봇 시스템 ID에 대응하는 챗봇 시스템에 대해 클라우드 인프라스트럭처 시스템(702)에 의해 선택된 챗봇 시스템을 식별하는 정보를 포함할 수 있다.

[0184] 클라우드 인프라스트럭처 시스템(702)은 다수의 고객들에게 서비스들을 제공할 수 있다. 각각의 고객에 대해, 클라우드 인프라스트럭처 시스템(702)은 고객으로부터 수신된 하나 이상의 가입 주문들과 관련된 정보를 관리하는 것, 주문과 관련된 고객 데이터를 관리하는 것, 및 요청된 서비스들을 고객에게 제공하는 것을 담당한다. 클라우드 인프라스트럭처 시스템(702)은 또한 가입된 서비스들의 고객의 사용에 관한 사용 통계자료를 수집할 수 있다. 예를 들어, 통계자료는, 사용되는 저장소의 양, 전송되는 데이터의 양, 사용자들의 수, 및 시스템 사용 시간(system up time)과 시스템 정지 시간(system down time)의 양, 및 유사한 것에 대해 수집될 수 있다. 이러한 사용 정보는 고객에게 요금을 청구하기 위해 사용될 수 있다. 요금 청구는, 예를 들어, 월 단위로 이루어질 수 있다.

[0185] 클라우드 인프라스트럭처 시스템(702)은 다수의 고객들에게 병렬로 서비스들을 제공할 수 있다. 클라우드 인프라스트럭처 시스템(702)은, 어쩌면 재산권적 정보를 포함하는, 이러한 고객들에 대한 정보를 저장할 수 있다. 특정 예들에서, 클라우드 인프라스트럭처 시스템(702)은, 고객 정보를 관리하고, 하나의 고객에 관한 정보가 다른 고객에 의해 액세스가능하지 않도록 관리되는 정보의 분리를 제공하도록 구성된 신원 관리 서브시스템(identity management subsystem; IMS)(728)을 포함한다. IMS(728)는, 다양한 보안-관련 서비스들 예컨대 신원 서비스들, 예컨대 정보 액세스 관리, 인증 및 인가 서비스들, 고객 신원들 및 역할들 및 관련된 기능들을 관리하기 위한 서비스들 및 유사한 것을 제공하도록 구성될 수 있다.

[0186] 도 8은 컴퓨터 시스템(800)의 일 예를 예시한다. 일부 예들에서, 컴퓨터 시스템(800)은 이상에서 설명된 분산형 환경, 및 다양한 서버들과 컴퓨터 시스템들 내에 디지털 어시스턴트 또는 챗봇 시스템들 중 임의의 것을 구현하기 위해 사용될 수 있다. 도 8에 도시된 바와 같이, 컴퓨터 시스템(800)은, 버스 서브시스템(802)을 통해 복수의 다른 서브시스템들과 통신하는 프로세싱 서브시스템(804)을 포함하는 다양한 서브시스템들을 포함한다. 이러한 다른 서브시스템들은 프로세싱 가속 유닛(806), I/O 서브시스템(808), 저장 서브시스템(818) 및 통신 서브시스템(824)을 포함할 수 있다. 저장 서브시스템(818)은 시스템 메모리(810) 및 저장 매체(822)를 포함하는 비-일시적 컴퓨터-판독가능 저장 매체를 포함할 수 있다.

[0187] 버스 서브시스템(802)은 컴퓨터 시스템(800)의 다양한 구성요소들 및 서브시스템들이 의도된 바와 같이 서로 통신하는 것을 가능하게 하기 위한 메커니즘을 제공한다. 버스 서브시스템(802)이 단일 버스로서 개략적으로 도시되었지만, 버스 서브시스템의 대안적인 예들은 복수의 버스들을 사용할 수 있다. 버스 서브시스템(802)은, 다양한 버스 아키텍처들 중 임의의 것을 사용하는 메모리 버스 또는 메모리 제어기, 주변기기 버스, 로컬 버스, 및 유사한 것을 포함하는 몇몇 유형들의 버스 구조들 중 임의의 버스 구조일 수 있다. 예를 들어, 이러한 아키텍처들은, 산업 표준 아키텍처(Industry Standard Architecture; ISA) 버스, 마이크로 채널 아키텍처(Micro Channel Architecture; MCA) 버스, 개량 ISA(Enhanced ISA; EISA) 버스, 비디오 전자공학 표준 위원회(Video Electronics Standards Association; VESA) 로컬 버스, 및 주변 구성요소 상호연결(Peripheral Component Interconnect; PCI) 버스를 포함할 수 있으며, 이들은 IEEE P1386.1 표준에 대하여 제조되는 메자닌 버스(Mezzanine bus) 및 유사한 것으로서 구현될 수 있다.

[0188] 프로세싱 서브시스템(804)은 컴퓨터 시스템(800)의 동작을 제어하며, 하나 이상의 프로세서들, 애플리케이션 특정 집적 회로(application specific integrated circuit; ASIC)들, 또는 필드 프로그램가능 게이트 어레이(field programmable gate array; FPGA)들을 포함할 수 있다. 프로세서들은 단일 코어 또는 다중코어 프로세서들을 포함할 수 있다. 컴퓨터 시스템(800)의 프로세싱 자원들은 하나 이상의 프로세싱 유닛들(832, 834), 등으로 조직될 수 있다. 프로세싱 유닛은 하나 이상의 프로세서들, 동일하거나 또는 상이한 프로세서들로부터의 하나 이상의 코어들, 코어들 및 프로세서들의 조합, 또는 코어들 및 프로세서들의 다른 조합들을 포함할 수 있다. 일부 예들에서, 프로세싱 서브시스템(804)은 하나 이상의 특수 목적 코-프로세서들, 예컨대 그래픽 프로세서들, 디지털 신호 프로세서(digital signal processor; DSP)들, 또는 유사한 것을 포함할 수 있다. 일부 예들에서, 프로세싱 서브시스템(804)의 프로세싱 유닛들 중 일부 또는 전부는, 애플리케이션 특정 집적 회로(application

specific integrated circuit; ASIC)들, 또는 필드 프로그램가능 게이트 어레이(field programmable gate array; FPGA)들과 같은 맞춤형 회로들을 사용하여 구현될 수 있다.

[0189] 일부 예들에서, 프로세싱 서브시스템(804) 내의 프로세싱 유닛들은 시스템 메모리(810) 내에 또는 컴퓨터 판독가능 저장 매체들(822) 상에 저장된 명령어들을 실행할 수 있다. 다양한 예들에서, 프로세싱 유닛들은 다양한 프로그램 프로그램들 또는 코들 명령어들을 실행할 수 있으며, 다수의 동시에 실행되는 프로그램들 또는 프로세스들을 유지할 수 있다. 임의의 주어진 시점에, 실행될 프로그램 코드의 전부 또는 일부가 잠재적으로 하나 이상의 저장 디바이스들 상에 상주하는 것을 포함하여 시스템 메모리(810) 및/또는 컴퓨터-판독가능 저장 매체(822)에 상주할 수 있다. 적절한 프로그래밍을 통하여, 프로세싱 서브시스템(804)는 이상에서 설명된 다양한 기능성들을 제공할 수 있다. 컴퓨터 시스템(800)이 하나 이상의 가상 머신들을 실행하는 경우들에서, 하나 이상의 프로세싱 유닛들이 각각의 가상 머신에 할당될 수 있다.

[0190] 특정 예들에서, 프로세싱 가속 유닛(806)은 선택적으로, 컴퓨터 시스템(800)에 의해 수행되는 전체 프로세싱을 가속하기 위해 프로세싱 서브시스템(804)에 의해 수행되는 프로세싱 중 일부를 오프-로딩(off-load)하기 위해 또는 맞춤화된 프로세싱을 수행하기 위해 제공될 수 있다.

[0191] I/O 서브시스템(808)은 컴퓨터 시스템(800)으로 정보를 입력하기 위한 및/또는 컴퓨터 시스템(800)으로부터 또는 이를 통해 정보를 출력하기 위한 디바이스들 및 메커니즘들을 포함할 수 있다. 일반적으로, 용어 입력 디바이스의 사용은 컴퓨터 시스템(800)으로 정보를 입력하기 위한 모든 가능한 유형들의 디바이스들 및 메커니즘들을 포함하도록 의도된다. 사용자 인터페이스 입력 디바이스들은, 예를 들어, 키보드, 포인팅 디바이스들 예컨대 마우스 또는 트랙볼, 터치패드 또는 디스플레이 내에 통합된 터치 스크린, 스크롤 휠, 클릭 휠, 다이얼, 버튼, 스위치, 키패드, 음성 명령 인식 시스템들을 가진 오디오 입력 디바이스들, 마이크들, 및 다른 유형들의 입력 디바이스들을 포함할 수 있다. 사용자 인터페이스 입력 디바이스들은 또한, 사용자들이 제스처들 및 구두(spoken) 명령들을 사용하는 입력을 수신하기 위한 인터페이스를 제공하는, 입력 디바이스, Microsoft Xbox® 360 게임 제어기, 디바이스들을 제어하고 이와 상호작용하는 것을 가능하게 하는 Microsoft Kinect® 모션 센서와 같은 모션 센싱 및/또는 제스처 인식 디바이스들을 포함할 수 있다. 사용자 인터페이스 입력 디바이스들은 또한, 사용자들로부터 눈 움직임(eye activity)(예를 들어, 사진을 찍는 동안의 및/또는 메뉴를 선택하는 동안의 '깜박임')을 검출하고 눈 제스처들을 입력 디바이스(예를 들어, Google Glass®)로의 입력으로서 변환하는 Google Glass® 눈 깜박임 검출기와 같은 눈 제스처 인식 디바이스들을 포함할 수 있다. 추가적으로, 사용자 인터페이스 입력 디바이스들은, 사용자들이 음성 명령들을 통하여 음성 인식 시스템(예를 들어, Siri® 네비게이터(navigator))과 상호작용하는 것을 가능하게 하는 음성 인식 센싱 디바이스들을 포함할 수 있다.

[0192] 사용자 인터페이스 입력 디바이스들의 다른 예들은, 비제한적으로, 3차원(3D) 마우스들, 조이스틱들 또는 포인팅 스틱들, 게임패드들 및 그래픽 태블릿들, 및 음향/시각 디바이스들 예컨대 스피커들, 디지털 카메라들, 디지털 캠코더들, 휴대용 매체 플레이어들, 웹캠들, 이미지 스캐너들, 핑거프린트 스캐너들, 바코드 리더 3D 스캐너들, 3D 프린터들, 레이저 거리계들, 및 시선 추적 디바이스들을 포함할 수 있다. 추가적으로, 사용자 인터페이스 입력 디바이스들은, 예를 들어, 의료 이미징 입력 디바이스들 예컨대 컴퓨터 단층촬영, 자기 공명 이미징, 양전자 방출 단층촬영, 및 의료 초음파 검사 디바이스들을 포함할 수 있다. 사용자 인터페이스 입력 디바이스들은 또한, 예를 들어, 오디오 입력 디바이스들 예컨대 MIDI 키보드들, 디지털 악기들, 및 유사한 것을 포함할 수 있다.

[0193] 일반적으로, 용어 출력 디바이스의 사용은 컴퓨터 시스템(800)으로부터 사용자 또는 다른 컴퓨터로 정보를 출력하기 위한 모든 가능한 유형들의 디바이스들 및 메커니즘들을 포함하도록 의도된다. 사용자 인터페이스 출력 디바이스들은 디스플레이 서브시스템, 표시등들, 또는 비-시각적 디스플레이들 예컨대 오디오 출력 디바이스들, 등을 포함할 수 있다. 디스플레이 서브시스템은 음극선관(cathode ray tube; CRT), 액정 디스플레이(liquid crystal display; LCD) 또는 플라즈마 디스플레이를 사용하는 것과 같은 평면-패널 디바이스, 프로젝션 디바이스, 터치 스크린, 및 유사한 것일 수 있다. 예를 들어, 사용자 인터페이스 출력 디바이스들은, 비제한적으로, 시각적으로 텍스트, 그래픽들 및 오디오/비디오 정보를 전달하는 다양한 디스플레이 디바이스들, 예컨대 모니터들, 프린터들, 스피커들, 헤드폰들, 자동차 네비게이션 시스템들, 플로터(plotter)들, 음성 출력 디바이스들, 및 모뎀들을 포함할 수 있다.

[0194] 저장 서브시스템(818)은 컴퓨터 시스템(800)에 의해 사용되는 정보 및 데이터를 저장하기 위한 데이터 스토어 또는 저장소를 제공한다. 저장 서브시스템(818)은 일부 예들의 기능성을 제공하는 기본 프로그래밍 및 데이터 구성물들을 저장하기 위한 유형적이고 비-일시적인 컴퓨터-판독가능 저장 매체를 제공한다. 저장 서브시스템

(818)은, 프로세싱 서브시스템(804)에 의해 실행될 때 이상에서 설명된 기능성을 제공하는 소프트웨어(예를 들어, 프로그램들, 코드 모듈들, 명령어들)를 저장할 수 있다. 소프트웨어는 프로세싱 서브시스템(804)의 하나 이상의 프로세싱 유닛들에 의해 실행될 수 있다. 저장 서브시스템(818)은 또한 본 개시내용의 교시들에 따라 인증을 제공할 수 있다.

[0195] 저장 서브시스템(818)은 휘발성 및 비-휘발성 메모리 디바이스들을 포함하는 하나 이상의 비-일시적 메모리 디바이스들을 포함할 수 있다. 도 8에 도시된 바와 같이, 저장 서브시스템(818)은 시스템 메모리(810) 및 컴퓨터-판독가능 저장 매체(822)를 포함한다. 시스템 메모리(810)는, 프로그램 실행 동안 명령어들 및 데이터의 저장을 위한 휘발성 메인 랜덤 액세스 메모리(random access memory; RAM) 및 그 내부에 고정된 명령어들이 저장되는 비-휘발성 판독 전용 메모리(read only memory; ROM) 또는 플래시 메모리를 포함하는 복수의 메모리들을 포함할 수 있다. 일부 구현예들에서, 예컨대 기동 동안에 컴퓨터 시스템(800) 내의 요소들 사이에서 정보를 전송하는 것을 돕는 기본 루틴들을 포함하는 기본 입력/출력 시스템(basic input/output system; BIOS)은 전형적으로 ROM에 저장될 수 있다. RAM은 전형적으로 프로세싱 서브시스템(804)에 의해 현재 실행되며 동작되고 있는 데이터 및/또는 프로그램 모듈들을 포함한다. 일부 구현예들에서, 시스템 메모리(810)는 복수의 상이한 유형들의 메모리들, 예컨대 정적 랜덤 액세스 메모리(static random access memory; SRAM), 동적 랜덤 액세스 메모리(dynamic random access memory; DRAM), 및 유사한 것을 포함할 수 있다.

[0196] 예로서 그리고 비제한적으로, 도 8에 도시된 바와 같이, 시스템 메모리(810)는, 웹 브라우저들, 중간-계층 애플리케이션들, 관계형 데이터 베이스 관리 시스템(relational database management system; RDBMS)들, 등을 포함할 수 있는 다양한 애플리케이션들을 포함할 수 있는 실행되는 애플리케이션 프로그램들(812), 프로그램 데이터(814), 및 운영 시스템(816)을 로딩할 수 있다. 예로서, 운영 시스템(816)은, 다양한 버전들의 Microsoft Windows®, Apple Macintosh®, 및/또는 리눅스 운영 시스템들, (비제한적으로 다양한 GNU/리눅스 운영 시스템들, Google Chrome® OS, 및 유사한 것을 포함하는) 다양한 상용-이용가능 UNIX® 또는 UNIX-유사 운영 시스템들 및/또는 모바일 운영 시스템들 예컨대 iOS, Windows® Phone, Android® OS, BlackBerry® OS, 및 Palm® OS 운영 시스템들, 및 다른 것들을 포함할 수 있다.

[0197] 컴퓨터-판독가능 저장 매체(822)는 일부 예들의 기능성을 제공하는 프로그래밍 및 데이터 구성물들을 저장할 수 있다. 컴퓨터-판독가능 매체(822)는 컴퓨터-판독가능 명령어들, 데이터 구조들, 프로그램 모듈들, 및 컴퓨터 시스템(800)에 대한 다른 데이터의 저장을 제공할 수 있다. 프로세싱 서브시스템(804)에 의해 실행될 때 이상에서 설명된 기능을 제공하는 소프트웨어(프로그램들, 코드 모듈들, 명령어들)가 저장 서브시스템(818)에 저장될 수 있다. 예로서, 컴퓨터-판독가능 저장 매체(822)는 비-휘발성 메모리 예컨대 하드 디스크 드라이브, 자기 디스크 드라이브, 광 디스크 드라이브 예컨대 CD ROM, DVD, Blu-Ray® 디스크, 또는 다른 광 매체를 포함할 수 있다. 컴퓨터-판독가능 저장 매체(822)는, 비제한적으로, Zip® 드라이브들, 플래시 메모리 카드들, 범용 직렬 버스(universal serial bus; USB) 플래시 드라이브들, 보안 디지털(secure digital; SD) 카드들, DVD 디스크들, 디지털 비디오 테이프, 및 유사한 것을 포함할 수 있다. 컴퓨터-판독가능 저장 매체(822)는 또한, 비-휘발성 메모리 기반 고체-상태 드라이브(solid-state drive; SSD)들 예컨대 플래시-메모리 기반 SSD들, 기업 플래시 드라이브들, 고체 상태 ROM, 및 유사한 것, 휘발성 메모리 기반 SSD들 예컨대 고체 상태 RAM, 동적 RAM, 정적 RAM, DRAM-기반 SSD들, 자기저항성 RAM(magnetoresistive RAM; MRAM) SSD들, 및 DRAM 및 플래시 메모리 기반 SSD들의 조합을 사용하는 하이브리드 SSD들을 포함할 수 있다.

[0198] 특정 예들에서, 저장 서브시스템(818)은 또한, 추가적으로 컴퓨터-판독가능 저장 매체(822)에 연결될 수 있는 컴퓨터-판독가능 저장 매체 리더(820)를 포함할 수 있다. 리더(820)는 디스크, 플래시 드라이브, 등과 같은 메모리 디바이스로부터 데이터를 수신하고 이를 판독하도록 구성될 수 있다.

[0199] 특정 예들에서, 컴퓨터 시스템(800)은, 비제한적으로 프로세싱 및 메모리 자원들의 가상화를 포함하는, 가상화 기술들을 지원할 수 있다. 예를 들어, 컴퓨터 시스템(800)은 하나 이상의 가상 머신들을 실행하기 위한 지원을 제공할 수 있다. 특정 예들에서, 컴퓨터 시스템(800)은 가상 머신들의 구성 및 관리를 가능하는 하이퍼바이저(hypervisor)로서 프로그램을 실행할 수 있다. 각각의 가상 머신은 할당된 메모리, 연산부(예를 들어, 프로세서들, 코어들), I/O, 및 네트워킹 자원들일 수 있다. 각각의 가상 머신은 일반적으로 다른 가상 머신들과 독립적으로 실행된다. 가상 머신은 전형적으로 자체적인 운영 시스템을 실행하며, 이는 컴퓨터 시스템(800)에 의해 실행되는 다른 가상 머신들에 의해 실행되는 운영 시스템들과 동일하거나 또는 상이할 수 있다. 따라서, 복수의 운영 시스템들이 잠재적으로 컴퓨터 시스템(800)에 의해 동시에 실행될 수 있다.

[0200] 통신 서브시스템(824)은 다른 컴퓨터 시스템들 및 네트워크들에 대한 인터페이스를 제공한다. 통신 서브시스템

(824)은 컴퓨터 시스템(800)으로부터 다른 시스템들로 데이터를 송신하고 이로부터 데이터를 수신하기 위한 인터페이스로서 역할한다. 예를 들어, 통신 서브시스템(824)은, 컴퓨터 시스템(800)이 인터넷을 통해 클라이언트 디바이스들로부터 정보를 수신하고 클라이언트 디바이스들로 정보를 전송하기 위한 하나 이상의 클라이언트 디바이스들로의 통신 채널을 수립하는 것을 가능하게 한다. 예를 들어, 컴퓨터 시스템(800)이 도 1에 도시된 봇 시스템(120)을 구현하기 위해 사용될 때, 통신 서브시스템은 애플리케이션에 대해 선택된 챗봇 시스템과 통신하기 위해 사용될 수 있다.

[0201] 통신 서브시스템(824)은 유선 및/또는 무선 통신 프로토콜들 둘 모두를 지원할 수 있다. 특정 예들에서, 통신 서브시스템(824)은 (예를 들어, 셀룰러 전화기 기술, 진보된 데이터 네트워크 기술, 예컨대 3G, 4G 또는 EDGE(enhanced data rates for global evolution), WiFi(IEEE 802.XX 패밀리 표준들, 또는 다른 모바일 통신 기술들, 또는 이들의 임의의 조합)을 사용하여) 무선 음성 및/또는 데이터 네트워크들에 액세스하기 위한 라디오 주파수(radio frequency; RF) 트랜시버 구성요소들, 위성 위치확인 시스템(global positioning system; GPS) 수신기 구성요소들, 및/또는 다른 구성요소들을 포함할 수 있다. 일부 예들에 있어서, 통신 서브시스템(824)은 무선 인터페이스에 더하여 또는 그 대신에 유선 네트워크 연결(예를 들어, 이더넷)을 제공할 수 있다.

[0202] 통신 서브시스템(824)는 다양한 형태들의 데이터를 수신하고 송신할 수 있다. 일부 예들에서, 다른 형태들에 더하여, 통신 서브시스템(824)은 구조화된 및/또는 구조화되지 않은 데이터 피드들(826), 이벤트 스트림들(828), 이벤트 업데이트들(830), 및 그와 유사한 것의 형태의 입력 통신을 수신할 수 있다. 예를 들어, 통신 서브시스템(824)은 소셜 네트워크들 및/또는 다른 통신 서비스들의 사용자들로부터의 실-시간 데이터 피드들(826) 예컨대 Twitter® 피드들, Facebook® 업데이트들, 웹 피드들 예컨대 리치 사이트 서머리(Rich Site Summary; RSS) 피드들, 및/또는 하나 이상의 제3자 정보 소스들로부터의 실-시간 업데이트들을 수신(또는 전송)하도록 구성될 수 있다.

[0203] 특정 예들에서, 통신 서브시스템(824)은, 사실상 명시적인 종료를 갖지 않는 제한이 없거나 또는 연속적일 수 있는 실-시간 이벤트들 및/또는 이벤트 업데이트들(830)의 이벤트 스트림들(828)을 포함할 수 있는 연속적인 데이터 스트림들의 형태로 데이터를 수신하도록 구성될 수 있다. 연속적인 데이터를 생성하는 애플리케이션들의 예들은, 예를 들어, 센서 데이터 애플리케이션들, 금융 시세표시기들, 네트워크 성능 측정 툴들(예를 들어, 네트워크 모니터링 및 트래픽 관리 애플리케이션들), 클릭스트림 분석 툴들, 자동차 트래픽 모니터링, 및 유사한 것을 포함할 수 있다.

[0204] 통신 서브시스템(824)은 또한 컴퓨터 시스템(800)으로부터 다른 컴퓨터 시스템들 또는 네트워크들로 데이터를 통신하도록 구성될 수 있다. 데이터는, 구조화되거나 및/또는 구조화되지 않은 데이터 피드들(826), 이벤트 스트림들(828), 이벤트 업데이트들(830), 및 유사한 것과 같은 다양한 형태들로, 컴퓨터 시스템(800)에 결합된 하나 이상의 스트리밍 데이터 소스 컴퓨터들과 통신하고 있을 수 있는 하나 이상의 데이터베이스들로 통신될 수 있다.

[0205] 컴퓨터 시스템(800)은, 핸드헬드 휴대용 디바이스(예를 들어, iPhone® 셀룰러 폰, iPad® 컴퓨팅 태블릿, PDA), 웨어러블 디바이스(예를 들어, Google Glass® 머리 착용형 디스플레이), 개인용 컴퓨터, 워크스테이션, 메인프레임, 키오스크, 서버 랙, 또는 임의의 다른 데이터 프로세싱 시스템을 포함하는 다양한 유형들 중 임의의 유형일 수 있다. 컴퓨터들 및 네트워크들의 계속해서 변화하는 성질에 기인하여, 도 8에 도시된 컴퓨터 시스템(800)의 설명은 오로지 특정한 일 예로서만 의도된다. 도 8에 도시된 시스템보다 더 많거나 또는 더 적은 구성요소들을 갖는 다수의 다른 구성들이 가능하다. 본원에 제공되는 개시내용 및 교시들에 기초하여, 다양한 예들을 구현하기 위한 다른 방식들 및/또는 방법들이 있다는 것이 이해되어야 한다.

[0206] **예**

[0207] **예 1**

[0208] BERT 모델 및 디코더 모델을 포함하는 기계-학습 모델이 구성되었다. BERT 모델은 12개의 양방향 셀프-어텐션 헤드들을 갖는 12-계층 양방향 트랜스포머 인코더(도 9a에 예시된 바와 같음)를 포함한다. 각각의 계층에서, 트랜스포머는 인코더를 포함한다. 도 9b에 표시된 바와 같이, 트랜스포머의 인코더는 다중-헤드 어텐션 계층을 포함하며, 이는 입력의 개별적인 값들뿐만 아니라, 다른 값들 및 다른 값들에 기울여지는 어텐션에 기초하여 입력들을 변환할 방법을 결정하기 위해 본 명세서에서 설명된 바와 같이 값들(V), 키들(K), 및 쿼리들(Q)을 사용한다. (도 9c 참조).

[0209] BERT 모델(예를 들어, 다국어 BERT 모델)은 트레이닝 데이터 세트를 사용하여 트레이닝되었다. 이러한 트레이닝

동안, 12개의 계층들 모두에 걸친 파라미터들의 값들이 학습되었다. 그런 다음, 다수의 미세-튜닝 프로세스들 각각이 모델을 미세-튜닝하기 위해 수행되었다. 각각의 프로세스는, 미세-튜닝에서 사용할 특정 데이터 세트 및 다중-헤드 셀프-어텐션 네트워크들에서 이에 대한 파라미터 값들이 미세-튜닝될 하나 이상의 계층들의 식별에 대응한다. 트레이닝된 BERT 모델은, 이로써 모든 목적들을 위해 참조로서 통합되는 것으로 식별되는 모델을 포함할 수 있다. 미세-튜닝 프로세스들에 걸쳐, 8개의 데이터 세트들 각각은 미세-튜닝을 위해 사용되었다. 이러한 8개의 데이터 세트들 각각은, 도 10에 도시된 표의 표제에서 식별된 공개 데이터 세트이다. 도 10에 도시된 표의 상이한 행들은 얼마나 많은 계층들에 대해 다중-헤드 셀프-어텐션 네트워크들이 미세-튜닝되었는지를 나타낸다. 이에 대해 다중-헤드 셀프-어텐션 네트워크들이 미세-튜닝되는 계층들의 수 및 미세-튜닝에 대해 사용되는 데이터 세트의 각각의 조합에 대해, 도 10에 도시된 표는, 대응하는 공개적으로 이용가능한 데이터세트의 테스트 데이터 및 미세-튜닝된 모델 예측된 라벨들에 의해 생성된 마이크로 F1 스코어들을 식별한다. 마지막 열은 데이터 세트들에 걸쳐 표시된 수의 계층들을 미세-튜닝하는 것에 대한 평균 마이크로 F1 스코어를 보여준다.

[0210] 도 10에 도시된 바와 같이, 12개 전부가 아니라 5개 내지 9개 사이의 계층들의 셀프-어텐션 네트워크들이 미세-튜닝되었을 때 가장 높은 정확도가 달성되었다. 평균적으로, 가장 높은 정확도는, 계층 8의 다중헤드 셀프 어텐션 네트워크만이 미세-튜닝되었을 때 달성되었다. 추가적으로, 선택된 계층 내의 트랜스포머 블록 내의 다중-헤드 어텐션 네트워크들만을 미세-튜닝하는 것은 미세-튜닝에 대한 시간을 75%만큼 감소시켰다.

[0211] **예 2**

[0212] 예 1에서 식별된 것과 동일한 아키텍처를 포함하는 기계-학습 모델(도 11에 예시된 바와 같은 계층들을 가짐)이 구성된다. 이러한 아키텍처는 12-계층 양방향 트랜스포머 인코더 및 활성화 계층(1110)을 갖는 BERT 모델(1105)을 포함한다. 계층들의 초기 세트(1115)(디코더 모델/활성화 계층 및 BERT 모델 내의 모든 계층들을 포함함)는 초기 트레이닝 데이터 세트를 사용하여 초기 트레이닝 스테이지 동안 트레이닝된다. 계층들의 미세-튜닝 세트(1120)는 계층들의 초기 세트(1115)의 불완전한 서브세트인 것으로 식별된다. 미세-튜닝 계층(1120) 내의 단일 계층의 부분(예를 들어, 각각의 다중-헤드 어텐션 구성요소)만이 상이한(예를 들어, 컨텍스트-특정) 트레이닝 세트를 사용하여 미세-튜닝된다. 그런 다음, 미세-튜닝된 모델은, 전체 기계-학습 모델이 컨텍스트-특정 데이터를 사용하여 트레이닝되었던 경우에 필요한 것보다 더 적은 컨텍스트-특정 데이터를 사용하면서 컨텍스트에 대응하는 데이터를 프로세싱하도록 구성된다.

[0213] **추가적인 고려사항들:**

[0214] 특정 세부사항들 및 예들이 설명되었지만, 다양한 수정예들, 대안예들, 대안적인 구성들, 및 등가물들이 또한 가능하다. 예들은 정확한 특정 데이터 프로세싱 환경들 내에서 동작하도록 제한되는 것이 아니라, 복수의 데이터 프로세싱 환경들 내에서 자유롭게 동작할 수 있다. 추가적으로, 특정 예들이 특정한 일련의 트랜잭션(transaction)들 및 단계들을 사용하여 설명되었지만, 이것이 제한적으로 의도되지 않는다는 것이 당업자들에게 명백할 것이다. 어떤 순서도들이 순차적인 프로세스로서 동작들을 설명할 수 있지만, 동작들 중 다수는 병렬로 또는 동시에 수행될 수 있다. 이에 더하여, 동작들의 순서는 재배열될 수 있다. 프로세스는 도면에 포함되지 않은 추가적인 단계들을 가질 수 있다. 이상에서 설명된 예들의 다양한 특징들 및 측면들이 개별적으로 또는 함께 사용될 수 있다.

[0215] 추가로, 특정 예들이 하드웨어 및 소프트웨어의 특정한 조합을 사용하여 설명되었지만, 하드웨어 및 소프트웨어의 다른 조합들이 또한 가능하다는 것이 이해되어야 한다. 특정 예들은 하드웨어로만, 또는 소프트웨어로만, 또는 이들의 조합들을 사용하여 구현될 수 있다. 본 명세서에서 설명된 다양한 프로세스들은 동일한 프로세서 또는 임의의 조합의 상이한 프로세서들 상에 구현될 수 있다.

[0216] 구성요소들 또는 모듈들이 특정 동작들 또는 기능들을 수행하도록 구성된 것으로 설명되는 경우, 이러한 구성은, 예를 들어, 동작을 수행하기 위한 전자 회로들을 설계함으로써, 동작을 수행하기 위하여 프로그램가능 전자 회로들(예컨대 마이크로프로세서들)을 프로그래밍함으로써 예컨대 컴퓨터 명령어들 또는 코드, 또는 비-일시적인 메모리 매체 상에 저장된 코드 또는 명령어들을 실행하도록 프로그래밍된 프로세서들 또는 코어들을 실행함으로써, 또는 이들의 임의의 조합에 의해 달성될 수 있다. 프로세스들은 비제한적으로 프로세스-간 통신을 위한 통상적을 기술을 포함하는 다양한 기술들을 사용하여 통신할 수 있으며, 프로세스들의 상이한 쌍들이 상이한 기술들을 사용할 수 있거나, 또는 프로세스들의 동일한 쌍이 상이한 시점에 상이한 기술들을 사용할 수 있다.

[0217] 다수의 특정 세부사항들이 예들의 완전한 이해를 제공하기 위하여 본 개시내용에서 주어진다. 그러나, 예들은

이러한 특정 세부사항들 없이 실시될 수 있다. 예를 들어, 잘 알려진 회로들, 프로세스들, 알고리즘들, 구조들, 및 기술들은 예들을 모호하게 하는 것을 피하기 위하여 불필요한 세부사항 없이 도시되었다. 이러한 설명은 오로지 예시적인 예들만을 제공하며, 다른 예들의 범위, 적용가능성, 또는 구성을 제한하도록 의도되지 않는다. 오히려, 예들의 이상의 설명은 당업자들에게 다양한 예들을 구현하기 위한 사용 가능한 설명을 제공할 것이다. 요소들의 기능 및 배열에 있어서 다양한 변화들이 이루어질 수 있다.

[0218] 따라서 본 명세서 및 도면은 제한적인 의미라기보다는 예시적인 의미로 간주되어야 한다. 그러나, 청구항들에 기술되는 바와 같은 광범위한 사상 및 범위로부터 벗어나지 않고 이에 대한 추가들, 대체들, 삭제들, 및 다른 수정들 및 변화들이 이루어질 수 있다는 것이 명백할 것이다. 따라서, 특정 예들이 설명되었지만, 이들은 제한적인 것으로 의도되지 않는다. 다양한 수정예들 및 등가물들이 다음의 청구항들의 범위 내에 속한다.

[0219] 이상의 명세서에서, 본 개시내용의 측면들이 본 개시의 특정 예들을 참조하여 설명되었지만, 당업자들은 본 개시가 이에 한정되지 않는다는 것을 인식할 것이다. 이상에서 설명된 개시내용의 다양한 특징들 및 측면들이 개별적으로 또는 함께 사용될 수 있다. 추가로, 예들은 본 명세서의 광범위한 사상 및 범위로부터 벗어나지 않고 본원에서 설명된 것들을 넘어 임의의 수의 환경들 및 애플리케이션들에서 사용될 수 있다. 이에 따라, 본 명세서 및 도면들은 제한적인 것이 아니라 예시적인 것으로 간주되어야 한다.

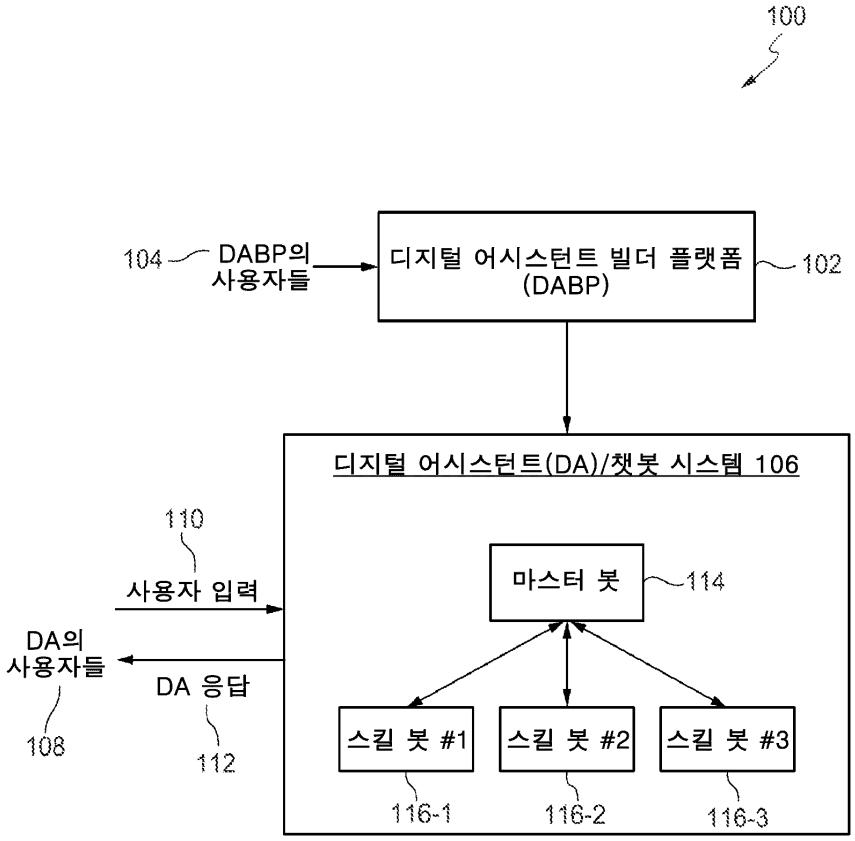
[0220] 이상의 설명서, 예시의 목적들을 위하여, 방법들이 특정한 순서로 설명되었다. 대안적인 예들에서, 방법들은 설명된 것과 상이한 순서로 수행될 수 있다는 것이 이해되어야 한다. 이상에서 설명된 방법들은 하드웨어 구성요소들에 위해 수행될 수 있거나 또는, 범용 또는 전용 프로세서 또는 로직 회로들과 같은 기계가 방법들을 수행하기 위한 명령어들로 프로그래밍되게끔 하기 위하여 사용될 수 있는 기계-실행가능 명령어들의 시퀀스들로 구현될 수 있다. 이러한 기계-실행가능 명령어들은 하나 이상의 기계 판독가능 매체들, 예컨대 CD-ROM들, 또는 다른 유형의 광 디스크들, 플로피 디스켓들, ROM들, RAM들, EPROM들, EEPROM들, 자기 또는 광 카드들, 플래시 메모리, 또는 전자적 명령어들을 저장하기에 적절한 다른 유형들의 기계-판독가능 매체들 상에 저장될 수 있다. 대안적으로, 방법들은 하드웨어 및 소프트웨어의 조합에 의해 수행될 수 있다.

[0221] 구성요소들이 특정 동작을 수행하도록 "구성된" 것으로 기술되는 경우, 이러한 구성은 예를 들어, 전자 회로 또는 다른 하드웨어를 설계하여 그 동작을 수행하는 것에 의해, 프로그래밍 가능한 전자 회로(예를 들어, 마이크로프로세서 또는 다른 적절한 전자 회로)를 프로그래밍하여 그 동작을 수행하는 것에 의해 또는 이들의 임의의 조합에 의해, 달성될 수 있다.

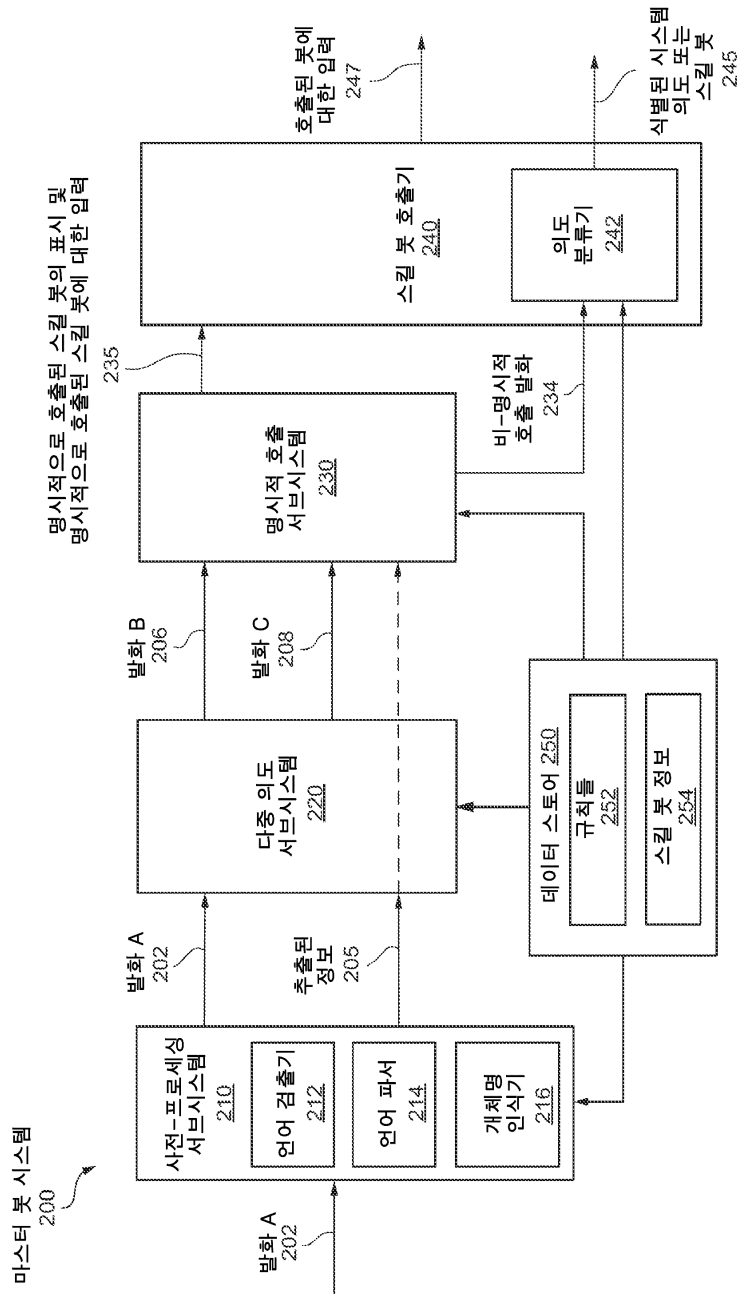
[0222] 본 출원의 예시적인 예들이 본 명세서에서 상세히 설명되었지만, 본 개시내용의 개념들은 다른 식으로 다양하게 구현 및 이용될 수 있고, 첨부된 청구항들은 종래 기술에 의해 제한된 것을 제외하면, 이러한 변형들을 포함하는 것으로 해석되어야 한다고 이해되어야 한다.

도면

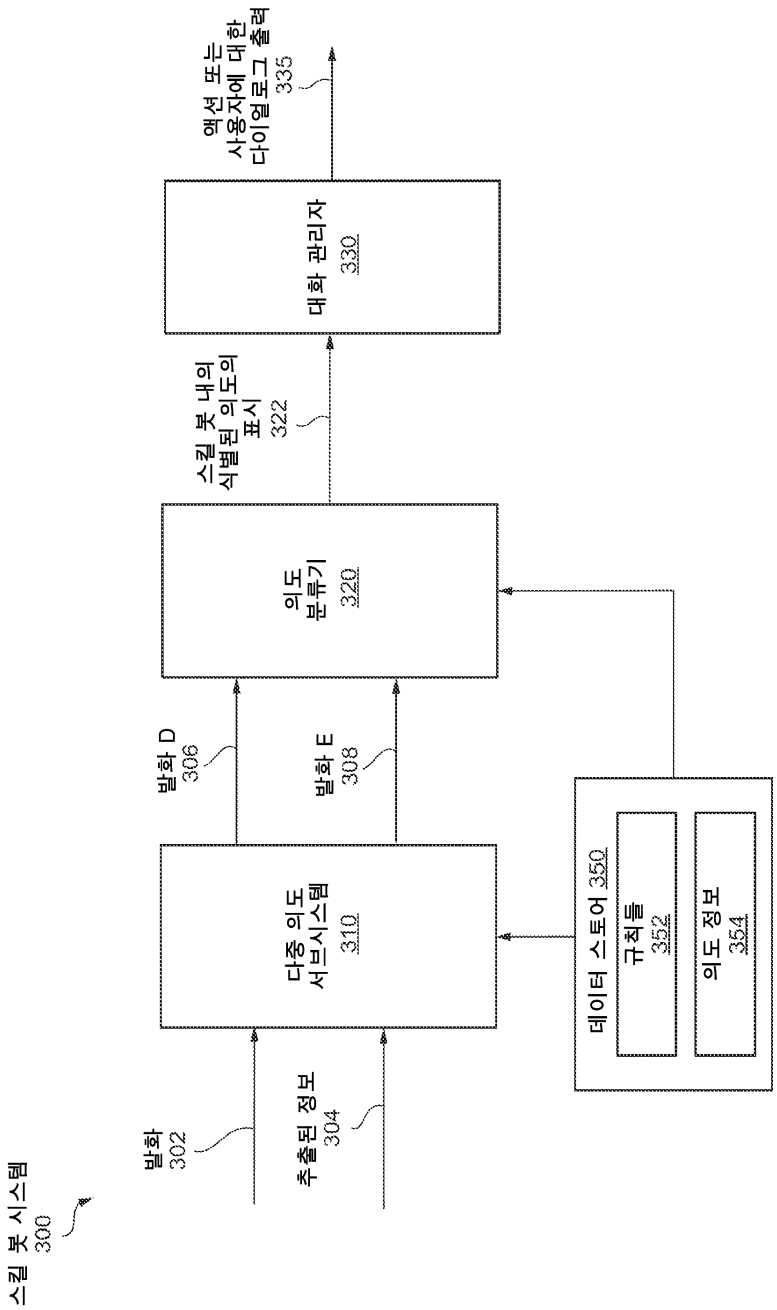
도면1


100
104
DABP의
사용자들
디지털 어시스턴트 빌더 플랫폼
(DABP)
102
디지털 어시스턴트(DA)/챗봇 시스템 106
110
사용자 입력
마스터 봇
114
DA의
사용자들
DA 응답
스킬 봇 #1
스킬 봇 #2
스킬 봇 #3
108
112
116-1
116-2
116-3

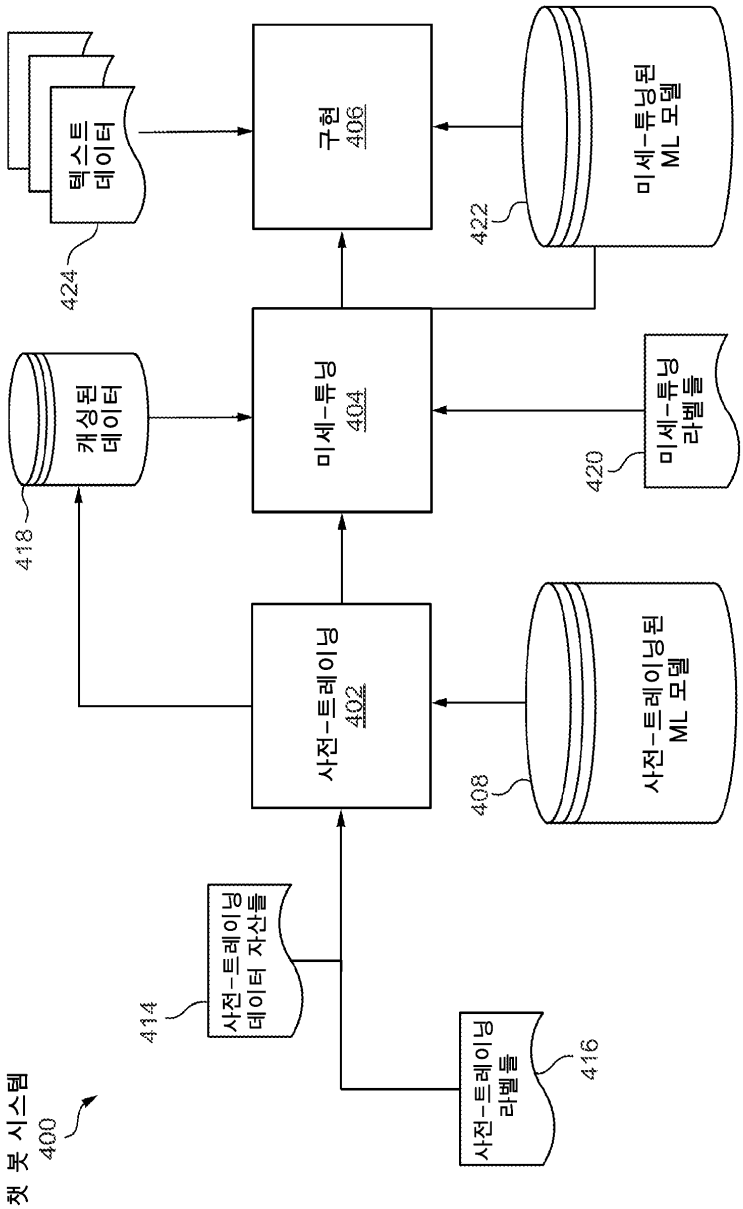
도면2


마스터 봇 시스템
200
발화 A
202
사전-프로세싱 서브시스템
210
언어 검출기
212
언어 파서
214
개체명 인식기
216
발화 A
202
추출된 정보
205
다중 의도 서브시스템
220
발화 B
206
발화 C
208
명시적 호출 서브시스템
230
명시적으로 호출된 스킬 봇의 표시 및 명시적으로 호출된 스킬 봇에 대한 입력
235
비-명시적 호출 발화
234
데이터 스토어 250
규칙들
252
스킬 봇 정보
254
스킬 봇 호출기
240
의도 분류기
242
호출된 봇에 대한 입력
247
식별된 시스템 의도 또는 스킬 봇
245

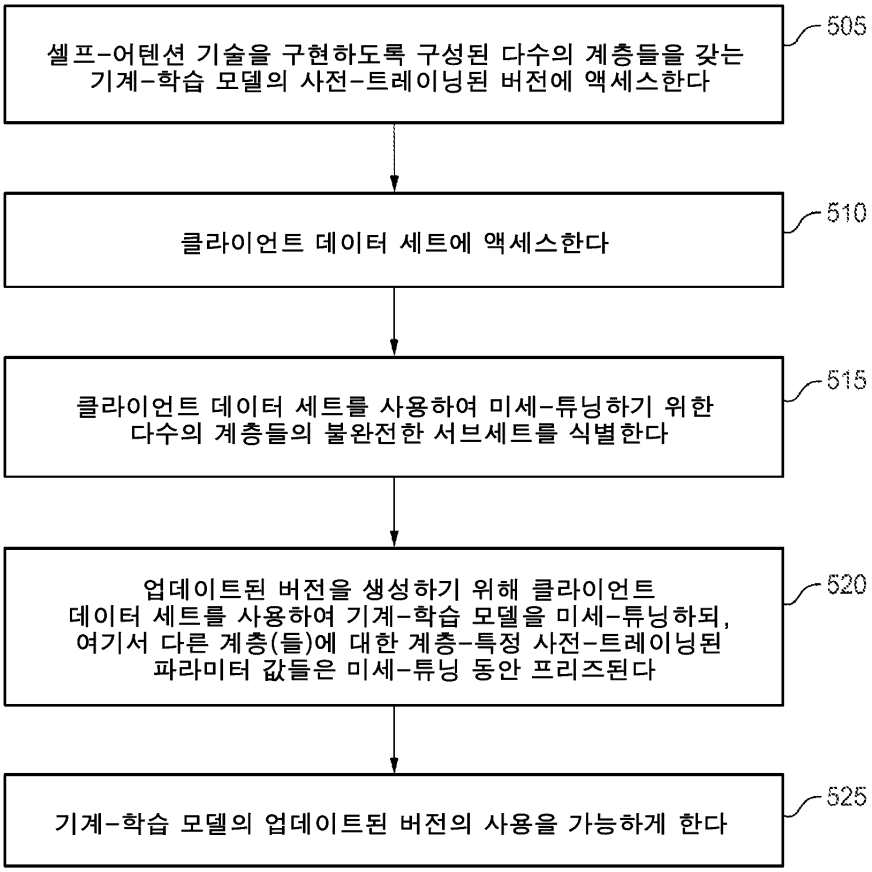
도면3


스킬 봇 시스템
300
발화
302
추출된 정보
304
다중 의도
서브시스템
310
발화 D
306
발화 E
308
의도
분류기
320
스킬 봇 내의
식별된 의도의
표시
322
대화 관리자
330
액션 또는
사용자에 대한
다이얼로그 출력
335
데이터 스토어 350
규칙들
352
의도 정보
354

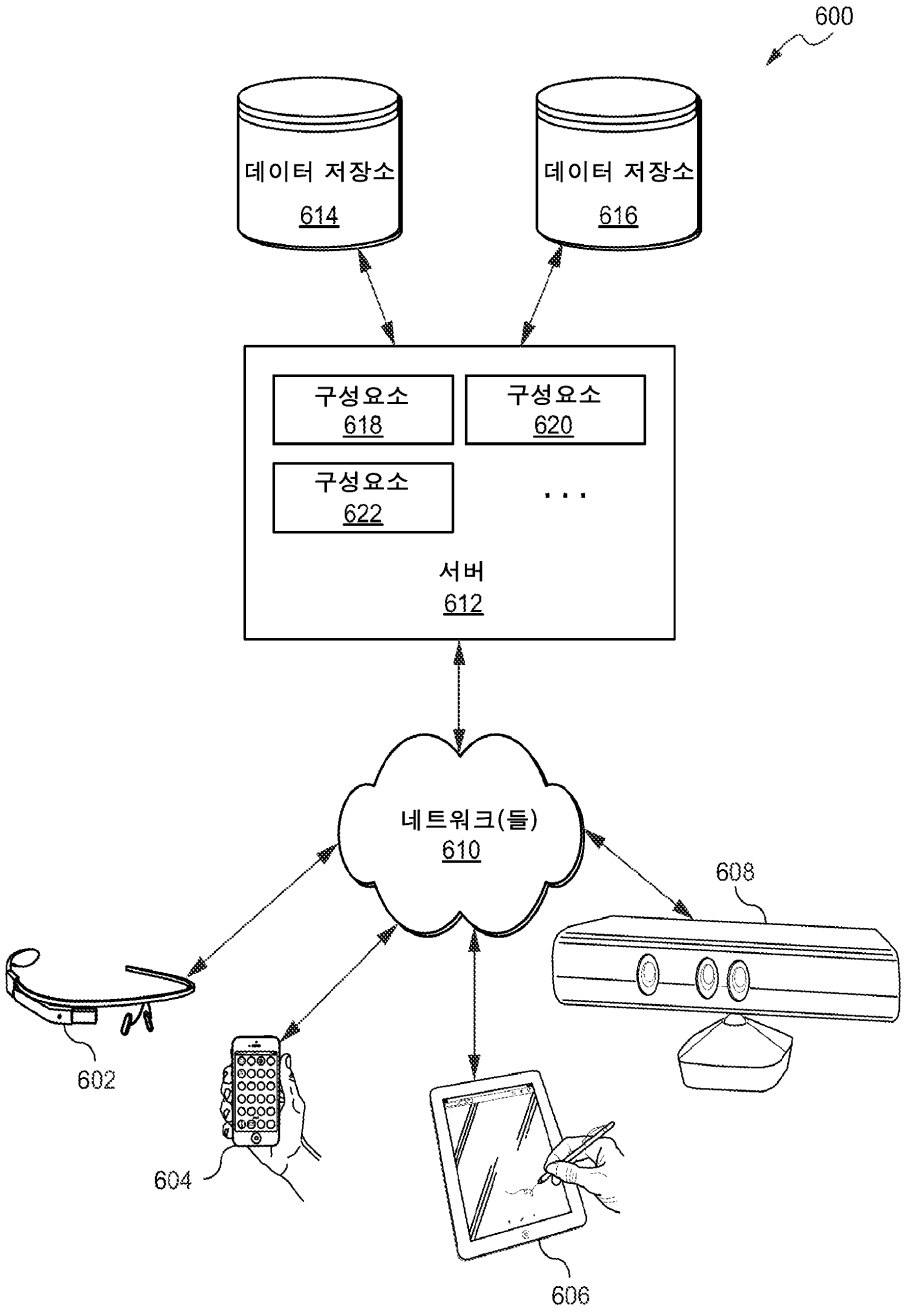
도면4


챗 봇 시스템
400
사전-트레이닝 데이터 자산들
414
사전-트레이닝 라벨들
416
사전-트레이닝 402
사전-트레이닝된 ML 모델
408
캐싱된 데이터
418
미세-튜닝 404
미세-튜닝 라벨들
420
구현 406
미세-튜닝된 ML 모델
422
텍스트 데이터
424

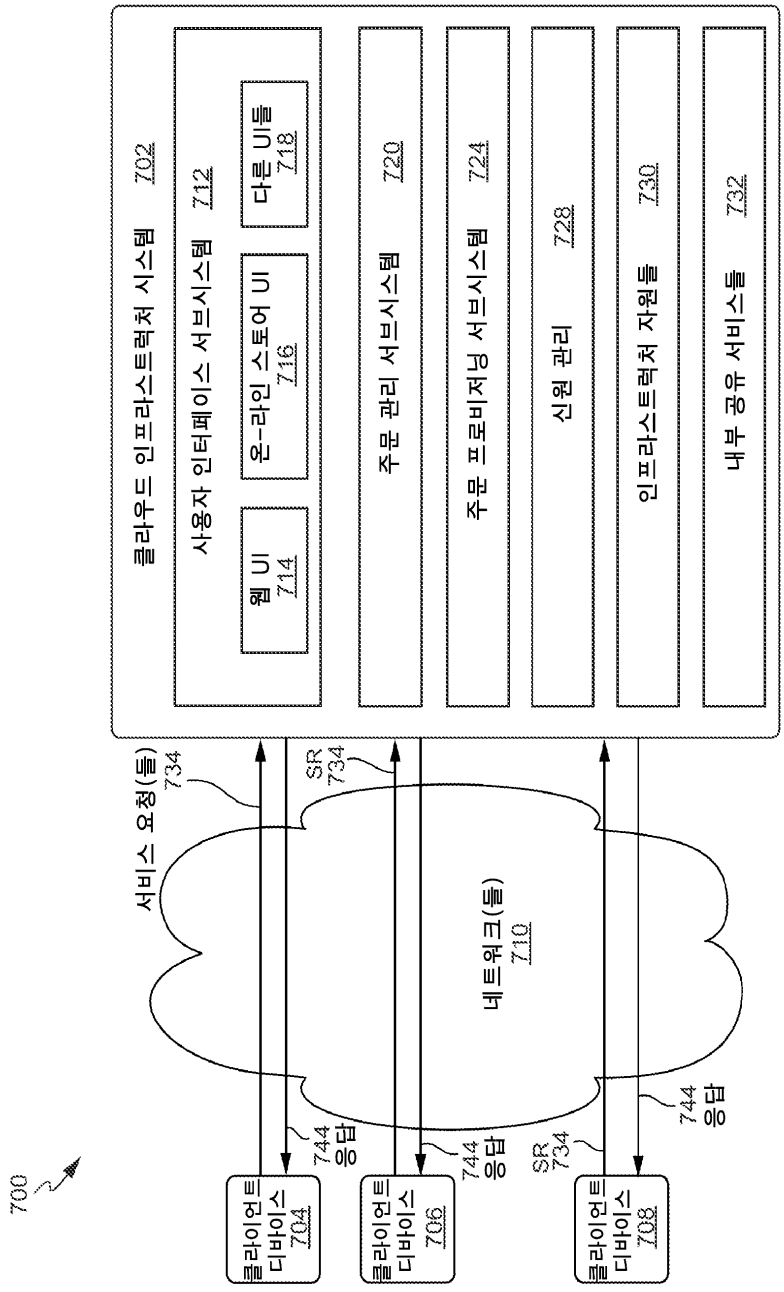
도면5


셀프-어텐션 기술을 구현하도록 구성된 다수의 계층들을 갖는 기계-학습 모델의 사전-트레이닝된 버전에 액세스한다
505
클라이언트 데이터 세트에 액세스한다
510
클라이언트 데이터 세트를 사용하여 미세-튜닝하기 위한 다수의 계층들의 불완전한 서브세트를 식별한다
515
업데이트된 버전을 생성하기 위해 클라이언트 데이터 세트를 사용하여 기계-학습 모델을 미세-튜닝하되, 여기서 다른 계층(들)에 대한 계층-특정 사전-트레이닝된 파라미터 값들은 미세-튜닝 동안 프리즈된다
520
기계-학습 모델의 업데이트된 버전의 사용을 가능하게 한다
525

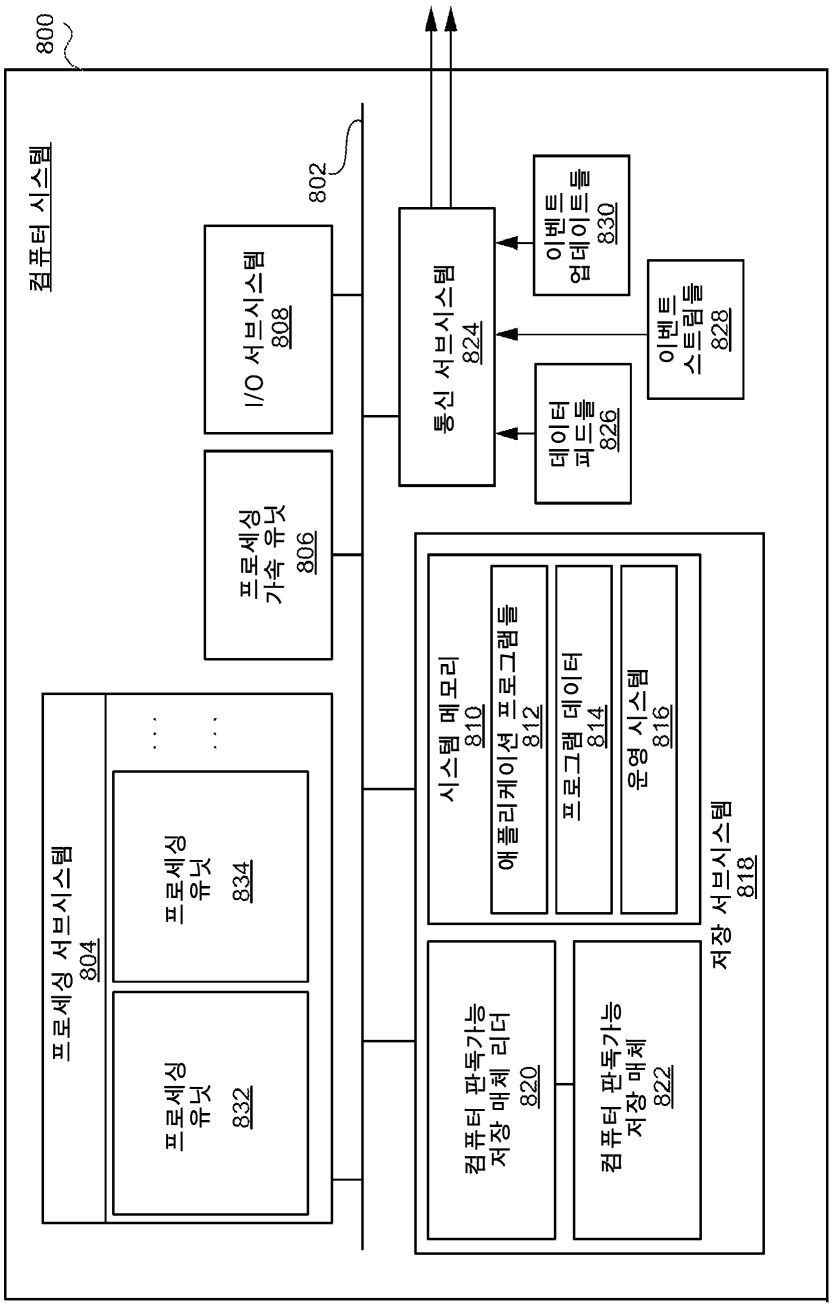
도면6


600
데이터 저장소
614
데이터 저장소
616
구성요소
618
구성요소
620
구성요소
622
. . .
서버
612
네트워크(들)
610
608
602
604
606

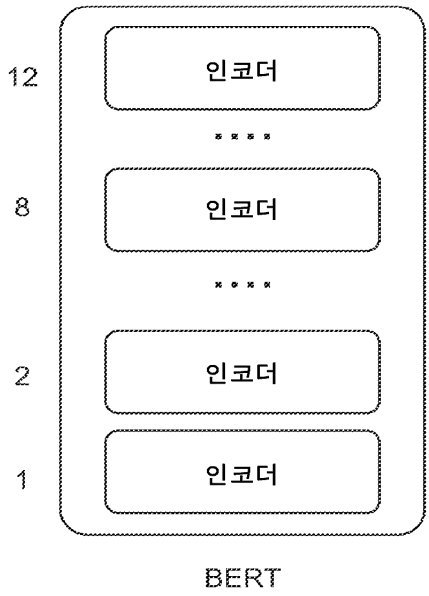
도면7


700
클라우드 인프라스트럭처 시스템 702
사용자 인터페이스 서브시스템 712
웹 UI 714
온-라인 스토어 UI 716
다른 UI들 718
주문 관리 서브시스템 720
주문 프로비저닝 서브시스템 724
신원 관리 728
인프라스트럭처 자원들 730
내부 공유 서비스들 732
서비스 요청(들) 734
SR 734
네트워크(들) 710
744 응답
744 응답
744 응답
SR 734
클라이언트 디바이스 704
클라이언트 디바이스 706
클라이언트 디바이스 708

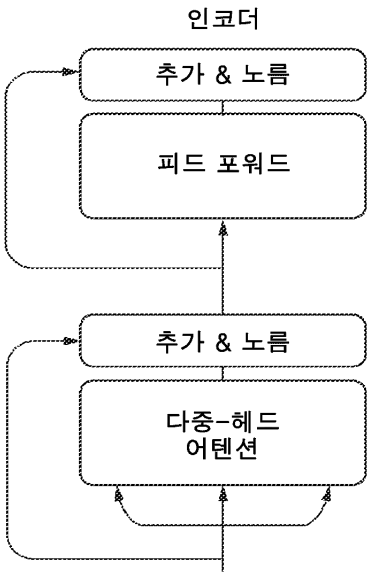
도면8


800
컴퓨터 시스템
802
I/O 서브시스템
808
프로세싱
가속 유닛
806
통신 서브시스템
824
이벤트
업데이트들
830
이벤트
스트림들
828
데이터
피드들
826
프로세싱 서브시스템
804
프로세싱
유닛
832
프로세싱
유닛
834
시스템 메모리
810
애플리케이션 프로그램들
812
프로그램 데이터
814
운영 시스템
816
컴퓨터 판독가능
저장 매체 리더
820
컴퓨터 판독가능
저장 매체
822
저장 서브시스템
818

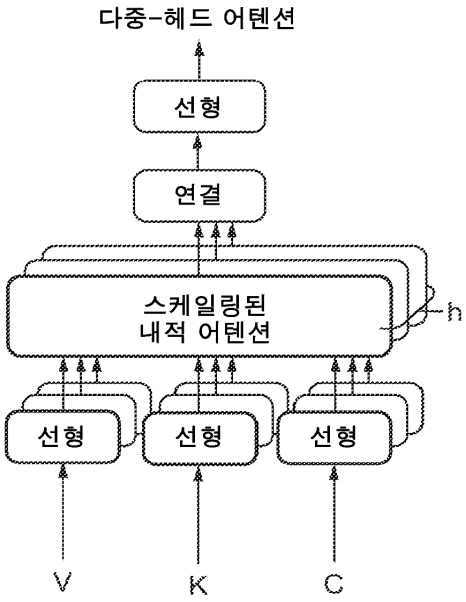
도면9a



도면9b



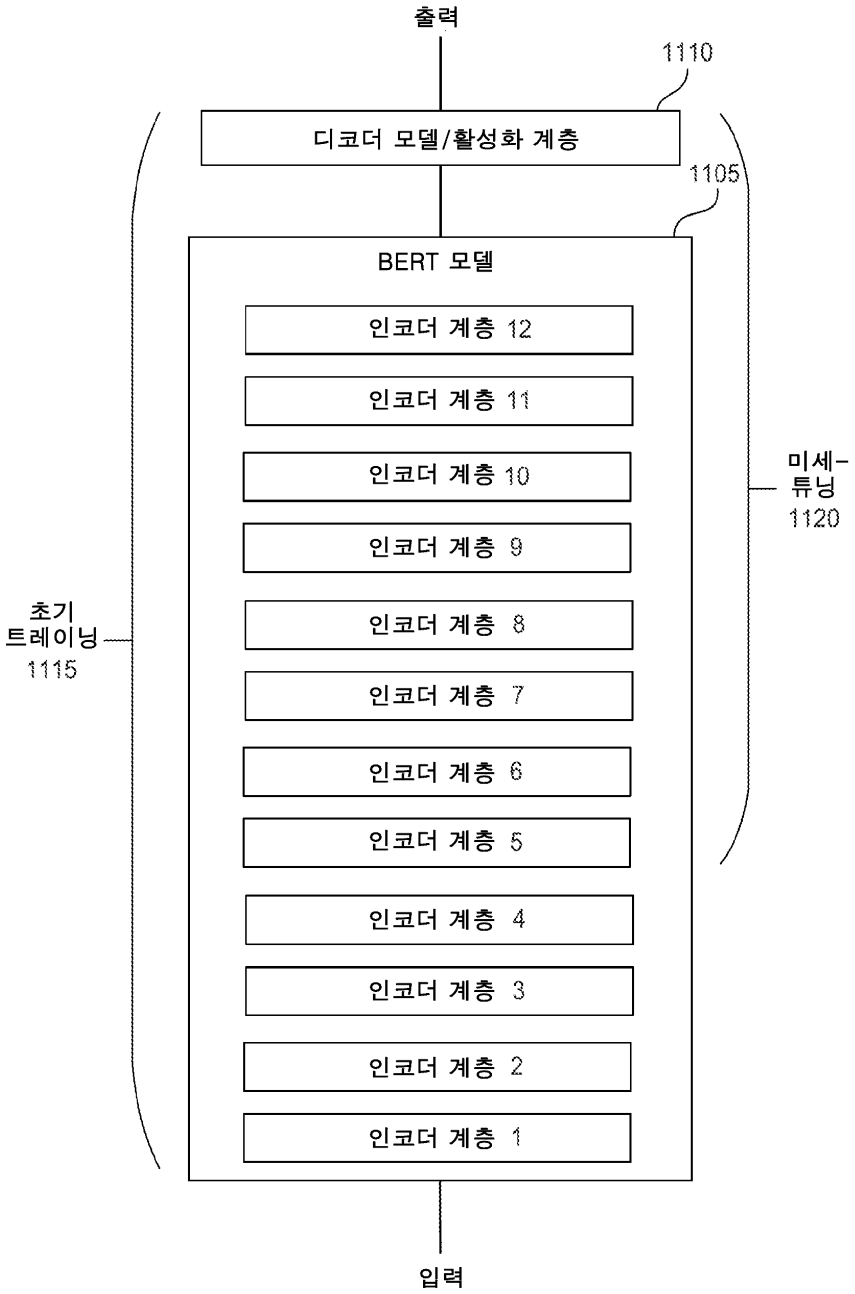
도면9c



도면10

	CoNLL2003	MITMovie	MIT 레스토랑	ATIS	MultiWOZ	SNIPS	OntoNotes_en	WikiGold	평균
n (MLP-CRF) + 미세튜닝 다중헤드 셀프 어텐션만 (1.7M개의 트레이닝가능 파라미터들)									
12	0.878	0.692	0.743	0.939	0.823	0.897	0.776	0.744	0.811
11	0.885	0.697	0.757	0.948	0.860	0.910	0.802	0.767	0.828
10	0.885	0.705	0.762	0.949	0.870	0.912	0.810	0.755	0.831
9	0.891	**0.707**	0.769	0.952	0.874	0.923	0.816	**0.783**	0.839
8	**0.892**	0.705	0.769	0.955	0.887	**0.927**	**0.817**	0.779	**0.841**
7	0.887	0.700	**0.771**	**0.956**	0.891	0.923	0.813	0.758	0.837
6	0.883	0.697	**0.771**	0.954	**0.894**	0.920	0.806	0.729	0.832
5	0.876	0.693	**0.771**	0.952	0.882	0.923	0.797	0.725	0.827
4	0.859	0.678	0.756	0.948	0.885	0.911	0.779	0.664	0.810
3	0.852	0.670	0.763	0.946	0.885	0.920	0.772	0.651	0.807
2	0.832	0.654	0.739	0.945	0.884	0.907	0.746	0.594	0.788
1	0.792	0.635	0.713	0.927	0.861	0.876	0.692	0.534	0.754
0 (임베딩 계층)	0.595	0.483	0.596	0.714	0.449	0.568	0.478	0.344	0.528

도면11


출력
1110
디코더 모델/활성화 계층
1105
BERT 모델
인코더 계층 12
인코더 계층 11
인코더 계층 10
인코더 계층 9
인코더 계층 8
인코더 계층 7
인코더 계층 6
인코더 계층 5
인코더 계층 4
인코더 계층 3
인코더 계층 2
인코더 계층 1
미세-튜닝
1120
초기 트레이닝
1115
입력