



(51) International Patent Classification:

G06F 16/332 (2019.01) G06N 20/00 (2019.01)
G06F 16/33 (2019.01)

(21) International Application Number:

PCT/US2022/012972

(22) International Filing Date:

19 January 2022 (19.01.2022)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/139,569 20 January 2021 (20.01.2021) US

(71) Applicant: **ORACLE INTERNATIONAL CORPORATION** [US/US]; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US).

(72) Inventors: **VU, Duy**; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US). **PHAM, Tuyen Quang**; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US). **HOANG, Cong Duy Vu**; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US). **GADDE, Srinivasa Phani Kumar**; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US). **DUONG, Thanh Long**; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US). **JOHNSON, Mark Edward**; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US). **VISHNOI, Vishal**; 500 Oracle Parkway M/S 50P7, Redwood Shores, California 94065 (US).

(74) Agent: **ROTHWELL, Rodney** et al.; 1100 Peachtree Street NE, Suite 2800, Atlanta, Georgia 30309 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,

(54) Title: CONTEXT TAG INTEGRATION WITH NAMED ENTITY RECOGNITION MODELS

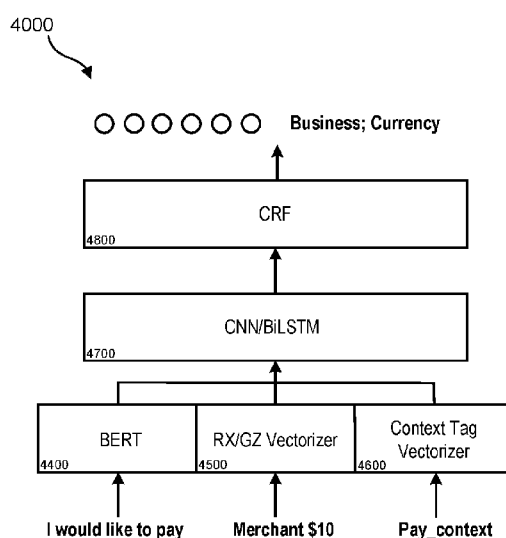


FIG. 4B

(57) **Abstract:** Techniques are provided for using context tags in named-entity recognition (NER) models. In one particular aspect, a method is provided that includes receiving an utterance, generating embeddings for words of the utterance, generating a regular expression and gazetteer feature vector for the utterance, generating a context tag distribution feature vector for the utterance, concatenating or interpolating the embeddings with the regular expression and gazetteer feature vector and the context tag distribution feature vector to generate a set of feature vectors, generating an encoded form of the utterance based on the set of feature vectors, generating log-probabilities based on the encoded form of the utterance, and identifying one or more constraints for the utterance.

CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN,
KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD,
ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO,
NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW,
SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

CONTEXT TAG INTEGRATION WITH NAMED ENTITY RECOGNITION MODELS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims the benefit and priority of U.S. Provisional Application No. 63/139,569, filed on January 20, 2021, the entire contents of which are incorporated herein
5 by reference for all purposes.

FIELD

[0002] The present disclosure relates generally to chatbot systems, and more particularly, to techniques for adding context tags to Named Entity Recognition (NER) models.

BACKGROUND

10 [0003] People around the world use instant messaging or chat platforms in order to get instant reactions. Organizations often use these instant messaging or chat platforms to engage with customers (or end users) in live conversations. However, it can be very costly for organizations to employ service people to engage in live communication with customers or end users. Chatbots or bots have been developed to simulate conversations with end users, especially over the
15 Internet. End users can communicate with such bots through messaging apps. An intelligent bot, generally a bot powered by artificial intelligence (AI), can communicate intelligently and contextually in live conversations with end users, which allows for a more natural conversation and an improved conversational experience. Instead of relying on a fixed set of keywords or commands, intelligent bots may be able to receive utterances of end users in natural language,
20 understand their intentions, and respond accordingly.

[0004] However, chatbots are difficult to build because they require specific knowledge in certain fields and the application of certain techniques that may be solely within the capabilities of specialized developers. To build these chatbots, developers seek to understand the needs of end users and build machine learning (ML) model(s) tailored to their needs. The task of building
25 ML model(s) typically involves developing and testing multiple models using unsupervised and/or supervised learning-based solutions. In some cases, building an ML model involves a training phase, an application (*i.e.*, inference) phase, and iterative loops between the training phase and the application phase. In some cases, precise training data is required to enable the

algorithms to understand and learn certain patterns or features such that the trained ML model(s) can predict a desired outcome (e.g., inference of an intent from an utterance).

BRIEF SUMMARY

[0005] Techniques are disclosed for adding context tags to NER models.

5 **[0006]** In various embodiments, a computer-implemented method comprises receiving, at a chatbot system comprising a processor, at least one utterance comprising one or more words, generating, by a transformer-based model of the chatbot system, a plurality of embeddings for the one or more words of the at least one utterance, generating, by a first vectorizer of the chatbot system, at least one regular expression and gazetteer feature vector for the at least one utterance, 10 generating, by a second vectorizer of the chatbot system, at least one context tag distribution feature vector for the at least one utterance, concatenating or interpolating the plurality of embeddings with the at least one regular expression and gazetteer feature vector and the at least one context tag distribution feature vector to generate a first set of feature vectors, generating, by a main sequence model of the chatbot system, an encoded form of the at least one utterance 15 based on the first set of feature vectors, generating, by a discriminative model of the chatbot system, a plurality of log-probabilities for candidate entities based on the encoded form of the at least one utterance, and identifying, using the plurality of log-probabilities, one or more constraints for the at least one utterance based on the candidate entities.

[0007] In some embodiments, the least one utterance comprises at least one of one or more queries of the chatbot system, one or more queries input to the chatbot system by a user, one or more responses provided by the user in response to the one or more queries of the chatbot system, or combination thereof. 20

[0008] In some embodiments, the transformer-based model of the chatbot system comprises of a bidirectional encoder representations from transformers model.

25 **[0009]** In some embodiments, the first vectorizer generates the at least one regular expression and gazetteer feature vector based on one or more regular expression patterns and one or more gazetteers.

[0010] In some embodiments, wherein the second vectorizer generates the at least one context tag distribution feature vector based on a context of at least one of one or more queries of the

chatbot system, one or more queries input to the chatbot system by a user, one or more responses provided by the user in response to the one or more queries of the chatbot system, or a combination thereof.

[0011] In some embodiments, the main sequence model of the chatbot system comprises a combined convolutional neural network/bidirectional long short-term memory model.

[0012] In some embodiments, the discriminative model of the chatbot system comprises a conditional random field model.

[0013] Some embodiments of the present disclosure include a system including one or more data processors and a non-transitory computer readable storage medium containing instructions which, when executed on the one or more data processors, cause the one or more data processors to perform part or all of one or more methods and/or part or all of one or more processes disclosed herein.

[0014] Some embodiments of the present disclosure include a computer-program product tangibly embodied in a non-transitory machine-readable storage medium, including instructions configured to cause one or more data processors to perform part or all of one or more methods and/or part or all of one or more processes disclosed herein.

[0015] The techniques described above and below may be implemented in a number of ways and in a number of contexts. Several example implementations and contexts are provided with reference to the following figures, as described below in more detail. However, the following implementations and contexts are but a few of many.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] FIG. 1 is a simplified block diagram of a distributed environment incorporating an exemplary embodiment.

[0017] FIG. 2 is a simplified block diagram of a computing system implementing a master bot according to certain embodiments.

[0018] FIG. 3 is a simplified block diagram of a computing system implementing a skill bot according to certain embodiments.

[0019] FIG. 4A is a simplified block diagram of a chatbot training and deployment system in accordance with various embodiments.

[0020] FIG. 4B is a simplified block diagram of a Named Entity Recognition (NER) architecture in accordance with various embodiments.

5 [0021] FIG. 5 illustrates a process flow for taking context into consideration for entity recognition in accordance with various embodiments.

[0022] FIG. 6 depicts a simplified diagram of a distributed system for implementing various embodiments.

10 [0023] FIG. 7 is a simplified block diagram of one or more components of a system environment by which services provided by one or more components of an embodiment system may be offered as cloud services, in accordance with various embodiments.

[0024] FIG. 8 illustrates an example computer system that may be used to implement various embodiments.

DETAILED DESCRIPTION

15 [0025] In the following description, for the purposes of explanation, specific details are set forth in order to provide a thorough understanding of certain embodiments. However, it will be apparent that various embodiments may be practiced without these specific details. The figures and description are not intended to be restrictive. The word “exemplary” is used herein to mean
20 “serving as an example, instance, or illustration.” Any embodiment or design described herein as “exemplary” is not necessarily to be construed as preferred or advantageous over other embodiments or designs.

Introduction

25 [0026] A digital assistant is an artificially intelligent driven interface that helps users accomplish a variety of tasks in natural language conversations. For each digital assistant, a customer may assemble one or more skills. Skills (also described herein as chatbots, bots, or skill bots) are individual bots that are focused on specific types of tasks, such as tracking inventory, submitting timecards, and creating expense reports. When an end user engages with the digital assistant, the digital assistant evaluates the end user input and routes the conversation to and

from the appropriate chatbot. The digital assistant can be made available to end users through a variety of channels such as FACEBOOK[®] Messenger, SKYPE MOBILE[®] messenger, or a Short Message Service (SMS). Channels carry the chat back and forth from end users on various messaging platforms to the digital assistant and its various bots. The channels may also support user agent escalation, event-initiated conversations, and testing.

[0027] Intents allow the chatbot to understand what the user wants the chatbot to do. Intents are the user's intention communicated to the chatbot via user requests and statements, which are also referred to as utterances (e.g., get account balance, make a purchase, etc.). As used herein, an utterance or a message may refer to a set of words (e.g., one or more sentences) exchanged during a conversation with a chatbot. Intents may be created by providing a name that illustrates some user action (e.g., order a pizza) and compiling a set of real-life user statements, or utterances that are commonly associated with triggering the action. Because the chatbot's cognition is derived from these intents, each intent may be created from a data set that is robust (one to two dozen utterances) and varied, so that the chatbot may interpret ambiguous user input. A rich set of utterances enables a chatbot to understand what the user wants when it receives messages like "Forget this order!" or "Cancel delivery!" – messages that mean the same thing but are expressed differently. Collectively, the intents, and the utterances that belong to them, make up a training corpus for the chatbot. By training an algorithm with the corpus, a customer turns that algorithm into a model that serves as a reference tool for resolving end user input(s) to a single intent. A customer can improve the acuity of the chatbot's cognition through rounds of intent testing and intent training.

[0028] However, building a chatbot that can determine end users' intents based upon the end users' utterances is a challenging task at least due to the subtleties and ambiguity of natural languages and the dimensions of the input/output space (e.g., possible user utterances, number of intents, etc.). An illustrative example of this difficulty arises from characteristics of natural language, such as employing euphemisms, synonyms, or ungrammatical speech to express intent. For example, an utterance may express an intent to order a pizza without explicitly mentioning the words pizza, ordering, or delivery. These characteristics of natural language give rise to uncertainty and result in chatbots using confidence as a parameter for prediction of user intents. As such, chatbots may need to be trained, monitored, debugged, and retrained in order to

improve the performance of the chatbot and user experience with the chatbot. In conventional spoken language understanding (SLU) and natural language processing (NLP) systems, training mechanisms are provided for training and retraining machine-learning algorithms of the digital assistant or chatbot included therein. Conventionally, these algorithms are trained with

5 “manufactured” utterances for any intent. For example, the utterance “Do you do price changes?” may be used to train a classification algorithm of a chatbot system to classify this type of utterance as the intent – “Do you offer a price match.” The training of algorithms with manufactured utterances helps initially train the chatbot system for providing services and re-train the chatbot system once it is deployed and receives utterances from users.

10 **[0029]** A user utterance may contain a name entity. In addition to intention, named entities further allow a chatbot to understand the meaning of the user’s utterances. Named entities modify an intent(s). For example, if a user types “show me yesterday’s financial news”, the named entities “yesterday” and “financial” assist the chatbot in understanding the user’s request. Entities may be categorized according to what they represent. For example, “yesterday”
15 may be categorized as “dateTime” and “financial” may be categorized as “newsType.” Entities are sometimes referred to as slots. Named Entity Recognition (NER) is a tool used by chatbot systems to automatically recognize and extract entities. NER typically involves named entity resolution and named entity disambiguation. Named entity resolution involves identifying named entities in a sequence of words and name entity disambiguation involves identifying the exact
20 referent of each named entity in the sequence of words. For example, with respect to the toponym: “Paris”. People will generally assume that the referent of this entity is the city of Paris in France since the city of Paris in France is widely known. However, there are other possible referents for the toponym “Paris” (e.g., referents may include Paris, Texas, USA; Paris, Ontario, Canada; Paris, Panama; Paris, Togo, etc.). Additionally, referents may include a person
25 named Paris or a commercial entity or business enterprise named Paris. Because the referent of a named entity may not always correspond to an obvious or popular referent, identifying the intended referent is challenging.

[0030] To overcome these challenges and others to correctly identify the intended referent for a particular named entity, the approaches described herein consider context of the intended
30 referent. In various embodiments, a method is provided that includes receiving at least one

utterance, generating embeddings for one or more words of the at least one utterance, generating at least one regular expression and gazetteer feature vector for the at least one utterance, generating at least one context tag distribution feature vector for the at least one utterance, concatenating or interpolating the embeddings with the at least one regular expression and
5 gazetteer feature vector and the at least one context tag distribution feature vector to generate a first set of feature vectors, generating an encoded form of the at least one utterance based on the first set of feature vectors, generating a plurality of log-probabilities for candidate entities based on the encoded form of the at least one utterance, and identifying, using the plurality of log-probabilities, one or more constraints for the at least one utterance based on the candidate
10 entities. Other features and advantages of the various embodiments are apparent throughout this disclosure.

Bot Systems

[0031] A bot (also referred to as a skill, chatbot, chatterbot, or talkbot) is a computer program that can perform conversations with end users. The bot can generally respond to natural-language
15 messages (e.g., questions or comments) through a messaging application that uses natural-language messages. Enterprises may use one or more bots to communicate with end users through a messaging application. The messaging application may include, for example, over-the-top (OTT) messaging channels (such as Facebook Messenger, Facebook WhatsApp, WeChat, Line, Kik, Telegram, Talk, Skype, Slack, or SMS), virtual private assistants (such as Amazon
20 Dot, Echo, or Show, Google Home, Apple HomePod, etc.), mobile and web app extensions that extend native or hybrid/responsive mobile apps or web applications with chat capabilities, or voice based input (such as devices or apps with interfaces that use Siri, Cortana, Google Voice, or other speech input for interaction).

[0032] In some examples, the bot may be associated with a Uniform Resource Identifier
25 (URI). The URI may identify the bot using a string of characters. The URI may be used as a webhook for one or more messaging application systems. The URI may include, for example, a Uniform Resource Locator (URL) or a Uniform Resource Name (URN). The bot may be designed to receive a message (e.g., a hypertext transfer protocol (HTTP) post call message) from a messaging application system. The HTTP post call message may be directed to the URI
30 from the messaging application system. In some examples, the message may be different from a

HTTP post call message. For example, the bot may receive a message from a Short Message Service (SMS). While discussion herein refers to communications that the bot receives as a message, it should be understood that the message may be an HTTP post call message, a SMS message, or any other type of communication between two systems.

5 **[0033]** End users interact with the bot through conversational interactions (sometimes referred to as a conversational user interface (UI)), just as end users interact with other people. In some cases, the conversational interactions may include the end user saying “Hello” to the bot and the bot responding with a “Hi” and asking the end user how it can help. End users also interact with the bot through other types of interactions, such as transactional interactions (e.g., with a banking
10 bot that is at least trained to transfer money from one account to another), informational interactions (e.g., with a human resources bot that is at least trained check the remaining vacation hours the user has), and/or retail interactions (e.g., with a retail bot that is at least trained for discussing returning purchased goods or seeking technical support).

[0034] In some examples, the bot may intelligently handle end user interactions without
15 intervention by an administrator or developer of the bot. For example, an end user may send one or more messages to the bot in order to achieve a desired goal. A message may include certain content, such as text, emojis, audio, image, video, or other method of conveying a message. In some examples, the bot may automatically convert content into a standardized form and generate a natural language response. The bot may also automatically prompt the end user for additional
20 input parameters or request other additional information. In some examples, the bot may also initiate communication with the end user, rather than passively responding to end user utterances.

[0035] A conversation with a bot may follow a specific conversation flow including multiple states. The flow may define what would happen next based on an input. In some examples, a
25 state machine that includes user defined states (e.g., end user intents) and actions to take in the states or from state to state may be used to implement the bot. A conversation may take different paths based on the end user input, which may impact the decision the bot makes for the flow. For example, at each state, based on the end user input or utterances, the bot may determine the end user’s intent in order to determine the appropriate next action to take. As used herein and in the

context of an utterance, the term “intent” refers to an intent of the user who provided the utterance. For example, the user may intend to engage the bot in a conversation to order pizza, where the user’s intent would be represented through the utterance “order pizza.” A user intent can be directed to a particular task that the user wishes the bot to perform on behalf of the user.

5 Therefore, utterances reflecting the user’s intent can be phrased as questions, commands, requests, and the like.

[0036] In the context of the configuration of the bot, the term “intent” is also used herein to refer to configuration information for mapping a user’s utterance to a specific task/action or category of task/action that the bot can perform. In order to distinguish between the intent of an utterance (i.e., a user intent) and the intent of the bot, the latter is sometimes referred to herein as

10 a “bot intent.” A bot intent may comprise a set of one or more utterances associated with the intent. For instance, an intent for ordering pizza can have various permutations of utterances that express a desire to place an order for pizza. These associated utterances can be used to train an intent classifier of the bot to enable the intent classifier to subsequently determine whether an

15 input utterance from a user matches the order pizza intent. Bot intents may be associated with one or more dialog flows for starting a conversation with the user and in a certain state. For example, the first message for the order pizza intent could be the question “What kind of pizza would you like?” In addition to associated utterances, bot intents may further comprise named entities that relate to the intent. For example, the order pizza intent could include variables or

20 parameters used to perform the task of ordering pizza (e.g., topping 1, topping 2, pizza type, pizza size, pizza quantity, and the like). The value of an entity is typically obtained through conversing with the user.

[0037] FIG. 1 is a simplified block diagram of an environment 100 incorporating a chatbot system according to certain embodiments. Environment 100 comprises a digital assistant builder platform (DABP) 102 that enables users 104 of DABP 102 to create and deploy digital assistants

25 or chatbot systems. DABP 102 can be used to create one or more digital assistants (or DAs) or chatbot systems. For example, as shown in FIG. 1, users 104 representing a particular enterprise can use DABP 102 to create and deploy a digital assistant 106 for users of the particular enterprise. For example, DABP 102 can be used by a bank to create one or more digital

30 assistants for use by the bank's customers. The same DABP 102 platform can be used by

multiple enterprises to create digital assistants. As another example, an owner of a restaurant (e.g., a pizza shop) may use DABP 102 to create and deploy a digital assistant that enables customers of the restaurant to order food (e.g., order pizza).

[0038] For purposes of this disclosure, a “digital assistant” is a tool that helps users of the digital assistant accomplish various tasks through natural language conversations. A digital assistant can be implemented using software only (e.g., the digital assistant is a digital tool implemented using programs, code, or instructions executable by one or more processors), using hardware, or using a combination of hardware and software. A digital assistant can be embodied or implemented in various physical systems or devices, such as in a computer, a mobile phone, a watch, an appliance, a vehicle, and the like. A digital assistant is also sometimes referred to as a chatbot system. Accordingly, for purposes of this disclosure, the terms digital assistant and chatbot system are interchangeable.

[0039] A digital assistant, such as digital assistant 106 built using DABP 102, can be used to perform various tasks via natural language-based conversations between the digital assistant and its users 108. As part of a conversation, a user may provide one or more user inputs 110 to digital assistant 106 and get responses 112 back from digital assistant 106. A conversation can include one or more of inputs 110 and responses 112. Via these conversations, a user can request one or more tasks to be performed by the digital assistant and, in response, the digital assistant is configured to perform the user-requested tasks and respond with appropriate responses to the user.

[0040] User inputs 110 are generally in a natural language form and are referred to as utterances. A user utterance 110 can be in text form, such as when a user types in a sentence, a question, a text fragment, or even a single word and provides it as input to digital assistant 106. In some examples, a user utterance 110 can be in audio input or speech form, such as when a user says or speaks something that is provided as input to digital assistant 106. The utterances are typically in a language spoken by the user. For example, the utterances may be in English, or some other language. When an utterance is in speech form, the speech input is converted to text form utterances in that particular language and the text utterances are then processed by digital assistant 106. Various speech-to-text processing techniques may be used to convert a speech or

audio input to a text utterance, which is then processed by digital assistant 106. In some examples, the speech-to-text conversion may be done by digital assistant 106 itself.

[0041] An utterance, which may be a text utterance or a speech utterance, can be a fragment, a sentence, multiple sentences, one or more words, one or more questions, combinations of the
5 aforementioned types, and the like. Digital assistant 106 is configured to apply natural language understanding (NLU) techniques to the utterance to understand the meaning of the user input. As part of the NLU processing for an utterance, digital assistant 106 is configured to perform processing to understand the meaning of the utterance, which involves identifying one or more
10 intents and one or more entities corresponding to the utterance. Upon understanding the meaning of an utterance, digital assistant 106 may perform one or more actions or operations responsive to the understood meaning or intents. For purposes of this disclosure, it is assumed that the utterances are text utterances that have been provided directly by a user of digital assistant 106 or are the results of conversion of input speech utterances to text form. This however is not intended to be limiting or restrictive in any manner.

[0042] For example, a user input may request a pizza to be ordered by providing an utterance such as “I want to order a pizza.” Upon receiving such an utterance, digital assistant 106 is configured to understand the meaning of the utterance and take appropriate actions. The appropriate actions may involve, for example, responding to the user with questions requesting
15 user input on the type of pizza the user desires to order, the size of the pizza, any toppings for the pizza, and the like. The responses provided by digital assistant 106 may also be in natural
20 language form and typically in the same language as the input utterance. As part of generating these responses, digital assistant 106 may perform natural language generation (NLG). For the user ordering a pizza, via the conversation between the user and digital assistant 106, the digital assistant may guide the user to provide all the requisite information for the pizza order, and then
25 at the end of the conversation cause the pizza to be ordered. Digital assistant 106 may end the conversation by outputting information to the user indicating that the pizza has been ordered.

[0043] At a conceptual level, digital assistant 106 performs various processing in response to an utterance received from a user. In some examples, this processing involves a series or pipeline of processing steps including, for example, understanding the meaning of the input utterance,
30 determining an action to be performed in response to the utterance, where appropriate causing

the action to be performed, generating a response to be output to the user responsive to the user utterance, outputting the response to the user, and the like. The NLU processing can include parsing the received input utterance to understand the structure and meaning of the utterance, refining and reforming the utterance to develop a better understandable form (e.g., logical form) or structure for the utterance. Generating a response may include using NLG techniques.

[0044] The NLU processing performed by a digital assistant, such as digital assistant 106, can include various NLP related tasks such as sentence parsing (e.g., tokenizing, lemmatizing, identifying part-of-speech tags for the sentence, identifying named entities in the sentence, generating dependency trees to represent the sentence structure, splitting a sentence into clauses, analyzing individual clauses, resolving anaphoras, performing chunking, and the like). In certain examples, the NLU processing is performed by digital assistant 106 itself. In some other examples, digital assistant 106 may use other resources to perform portions of the NLU processing. For example, the syntax and structure of an input utterance sentence may be identified by processing the sentence using a parser, a part-of-speech tagger, and/or a NER. In one implementation, for the English language, a parser, a part-of-speech tagger, and a named entity recognizer such as ones provided by the Stanford NLP Group are used for analyzing the sentence structure and syntax. These are provided as part of the Stanford CoreNLP toolkit.

[0045] While the various examples provided in this disclosure show utterances in the English language, this is meant only as an example. In certain examples, digital assistant 106 is also capable of handling utterances in languages other than English. Digital assistant 106 may provide subsystems (e.g., components implementing NLU functionality) that are configured for performing processing for different languages. These subsystems may be implemented as pluggable units that can be called using service calls from an NLU core server. This makes the NLU processing flexible and extensible for each language, including allowing different orders of processing. A language pack may be provided for individual languages, where a language pack can register a list of subsystems that can be served from the NLU core server.

[0046] A digital assistant, such as digital assistant 106 depicted in FIG. 1, can be made available or accessible to its users 108 through a variety of different channels, such as but not limited to, via certain applications, via social media platforms, via various messaging services and applications, and other applications or channels. A single digital assistant can have several

channels configured for it so that it can be run on and be accessed by different services simultaneously.

[0047] A digital assistant or chatbot system generally contains or is associated with one or more skills. In certain embodiments, these skills are individual chatbots (referred to as skill bots) that are configured to interact with users and fulfill specific types of tasks, such as tracking inventory, submitting timecards, creating expense reports, ordering food, checking a bank account, making reservations, buying a widget, and the like. For example, for the embodiment depicted in FIG. 1, digital assistant or chatbot system 106 includes skills 116-1, 116-2, 116-3, and so on. For purposes of this disclosure, the terms “skill” and “skills” are used synonymously with the terms “skill bot” and “skill bots,” respectively.

[0048] Each skill associated with a digital assistant helps a user of the digital assistant complete a task through a conversation with the user, where the conversation can include a combination of text or audio inputs provided by the user and responses provided by the skill bots. These responses may be in the form of text or audio messages to the user and/or using simple user interface elements (e.g., select lists) that are presented to the user for the user to make selections.

[0049] There are various ways in which a skill or skill bot can be associated or added to a digital assistant. In some instances, a skill bot can be developed by an enterprise and then added to a digital assistant using DABP 102. In other instances, a skill bot can be developed and created using DABP 102 and then added to a digital assistant created using DABP 102. In yet other instances, DABP 102 provides an online digital store (referred to as a “skills store”) that offers multiple skills directed to a wide range of tasks. The skills offered through the skills store may also expose various cloud services. In order to add a skill to a digital assistant being generated using DABP 102, a user of DABP 102 can access the skills store via DABP 102, select a desired skill, and indicate that the selected skill is to be added to the digital assistant created using DABP 102. A skill from the skills store can be added to a digital assistant as is or in a modified form (for example, a user of DABP 102 may select and clone a particular skill bot provided by the skills store, make customizations or modifications to the selected skill bot, and then add the modified skill bot to a digital assistant created using DABP 102).

[0050] Various different architectures may be used to implement a digital assistant or chatbot system. For example, in certain embodiments, the digital assistants created and deployed using DABP 102 may be implemented using a master bot/child (or sub) bot paradigm or architecture. According to this paradigm, a digital assistant is implemented as a master bot that interacts with one or more child bots that are skill bots. For example, in the embodiment depicted in FIG. 1, digital assistant 106 comprises a master bot 114 and skill bots 116-1, 116-2, etc. that are child bots of master bot 114. In certain examples, digital assistant 106 is itself considered to act as the master bot.

[0051] A digital assistant implemented according to the master-child bot architecture enables users of the digital assistant to interact with multiple skills through a unified user interface, namely via the master bot. When a user engages with a digital assistant, the user input is received by the master bot. The master bot then performs processing to determine the meaning of the user input utterance. The master bot then determines whether the task requested by the user in the utterance can be handled by the master bot itself, else the master bot selects an appropriate skill bot for handling the user request and routes the conversation to the selected skill bot. This enables a user to converse with the digital assistant through a common single interface and still provide the capability to use several skill bots configured to perform specific tasks. For example, for a digital assistance developed for an enterprise, the master bot of the digital assistant may interface with skill bots with specific functionalities, such as a customer relationship management (CRM) bot for performing functions related to customer relationship management , an enterprise resource planning (ERP) bot for performing functions related to enterprise resource planning, a human capital management (HCM) bot for performing functions related to human capital management, etc. This way the end user or consumer of the digital assistant need only know how to access the digital assistant through the common master bot interface and behind the scenes multiple skill bots are provided for handling the user request.

[0052] In certain examples, in a master bot/child bots infrastructure, the master bot is configured to be aware of the available list of skill bots. The master bot may have access to metadata that identifies the various available skill bots, and for each skill bot, the capabilities of the skill bot including the tasks that can be performed by the skill bot. Upon receiving a user request in the form of an utterance, the master bot is configured to, from the multiple available

skill bots, identify or predict a specific skill bot that can best serve or handle the user request. The master bot then routes the utterance (or a portion of the utterance) to that specific skill bot for further handling. Control thus flows from the master bot to the skill bots. The master bot can support multiple input and output channels. In certain examples, routing may be performed with the aid of processing performed by one or more available skill bots. For example, as discussed below, a skill bot can be trained to infer an intent for an utterance and to determine whether the inferred intent matches an intent with which the skill bot is configured. Thus, the routing performed by the master bot can involve the skill bot communicating to the master bot an indication of whether the skill bot has been configured with an intent suitable for handling the utterance.

[0053] While the embodiment in FIG. 1 shows digital assistant 106 comprising a master bot 114 and skill bots 116-1, 116-2, and 116-3, this is not intended to be limiting. A digital assistant can include various other components (e.g., other systems and subsystems) that provide the functionalities of the digital assistant. These systems and subsystems may be implemented only in software (e.g., code, instructions stored on a computer-readable medium and executable by one or more processors), in hardware only, or in implementations that use a combination of software and hardware.

[0054] DABP 102 provides an infrastructure and various services and features that enable a user of DABP 102 to create a digital assistant including one or more skill bots associated with the digital assistant. In some instances, a skill bot can be created by cloning an existing skill bot, for example, cloning a skill bot provided by the skills store. As previously indicated, DABP 102 provides a skills store or skills catalog that offers multiple skill bots for performing various tasks. A user of DABP 102 can clone a skill bot from the skills store. As needed, modifications or customizations may be made to the cloned skill bot. In some other instances, a user of DABP 102 created a skill bot from scratch using tools and services offered by DABP 102. As previously indicated, the skills store or skills catalog provided by DABP 102 may offer multiple skill bots for performing various tasks.

[0055] In certain examples, at a high level, creating or customizing a skill bot involves the following steps:

- (1) Configuring settings for a new skill bot

- (2) Configuring one or more intents for the skill bot
- (3) Configuring one or more entities for one or more intents
- (4) Training the skill bot
- (5) Creating a dialog flow for the skill bot
- 5 (6) Adding custom components to the skill bot as needed
- (7) Testing and deploying the skill bot

Each of the above steps is briefly described below.

[0056] (1) Configuring settings for a new skill bot – Various settings may be configured for the skill bot. For example, a skill bot designer can specify one or more invocation names for the skill bot being created. These invocation names can then be used by users of a digital assistant to explicitly invoke the skill bot. For example, a user can input an invocation name in the user's utterance to explicitly invoke the corresponding skill bot.

[0057] (2) Configuring one or more intents and associated example utterances for the skill bot– The skill bot designer specifies one or more intents (also referred to as bot intents) for a skill bot being created. The skill bot is then trained based upon these specified intents. These intents represent categories or classes that the skill bot is trained to infer for input utterances. Upon receiving an utterance, a trained skill bot infers an intent for the utterance, where the inferred intent is selected from the predefined set of intents used to train the skill bot. The skill bot then takes an appropriate action responsive to an utterance based upon the intent inferred for that utterance. In some instances, the intents for a skill bot represent tasks that the skill bot can perform for users of the digital assistant. Each intent is given an intent identifier or intent name. For example, for a skill bot trained for a bank, the intents specified for the skill bot may include "CheckBalance," "TransferMoney," "DepositCheck," and the like.

[0058] For each intent defined for a skill bot, the skill bot designer may also provide one or more example utterances that are representative of and illustrate the intent. These example utterances are meant to represent utterances that a user may input to the skill bot for that intent. For example, for the CheckBalance intent, example utterances may include "What's my savings account balance?", "How much is in my checking account?", "How much money do I have in my

account," and the like. Accordingly, various permutations of typical user utterances may be specified as example utterances for an intent.

[0059] The intents and their associated example utterances are used as training data to train the skill bot. Various different training techniques may be used. As a result of this training, a predictive model is generated that is configured to take an utterance as input and output an intent inferred for the utterance by the predictive model. In some instances, input utterances are provided to an intent analysis engine, which is configured to use the trained model to predict or infer an intent for the input utterance. The skill bot may then take one or more actions based upon the inferred intent.

[0060] (3) Configuring entities for one or more intents of the skill bot – In some instances, additional context may be needed to enable the skill bot to properly respond to a user utterance. For example, there may be situations where a user input utterance resolves to the same intent in a skill bot. For instance, in the above example, utterances “What’s my savings account balance?” and “How much is in my checking account?” both resolve to the same CheckBalance intent, but these utterances are different requests asking for different things. To clarify such requests, one or more entities are added to an intent. Using the banking skill bot example, an entity called AccountType, which defines values called “checking” and “saving” may enable the skill bot to parse the user request and respond appropriately. In the above example, while the utterances resolve to the same intent, the value associated with the AccountType entity is different for the two utterances. This enables the skill bot to perform possibly different actions for the two utterances in spite of them resolving to the same intent. One or more entities can be specified for certain intents configured for the skill bot. Entities are thus used to add context to the intent itself. Entities help describe an intent more fully and enable the skill bot to complete a user request.

[0061] In certain examples, there are two types of entities: (a) built-in entities provided by DABP 102, and (2) custom entities that can be specified by a skill bot designer. Built-in entities are generic entities that can be used with a wide variety of bots. Examples of built-in entities include, without limitation, entities related to time, date, addresses, numbers, email addresses, duration, recurring time periods, currencies, phone numbers, URLs, and the like. Custom entities are used for more customized applications. For example, for a banking skill, an AccountType

entity may be defined by the skill bot designer that enables various banking transactions by checking the user input for keywords like checking, savings, and credit cards, etc.

[0062] (4) Training the skill bot – A skill bot is configured to receive user input in the form of utterances parse or otherwise process the received input and identify or select an intent that is relevant to the received user input. As indicated above, the skill bot has to be trained for this. In certain embodiments, a skill bot is trained based upon the intents configured for the skill bot and the example utterances associated with the intents (collectively, the training data), so that the skill bot can resolve user input utterances to one of its configured intents. In certain examples, the skill bot uses a predictive model that is trained using the training data and allows the skill bot to discern what users say (or in some cases, are trying to say). DABP 102 provides various different training techniques that can be used by a skill bot designer to train a skill bot, including various machine-learning based training techniques, rules-based training techniques, and/or combinations thereof. In certain examples, a portion (e.g., 80%) of the training data is used to train a skill bot model and another portion (e.g., the remaining 20%) is used to test or verify the model. Once trained, the trained model (also sometimes referred to as the trained skill bot) can then be used to handle and respond to user utterances. In certain cases, a user's utterance may be a question that requires only a single answer and no further conversation. In order to handle such situations, a Q&A (question-and-answer) intent may be defined for a skill bot. This enables a skill bot to output replies to user requests without having to update the dialog definition. Q&A intents are created in a similar manner as regular intents. The dialog flow for Q&A intents can be different from that for regular intents.

[0063] (5) Creating a dialog flow for the skill bot -- A dialog flow specified for a skill bot describes how the skill bot reacts as different intents for the skill bot are resolved responsive to received user input. The dialog flow defines operations or actions that a skill bot will take, e.g., how the skill bot responds to user utterances, how the skill bot prompts users for input, how the skill bot returns data. A dialog flow is like a flowchart that is followed by the skill bot. The skill bot designer specifies a dialog flow using a language, such as markdown language. In certain embodiments, a version of YAML called OBotML may be used to specify a dialog flow for a skill bot. The dialog flow definition for a skill bot acts as a model for the conversation itself, one

that lets the skill bot designer choreograph the interactions between a skill bot and the users that the skill bot services.

[0064] In certain examples, the dialog flow definition for a skill bot contains three sections:

(a) a context section

5 (b) a default transitions section

(c) a states section

[0065] Context section – The skill bot designer can define variables that are used in a conversation flow in the context section. Other variables that may be named in the context section include, without limitation: variables for error handling, variables for built-in or custom entities, user variables that enable the skill bot to recognize and persist user preferences, and the like.

[0066] Default transitions section – Transitions for a skill bot can be defined in the dialog flow states section or in the default transitions section. The transitions defined in the default transition section act as a fallback and get triggered when there are no applicable transitions defined within a state, or the conditions required to trigger a state transition cannot be met. The default transitions section can be used to define routing that allows the skill bot to gracefully handle unexpected user actions.

[0067] States section – A dialog flow and its related operations are defined as a sequence of transitory states, which manage the logic within the dialog flow. Each state node within a dialog flow definition names a component that provides the functionality needed at that point in the dialog. States are thus built around the components. A state contains component-specific properties and defines the transitions to other states that get triggered after the component executes.

[0068] Special case scenarios may be handled using the states sections. For example, there might be times when you want to provide users the option to temporarily leave a first skill, they are engaged with to do something in a second skill within the digital assistant. For example, if a user is engaged in a conversation with a shopping skill (e.g., the user has made some selections for purchase), the user may want to jump to a banking skill (e.g., the user may want to ensure

that he/she has enough money for the purchase), and then return to the shopping skill to complete the user's order. To address this, an action in the first skill can be configured to initiate an interaction with the second different skill in the same digital assistant and then return to the original flow.

5 **[0069]** (6) Adding custom components to the skill bot – As described above, states specified in a dialog flow for skill bot name components that provide the functionality needed corresponding to the states. Components enable a skill bot to perform functions. In certain embodiments, DABP 102 provides a set of preconfigured components for performing a wide range of functions. A skill bot designer can select one or more of these preconfigured components and associate them
10 with states in the dialog flow for a skill bot. The skill bot designer can also create custom or new components using tools provided by DABP 102 and associate the custom components with one or more states in the dialog flow for a skill bot.

[0070] (7) Testing and deploying the skill bot – DABP 102 provides several features that enable the skill bot designer to test a skill bot being developed. The skill bot can then be
15 deployed and included in a digital assistant.

[0071] While the description above describes how to create a skill bot, similar techniques may also be used to create a digital assistant (or the master bot). At the master bot or digital assistant level, built-in system intents may be configured for the digital assistant. These built-in system intents are used to identify general tasks that the digital assistant itself (i.e., the master bot) can
20 handle without invoking a skill bot associated with the digital assistant. Examples of system intents defined for a master bot include: (1) Exit: applies when the user signals the desire to exit the current conversation or context in the digital assistant; (2) Help: applies when the user asks for help or orientation; and (3) Unresolved Intent: applies to user input that doesn't match well with the exit and help intents. The digital assistant also stores information about the one or more
25 skill bots associated with the digital assistant. This information enables the master bot to select a particular skill bot for handling an utterance.

[0072] At the master bot or digital assistant level, when a user inputs a phrase or utterance to the digital assistant, the digital assistant is configured to perform processing to determine how to route the utterance and the related conversation. The digital assistant determines this using a
30 routing model, which can be rules-based, AI-based, or a combination thereof. The digital

assistant uses the routing model to determine whether the conversation corresponding to the user input utterance is to be routed to a particular skill for handling, is to be handled by the digital assistant or master bot itself per a built-in system intent or is to be handled as a different state in a current conversation flow.

5 **[0073]** In certain embodiments, as part of this processing, the digital assistant determines if the user input utterance explicitly identifies a skill bot using its invocation name. If an invocation name is present in the user input, then it is treated as explicit invocation of the skill bot corresponding to the invocation name. In such a scenario, the digital assistant may route the user input to the explicitly invoked skill bot for further handling. If there is no specific or explicit
10 invocation, in certain embodiments, the digital assistant evaluates the received user input utterance and computes confidence scores for the system intents and the skill bots associated with the digital assistant. The score computed for a skill bot or system intent represents how likely the user input is representative of a task that the skill bot is configured to perform or is representative of a system intent. Any system intent or skill bot with an associated computed
15 confidence score exceeding a threshold value (e.g., a Confidence Threshold routing parameter) is selected as a candidate for further evaluation. The digital assistant then selects, from the identified candidates, a particular system intent or a skill bot for further handling of the user input utterance. In certain embodiments, after one or more skill bots are identified as candidates, the intents associated with those candidate skills are evaluated (according to the intent model for
20 each skill) and confidence scores are determined for each intent. In general, any intent that has a confidence score exceeding a threshold value (e.g., 70%) is treated as a candidate intent. If a particular skill bot is selected, then the user utterance is routed to that skill bot for further processing. If a system intent is selected, then one or more actions are performed by the master bot itself according to the selected system intent.

25 **[0074]** FIG. 2 is a simplified block diagram of a master bot (MB) system 200 according to certain embodiments. MB system 200 can be implemented in software only, hardware only, or a combination of hardware and software. MB system 200 includes a pre-processing subsystem 210, a multiple intent subsystem (MIS) 220, an explicit invocation subsystem (EIS) 230, a skill bot invoker 240, and a data store 250. MB system 200 depicted in FIG. 2 is merely an example
30 of an arrangement of components in a master bot. One of ordinary skill in the art would

recognize many possible variations, alternatives, and modifications. For example, in some implementations, MB system 200 may have more or fewer systems or components than those shown in FIG. 2, may combine two or more subsystems, or may have a different configuration or arrangement of subsystems.

5 **[0075]** Pre-processing subsystem 210 receives an utterance “A” 202 from a user and processes the utterance through a language detector 212 and a language parser 214. As indicated above, an utterance can be provided in various ways including audio or text. The utterance 202 can be a sentence fragment, a complete sentence, multiple sentences, and the like. Utterance 202 can include punctuation. For example, if the utterance 202 is provided as audio, the pre-processing
10 subsystem 210 may convert the audio to text using a speech-to-text converter (not shown) that inserts punctuation marks into the resulting text, e.g., commas, semicolons, periods, etc.

15 **[0076]** Language detector 212 detects the language of the utterance 202 based on the text of the utterance 202. The manner in which the utterance 202 is handled depends on the language since each language has its own grammar and semantics. Differences between languages are taken into consideration when analyzing the syntax and structure of an utterance.

20 **[0077]** Language parser 214 parses the utterance 202 to extract part of speech (POS) tags for individual linguistic units (e.g., words) in the utterance 202. POS tags include, for example, noun (NN), pronoun (PN), verb (VB), and the like. Language parser 214 may also tokenize the linguistic units of the utterance 202 (e.g., to convert each word into a separate token) and
25 lemmatize words. A lemma is the main form of a set of words as represented in a dictionary (e.g., “run” is the lemma for run, runs, ran, running, etc.). Other types of pre-processing that the language parser 214 can perform include chunking of compound expressions, e.g., combining “credit” and “card” into a single expression “credit card.” Language parser 214 may also identify relationships between the words in the utterance 202. For example, in some embodiments, the
30 language parser 214 generates a dependency tree that indicates which part of the utterance (e.g., a particular noun) is a direct object, which part of the utterance is a preposition, and so on. The results of the processing performed by the language parser 214 form extracted information 205 and are provided as input to MIS 220 together with the utterance 202 itself.

30 **[0078]** As indicated above, the utterance 202 can include more than one sentence. For purposes of detecting multiple intents and explicit invocation, the utterance 202 can be treated as

a single unit even if it includes multiple sentences. However, in certain embodiments, pre-processing can be performed, e.g., by the pre-processing subsystem 210, to identify a single sentence among multiple sentences for multiple intents analysis and explicit invocation analysis. In general, the results produced by MIS 220 and EIS 230 are substantially the same regardless of whether the utterance 202 is processed at the level of an individual sentence or as a single unit comprising multiple sentences.

[0079] MIS 220 determines whether the utterance 202 represents multiple intents. Although MIS 220 can detect the presence of multiple intents in the utterance 202, the processing performed by MIS 220 does not involve determining whether the intents of the utterance 202 match to any intents that have been configured for a bot. Instead, processing to determine whether an intent of the utterance 202 matches a bot intent can be performed by an intent classifier 242 of the MB system 200 or by an intent classifier of a skill bot (e.g., as shown in FIG. 3). The processing performed by MIS 220 assumes that there exists a bot (e.g., a particular skill bot or the master bot itself) that can handle the utterance 202. Therefore, the processing performed by MIS 220 does not require knowledge of what bots are in the chatbot system (e.g., the identities of skill bots registered with the master bot) or knowledge of what intents have been configured for a particular bot.

[0080] To determine that the utterance 202 includes multiple intents, the MIS 220 applies one or more rules from a set of rules 252 in the data store 250. The rules applied to the utterance 202 depend on the language of the utterance 202 and may include sentence patterns that indicate the presence of multiple intents. For example, a sentence pattern may include a coordinating conjunction that joins two parts (e.g., conjuncts) of a sentence, where both parts correspond to a separate intent. If the utterance 202 matches the sentence pattern, it can be inferred that the utterance 202 represents multiple intents. It should be noted that an utterance with multiple intents does not necessarily have different intents (e.g., intents directed to different bots or to different intents within the same bot). Instead, the utterance could have separate instances of the same intent (e.g. “Place a pizza order using payment account X, then place a pizza order using payment account Y”).

[0081] As part of determining that the utterance 202 represents multiple intents, the MIS 220 also determines what portions of the utterance 202 are associated with each intent. MIS 220

constructs, for each intent represented in an utterance containing multiple intents, a new utterance for separate processing in place of the original utterance, e.g., an utterance “B” 206 and an utterance “C” 208, as depicted in FIG. 2. Thus, the original utterance 202 can be split into two or more separate utterances that are handled one at a time. MIS 220 determines, using the
5 extracted information 205 and/or from analysis of the utterance 202 itself, which of the two or more utterances should be handled first. For example, MIS 220 may determine that the utterance 202 contains a marker word indicating that a particular intent should be handled first. The newly formed utterance corresponding to this particular intent (e.g., one of utterance 206 or utterance 208) will be the first to be sent for further processing by EIS 230. After a conversation triggered
10 by the first utterance has ended (or has been temporarily suspended), the next highest priority utterance (e.g., the other one of utterance 206 or utterance 208) can then be sent to the EIS 230 for processing.

[0082] EIS 230 determines whether the utterance that it receives (e.g., utterance 206 or utterance 208) contains an invocation name of a skill bot. In certain embodiments, each skill bot
15 in a chatbot system is assigned a unique invocation name that distinguishes the skill bot from other skill bots in the chatbot system. A list of invocation names can be maintained as part of skill bot information 254 in data store 250. An utterance is deemed to be an explicit invocation when the utterance contains a word match to an invocation name. If a bot is not explicitly invoked, then the utterance received by the EIS 230 is deemed a non-explicitly invoking
20 utterance 234 and is input to an intent classifier (e.g., intent classifier 242) of the master bot to determine which bot to use for handling the utterance. In some instances, the intent classifier 242 will determine that the master bot should handle a non-explicitly invoking utterance. In other instances, the intent classifier 242 will determine a skill bot to route the utterance to for handling.

[0083] The explicit invocation functionality provided by the EIS 230 has several advantages. It
25 can reduce the amount of processing that the master bot has to perform. For example, when there is an explicit invocation, the master bot may not have to do any intent classification analysis (e.g., using the intent classifier 242), or may have to do reduced intent classification analysis for selecting a skill bot. Thus, explicit invocation analysis may enable selection of a particular skill bot without resorting to intent classification analysis.

[0084] Also, there may be situations where there is an overlap in functionalities between multiple skill bots. This may happen, for example, if the intents handled by the two skill bots overlap or are very close to each other. In such a situation, it may be difficult for the master bot to identify which of the multiple skill bots to select based upon intent classification analysis alone. In such scenarios, the explicit invocation disambiguates the particular skill bot to be used.

[0085] In addition to determining that an utterance is an explicit invocation, the EIS 230 is responsible for determining whether any portion of the utterance should be used as input to the skill bot being explicitly invoked. In particular, EIS 230 can determine whether part of the utterance is not associated with the invocation. The EIS 230 can perform this determination through analysis of the utterance and/or analysis of the extracted information 205. EIS 230 can send the part of the utterance not associated with the invocation to the invoked skill bot in lieu of sending the entire utterance that was received by the EIS 230. In some instances, the input to the invoked skill bot is formed simply by removing any portion of the utterance associated with the invocation. For example, “I want to order pizza using Pizza Bot” can be shortened to “I want to order pizza” since “using Pizza Bot” is relevant to the invocation of the pizza bot, but irrelevant to any processing to be performed by the pizza bot. In some instances, EIS 230 may reformat the part to be sent to the invoked bot, e.g., to form a complete sentence. Thus, the EIS 230 determines not only that there is an explicit invocation, but also what to send to the skill bot when there is an explicit invocation. In some instances, there may not be any text to input to the bot being invoked. For example, if the utterance was “Pizza Bot”, then the EIS 230 could determine that the pizza bot is being invoked, but there is no text to be processed by the pizza bot. In such scenarios, the EIS 230 may indicate to the skill bot invoker 240 that there is nothing to send.

[0086] Skill bot invoker 240 invokes a skill bot in various ways. For instance, skill bot invoker 240 can invoke a bot in response to receiving an indication 235 that a particular skill bot has been selected as a result of an explicit invocation. The indication 235 can be sent by the EIS 230 together with the input for the explicitly invoked skill bot. In this scenario, the skill bot invoker 240 will turn control of the conversation over to the explicitly invoked skill bot. The explicitly invoked skill bot will determine an appropriate response to the input from the EIS 230 by treating the input as a stand-alone utterance. For example, the response could be to perform a

specific action or to start a new conversation in a particular state, where the initial state of the new conversation depends on the input sent from the EIS 230.

[0087] Another way in which skill bot invoker 240 can invoke a skill bot is through implicit invocation using the intent classifier 242. The intent classifier 242 can be trained, using machine-learning and/or rules-based training techniques, to determine a likelihood that an utterance is representative of a task that a particular skill bot is configured to perform. The intent classifier 242 is trained on different classes, one class for each skill bot. For instance, whenever a new skill bot is registered with the master bot, a list of example utterances associated with the new skill bot can be used to train the intent classifier 242 to determine a likelihood that a particular utterance is representative of a task that the new skill bot can perform. The parameters produced as result of this training (e.g., a set of values for parameters of a machine-learning model) can be stored as part of skill bot information 254.

[0088] In certain embodiments, the intent classifier 242 is implemented using a machine-learning model, as described in further detail herein. Training of the machine-learning model may involve inputting at least a subset of utterances from the example utterances associated with various skill bots to generate, as an output of the machine-learning model, inferences as to which bot is the correct bot for handling any particular training utterance. For each training utterance, an indication of the correct bot to use for the training utterance may be provided as ground truth information. The behavior of the machine-learning model can then be adapted (e.g., through back-propagation) to minimize the difference between the generated inferences and the ground truth information.

[0089] In certain embodiments, the intent classifier 242 determines, for each skill bot registered with the master bot, a confidence score indicating a likelihood that the skill bot can handle an utterance (e.g., the non-explicitly invoking utterance 234 received from EIS 230). The intent classifier 242 may also determine a confidence score for each system level intent (e.g., help, exit) that has been configured. If a particular confidence score meets one or more conditions, then the skill bot invoker 240 will invoke the bot associated with the particular confidence score. For example, a threshold confidence score value may need to be met. Thus, an output 245 of the intent classifier 242 is either an identification of a system intent or an identification of a particular skill bot. In some embodiments, in addition to meeting a threshold

confidence score value, the confidence score must exceed the next highest confidence score by a certain win margin. Imposing such a condition would enable routing to a particular skill bot when the confidence scores of multiple skill bots each exceed the threshold confidence score value.

5 **[0090]** After identifying a bot based on evaluation of confidence scores, the skill bot invoker 240 hands over processing to the identified bot. In the case of a system intent, the identified bot is the master bot. Otherwise, the identified bot is a skill bot. Further, the skill bot invoker 240 will determine what to provide as input 247 for the identified bot. As indicated above, in the case of an explicit invocation, the input 247 can be based on a part of an utterance that is not
10 associated with the invocation, or the input 247 can be nothing (e.g., an empty string). In the case of an implicit invocation, the input 247 can be the entire utterance.

[0091] Data store 250 comprises one or more computing devices that store data used by the various subsystems of the master bot system 200. As explained above, the data store 250 includes rules 252 and skill bot information 254. The rules 252 include, for example, rules for
15 determining, by MIS 220, when an utterance represents multiple intents and how to split an utterance that represents multiple intents. The rules 252 further include rules for determining, by EIS 230, which parts of an utterance that explicitly invokes a skill bot to send to the skill bot. The skill bot information 254 includes invocation names of skill bots in the chatbot system, e.g., a list of the invocation names of all skill bots registered with a particular master bot. The skill bot
20 information 254 can also include information used by intent classifier 242 to determine a confidence score for each skill bot in the chatbot system, e.g., parameters of a machine-learning model.

[0092] FIG. 3 is a simplified block diagram of a skill bot system 300 according to certain embodiments. Skill bot system 300 is a computing system that can be implemented in software
25 only, hardware only, or a combination of hardware and software. In certain embodiments such as the embodiment depicted in FIG. 1, skill bot system 300 can be used to implement one or more skill bots within a digital assistant.

[0093] Skill bot system 300 includes an MIS 310, an intent classifier 320, and a conversation manager 330. The MIS 310 is analogous to the MIS 220 in FIG. 2 and provides similar
30 functionality, including being operable to determine, using rules 352 in a data store 350: (1)

whether an utterance represents multiple intents and, if so, (2) how to split the utterance into a separate utterance for each intent of the multiple intents. In certain embodiments, the rules applied by MIS 310 for detecting multiple intents and for splitting an utterance are the same as those applied by MIS 220. The MIS 310 receives an utterance 302 and extracted information 5 304. The extracted information 304 is analogous to the extracted information 205 in FIG. 1 and can be generated using the language parser 214 or a language parser local to the skill bot system 300.

[0094] Intent classifier 320 can be trained in a similar manner to the intent classifier 242 discussed above in connection with the embodiment of FIG. 2 and as described in further detail 10 herein. For instance, in certain embodiments, the intent classifier 320 is implemented using a machine-learning model. The machine-learning model of the intent classifier 320 is trained for a particular skill bot, using at least a subset of example utterances associated with that particular skill bot as training utterances. The ground truth for each training utterance would be the particular bot intent associated with the training utterance.

15 **[0095]** The utterance 302 can be received directly from the user or supplied through a master bot. When the utterance 302 is supplied through a master bot, e.g., as a result of processing through MIS 220 and EIS 230 in the embodiment depicted in FIG. 2, the MIS 310 can be bypassed so as to avoid repeating processing already performed by MIS 220. However, if the utterance 302 is received directly from the user, e.g., during a conversation that occurs after 20 routing to a skill bot, then MIS 310 can process the utterance 302 to determine whether the utterance 302 represents multiple intents. If so, then MIS 310 applies one or more rules to split the utterance 302 into a separate utterance for each intent, e.g., an utterance “D” 306 and an utterance “E” 308. If utterance 302 does not represent multiple intents, then MIS 310 forwards the utterance 302 to intent classifier 320 for intent classification and without splitting the 25 utterance 302.

[0096] Intent classifier 320 is configured to match a received utterance (e.g., utterance 306 or 308) to an intent associated with skill bot system 300. As explained above, a skill bot can be configured with one or more intents, each intent including at least one example utterance that is associated with the intent and used for training a classifier. In the embodiment of FIG. 2, the 30 intent classifier 242 of the master bot system 200 is trained to determine confidence scores for

individual skill bots and confidence scores for system intents. Similarly, intent classifier 320 can be trained to determine a confidence score for each intent associated with the skill bot system 300. Whereas the classification performed by intent classifier 242 is at the bot level, the classification performed by intent classifier 320 is at the intent level and therefore finer grained.

5 The intent classifier 320 has access to intents information 354. The intents information 354 includes, for each intent associated with the skill bot system 300, a list of utterances that are representative of and illustrate the meaning of the intent and are typically associated with a task performable by that intent. The intents information 354 can further include parameters produced as a result of training on this list of utterances.

10 **[0097]** Conversation manager 330 receives, as an output of intent classifier 320, an indication 322 of a particular intent, identified by the intent classifier 320, as best matching the utterance that was input to the intent classifier 320. In some instances, the intent classifier 320 is unable to determine any match. For example, the confidence scores computed by the intent classifier 320 could fall below a threshold confidence score value if the utterance is directed to a system intent
15 or an intent of a different skill bot. When this occurs, the skill bot system 300 may refer the utterance to the master bot for handling, e.g., to route to a different skill bot. However, if the intent classifier 320 is successful in identifying an intent within the skill bot, then the conversation manager 330 will initiate a conversation with the user.

[0098] The conversation initiated by the conversation manager 330 is a conversation specific
20 to the intent identified by the intent classifier 320. For instance, the conversation manager 330 may be implemented using a state machine configured to execute a dialog flow for the identified intent. The state machine can include a default starting state (e.g., for when the intent is invoked without any additional input) and one or more additional states, where each state has associated with it actions to be performed by the skill bot (e.g., executing a purchase transaction) and/or
25 dialog (e.g., questions, responses) to be presented to the user. Thus, the conversation manager 330 can determine an action/dialog 335 upon receiving the indication 322 identifying the intent and can determine additional actions or dialog in response to subsequent utterances received during the conversation.

[0099] Data store 350 comprises one or more computing devices that store data used by the
30 various subsystems of the skill bot system 300. As depicted in FIG. 3, the data store 350 includes

the rules 352 and the intents information 354. In certain embodiments, data store 350 can be integrated into a data store of a master bot or digital assistant, e.g., the data store 250 in FIG. 2.

Context Tag Integration

[0100] Intent prediction and entity extraction, which are the two major components of natural language processing, help chatbot systems understand user queries and user utterances with respect to the domain of a given service or set of services. Intent prediction determines an objective (i.e., intent) of the user's query or utterance. Entity extraction determines one or more constraints, if any, of the user's query or utterance. For example, for a user inquiry regarding "the weather on Wednesday in the Poconos," intent prediction determines that the user's intent is to learn about the "weather" and entity extraction determines that "Wednesday" and "Poconos" are constraints that focus the user's intent to a particular day and geographical location. Entity extraction can involve matching, where the words of the user's query are confirmed as entities by matching words to a list of pre-defined entities. However, because the subject of an entity may not always correspond to an obvious or popular referent, matching often fails to identify the referent intended by the user. Matching is even more challenging when a user's utterance is limited to one or two words per utterance or when a user's utterance contains more information than is necessary. For example, in the case of a limited utterance, a user utterance involving just the word "2020" may refer to a specific calendar year, specific cost of an item, quantity of an item, etc., depending on the context in other user utterances or system queries associated with user's utterance. In the case of a user utterance with additional information, a user utterance reciting "it is 2020 for 20 people costing 2000" may or may not correspond to a year entity, a cost entity, or quantity entity depending on the context of the user's utterance. As can be seen from above examples, unless context is considered, the referent intended by the user in an utterance cannot be accurately determined. Features of the present disclosure, as described herein, overcome these challenges by evaluating the distribution of context tags within a group of entities associated with a system query or a user's utterance(s).

[0101] FIG. 4A shows a block diagram illustrating aspects of a chatbot system 400 configured to train and utilize NER models within intent classifiers (e.g., the intent classifier 242 of FIG. 2 or intent classifier 320 of FIG. 3). As shown in FIG. 4A, chatbot system 400 may include a prediction model training stage 410, a skill bot invocation stage 415 configured to determine a

likelihood that an utterance is representative of a task that a particular skill bot is configured to perform, an intent prediction stage 420 configured to classify utterances as one or more intents, and an entity detection stage 422 configured to determine one or more constraints 480 of the utterances. The prediction model training stage 410 may be configured to build and train one or more prediction models 425a-425n (which may be referred to herein individually as a prediction model or collectively as the prediction models) to be used by the other stages. In some examples, the prediction models can include a model for determining a likelihood that an utterance is representative of a task that a particular skill bot is configured to perform, a model for predicting an intent from an utterance for a first type of skill bot, a model for predicting an intent from an utterance for a second type of skill bot, and a model for identifying mentions of conceptual entities in text and classifying them according to a given set of categories. Still other types of prediction models may be implemented in other examples according to this disclosure.

[0102] A prediction model can be a machine-learning (ML) model, such as a convolutional neural network (CNN), *e.g.* an inception neural network, a residual neural network (Resnet), or a recurrent neural network, *e.g.* long short-term memory (LSTM) models, a bidirectional LSTM (BiLSTM) or gated recurrent units (GRU) models, other variants of deep neural networks (DNN). A prediction model can also be any other suitable ML model trained for natural language processing, such as a bidirectional encoder representations from transformers (BERT) model, naive bayes classifier, linear classifier, support vector machine, conditional random field model, random forest model, boosting models, shallow neural networks, or combinations of one or more of such techniques (*e.g.*, CNN-HMM or MCNN). The chatbot system 400 may employ the same type of prediction model or different types of prediction models for determining a likelihood of a task that a particular skill bot is configured to perform, predicting an intent from an utterance for a first type of skill bot, predicting an intent from an utterance for a second type of skill bot, and identifying mentions of conceptual entities in text and classifying them according to a given set of categories. Still other types of prediction models may be implemented in other examples according to this disclosure.

[0103] As further shown in FIG. 4A, prediction model training stage 410 may include dataset preparation 430, feature engineering 435, and model training 440. Dataset preparation 430 may be configured to process input data assets 445 into separate training and validation sets 445 a-n

for each prediction model. Data assets 445 may include at least a subset of utterances from example utterances associated with various skill bots. As previously described, an utterance can be provided in various ways including audio or text. The utterance can be a sentence fragment, a complete sentence, multiple sentences, and the like. If, for example, the utterance is provided as audio, data preparation 430 may convert the audio to text using a speech-to-text converter (not shown) that inserts punctuation marks into the resulting text (e.g., commas, semicolons, periods, etc.). In some instances, the example utterances are provided by a client or customer. In other instances, the example utterances are automatically generated from prior libraries of utterances (e.g., identifying utterances from a library that are specific to a skill that a chatbot is to learn). In some examples, data assets 445 for a prediction model can include input text or audio (or input features of text or audio frames) and corresponding labels 450 for the input text or audio (or input features) as a matrix or table of values. For example, for each training utterance, an indication of the correct bot to use for the training utterance may be provided as ground truth information for labels 450. The behavior of a respective prediction model can then be adapted (e.g., through back-propagation) to minimize the difference between the generated inferences and the ground truth information. Alternatively, a prediction model may be trained for a particular skill bot, using at least a subset of example utterances associated with that particular skill bot as training utterances. The ground truth information for labels 450 for each training utterance would be the particular bot intent associated with the training utterance. Alternatively, a prediction model may be trained for a particular skill bot, using at least a subset of example utterances associated with that particular skill bot as training utterances. The ground truth information for labels 450 for each training utterance would be the particular bot intent associated with the training utterance.

[0104] In some instances, augmentation may be applied to the data assets 445. For example, Easy Data Augmentation (EDA) techniques may be used for boosting performance on text classification tasks. EDA includes four operations: synonym replacement, ransom insertion, random swap, and random deletion that prevent overfitting and helping train more robust models. Note that the EDA operations in general: (i) obtain words from the original text, and (ii) incorporate the words within each data asset 445 relative to the original text. For example, synonym replacement operation includes randomly selecting n words from the original sentence (e.g., utterance) that are not stop words, and replacing each of these words with one of its

synonyms chosen at random. The random insertion operation includes - n times - finding a random synonym of a random word in the original sentence that is not a stop word and inserting that synonym into a random position in the sentence. The random swap operation includes - n times - randomly choosing two words in the sentence and swapping their positions. The random
5 deletion operation includes randomly removing each word in the sentence with probability p .

[0105] In some examples, feature engineering 435 may include transforming data assets 445 into feature vectors and/or creating new features will be created using the data assets 445. The feature vectors may include count vectors as features, TF-IDF vectors as features such as word level, n -gram level or character level, word embedding as features, text/NLP as features, topic
10 models as features, or a combination thereof. Count vector is a matrix notation of the data assets 445 in which every row represents an utterance, every column represents a term from the utterance, and every cell represents the frequency count of a particular term in a utterance. TF-IDF score represents the relative importance of a term in the utterance. A word embedding is a form of representing words and utterances using a dense vector representation. The position of a
15 word within the vector space is learned from text and is based on the words that surround the word when it is used. Text/NLP based features may include word count in the utterance, character count in the utterance, average word density, punctuation count, upper case count, title word count, frequency distribution of part of speech tags (e.g., nouns and verbs), or any combination thereof. Topic modelling is a technique to identify the groups of words (called a
20 topic) from a collection of utterances that contains best information in the collection.

[0106] In some examples, model training 440 may include training a classifier using the feature vectors and/or new features created in feature engineering 435. In some instances, the training process includes iterative operations to find a set of parameters for the prediction model that minimizes a loss or error function for the prediction models. Each iteration can involve
25 finding a set of parameters for the prediction model so that the value of the loss or error function using the set of parameters is smaller than the value of the loss or error function using another set of parameters in a previous iteration. The loss or error function can be constructed to measure the difference between the outputs predicted using the prediction models and the labels 450 contained in the data assets 445. Once the set of parameters are identified, the prediction model
30 has been trained and can be utilized for prediction as designed.

[0107] In addition to the data assets 445, labels 450, the feature vectors and/or new features, other techniques and information can also be employed to refine the training process of the prediction models. For example, the feature vectors and/or new features may be combined together to help to improve the accuracy of the classifier or model. Additionally, or alternatively, the hyperparameters may be tuned or optimized, for example, a number of parameters such as tree length, leaves, network parameters etc. can be fine-tuned to get a best fit model. Although the training mechanisms described herein mainly focus on training a prediction model. These training mechanisms can also be utilized to fine tune existing prediction models trained from other data assets. For example, in some cases, a prediction model might have been pre-trained using utterance specific to another skill bot. In those cases, the prediction models can be retrained using the data assets 445 as discussed herein.

[0108] In some examples, the prediction model training stage 410 may output trained prediction models including the task prediction models 460, intent prediction models 465, and entity extraction models 467. The task prediction models 460 may be used in the skill bot invocation stage 415 to determine a likelihood that an utterance is representative of a task that a particular skill bot is configured to perform 470, the intent prediction models 465 may be used in the intent prediction stage 420 for classifying utterances as one or more intents 475, and the entity extraction models 467 may be used in the entity detection stage 422 for extracting and classifying entities as one or more constraints 480. In some instances, the skill bot invocation stage 415, the intent prediction stage 420, and the entity detection stage 422 may proceed independently in some examples with separate models. For example, the trained intent prediction models 465 may be used in the intent prediction stage 420 to predict intents for skill bots and/or the trained entity extraction models 467 may be used in the entity detection stage 422 to extract and classify entities for skill bots without first identifying the skill bots in the skill bot invocation stage 415. Similarly, the task prediction models 460 may be used in the skill bot invocation stage 415 to predict tasks or skill bots to be used for utterances without identifying the intent and/or entities of the utterances in the intent prediction stage 420 and/or the entity detection stage 422.

[0109] Alternatively, the skill bot invocation stage 415, the intent prediction stage 420, and the entity detection stage 422 may be conducted sequentially with one stage using the outputs of the other as inputs or one stage being invokes in a particular manner for a specific skill bot based on

the outputs of the other. For instance, for a given text data 405, a skill bot invoker can invoke a skill bot through implicit invocation using the skill bot invocation stage 415 and the task prediction models 460. The task prediction models 460 can be trained, using machine-learning and/or rules-based training techniques, to determine a likelihood that an utterance is

5 representative of a task that a particular skill bot 470 is configured to perform. Then for an identified or invoked skill bot and a given text data 405, the intent prediction stage 420 and intent prediction models 465 and/or the entity detection stage 422 and the entity extraction models 467 can be used to match a received utterance (e.g., utterance within given data asset 445) to an intent 475 associated with skill bot. As explained herein, a skill bot can be configured with one

10 or more intents, each intent including at least one example utterance that is associated with the intent and used for training a classifier. In some embodiments, the skill bot invocation stage 415, the task prediction models 460, and the entity extraction models 467 for the master bot system are trained to determine confidence scores for individual skill bots and confidence scores for system intents. Similarly, the intent prediction stage 420 and intent prediction models 465 and/or

15 the entity detection stage 422 and the entity extraction models 467 can be trained to determine a confidence score for each intent associated with the skill bot system. Whereas the classification performed by the skill bot invocation stage 415, the task prediction models 460, and the entity extraction models 467 are at the bot level, the classification performed by the intent prediction stage 420 and intent prediction models 465 and/or the entity detection stage 422 and the entity

20 extraction models 467 are at the intent level and therefore finer grained.

[0110] In some examples, entities associated with entity extraction models 467 are included in a group of entities defined by a gazetteer. For a bot trained with a particular skill, a gazetteer for that skill includes a group of entities related to that skill. For example, for a bot trained with a banking skill, a gazetteer for that skill may include a group of entities comprising a number

25 entity (number tag), a date and time entity (date and time tag), a currency entity (currency tag), a person entity (person tag), and a location entity (location tag). In another example, for a bot trained with a food ordering skill, a gazetteer for that skill includes a group of entities comprising a quantity entity (quantity tag), a type of food entity (food type tag), a time entity (time tag), a person entity (person tag), and a location entity (location tag). In some embodiments, entity

30 detection stage 422 identifies a particular skill of the bot, extracts one or more groups of entities pertaining to that particular skill from one or more gazetteers, detects entities within the one or

more groups of entities in one or more system queries and/or one or more user utterances, labels the detected entities with one or more context tags, determines a confidence score for each context tag (context tag distribution), and identifies one or more constraints 480 based on the context tag distribution for the one or more groups of entities.

- 5 **[0111]** In some examples, the context tag distribution is determined based on the context of one or more system queries, one or more user utterances, and/or an entire interaction between the system and the user. In some examples, a context tag for a detected entity may be given a high confidence score relative to other entities within the group of entities or a lower confidence score relative to other entities within the group of entities based on the context of one or more system
- 10 queries, one or more user utterances, and/or an entire interaction between the system and the user. In some examples, if a bot queries a user about one or more specific entities, one or more entities in the user's responsive utterance(s) pertaining to the one or more specific entities queried by the bot will be given a high confidence score relative to other entities detected in the user's responsive utterance(s). For example, a bot relating to a particular skill (e.g., banking)
- 15 may query a user about a specific entity (e.g., a quantity) within a group of entities corresponding to that particular skill (e.g., quantity, date and time, location, person, and transaction type). A user's responsive utterance(s) may include one or more detected entities including the specific entity. In this case, the detected quantity entity (i.e., quantity tag) in the user's responsive utterance(s) will be given a high confidence score and the other detected
- 20 entities (e.g., date and time, location, person, and transaction type) within the group of entities for the particular skill will be given lower confidence scores. Similarly, if multiple occurrences of the same entity are detected in the bot's query and/or user's utterance(s), then that entity will be given a high confidence score relative to confidence scores given for the other detected entities (i.e., those occurring with less frequency).
- 25 **[0112]** On some occasions, all detected entities will be given the same confidence score. For example, if a bot does not query a user about a specific entity and the user's responsive utterance(s) includes different entities, each detected entity in the user's responsive utterance(s) may be given the same confidence score. Similarly, if the bot's query and/or the user's responsive utterance(s) does not include multiple occurrences of the same entity, each detected
- 30 entity may be given the same confidence score. In some cases, one or more detected entities of

the group of entities may be given a first confidence score, one or more detected entities of the group of entities may be given a second confidence score, different from the first confidence score, and one or more detected entities of the group of entities may be given a third confidence score, different from the first and second confidence scores. The foregoing discussion is merely
5 exemplary and not limited to determining confidence scores based on specific entity inclusion and entity occurrence frequency. Other metrics for determining which detected entities within a group of entities are contextually relevant. For example, one or more detected entities of a group of entities may be considered by the bot to be more contextually relevant than one or more other detected entities of the group of entities.

10 **[0113]** In some examples, the confidence scores for detected entities in the bot's query and/or user's utterance(s) form a context tag distribution. In some examples, the context tag distribution is a vector of confidence scores for all the detected entities within a group of entities for a bot's query and/or user's utterance(s). Accordingly, by considering the context of one or more system queries, one or more user utterances, and/or an entire interaction between the system and the
15 user, features of the present disclosure determine how contextually relevant one or more detected entities within a group of entities may be to a system query and/or a user's utterance(s).

[0114] In some examples, entity detection stage 422 includes a baseline model to predict one or more constraints 480 based on the context tag distribution. In some examples, the baseline model may be pre-trained. Alternatively, in some examples, the baseline model may be initially
20 trained and updated. In some examples, the baseline model may be trained based on a crowdsourced dataset. In some examples, the baseline model may be trained based on the Conference on Computational Natural Language Learning (CoNLL) dataset. In some examples, the baseline model 4000 is configured as shown in FIG. 4B. As shown in FIG. 4B, the baseline model 4000 is configured with a transformer-based model, such as a bidirectional encoder
25 representations from transformers (BERT) model 4400, a regular expression (RX)/gazetteer (GZ) vectorizer 4500, a context tag vectorizer 4600, a sequence processing model, such as a combined convolutional neural network/bidirectional long short-term memory (CNN/BiLSTM) model 4700, and a discriminative model, such as conditional random field (CRF) model 4800.

[0115] The BERT model 4400 is a pre-trained algorithm that accepts one or more sequences of
30 words from a user utterance(s) or system query(ies) as an input and generates one or more

feature vectors (word embeddings) for each of the one or more words of the one or more sequences of words. For example, as shown in FIG. 4B, for an input sequence of words “I would like to pay Merchant \$10,” BERT model 4400 generates a separate word embedding for each individual word of the sequence (“I,” “would,” “like,” “to,” “pay,” “Merchant,” and “\$10”). In some examples, BERT model 4400 includes at least one transformer layer for receiving the input sequence of words. In some examples, the at least one transformer layer includes a plurality of encoders. In some examples, each encoder includes a plurality of attention mechanisms and a plurality of feed-forward networks. In some examples, the input sequence of words is tokenized to generate a plurality of word tokens. In some examples, the plurality of attention mechanisms operates directly on the words of the input of sequence of words. In some examples, the plurality of attention mechanisms operates on the plurality of word tokens. In some examples, the plurality of attention mechanisms generates an attention score for each word of the input sequence of words or each token of the plurality of word tokens. In some examples, the input sequence of words (or the plurality of word tokens) and the attention scores are input into the plurality of feed-forward networks. In some examples, the plurality of feed-forward networks encodes the input sequence of words (or the plurality of word tokens) into a plurality of word embeddings.

[0116] The RX/GZ vectorizer 4500 generates one or more feature vectors based on one or more words of the input sequence of words that match(es) one or more known regular expression patterns in one or more known gazetteers. For example, as shown in FIG. 4B, for the input sequence of words “I would like to pay Merchant \$10,” RX/GZ vectorizer 4500 determines one or more words of the input sequence of words (“Merchant,” “\$10”) that match one or more regular expression patterns (e.g., Merchants, ten dollars) of words listed in one or more gazetteers and extracts a pre-defined vector for each of the one or more matched words. In some examples, the one or more gazetteers is automatically selected based on the skill of the chatbot. In some examples, the one or more gazetteers is selected by a user of the chatbot. In some examples, the one or more gazetteers is selected by a user of the digital assistant that includes the chatbot. In some examples, a group of entities for each skill of a plurality of chatbot skills is defined by one or more gazetteers. For example, for a chatbot trained with a banking skill, a gazetteer for that chatbot skill includes a group of entities including a number entity (number tag), a date and time entity (date and time tag), a currency entity (currency tag), a person entity

(person tag), and a location entity (location tag). In another example, for a chatbot trained with a pizza ordering skill, a gazetteer for that chatbot skill includes a group of entities including a quantity entity (quantity tag), a type entity (type tag), a toppings entity (toppings tag), an address entity (address tag), and cost entity (cost tag). In some examples, entities within a group of entities defined by one or more gazetteers pertaining to a particular skill of the chatbot are matched to one or more words of the input sequence of words. In some examples, a regular expression algorithm matches regular patterns of one or more of the words of the input sequence of words to regular expression patterns of the one or more entities in the one or more groups of entities listed in the one or more gazetteers. In some examples, the one or more gazetteers include a pre-defined vector for each entity and its associated regular expression patterns. In some examples, for every matched entity and matched regular expression pattern, a pre-defined vector is extracted. For example, for the input sequence of "I would like to pay Merchant \$10," RX/GZ vectorizer 4500 extracts a pre-defined vector for the Merchant entity and a pre-defined vector for the \$10 entity. In some examples, the plurality of word embeddings output from the BERT model 4400 are concatenated and/or interpolated with one or more of the pre-defined vectors extracted from the RX/GZ vectorizer 4500 to generate a concatenated and/or interpolated set of vectors.

[0117] The context tag vectorizer 4600 generates one or more vectors based on the context tag distribution for the input sequence of words. The process for determining a context tag distribution for one or more system queries and/or one or more user utterance(s) is described above and not repeated herein. However, to illustrate, as shown in FIG. 4B, for the input sequence of words "I would like to pay Merchant \$10," a context tag distribution may be determined to be 0.5, 0.0, 0.0, 0.0, 0.8, 0.8, 0.8 for detected entities within a group of entities listed in a gazetteer for a particular skill of the chatbot. In some examples, the context tag vectorizer 4600 transforms the context tag distribution into one or more vectors. In some examples, one or more vectors generated by the context tag vectorizer 4600 are concatenated and/or interpolated with the concatenated and/or interpolated set of vectors to generate a first set of vector representations. In some embodiments, as described below, one or more vectors generated by the context tag vectorizer 4600 are concatenated and/or interpolated with one or more sentence-level vector representations generated by the CNN/BiLSTM model 4700.

[0118] In some examples, the first set of vector representations is input into the CNN/BiLSTM model 4700. Based on the first set of vector representations, the CNN of the CNN/BiLSTM model 4700 generates one or more character-level vector representations for each character of each word of the input sequence of words. The one or more character-level vector representations is then concatenated and/or interpolated with the first set of vector representations and input into the BiLSTM network to generate one or more sentence-level vector representations for the input sequence of words. In some examples, the one or more sentence-level vector representations represents named entity tag scores. In some examples, one or more vectors generated by the context tag vectorizer 4600 are concatenated and/or interpolated with one or more sentence-level vector representations generated by the CNN/BiLSTM model 4700 to generate a second set of vector representations. In some examples, the second set of vector representations represent named entity tag scores. In some examples, the named entity tag scores are decoded into named entities using CRF model 4800. For example, as shown in FIG. 4B, CRF model 4800 identifies “Business” and “Currency” as two named entities for the input sequence of words “I would like to pay Merchant \$10” based on the one or more sentence-level vector representations and/or the second set of vector representations for the input sequence words.

[0119] FIG. 5 is a flowchart illustrating a process 500 for taking context into consideration for entity recognition (entity extraction and classification) according to certain embodiments. The processing depicted in FIG. 5 may be implemented in software (e.g., code, instructions, program) executed by one or more processing units (e.g., processors, cores) of the respective systems, hardware, or combinations thereof. The software may be stored on a non-transitory storage medium (e.g., on a memory device). The method presented in FIG. 5 and described below is intended to be illustrative and non-limiting. Although FIG. 5 depicts the various processing steps occurring in a particular sequence or order, this is not intended to be limiting. In certain alternative embodiments, the steps may be performed in some different order or some steps may also be performed in parallel. In certain embodiments, such as in the embodiment depicted in FIGS. 1-4B, the processing depicted in FIG. 5 may be performed by a pre-processing subsystem (e.g., pre-processing subsystem 210 or prediction model training stage 410) to generate training sets with context labels for training by one or more predictive models (e.g., entity extraction models 467).

[0120] At step 505, an utterance is received by chatbot system 400 (FIG. 4). In some examples, the utterance may be received in response to a system query. In some examples, the utterance corresponds to one or more user utterance(s) in response to one or more system queries. In some examples, the utterance corresponds to one or more user interactions with the chatbot. In some examples, the utterance corresponds an input sequence of words.

[0121] At step 510, embeddings for words of the utterance are generated. In some examples, an embedding is generated for each word of the utterance. In some examples, embeddings are generated using a transformer-based model, such as BERT model 4400 of FIG. 4B. Features and operations of BERT model 4400 have been described above and are not repeated herein.

[0122] At step 515, RX/GZ feature vectors are generated and concatenated and/or interpolated with the embeddings to generate a concatenated and/or interpolated set of vectors. In some examples, RX/GZ feature vectors are generated using the RX/GZ vectorizer 4500 of FIG. 4B. In some examples, the RX/GZ vectorizer 4500 generates one or more feature vectors based on one or more words of the input sequence of words that match(es) one or more known regular expression patterns in one or more known gazetteers. In some examples, the one or more gazetteers is automatically selected based on the skill of the chatbot. In some examples, the one or more gazetteers is selected by a user of the chatbot. In some examples, the one or more gazetteers is selected by a user of the digital assistant that includes the chatbot. In some examples, a group of entities for each skill of a plurality of chatbot skills is defined by one or more gazetteers. In some examples, entities within a group of entities defined by one or more gazetteers pertaining to a particular skill of the chatbot are matched to one or more words of the input sequence of words. In some examples, a regular expression algorithm matches regular patterns of one or more of the words of the input sequence of words to regular expression patterns of the one or more entities in the one or more groups of entities listed in the one or more gazetteers. In some examples, the one or more gazetteers include a pre-defined vector for each entity and its associated regular expression patterns. In some examples, for every matched entity and matched regular expression pattern, a pre-defined vector is extracted. Other features and operations of RX/GZ vectorizer 4500 have been described above and are not repeated herein. In some examples, the plurality of word embeddings output from the BERT model 4400 are

concatenated and/or interpolated with the RXGZ feature vectors extracted from the RX/GZ vectorizer 4500 to generate a concatenated and/or interpolated set of vectors.

[0123] At optional step 520, context tag distribution feature vectors are generated for the received utterance and concatenated and/or interpolated with the concatenated and/or

5 interpolated RX/GZ feature vectors and embeddings to form a first set of feature vectors. In some examples, context tag distribution feature vectors are generated using the context tag vectorizer 4600 of FIG. 4B. In some examples, context tag distribution is determined based on the context of one or more system queries, one or more user utterances, and/or an entire interaction between the system and the user. In some examples, a context tag for a detected entity
10 may be given a high confidence score relative to other entities within a group of entities or a lower confidence score relative to other entities within the group of entities based on the context of one or more system queries, one or more user utterances, and/or an entire interaction between the system and the user. In some examples, if a bot queries a user about one or more specific entities, one or more entities in the user's responsive utterance(s) pertaining to the one or more
15 specific entities queried by the bot will be given a high confidence score relative to other entities detected in the user's responsive utterance(s). Similarly, if multiple occurrences of the same entity are detected in the bot's query and/or user's utterance(s), then that entity will be given a high confidence score relative to confidence scores given for the other detected entities (i.e., those occurring with less frequency). On some occasions, all detected entities will be given the
20 same confidence score. For example, if a bot does not query a user about a specific entity and the user's responsive utterance(s) includes different entities, each detected entity in the user's responsive utterance(s) may be given the same confidence score. Similarly, if the bot's query and/or the user's responsive utterance(s) does not include multiple occurrences of the same entity, each detected entity may be given the same confidence score. In some cases, one or more
25 detected entities of the group of entities may be given a first confidence score, one or more detected entities of the group of entities may be given a second confidence score, different from the first confidence score, and one or more detected entities of the group of entities may be given a third confidence score, different from the first and second confidence scores.

[0124] The foregoing discussion is merely exemplary and not limited to determining
30 confidence scores based on specific entity inclusion and entity occurrence frequency. Other

metrics for determining which detected entities within a group of entities are contextually relevant. For example, one or more detected entities of a group of entities may be considered by the bot to be more contextually relevant than one or more other detected entities of the group of entities. In some examples, the confidence scores for detected entities in the bot's query and/or user's utterance(s) form a context tag distribution. In some examples, the context tag distribution is a vector of confidence scores for all the detected entities within a group of entities for a bot's query and/or user's utterance(s). In some examples, the context tag vectorizer 4600 generates one or more vectors based on the context tag distribution. In some examples, the context tag vectorizer 4600 transforms the context tag distribution into one or more context tags distribution vectors. In some examples, the context tag distribution vectors generated by the context tag vectorizer 4600 are concatenated and/or interpolated with the concatenated and/or interpolated set of vectors to generate the first set of feature vectors. In some embodiments, as described below, context tag distribution vectors generated by the context tag vectorizer 4600 are concatenated and/or interpolated with the encoded form of the utterance generated by the CNN/BiLSTM model 4700.

[0125] At step 525, an encoded form of the utterance is generated based on the concatenated and/or interpolated set of vectors. In some examples, if the optional step 520 is included, an encoded form of the utterance is generated based on the first set of feature vectors. The encoded form of the utterance is generated using a sequence processing model, such as the CNN/BiLSTM model 4700 of FIG. 4B. In some examples, the concatenated and/or interpolated set of vectors and/or the first set of feature vectors is input into the CNN/BiLSTM model 4700. Based on the inputted vectors, the CNN of the CNN/BiLSTM model 4700 generates one or more character-level vector representations for each character of each word of the utterance. The one or more character-level vector representations is then concatenated and/or interpolated with the concatenated and/or interpolated set of vectors and/or the first set of feature vectors and input into the BiLSTM network to generate one or more sentence-level vector representations for utterance. In some examples, the encoded form of the utterance comprises the one or more sentence-level vector representations. In some examples, the encoded form of the utterance represents named entity tag scores.

[0126] At optional step 530, context tag distribution feature vectors are generated for the received utterance and concatenated and/or interpolated with the encoded form of the utterance. In some examples, context tax distribution feature vectors are generated using the context tag vectorizer 4600 of FIG. 4B. Features and operations of context tag vectorizer 4600 and context tag distribution feature vector generation have been described above and are not repeated herein. In some examples, the encoded form of the utterance concatenated and/or interpolated with the context tag distribution feature vectors represents name entity tag score.

[0127] At step 535, log-probabilities for candidate entities are generated based on the encoded form of the utterance. Log-probabilities for candidate entities may be generated using a discriminative model, such as the CRF model 4800 of FIG. 4B. In some examples, the log-probabilities are decoded by CRF model 4800 into named entities.

[0128] At step 540, the log-probabilities are used to identify one or more constraints 480 for the received utterance. In some examples, entity detection stage 422 (FIG. 4A) uses the decoded named entities to identify one or more constraints 480 for the received utterance.

[0129] At optional step 545, the one or more constraints 480 and one or more intent predictions generated by the intent prediction stage 420 (FIG. 4A) for the received utterance are matched to an intent 475 (FIG. 4A) associated with a skill bot.

[0130] Features of the present disclosure improve user interactions with digital assistants and/or chatbot systems. For example, a user may interact with digital assistant/chatbot system 106, as shown in FIG. 1, to order a to conduct a banking transaction. A skill bot pertaining to the banking transaction, such as Skill Bot #1 116-1, as shown in FIG. 1, may be invoked at Skill Bot Invocation stage 415. The interaction includes one or more user utterances and one or system queries. For a banking transaction, the skill bot may inquire the user about specific banking tasks, such as balance checks, deposit checks, transfers, etc., the user wishes to be performed by the skill bot and/or digital assistant. In response, the user may utter one or more utterances pertaining to the user's banking intent. In one example, the user may utter "please deposit 20 in my account 20." Based on the system, methods, and examples provided herein, as shown in FIG. 1-5 and described throughout, the digital assistant and/or chatbot system will be able to correctly identify "deposit 20" as pertaining to the amount the user wishes to deposit and "account 20" as pertaining to the account the user wishes to make the deposit along with the user's intent to

deposit money into the user's account. As discussed throughout, by considering the context of one or more system queries, one or more user utterances, and/or an entire interaction between the system and the user, features of the present disclosure determine how contextually relevant one or more detected entities within a group of entities may be to a system query and/or a user's utterance(s), features of the present disclosure are able to correctly identify the intended referent for a particular named entity and improve the user's interaction with the digital assistant and/or chatbot system.

Illustrative Systems

[0131] FIG. 6 depicts a simplified diagram of a distributed system 600. In the illustrated example, distributed system 600 includes one or more client computing devices 602, 604, 606, and 608, coupled to a server 612 via one or more communication networks 610. Clients computing devices 602, 604, 606, and 608 may be configured to execute one or more applications.

[0132] In various examples, server 612 may be adapted to run one or more services or software applications that enable one or more embodiments described in this disclosure. In certain examples, server 612 may also provide other services or software applications that may include non-virtual and virtual environments. In some examples, these services may be offered as web-based or cloud services, such as under a Software as a Service (SaaS) model to the users of client computing devices 602, 604, 606, and/or 608. Users operating client computing devices 602, 604, 606, and/or 608 may in turn utilize one or more client applications to interact with server 612 to utilize the services provided by these components.

[0133] In the configuration depicted in FIG. 6, server 612 may include one or more components 618, 620 and 622 that implement the functions performed by server 612. These components may include software components that may be executed by one or more processors, hardware components, or combinations thereof. It should be appreciated that various different system configurations are possible, which may be different from distributed system 600. The example shown in FIG. 6 is thus one example of a distributed system for implementing an example system and is not intended to be limiting.

[0134] Users may use client computing devices 602, 604, 606, and/or 608 to execute one or more applications, models or chatbots, which may generate one or more events or models that may then be implemented or serviced in accordance with the teachings of this disclosure. A client device may provide an interface that enables a user of the client device to interact with the client device. The client device may also output information to the user via this interface. Although FIG. 6 depicts only four client computing devices, any number of client computing devices may be supported.

[0135] The client devices may include various types of computing systems such as portable handheld devices, general purpose computers such as personal computers and laptops, workstation computers, wearable devices, gaming systems, thin clients, various messaging devices, sensors or other sensing devices, and the like. These computing devices may run various types and versions of software applications and operating systems (e.g., Microsoft Windows®, Apple Macintosh®, UNIX® or UNIX-like operating systems, Linux or Linux-like operating systems such as Google Chrome™ OS) including various mobile operating systems (e.g., Microsoft Windows Mobile®, iOS®, Windows Phone®, Android™, BlackBerry®, Palm OS®). Portable handheld devices may include cellular phones, smartphones, (e.g., an iPhone®), tablets (e.g., iPad®), personal digital assistants (PDAs), and the like. Wearable devices may include Google Glass® head mounted display, and other devices. Gaming systems may include various handheld gaming devices, Internet-enabled gaming devices (e.g., a Microsoft Xbox® gaming console with or without a Kinect® gesture input device, Sony PlayStation® system, various gaming systems provided by Nintendo®, and others), and the like. The client devices may be capable of executing various different applications such as various Internet-related apps, communication applications (e.g., E-mail applications, short message service (SMS) applications) and may use various communication protocols.

[0136] Network(s) 610 may be any type of network familiar to those skilled in the art that may support data communications using any of a variety of available protocols, including without limitation TCP/IP (transmission control protocol/Internet protocol), SNA (systems network architecture), IPX (Internet packet exchange), AppleTalk®, and the like. Merely by way of example, network(s) 610 may be a local area network (LAN), networks based on Ethernet, Token-Ring, a wide-area network (WAN), the Internet, a virtual network, a virtual private

network (VPN), an intranet, an extranet, a public switched telephone network (PSTN), an infra-red network, a wireless network (e.g., a network operating under any of the Institute of Electrical and Electronics (IEEE) 1002.11 suite of protocols, Bluetooth[®], and/or any other wireless protocol), and/or any combination of these and/or other networks.

5 **[0137]** Server 612 may be composed of one or more general purpose computers, specialized server computers (including, by way of example, PC (personal computer) servers, UNIX[®] servers, mid-range servers, mainframe computers, rack-mounted servers, etc.), server farms, server clusters, or any other appropriate arrangement and/or combination. Server 612 may include one or more virtual machines running virtual operating systems, or other computing
10 architectures involving virtualization such as one or more flexible pools of logical storage devices that may be virtualized to maintain virtual storage devices for the server. In various examples, server 612 may be adapted to run one or more services or software applications that provide the functionality described in the foregoing disclosure.

[0138] The computing systems in server 612 may run one or more operating systems including
15 any of those discussed above, as well as any commercially available server operating system. Server 612 may also run any of a variety of additional server applications and/or mid-tier applications, including HTTP (hypertext transport protocol) servers, FTP (file transfer protocol) servers, CGI (common gateway interface) servers, JAVA[®] servers, database servers, and the like. Exemplary database servers include without limitation those commercially available from
20 Oracle[®], Microsoft[®], Sybase[®], IBM[®] (International Business Machines), and the like.

[0139] In some implementations, server 612 may include one or more applications to analyze and consolidate data feeds and/or event updates received from users of client computing devices 602, 604, 606, and 608. As an example, data feeds and/or event updates may include, but are not limited to, Twitter[®] feeds, Facebook[®] updates or real-time updates received from one or more
25 third party information sources and continuous data streams, which may include real-time events related to sensor data applications, financial tickers, network performance measuring tools (e.g., network monitoring and traffic management applications), clickstream analysis tools, automobile traffic monitoring, and the like. Server 612 may also include one or more applications to display the data feeds and/or real-time events via one or more display devices of client computing
30 devices 602, 604, 606, and 608.

[0140] Distributed system 600 may also include one or more data repositories 614, 616. These data repositories may be used to store data and other information in certain examples. For example, one or more of the data repositories 614, 616 may be used to store information such as information related to chatbot performance or generated models for use by chatbots used by server 612 when performing various functions in accordance with various embodiments. Data repositories 614, 616 may reside in a variety of locations. For example, a data repository used by server 612 may be local to server 612 or may be remote from server 612 and in communication with server 612 via a network-based or dedicated connection. Data repositories 614, 616 may be of different types. In certain examples, a data repository used by server 612 may be a database, for example, a relational database, such as databases provided by Oracle Corporation® and other vendors. One or more of these databases may be adapted to enable storage, update, and retrieval of data to and from the database in response to SQL-formatted commands.

[0141] In certain examples, one or more of data repositories 614, 616 may also be used by applications to store application data. The data repositories used by applications may be of different types such as, for example, a key-value store repository, an object store repository, or a general storage repository supported by a file system.

[0142] In certain examples, the functionalities described in this disclosure may be offered as services via a cloud environment. FIG. 7 is a simplified block diagram of a cloud-based system environment in which various services may be offered as cloud services in accordance with certain examples. In the example depicted in FIG. 7, cloud infrastructure system 702 may provide one or more cloud services that may be requested by users using one or more client computing devices 704, 706, and 708. Cloud infrastructure system 702 may comprise one or more computers and/or servers that may include those described above for server 612. The computers in cloud infrastructure system 702 may be organized as general-purpose computers, specialized server computers, server farms, server clusters, or any other appropriate arrangement and/or combination.

[0143] Network(s) 710 may facilitate communication and exchange of data between clients 704, 706, and 708 and cloud infrastructure system 702. Network(s) 710 may include one or more networks. The networks may be of the same or different types. Network(s) 710 may support one

or more communication protocols, including wired and/or wireless protocols, for facilitating the communications.

[0144] The example depicted in FIG. 7 is only one example of a cloud infrastructure system and is not intended to be limiting. It should be appreciated that, in some other examples, cloud infrastructure system 702 may have more or fewer components than those depicted in FIG. 7, may combine two or more components, or may have a different configuration or arrangement of components. For example, although FIG. 7 depicts three client computing devices, any number of client computing devices may be supported in alternative examples.

[0145] The term cloud service is generally used to refer to a service that is made available to users on demand and via a communication network such as the Internet by systems (e.g., cloud infrastructure system 702) of a service provider. Typically, in a public cloud environment, servers and systems that make up the cloud service provider's system are different from the customer's own on-premise servers and systems. The cloud service provider's systems are managed by the cloud service provider. Customers may thus avail themselves of cloud services provided by a cloud service provider without having to purchase separate licenses, support, or hardware and software resources for the services. For example, a cloud service provider's system may host an application, and a user may, via the Internet, on demand, order and use the application without the user having to buy infrastructure resources for executing the application. Cloud services are designed to provide easy, scalable access to applications, resources and services. Several providers offer cloud services. For example, several cloud services are offered by Oracle Corporation® of Redwood Shores, California, such as middleware services, database services, Java cloud services, and others.

[0146] In certain examples, cloud infrastructure system 702 may provide one or more cloud services using different models such as under a Software as a Service (SaaS) model, a Platform as a Service (PaaS) model, an Infrastructure as a Service (IaaS) model, and others, including hybrid service models. Cloud infrastructure system 702 may include a suite of applications, middleware, databases, and other resources that enable provision of the various cloud services.

[0147] A SaaS model enables an application or software to be delivered to a customer over a communication network like the Internet, as a service, without the customer having to buy the hardware or software for the underlying application. For example, a SaaS model may be used to

provide customers access to on-demand applications that are hosted by cloud infrastructure system 702. Examples of SaaS services provided by Oracle Corporation® include, without limitation, various services for human resources/capital management, customer relationship management (CRM), enterprise resource planning (ERP), supply chain management (SCM),
5 enterprise performance management (EPM), analytics services, social applications, and others.

[0148] An IaaS model is generally used to provide infrastructure resources (e.g., servers, storage, hardware and networking resources) to a customer as a cloud service to provide elastic compute and storage capabilities. Various IaaS services are provided by Oracle Corporation®.

[0149] A PaaS model is generally used to provide, as a service, platform and environment
10 resources that enable customers to develop, run, and manage applications and services without the customer having to procure, build, or maintain such resources. Examples of PaaS services provided by Oracle Corporation® include, without limitation, Oracle Java Cloud Service (JCS), Oracle Database Cloud Service (DBCS), data management cloud service, various application development solutions services, and others.

[0150] Cloud services are generally provided on an on-demand self-service basis, subscription-based, elastically scalable, reliable, highly available, and secure manner. For example, a customer, via a subscription order, may order one or more services provided by cloud infrastructure system 702. Cloud infrastructure system 702 then performs processing to provide the services requested in the customer's subscription order. For example, a user may use
20 utterances to request the cloud infrastructure system to take a certain action (e.g., an intent), as described above, and/or provide services for a chatbot system as described herein. Cloud infrastructure system 702 may be configured to provide one or even multiple cloud services.

[0151] Cloud infrastructure system 702 may provide the cloud services via different deployment models. In a public cloud model, cloud infrastructure system 702 may be owned by
25 a third party cloud services provider and the cloud services are offered to any general public customer, where the customer may be an individual or an enterprise. In certain other examples, under a private cloud model, cloud infrastructure system 702 may be operated within an organization (e.g., within an enterprise organization) and services provided to customers that are within the organization. For example, the customers may be various departments of an enterprise
30 such as the Human Resources department, the Payroll department, etc. or even individuals within

the enterprise. In certain other examples, under a community cloud model, the cloud infrastructure system 702 and the services provided may be shared by several organizations in a related community. Various other models such as hybrids of the above mentioned models may also be used.

5 **[0152]** Client computing devices 704, 706, and 708 may be of different types (such as client computing devices 602, 604, 606, and 608 depicted in FIG. 6) and may be capable of operating one or more client applications. A user may use a client device to interact with cloud infrastructure system 702, such as to request a service provided by cloud infrastructure system 702. For example, a user may use a client device to request information or action from a chatbot
10 as described in this disclosure.

[0153] In some examples, the processing performed by cloud infrastructure system 702 for providing services may involve model training and deployment. This analysis may involve using, analyzing, and manipulating data sets to train and deploy one or more models. This analysis may be performed by one or more processors, possibly processing the data in parallel, performing
15 simulations using the data, and the like. For example, big data analysis may be performed by cloud infrastructure system 702 for generating and training one or more models for a chatbot system. The data used for this analysis may include structured data (e.g., data stored in a database or structured according to a structured model) and/or unstructured data (e.g., data blobs (binary large objects)).

20 **[0154]** As depicted in the example in FIG. 7, cloud infrastructure system 702 may include infrastructure resources 730 that are utilized for facilitating the provision of various cloud services offered by cloud infrastructure system 702. Infrastructure resources 730 may include, for example, processing resources, storage or memory resources, networking resources, and the like. In certain examples, the storage virtual machines that are available for servicing storage
25 requested from applications may be part of cloud infrastructure system 702. In other examples, the storage virtual machines may be part of different systems.

[0155] In certain examples, to facilitate efficient provisioning of these resources for supporting the various cloud services provided by cloud infrastructure system 702 for different customers, the resources may be bundled into sets of resources or resource modules (also referred to as
30 "pods"). Each resource module or pod may comprise a pre-integrated and optimized combination

of resources of one or more types. In certain examples, different pods may be pre-provisioned for different types of cloud services. For example, a first set of pods may be provisioned for a database service, a second set of pods, which may include a different combination of resources than a pod in the first set of pods, may be provisioned for Java service, and the like. For some
5 services, the resources allocated for provisioning the services may be shared between the services.

[0156] Cloud infrastructure system 702 may itself internally use services 732 that are shared by different components of cloud infrastructure system 702 and which facilitate the provisioning of services by cloud infrastructure system 702. These internal shared services may include,
10 without limitation, a security and identity service, an integration service, an enterprise repository service, an enterprise manager service, a virus scanning and whitelist service, a high availability, backup and recovery service, service for enabling cloud support, an email service, a notification service, a file transfer service, and the like.

[0157] Cloud infrastructure system 702 may comprise multiple subsystems. These subsystems
15 may be implemented in software, or hardware, or combinations thereof. As depicted in FIG. 7, the subsystems may include a user interface subsystem 712 that enables users or customers of cloud infrastructure system 702 to interact with cloud infrastructure system 702. User interface subsystem 712 may include various different interfaces such as a web interface 714, an online store interface 716 where cloud services provided by cloud infrastructure system 702 are
20 advertised and are purchasable by a consumer, and other interfaces 718. For example, a customer may, using a client device, request (service request 734) one or more services provided by cloud infrastructure system 702 using one or more of interfaces 714, 716, and 718. For example, a customer may access the online store, browse cloud services offered by cloud infrastructure system 702, and place a subscription order for one or more services offered by cloud
25 infrastructure system 702 that the customer wishes to subscribe to. The service request may include information identifying the customer and one or more services that the customer desires to subscribe to. For example, a customer may place a subscription order for a service offered by cloud infrastructure system 702. As part of the order, the customer may provide information identifying a chatbot system for which the service is to be provided and optionally one or more
30 credentials for the chatbot system.

[0158] In certain examples, such as the example depicted in FIG. 7, cloud infrastructure system 702 may comprise an order management subsystem (OMS) 720 that is configured to process the new order. As part of this processing, OMS 720 may be configured to: create an account for the customer, if not done already; receive billing and/or accounting information from the customer that is to be used for billing the customer for providing the requested service to the customer; verify the customer information; upon verification, book the order for the customer; and orchestrate various workflows to prepare the order for provisioning.

[0159] Once properly validated, OMS 720 may then invoke the order provisioning subsystem (OPS) 724 that is configured to provision resources for the order including processing, memory, and networking resources. The provisioning may include allocating resources for the order and configuring the resources to facilitate the service requested by the customer order. The manner in which resources are provisioned for an order and the type of the provisioned resources may depend upon the type of cloud service that has been ordered by the customer. For example, according to one workflow, OPS 724 may be configured to determine the particular cloud service being requested and identify a number of pods that may have been pre-configured for that particular cloud service. The number of pods that are allocated for an order may depend upon the size/amount/level/scope of the requested service. For example, the number of pods to be allocated may be determined based upon the number of users to be supported by the service, the duration of time for which the service is being requested, and the like. The allocated pods may then be customized for the particular requesting customer for providing the requested service.

[0160] In certain examples, setup phase processing, as described above, may be performed by cloud infrastructure system 702 as part of the provisioning process. Cloud infrastructure system 702 may generate an application ID and select a storage virtual machine for an application from among storage virtual machines provided by cloud infrastructure system 702 itself or from storage virtual machines provided by other systems other than cloud infrastructure system 702.

[0161] Cloud infrastructure system 702 may send a response or notification 744 to the requesting customer to indicate when the requested service is now ready for use. In some instances, information (e.g., a link) may be sent to the customer that enables the customer to start using and availing the benefits of the requested services. In certain examples, for a customer

requesting the service, the response may include a chatbot system ID generated by cloud infrastructure system 702 and information identifying a chatbot system selected by cloud infrastructure system 702 for the chatbot system corresponding to the chatbot system ID.

[0162] Cloud infrastructure system 702 may provide services to multiple customers. For each customer, cloud infrastructure system 702 is responsible for managing information related to one or more subscription orders received from the customer, maintaining customer data related to the orders, and providing the requested services to the customer. Cloud infrastructure system 702 may also collect usage statistics regarding a customer's use of subscribed services. For example, statistics may be collected for the amount of storage used, the amount of data transferred, the number of users, and the amount of system up time and system down time, and the like. This usage information may be used to bill the customer. Billing may be done, for example, on a monthly cycle.

[0163] Cloud infrastructure system 702 may provide services to multiple customers in parallel. Cloud infrastructure system 702 may store information for these customers, including possibly proprietary information. In certain examples, cloud infrastructure system 702 comprises an identity management subsystem (IMS) 728 that is configured to manage customer information and provide the separation of the managed information such that information related to one customer is not accessible by another customer. IMS 728 may be configured to provide various security-related services such as identity services, such as information access management, authentication and authorization services, services for managing customer identities and roles and related capabilities, and the like.

[0164] FIG. 8 illustrates an example of computer system 800. In some examples, computer system 800 may be used to implement any of the digital assistant or chatbot systems within a distributed environment, and various servers and computer systems described above. As shown in FIG. 8, computer system 800 includes various subsystems including a processing subsystem 804 that communicates with a number of other subsystems via a bus subsystem 802. These other subsystems may include a processing acceleration unit 806, an I/O subsystem 808, a storage subsystem 818, and a communications subsystem 824. Storage subsystem 818 may include non-transitory computer-readable storage media including storage media 822 and a system memory 810.

[0165] Bus subsystem 802 provides a mechanism for letting the various components and subsystems of computer system 800 communicate with each other as intended. Although bus subsystem 802 is shown schematically as a single bus, alternative examples of the bus subsystem may utilize multiple buses. Bus subsystem 802 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, a local bus using any of a variety of bus architectures, and the like. For example, such architectures may include an Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus, which may be implemented as a Mezzanine bus manufactured to the IEEE P1386.1 standard, and the like.

[0166] Processing subsystem 804 controls the operation of computer system 800 and may comprise one or more processors, application specific integrated circuits (ASICs), or field programmable gate arrays (FPGAs). The processors may include be single core or multicore processors. The processing resources of computer system 800 may be organized into one or more processing units 832, 834, etc. A processing unit may include one or more processors, one or more cores from the same or different processors, a combination of cores and processors, or other combinations of cores and processors. In some examples, processing subsystem 804 may include one or more special purpose co-processors such as graphics processors, digital signal processors (DSPs), or the like. In some examples, some or all of the processing units of processing subsystem 804 may be implemented using customized circuits, such as application specific integrated circuits (ASICs), or field programmable gate arrays (FPGAs).

[0167] In some examples, the processing units in processing subsystem 804 may execute instructions stored in system memory 810 or on computer readable storage media 822. In various examples, the processing units may execute a variety of programs or code instructions and may maintain multiple concurrently executing programs or processes. At any given time, some or all of the program code to be executed may be resident in system memory 810 and/or on computer-readable storage media 822 including potentially on one or more storage devices. Through suitable programming, processing subsystem 804 may provide various functionalities described above. In instances where computer system 800 is executing one or more virtual machines, one or more processing units may be allocated to each virtual machine.

[0168] In certain examples, a processing acceleration unit 806 may optionally be provided for performing customized processing or for off-loading some of the processing performed by processing subsystem 804 so as to accelerate the overall processing performed by computer system 800.

5 **[0169]** I/O subsystem 808 may include devices and mechanisms for inputting information to computer system 800 and/or for outputting information from or via computer system 800. In general, use of the term input device is intended to include all possible types of devices and mechanisms for inputting information to computer system 800. User interface input devices may include, for example, a keyboard, pointing devices such as a mouse or trackball, a touchpad or
10 touch screen incorporated into a display, a scroll wheel, a click wheel, a dial, a button, a switch, a keypad, audio input devices with voice command recognition systems, microphones, and other types of input devices. User interface input devices may also include motion sensing and/or gesture recognition devices such as the Microsoft Kinect[®] motion sensor that enables users to control and interact with an input device, the Microsoft Xbox[®] 360 game controller, devices that
15 provide an interface for receiving input using gestures and spoken commands. User interface input devices may also include eye gesture recognition devices such as the Google Glass[®] blink detector that detects eye activity (e.g., "blinking" while taking pictures and/or making a menu selection) from users and transforms the eye gestures as inputs to an input device (e.g., Google Glass[®]). Additionally, user interface input devices may include voice recognition sensing devices
20 that enable users to interact with voice recognition systems (e.g., Siri[®] navigator) through voice commands.

[0170] Other examples of user interface input devices include, without limitation, three dimensional (3D) mice, joysticks or pointing sticks, gamepads and graphic tablets, and audio/visual devices such as speakers, digital cameras, digital camcorders, portable media
25 players, webcams, image scanners, fingerprint scanners, barcode reader 3D scanners, 3D printers, laser rangefinders, and eye gaze tracking devices. Additionally, user interface input devices may include, for example, medical imaging input devices such as computed tomography, magnetic resonance imaging, position emission tomography, and medical ultrasonography devices. User interface input devices may also include, for example, audio input devices such as
30 MIDI keyboards, digital musical instruments and the like.

[0171] In general, use of the term output device is intended to include all possible types of devices and mechanisms for outputting information from computer system 800 to a user or other computer. User interface output devices may include a display subsystem, indicator lights, or non-visual displays such as audio output devices, etc. The display subsystem may be a cathode ray tube (CRT), a flat-panel device, such as that using a liquid crystal display (LCD) or plasma display, a projection device, a touch screen, and the like. For example, user interface output devices may include, without limitation, a variety of display devices that visually convey text, graphics and audio/video information such as monitors, printers, speakers, headphones, automotive navigation systems, plotters, voice output devices, and modems.

[0172] Storage subsystem 818 provides a repository or data store for storing information and data that is used by computer system 800. Storage subsystem 818 provides a tangible non-transitory computer-readable storage medium for storing the basic programming and data constructs that provide the functionality of some examples. Storage subsystem 818 may store software (e.g., programs, code modules, instructions) that when executed by processing subsystem 804 provides the functionality described above. The software may be executed by one or more processing units of processing subsystem 804. Storage subsystem 818 may also provide authentication in accordance with the teachings of this disclosure.

[0173] Storage subsystem 818 may include one or more non-transitory memory devices, including volatile and non-volatile memory devices. As shown in FIG. 8, storage subsystem 818 includes a system memory 810 and a computer-readable storage media 822. System memory 810 may include a number of memories including a volatile main random access memory (RAM) for storage of instructions and data during program execution and a non-volatile read only memory (ROM) or flash memory in which fixed instructions are stored. In some implementations, a basic input/output system (BIOS), containing the basic routines that help to transfer information between elements within computer system 800, such as during start-up, may typically be stored in the ROM. The RAM typically contains data and/or program modules that are presently being operated and executed by processing subsystem 804. In some implementations, system memory 810 may include multiple different types of memory, such as static random access memory (SRAM), dynamic random access memory (DRAM), and the like.

[0174] By way of example, and not limitation, as depicted in FIG. 8, system memory 810 may load application programs 812 that are being executed, which may include various applications such as Web browsers, mid-tier applications, relational database management systems (RDBMS), etc., program data 814, and an operating system 816. By way of example, operating system 816 may include various versions of Microsoft Windows[®], Apple Macintosh[®], and/or Linux operating systems, a variety of commercially-available UNIX[®] or UNIX-like operating systems (including without limitation the variety of GNU/Linux operating systems, the Google Chrome[®] OS, and the like) and/or mobile operating systems such as iOS, Windows[®] Phone, Android[®] OS, BlackBerry[®] OS, Palm[®] OS operating systems, and others.

[0175] Computer-readable storage media 822 may store programming and data constructs that provide the functionality of some examples. Computer-readable media 822 may provide storage of computer-readable instructions, data structures, program modules, and other data for computer system 800. Software (programs, code modules, instructions) that, when executed by processing subsystem 804 provides the functionality described above, may be stored in storage subsystem 818. By way of example, computer-readable storage media 822 may include non-volatile memory such as a hard disk drive, a magnetic disk drive, an optical disk drive such as a CD ROM, DVD, a Blu-Ray[®] disk, or other optical media. Computer-readable storage media 822 may include, but is not limited to, Zip[®] drives, flash memory cards, universal serial bus (USB) flash drives, secure digital (SD) cards, DVD disks, digital video tape, and the like. Computer-readable storage media 822 may also include, solid-state drives (SSD) based on non-volatile memory such as flash-memory based SSDs, enterprise flash drives, solid state ROM, and the like, SSDs based on volatile memory such as solid state RAM, dynamic RAM, static RAM, DRAM-based SSDs, magnetoresistive RAM (MRAM) SSDs, and hybrid SSDs that use a combination of DRAM and flash memory based SSDs.

[0176] In certain examples, storage subsystem 818 may also include a computer-readable storage media reader 820 that may further be connected to computer-readable storage media 822. Reader 820 may receive and be configured to read data from a memory device such as a disk, a flash drive, etc.

[0177] In certain examples, computer system 800 may support virtualization technologies, including but not limited to virtualization of processing and memory resources. For example,

computer system 800 may provide support for executing one or more virtual machines. In certain examples, computer system 800 may execute a program such as a hypervisor that facilitated the configuring and managing of the virtual machines. Each virtual machine may be allocated memory, compute (e.g., processors, cores), I/O, and networking resources. Each virtual machine
5 generally runs independently of the other virtual machines. A virtual machine typically runs its own operating system, which may be the same as or different from the operating systems executed by other virtual machines executed by computer system 800. Accordingly, multiple operating systems may potentially be run concurrently by computer system 800.

[0178] Communications subsystem 824 provides an interface to other computer systems and
10 networks. Communications subsystem 824 serves as an interface for receiving data from and transmitting data to other systems from computer system 800. For example, communications subsystem 824 may enable computer system 800 to establish a communication channel to one or more client devices via the Internet for receiving and sending information from and to the client devices. For example, when computer system 800 is used to implement bot system 120 depicted
15 in FIG. 1, the communication subsystem may be used to communicate with a chatbot system selected for an application.

[0179] Communication subsystem 824 may support both wired and/or wireless communication protocols. In certain examples, communications subsystem 824 may include radio frequency (RF) transceiver components for accessing wireless voice and/or data networks (e.g., using
20 cellular telephone technology, advanced data network technology, such as 3G, 4G or EDGE (enhanced data rates for global evolution), Wi-Fi (IEEE 802.XX family standards, or other mobile communication technologies, or any combination thereof), global positioning system (GPS) receiver components, and/or other components. In some examples, communications subsystem 824 may provide wired network connectivity (e.g., Ethernet) in addition to or instead
25 of a wireless interface.

[0180] Communication subsystem 824 may receive and transmit data in various forms. In some examples, in addition to other forms, communications subsystem 824 may receive input communications in the form of structured and/or unstructured data feeds 826, event streams 828, event updates 830, and the like. For example, communications subsystem 824 may be configured
30 to receive (or send) data feeds 826 in real-time from users of social media networks and/or other

communication services such as Twitter[®] feeds, Facebook[®] updates, web feeds such as Rich Site Summary (RSS) feeds, and/or real-time updates from one or more third party information sources.

[0181] In certain examples, communications subsystem 824 may be configured to receive data in the form of continuous data streams, which may include event streams 828 of real-time events and/or event updates 830, that may be continuous or unbounded in nature with no explicit end. Examples of applications that generate continuous data may include, for example, sensor data applications, financial tickers, network performance measuring tools (e.g., network monitoring and traffic management applications), clickstream analysis tools, automobile traffic monitoring, and the like.

[0182] Communications subsystem 824 may also be configured to communicate data from computer system 800 to other computer systems or networks. The data may be communicated in various different forms such as structured and/or unstructured data feeds 826, event streams 828, event updates 830, and the like to one or more databases that may be in communication with one or more streaming data source computers coupled to computer system 800.

[0183] Computer system 800 may be one of various types, including a handheld portable device (e.g., an iPhone[®] cellular phone, an iPad[®] computing tablet, a PDA), a wearable device (e.g., a Google Glass[®] head mounted display), a personal computer, a workstation, a mainframe, a kiosk, a server rack, or any other data processing system. Due to the ever-changing nature of computers and networks, the description of computer system 800 depicted in FIG. 8 is intended only as a specific example. Many other configurations having more or fewer components than the system depicted in FIG. 8 are possible. Based on the disclosure and teachings provided herein, it should be appreciated there are other ways and/or methods to implement the various examples.

[0184] Although specific examples have been described, various modifications, alterations, alternative constructions, and equivalents are possible. Examples are not restricted to operation within certain specific data processing environments but are free to operate within a plurality of data processing environments. Additionally, although certain examples have been described using a particular series of transactions and steps, it should be apparent to those skilled in the art that this is not intended to be limiting. Although some flowcharts describe operations as a

sequential process, many of the operations may be performed in parallel or concurrently. In addition, the order of the operations may be rearranged. A process may have additional steps not included in the figure. Various features and aspects of the above-described examples may be used individually or jointly.

5 **[0185]** Further, while certain examples have been described using a particular combination of hardware and software, it should be recognized that other combinations of hardware and software are also possible. Certain examples may be implemented only in hardware, or only in software, or using combinations thereof. The various processes described herein may be implemented on the same processor or different processors in any combination.

10 **[0186]** Where devices, systems, components or modules are described as being configured to perform certain operations or functions, such configuration may be accomplished, for example, by designing electronic circuits to perform the operation, by programming programmable electronic circuits (such as microprocessors) to perform the operation such as by executing computer instructions or code, or processors or cores programmed to execute code or instructions
15 stored on a non-transitory memory medium, or any combination thereof. Processes may communicate using a variety of techniques including but not limited to conventional techniques for inter-process communications, and different pairs of processes may use different techniques, or the same pair of processes may use different techniques at different times.

[0187] Specific details are given in this disclosure to provide a thorough understanding of the
20 examples. However, examples may be practiced without these specific details. For example, well-known circuits, processes, algorithms, structures, and techniques have been shown without unnecessary detail in order to avoid obscuring the examples. This description provides example examples only, and is not intended to limit the scope, applicability, or configuration of other examples. Rather, the preceding description of the examples will provide those skilled in the art
25 with an enabling description for implementing various examples. Various changes may be made in the function and arrangement of elements.

[0188] The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense. It will, however, be evident that additions, subtractions, deletions, and other modifications and changes may be made thereunto without departing from the broader
30 spirit and scope as set forth in the claims. Thus, although specific examples have been described,

these are not intended to be limiting. Various modifications and equivalents are within the scope of the following claims.

[0189] In the foregoing specification, aspects of the disclosure are described with reference to specific examples thereof, but those skilled in the art will recognize that the disclosure is not limited thereto. Various features and aspects of the above-described disclosure may be used individually or jointly. Further, examples may be utilized in any number of environments and applications beyond those described herein without departing from the broader spirit and scope of the specification. The specification and drawings are, accordingly, to be regarded as illustrative rather than restrictive.

[0190] In the foregoing description, for the purposes of illustration, methods were described in a particular order. It should be appreciated that in alternate examples, the methods may be performed in a different order than that described. It should also be appreciated that the methods described above may be performed by hardware components or may be embodied in sequences of machine-executable instructions, which may be used to cause a machine, such as a general-purpose or special-purpose processor or logic circuits programmed with the instructions to perform the methods. These machine-executable instructions may be stored on one or more machine readable mediums, such as CD-ROMs or other type of optical disks, floppy diskettes, ROMs, RAMs, EPROMs, EEPROMs, magnetic or optical cards, flash memory, or other types of machine-readable mediums suitable for storing electronic instructions. Alternatively, the methods may be performed by a combination of hardware and software.

[0191] Where components are described as being configured to perform certain operations, such configuration may be accomplished, for example, by designing electronic circuits or other hardware to perform the operation, by programming programmable electronic circuits (e.g., microprocessors, or other suitable electronic circuits) to perform the operation, or any combination thereof.

[0192] While illustrative examples of the application have been described in detail herein, it is to be understood that the inventive concepts may be otherwise variously embodied and employed, and that the appended claims are intended to be construed to include such variations, except as limited by the prior art.

WHAT IS CLAIMED IS:

1. A method, comprising:

receiving, at a chatbot system comprising a processor, at least one utterance comprising one or more words;

5 generating, by a transformer-based model of the chatbot system, a plurality of embeddings for the one or more words of the at least one utterance;

generating, by a first vectorizer of the chatbot system, at least one regular expression and gazetteer feature vector for the at least one utterance;

10 generating, by a second vectorizer of the chatbot system, at least one context tag distribution feature vector for the at least one utterance;

concatenating or interpolating the plurality of embeddings with the at least one regular expression and gazetteer feature vector and the at least one context tag distribution feature vector to generate a first set of feature vectors;

15 generating, by a main sequence model of the chatbot system, an encoded form of the at least one utterance based on the first set of feature vectors;

generating, by a discriminative model of the chatbot system, a plurality of log-probabilities for candidate entities based on the encoded form of the at least one utterance; and

20 identifying, using the plurality of log-probabilities, one or more constraints for the at least one utterance based on the candidate entities.

2. The method of claim 1, wherein the least one utterance comprises at least one of one or more queries of the chatbot system, one or more queries input to the chatbot system by a user, one or more responses provided by the user in response to the one or more queries of the chatbot system, or combination thereof.

25 3. The method of claim 1, wherein the transformer-based model of the chatbot system comprises of a bidirectional encoder representations from transformers model.

30 4. The method of claim 1, wherein the first vectorizer generates the at least one regular expression and gazetteer feature vector based on one or more regular expression patterns and one or more gazetteers.

5. The method of claim 1, wherein the second vectorizer generates the at least one context tag distribution feature vector based on a context of at least one of one or more queries of the chatbot system, one or more queries input to the chatbot system by a user, one or more
5 responses provided by the user in response to the one or more queries of the chatbot system, or a combination thereof.

6. The method of claim 1, wherein the main sequence model of the chatbot system comprises a combined convolutional neural network/bidirectional long short-term memory
10 model.

7. The method of claim 1, wherein the discriminative model of the chatbot system comprises a conditional random field model.

8. A chatbot system comprising:
one or more processors; and
a memory coupled to the one or more processors, the memory storing a plurality of instructions executable by the one or more processors, the plurality of instructions comprising instructions that when executed by the one or more processors cause the one or more processors
20 to:
receive, at the chatbot system, at least one utterance comprising one or more words;
generate, with a transformer-based model, a plurality of embeddings for the one or more words of the at least one utterance;
generate, with a first vectorizer, at least one regular expression and gazetteer feature
25 vector for the at least one utterance;
generate, with a second vectorizer, at least one context tag distribution feature vector for the at least one utterance;
concatenate or interpolate the plurality of embeddings with the at least one regular expression and gazetteer feature vector and the at least one context tag distribution feature vector
30 to generate a first set of feature vectors;

generate, with a main sequence model, an encoded form of the at least one utterance based on the first set of feature vectors;

generate, with a discriminative model, a plurality of log-probabilities for candidate entities based on the encoded form of the at least one utterance; and

5 identify, using the plurality of log-probabilities, one or more constraints for the at least one utterance based on the candidate entities.

9. The chatbot system of claim 8, wherein the least one utterance comprises at least one of one or more queries of the chatbot system, one or more queries input to the chatbot system by
10 a user, one or more responses provided by the user in response to the one or more queries of the chatbot system, or combination thereof.

10. The chatbot system of claim 8, wherein the transformer-based model comprises a bidirectional encoder representations from transformers model.

15

11. The chatbot system of claim 8, wherein the first vectorizer generates the at least one regular expression and gazetteer feature vector based on one or more regular expression patterns and one or more gazetteers.

20 12. The chatbot system of claim 8, wherein the second vectorizer generates the at least one context tag distribution feature vector based on a context of at least one of one or more queries of the chatbot system, one or more queries input to the chatbot system by a user, one or more responses provided by the user in response to the one or more queries of the chatbot system, or a combination thereof.

25

13. The chatbot system of claim 8, wherein the main sequence model comprises a combined convolutional neural network/bidirectional long short-term memory model.

30 14. The chatbot system of claim 8, wherein the discriminative model comprises a conditional random field model.

15. A non-transitory computer-readable memory storing a plurality of instructions executable by one or more processors, the plurality of instructions comprising instructions that when executed by the one or more processors cause the one or more processors to:

receive, at a chatbot system, at least one utterance comprising one or more words;

5 generate, with a transformer-based model of the chatbot system, a plurality of embeddings for the one or more words of the at least one utterance;

generate, with a first vectorizer of the chatbot system, at least one regular expression and gazetteer feature vector for the at least one utterance;

10 generate, with a second vectorizer of the chatbot system, at least one context tag distribution feature vector for the at least one utterance;

concatenate or interpolate the plurality of embeddings with the at least one regular expression and gazetteer feature vector and the at least one context tag distribution feature vector to generate a first set of feature vectors;

15 generate, with a main sequence model of the chatbot system, an encoded form of the at least one utterance based on the first set of feature vectors;

generate, with a discriminative model of the chatbot system, a plurality of log-probabilities for candidate entities based on the encoded form of the at least one utterance; and

20 identify, using the plurality of log-probabilities, one or more constraints for the at least one utterance based on the candidate entities.

16. The non-transitory computer-readable memory of claim 15, wherein the transformer-based model of the chatbot system comprises a bidirectional encoder representations from transformers model.

25 17. The non-transitory computer-readable memory of claim 15, wherein the first vectorizer of the chatbot system generates the at least one regular expression and gazetteer feature vector based on one or more regular expression patterns and one or more gazetteers.

30 18. The non-transitory computer-readable memory of claim 15, wherein the second vectorizer of the chatbot system generates the at least one context tag distribution feature vector based on a context of at least one of one or more queries of the chatbot system, one or more

queries input to the chatbot system by a user, one or more responses provided by the user in response to the one or more queries of the chatbot system, or a combination thereof.

19. The non-transitory computer-readable memory of claim 15, wherein the main
5 sequence model of the chatbot system comprises a combined convolutional neural
network/bidirectional long short-term memory model.

20. The non-transitory computer-readable memory of claim 15, wherein the
discriminative model of the chatbot system comprises a conditional random field model.
10

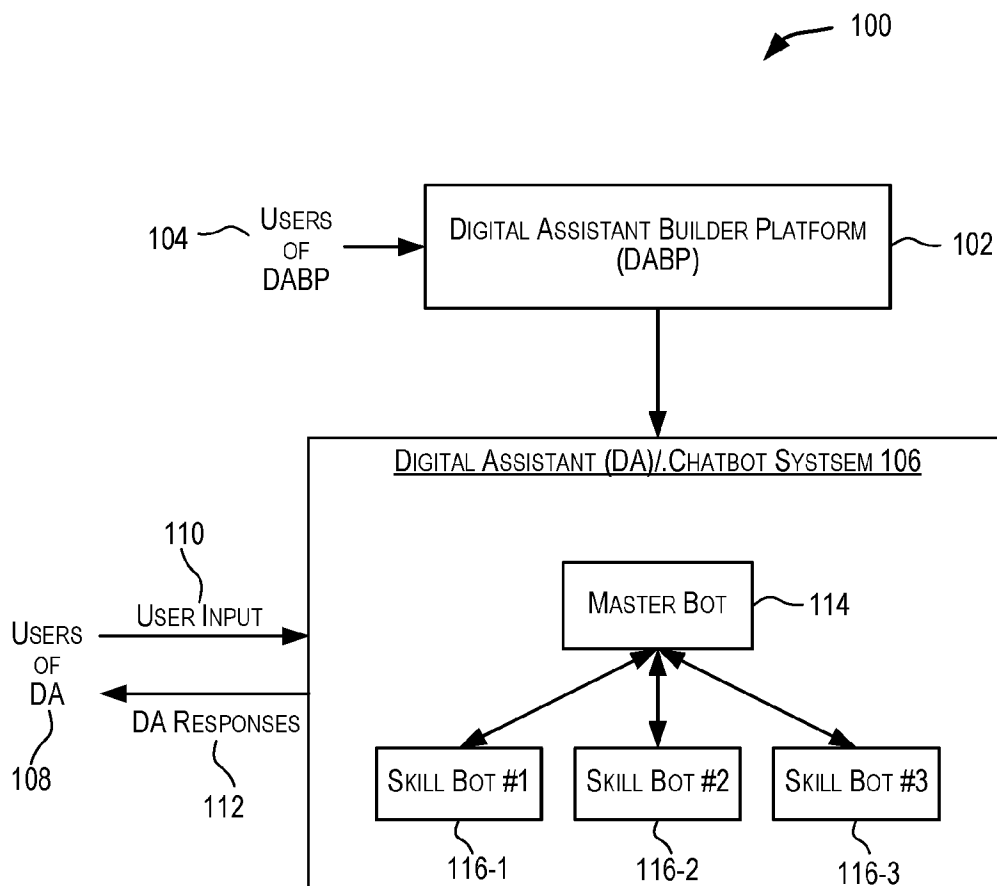


FIG. 1

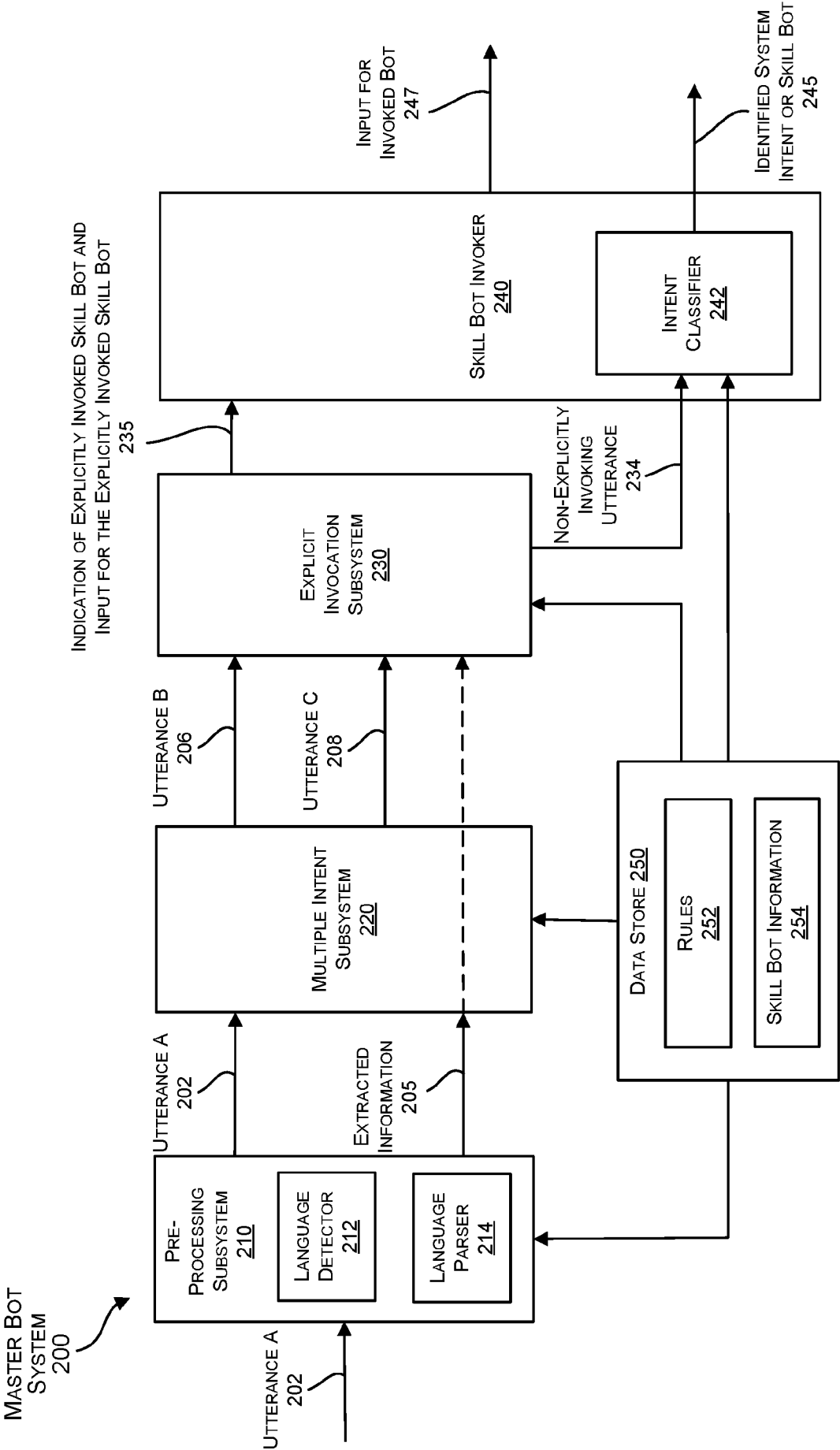


FIG. 2

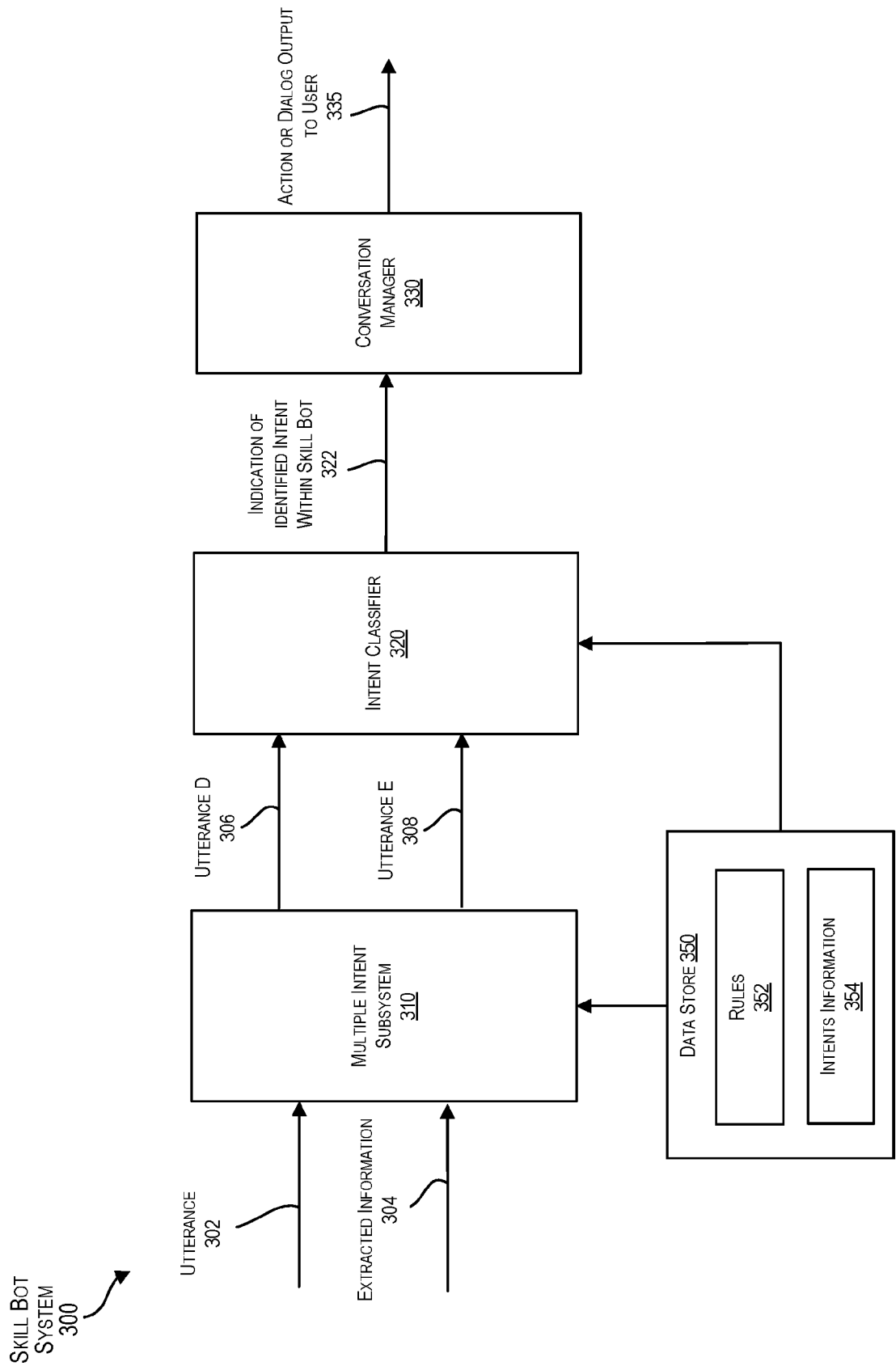


FIG. 3

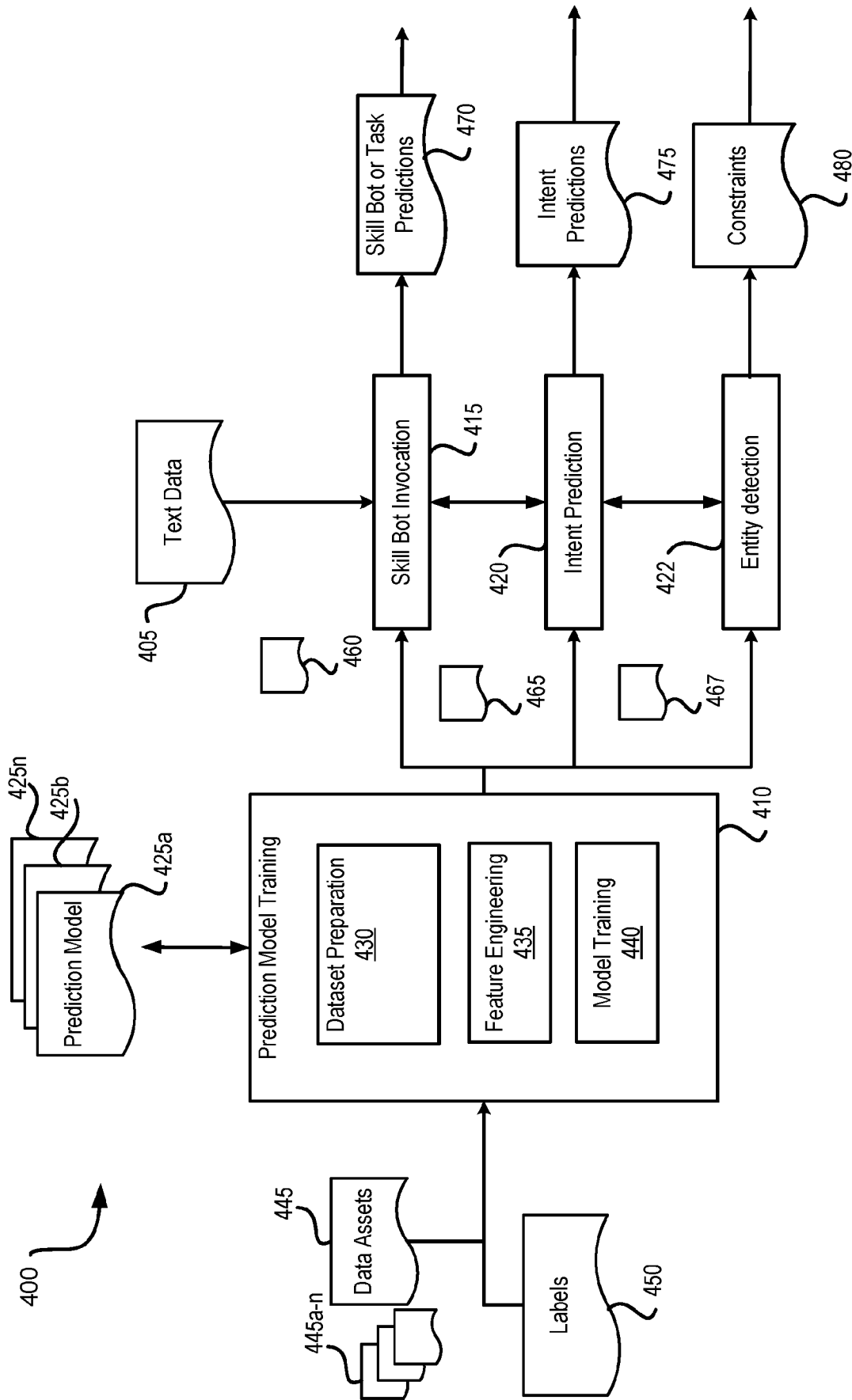


FIG. 4A

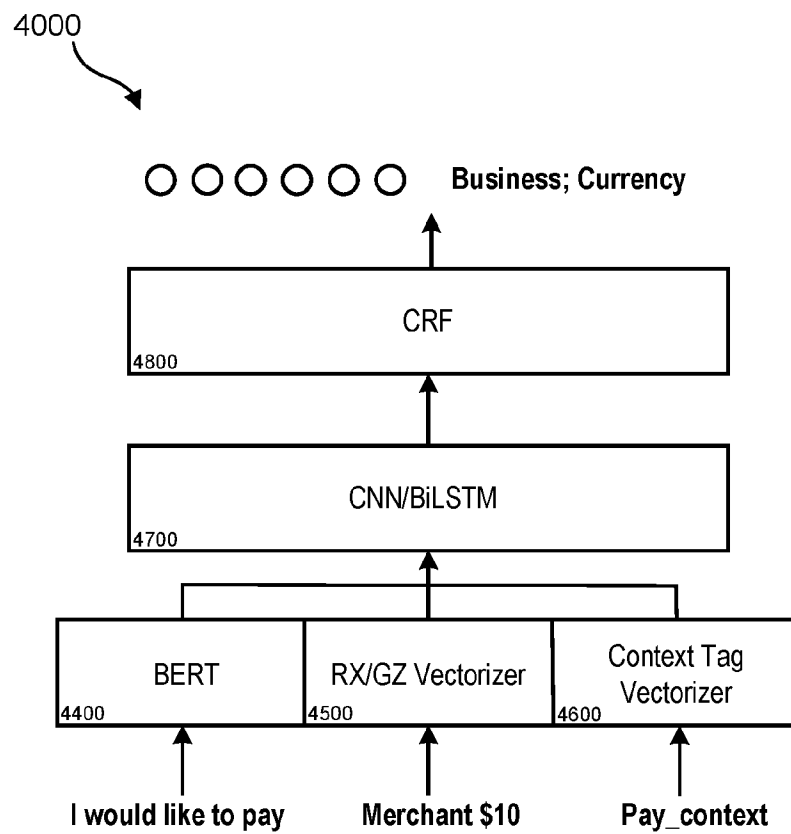


FIG. 4B

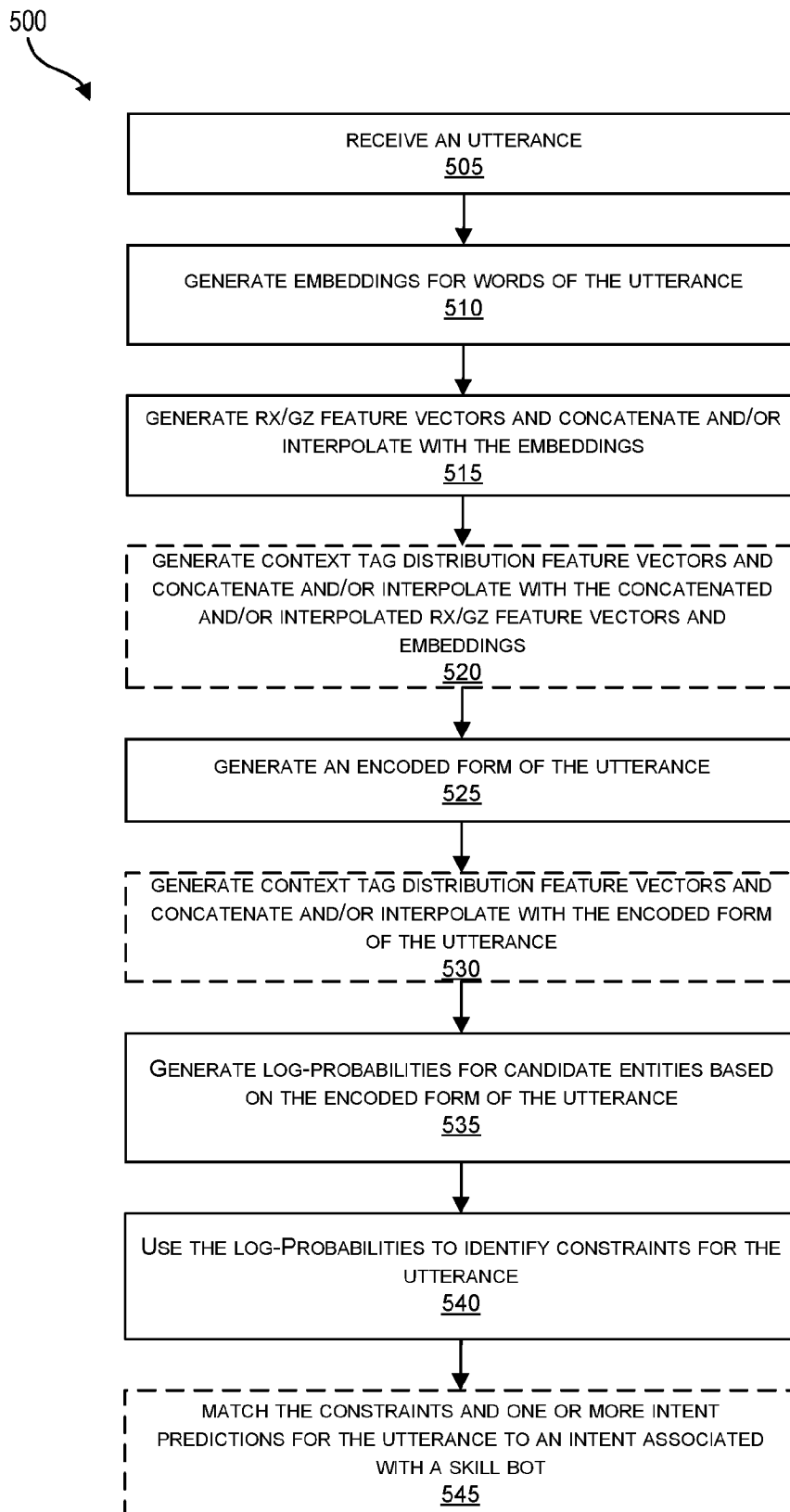


FIG. 5

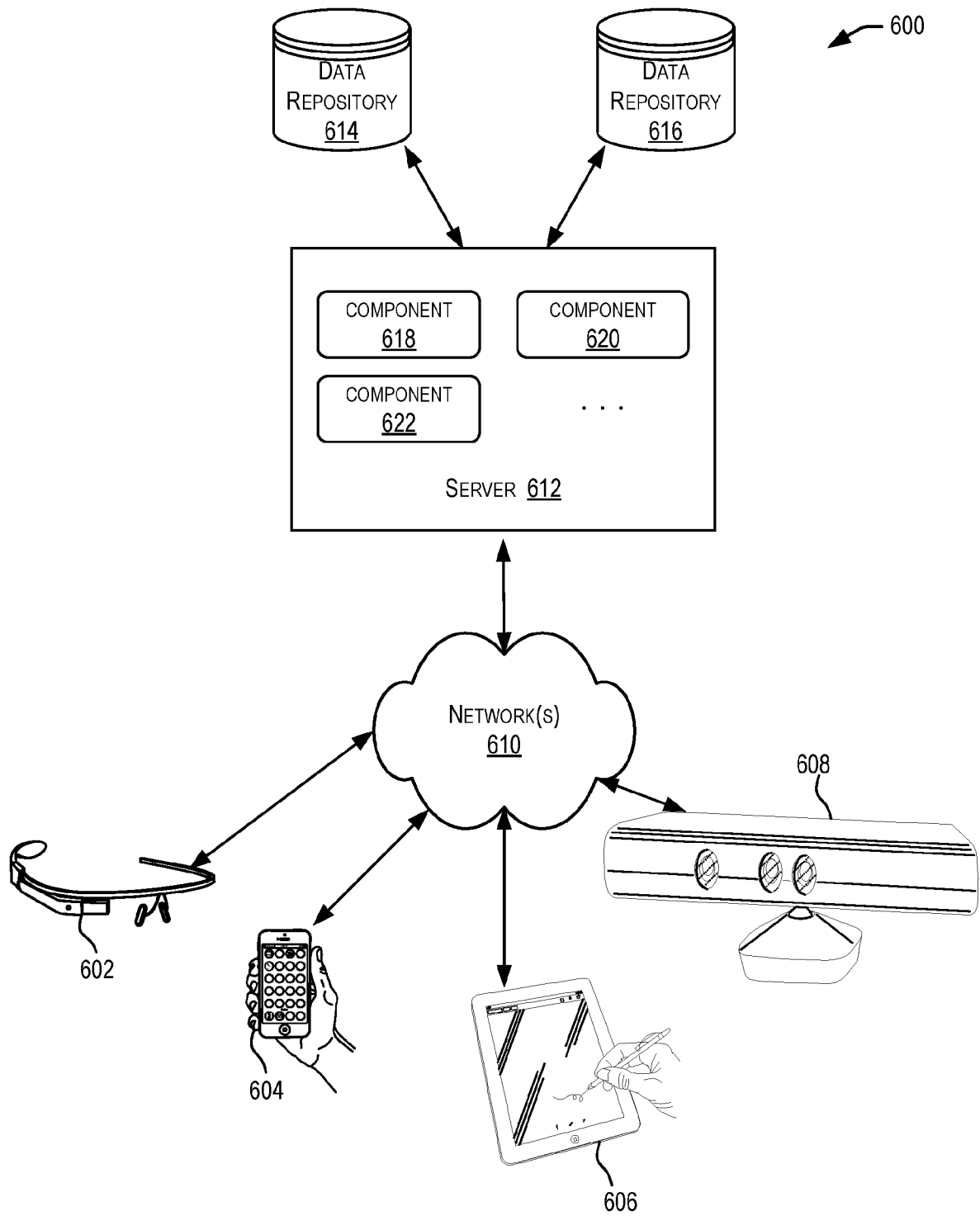


FIG. 6

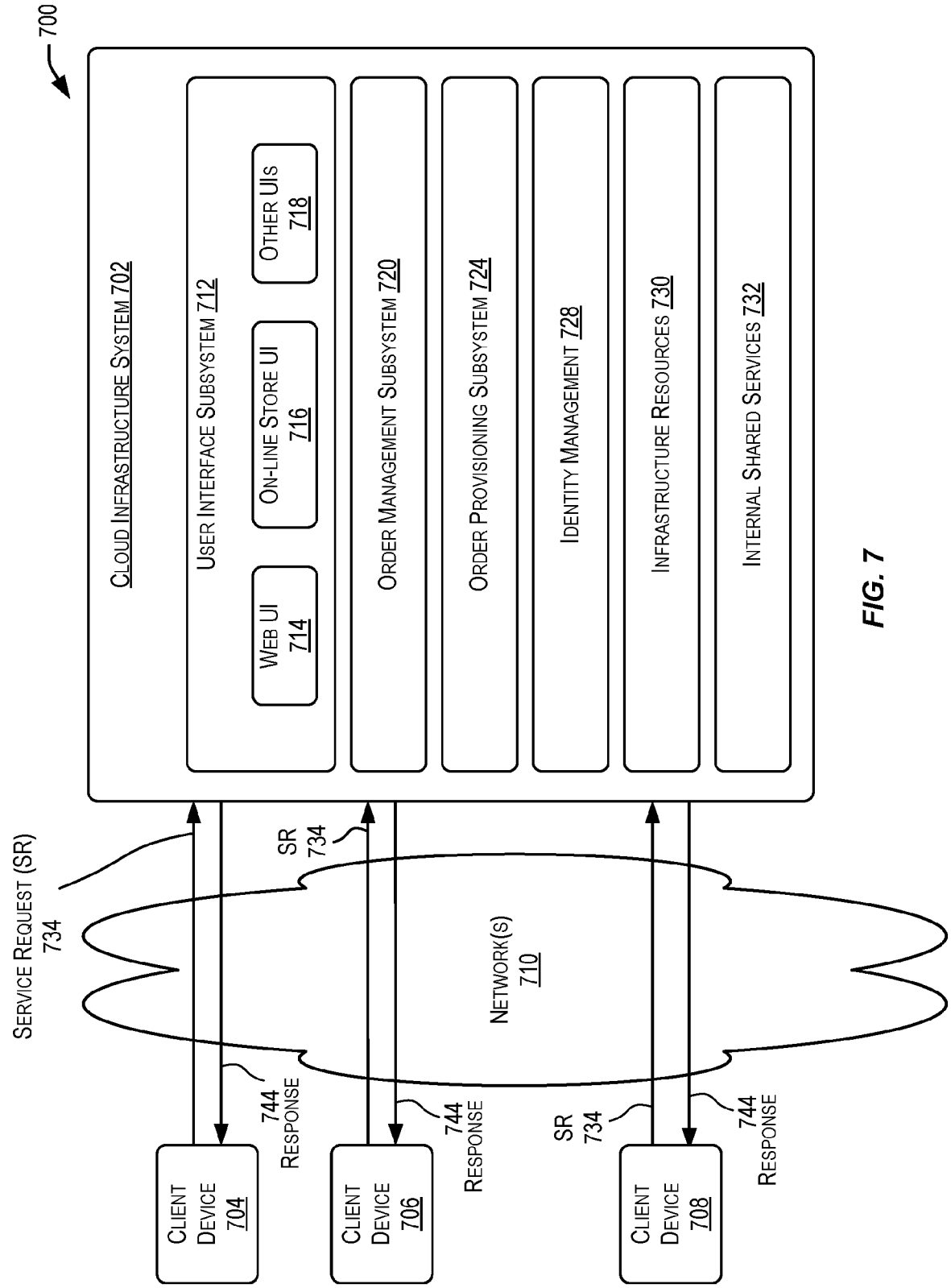


FIG. 7

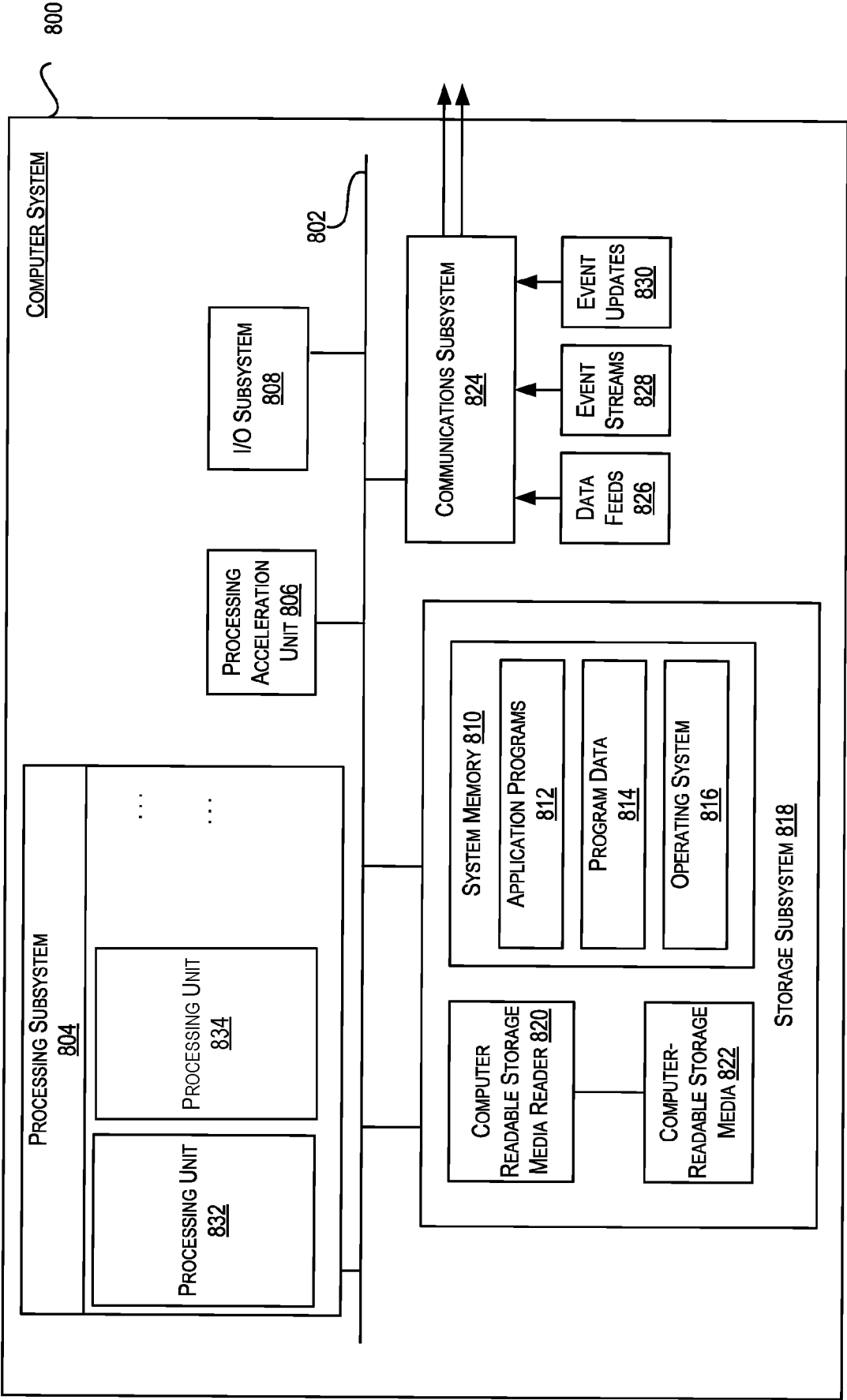


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2022/012972

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/332(2019.01); G06F 16/33(2019.01); G06N 20/00(2019.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 16/332(2019.01); G06F 17/30(2006.01); G10L 15/22(2006.01); H04L 12/58(2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: chatbot, feature vector, gazetteer, utterance, candidate, context, tag

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 10860629 B1 (AMAZON TECHNOLOGIES, INC.) 08 December 2020 (2020-12-08) column 3, lines 56-58; column 8, lines 55-61; claims 1-12; and figures 1-6	1-20
A	US 10224030 B1 (AMAZON TECHNOLOGIES, INC.) 05 March 2019 (2019-03-05) claims 1-4; and figures 4-8	1-20
A	CHUHAN WU et al., 'Detecting Entities of Works for Chinese Chatbot', ACM Transactions on Asian and Low-Resource Language Information Processing, November 2020, Volume 19, Issue 6, Article 88, pp. 1-13 [retrieved on 2022.04.18]. Retrieved from <https://dl.acm.org/doi/abs/10.1145/3414901> pages 1-11	1-20
A	SIMONE MAGNOLINI et al., 'How to Use Gazetteers for Entity Recognition with Neural Models', In : International Joint Conference on Artificial Intelligence (IJCAI), Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5), 12 August 2019, pp. 24-33 [retrieved on 2022.04.18]. Retrieved from <https://aclanthology.org/W19-58.pdf> pages 24-32	1-20
A	US 2019-0205463 A1 (FRESHWORKS INC.) 04 July 2019 (2019-07-04) claims 1-10; and figures 1-2	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

29 April 2022

Date of mailing of the international search report

03 May 2022

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon
35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

YANG, JEONG ROK

Telephone No. +82-42-481-5709

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US2022/012972

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
US	10860629	B1	08 December 2020	None		
US	10224030	B1	05 March 2019	US	11189277 B2	30 November 2021
				US	2019-0318737 A1	17 October 2019
US	2019-0205463	A1	04 July 2019	US	11086913 B2	10 August 2021