sensors

MDPI

Review

Conversational Agents: Goals, Technologies, Vision, and Challenges

Merav Allouch [1], Amos Azaria [1] and Rina Azoulay[2,*]

[1] Computer Science Department, Ariel University, Ariel, 40700, Israel; dmeravall@gmail.com (M.A.); amos.azaria@ariel.ac.il (A.A.)
[2] Department of Computer Science, Jerusalem College of Technology , Jerusalem, 9116001, Israel
* Correspondence: azrina@g.jct.ac.il

Abstract: In recent years, conversational agents (CAs) have become ubiquitous and are a presence in our daily routines. It seems that the technology has finally ripened to advance the use of CAs in various domains, including commercial, healthcare, educational, political, industrial, and personal domains. In this study, the main areas in which CAs are successful are described along with the main technologies that enable the creation of CAs. Capable of conducting ongoing communication with humans, CAs are encountered in natural-language processing, deep learning, and technologies that integrate emotional aspects. The technologies used for the evaluation of CAs and publicly available datasets are outlined. In addition, several areas for future research are identified to address moral and security issues, given the current state of CA-related technological developments. The uniqueness of our review is that an overview of the concepts and building blocks of CAs is provided, and CAs are categorized according to their abilities and main application domains. In addition, the primary tools and datasets that may be useful for the development and evaluation of CAs of different categories are described. Finally, some thoughts and directions for future research are provided, and domains that may benefit from conversational agents are introduced.

Keywords: smart environments; human–agent interaction; conversational agents

Citation: . *Sensors* **2021**, *1*, 0. https://doi.org/

Academic Editor: Carina Soledad González González

Received: 19 November 2021
Accepted: 10 December 2021
Published:

1. Introduction

Conversational agents (CA) are agents that interact with users via written or spoken natural language. CAs accept as input natural language as speech, text, or video; in addition, they may receive input from several different sensors. CAs are required to process the input and provide relevant advice or feedback in a form of text or speech or by manipulating a physical or a virtual body. Some CAs are capable of taking specific actions either in the real world or in the virtual world. Most CAs use natural-language processing to understand and generate speech, and some may also have engagement and personalization abilities. The rapidly growing abilities introduced by modern machine learning techniques facilitate the development of CAs capable of carrying out meaningful conversations with humans, learning to generate better and more relevant responses, expanding their knowledge-base, and performing actions beneficial to their users.

Current technological development enables the increasing use of CAs in several domains, such as assistance agents in the educational domain and health system, customer support agents in the commercial domain, and influence bots in the political domain. Commercial CAs for personal use, such as Siri [1] of Apple, Meena [2] of Google, and Cortana [3] of Microsoft, are widely used around the world. The aim of our study was to outline the principles behind the development of CAs and to survey the main domains in which conversational agents are successfully used.

Several recent studies have been carried out over the last years on CAs and, in particular, on text-based CAs that are called chatbots (as defined in Section 2). Some

studies concentrate on the technologies behind the development of CAs, and other studies examine their impact on people, i.e., the way people interact with them and perceive them.

Several recent reviews survey CA development and usage, at times referring to them as chatbots. Adamopoulou and Moussiades [4] provide a historical perspective of the chatbot development process, present a complete chatbot-categorization system, and analyze the two main approaches in chatbot development: pattern matching and machine learning. They mention two limitations of the current generation chatbots in understanding and producing natural speech, and they also point out that today's technology aims to build chatbots that can learn to talk but that cannot learn to think.

In another study, Adamopoulou and Moussiades [5] present an overview of the evolution of the international community's interest in chatbots and discuss the motivations that drive the use of chatbots and their usefulness in a variety of areas. They clarify the technological concepts and classify them based on various criteria, such as the area of knowledge and the need they serve. Furthermore, they present the general architecture of modern chatbots while also mentioning the main platforms they were created for. In another study, Nuruzzaman et al. [6] present a survey on commonly used chatbots and the underlying techniques. They focus on response-generating chatbots. In this category, the various response models can be categorized into four groups: template-based, generative, retrieval-based, and search engines. They compare the 11 most-popular chatbot application systems and present the similarities, differences, and limitations. They conclude that despite recent technological advances, chatbots conversing in a human-like manner are still hard to achieve.

Another survey concentrating on the technologies used by CAs is that of Borah et al. [7]. They describe the overall architecture of CAs, concentrating on the machine learning layer and analyze the recent development of text-based CAs. Chen et al. [8] describe the technology behind CAs and dialogue systems in real-world applications and discuss the effect of recent advances in deep learning on CA development. They emphasize that "big data" available from conversations on social media can be useful in building data-driven, open-domain CAs capable of responding to nearly any query. They further state that deep learning technologies can be used to leverage the massive amount of data to advance CAs from different perspectives. Gao et al. [9] concentrate on deep learning based CAs. They group the conversational agents into three categories: question-answering agents, task-oriented dialogue agents, and chatbots. For each category, they present a review of state-of-the-art neural approaches, draw the connection between neural and traditional approaches, and discuss the progress that has been made and challenges still being faced using specific systems and models as case studies.

Diederich et al. [10] review 36 studies on CAs in information systems (IS). They classify the literature along five dimensions. Three dimensions are related to CAs: the mode of communication, the context, and embodiment; and the other two dimensions are related to IS: the theory type and the research method. Wolff et al. [11] define a set of criteria to categorize chatbot applications. They review 52 articles describing chatbots. Most of the articles focus on customer-support chatbots, e.g., chatbots used to acquire information on specific services or products. In this article, we provide an overview of the concepts and building blocks of CAs and categorized them according to their abilities as well as the main domains of application. We emphasize the challenges and issues related to CA development for each domain while describing the tools and datasets useful for the development and evaluation of CAs of different categories. Finally, we provide some thoughts and directions for future studies and introduce domains that may benefit from conversational agents. For each of the topics in this survey, we focus on studies from the recent five years, though we also include earlier seminal studies as well as classical evaluation methods. In addition, the datasets provided in Section 8 include any relevant dataset that we found and are not limited to recent datasets.

The remainder of this article is organized as follows. Section 2 provides the terms and concepts used in the domain of conversational agents and defines the terms used in this

study. Section 3 describes the design components of primary CA types. Sections 4 and 5 survey the main technologies used for conversational software development, including machine learning (ML) methods and advanced technologies that enhance emotional abilities. Section 6 surveys recent CA applications, including personal assistants, healthcare agents, e-learning agents, and customer-support chatbots. The second part of this review focuses on technological issues. Sections 7 and 8 review commonly used datasets for CA development and testing and the technologies used to evaluate CAs. Finally, Section 9 concludes by providing ideas and directions for future developments.

2. Related Definitions and Terms

Conversational agents are highly referenced in the literature by numerous sources, including research articles, industry documentations, and internet blogs. Unfortunately, there exist inconsistencies in the references with respect to several central concepts related to conversational agents. Therefore, the aim of this section is to improve clarity, by providing definitions for the main relevant concepts currently in use, such as conversational agents, dialogue systems, chatbots, and virtual assistants.

It was observed that there are two terms that are sometimes used interchangeably: the term *conversational agent* and the term *chatbot*. There have been several attempts to define the distinction between the two terms. According to Vishnoi's definition [12], chatbots are software components that are designed to respond to human statements with a specific set of predefined replies. However, conversational agents are more contextual than chatbots and use more-advanced technologies such as deep learning methods and natural language understanding (NLU).

According to Nuseibeh [13], conversational agents are all types of software programs that interpret and respond to statements made by users in natural language. Chatbots, according to this definition, are a type of CA designed to simulate conversations with human users. Other types of CAs are programs designed to perform a particular goal, such as vacation planning and booking. CAs of this type are called *goal-oriented conversational agents*.

Radziwill and Benton [14] define conversational agents as software systems that mimic interactions with real people. They define chatbots as CAs that are implemented using a text-based interface.

Hussain et al. [15] classify chatbots into two main categories: task-oriented chatbots and non-task-oriented chatbots. According to Hussain et al., task-oriented chatbots are designed to accomplish specific goals such as ordering a pizza, guiding a user on social media, etc. The non-task-oriented chatbots for entertainment converse with users in an open domain. Masche and Le [16] categorize conversational systems into chatbots and dialogue systems. According to their definition, chatbots are systems mainly based on pattern matching, while dialogue systems are based on theoretically motivated techniques that enable conversations. Nimavat and Champaneria [17] distinguish between four criteria that can be used to classify chatbots: the knowledge domain, the type of service provided, the chatbot goal, and the the response-generation method. They define conversational bots as bots that talk to the user like another human being, in an open domain. It is worth noting that due to the ambiguity in the related terms and definitions, and the lack of a commonly agreed upon standard on the meaning of chatbot, the Alexa prize competition, set up with the goal of furthering conversational AI, uses the term *socialbot* to describe the conversational agents. These agents are intended to interact on a range of open-domain conversational topics [18].

In this review, our own definition for CA is provided, which is built upon the definitions provided in previous studies. To properly define CA, the more general concept of dialogue systems is introduced first. A *dialogue system* is a human–computer interaction system that uses natural language to communicate with the user. A *conversational agent* is a dialogue system that can also understand and generate natural language content, using text, voice, or hand gestures, such as sign language. Thus, to be categorized as CA, the

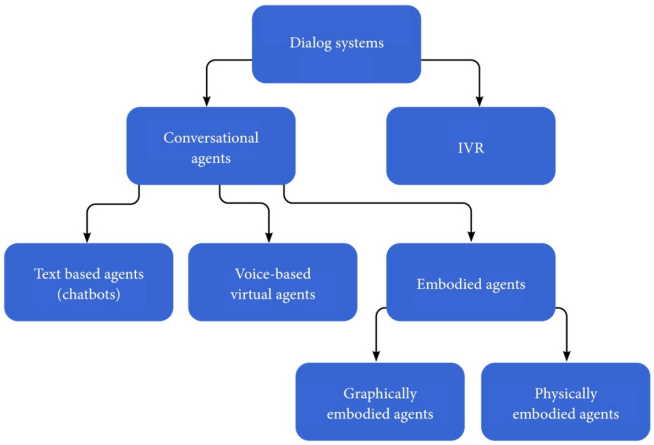

Figure 1. Conversational agents and chatbots: the definitions used in this article.

condition is, according to our definition, being able to understand and produce *sentences* in natural language. As a result, a CA is required to handle natural language that is not limited to a predetermined set of words (e.g., only numbers or a set of keywords) or a limited sentence structure.

The following examples cannot be considered CAs: (a) An interactive voice response (IVR) system in which the user is instructed to press a number on a keypad or say a specific word in order to advance to the next menu (e.g., "Press or Say 1 for English") is not considered a CA, since the user response does not include natural language *sentences*. (b) An embedded system in which a user provides voice commands (e.g., "Turn on the lights" or "Set the temperature to 25 degrees") and the system executes them without invoking any natural language response.

There are different criteria for categorizing CAs: the mode of communication, the action capabilities, and the domain/application in which the CA operates. First, our definition of conversational agents is refined according to the mode of communication between the CA and the human user. Here, a *chatbot* is defined as a CA that interacts with the user only by text and not by any other means of communication, for example, the ELIZA chatbot [19], or chatbots available on service platforms, such as banks, booking, and other e-commerce domains. Voice-based virtual agents are CAs that interact with the users by voice, for example, Siri, Google Now, Cortana, etc. Graphically embodied agents are virtual agents that have a virtual body as well as voice-understanding and speech-generation abilities. Their virtual body enables them to provide an additional means of communication through gestures. Finally, physical-based embodied agents are CAs that have a physical body, such as social robots, e.g., JIBO [20]. Both graphical and physical agents are called embodied CAs (ECAs). The above definitions are used throughout this article and are summarized in Figure 1.

CAs can also be classified according to their effector capabilities and actions. Communication-only agents merely communicate with a user and do not execute any action, e.g., ELIZA [19], Cleverbot [21,22] or CAs used only to answer questions. Other CAs, known as virtual or personal assistants, e.g., Alexa [23], are capable of executing physical or virtual actions, such as turning on an AC or booking a flight (see Figure 2).

Finally, CAs can be classified according to the application: (a) Open domain/general purpose CAs are mainly used to answer questions in various domains or in entertainment

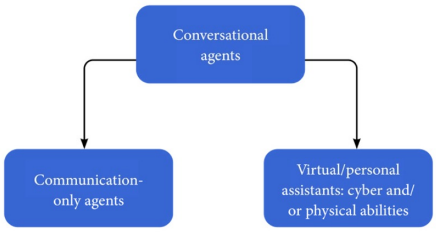

Figure 2. Conversational-agent classification according to action capabilities.

and are mostly communication-only agents. (b) Goal-oriented CAs assist users in completing tasks requiring multiple steps and decisions. Goal-oriented CAs are also task-oriented dialogue systems [24] and are referred to as taskbots according to the Alexa Prize competition [25]. These agents may be used both in the business domain or as personal assistants. In the business domain, they operate as customer-service and sales representatives. As personal support agents, they can assist the user in particular tasks, such as driving, vacation planning, or trip management. (c) Social-supporting agents can support patients in medical conditions or support students in the learning process. (d) Social-network bots, also known as influence agents, are intelligent CAs acting in social media to advertise a product or to influence opinions (see Figure 3). The rest of the article uses the terms defined in Figure 1 while considering various CA applications, as detailed in Figure 3. A detailed survey on CA usage in various domains is provided in Section 6.

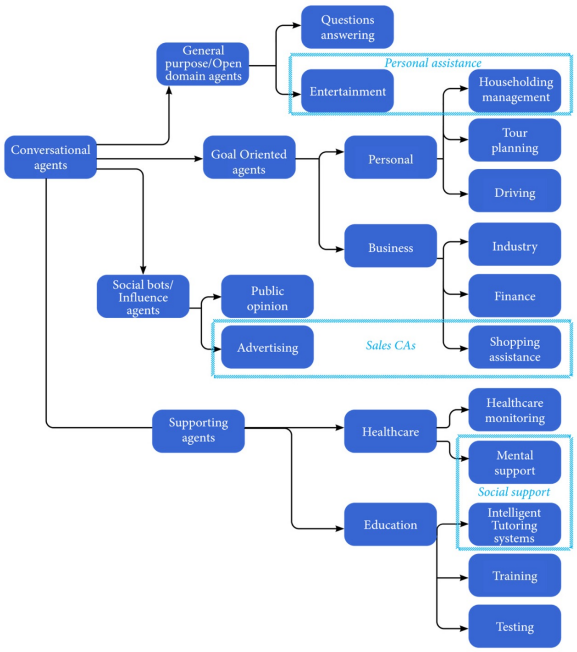

Figure 3. Conversational-agent applications.

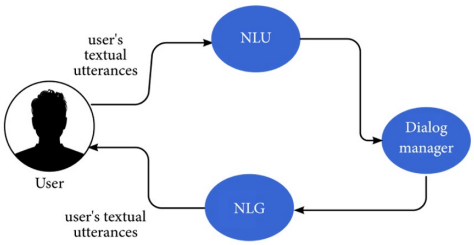

Figure 4. The textual components of CAs.

3. CA's Design Issues

This section describes the different components related to CA design. CA design is divided into four classes: text components for chatbots; CA components related to voice-based virtual agents; physical-related components for goal-oriented CAs or for embodied agents; and task-performance components for goal oriented CAs. For each of the four classes, the general goal is provided, the main components are detailed, and the relations between these components are described.

3.1. Text Related Components

The two main abilities required of CAs are the ability to logically understand the user's utterance and the ability to correctly reply to it. Overcoming these challenges require research in the fields of natural-language processing (NLP), information retrieval (IR), and machine learning (ML) [9].

Text-related components are used by most CAs, including embodied CAs and voice-based CAs, since voice-based virtual agents usually translate human speech to text, analyze the text, generate text responses, and then produce the speech signals. Therefore, in our design description, text-related components are discussed first.

CAs are commonly partitioned into components based on a pipeline determined by the order in which the component is used [26,27]. The most-common components are

- The natural-language-understanding (NLU) component: interprets the words into an internal computer language, called a logical form, which represents the meaning of the text.
- The dialogue manager component: receives the logical form and decides on how to respond. The dialogue manager may also include a module that assists with long-term conversations.
- The natural-language-generation (NLG) component: converts the answer into a text sequence in natural human language.

A schematic description of the textual processing components is provided in Figure 4.

Masche and Le [16] use a similar categorization, with an additional preprocessing component. They provide an alternative hierarchical approach to define text-related components by dividing the components into those responsible for text understanding, text processing, and text producing, as defined by Stoner et al. [28], as follows:

- Responder—the interface between the user and the CA: transfers and monitors the inputs and the outputs.
- Classifier—the interface between the responder and the graphmaster: normalizes and filters user inputs and processes the graphmaster output.
- Graphmaster—the brain behind the CA: manages the high-level algorithms.

According to this approach, the responder component includes parts from both NLU and NLG, while the dialogue manager component has parts from both the classifier and the graphmaster.

Abdul-Kader et al. [29] survey the techniques used to design CAs and describe the main techniques used by pattern-matching-based CAs, which are: (a) Parsing: manipulation of the input text using NLU functionality. (b) Pattern matching: analyzing user input and collecting relevant data, especially used by question-answering systems. (c) Chat script: used when no matches occur. (d) History database: used to enable the chatbot to remember previous conversations. (e) Markov Chain: enables probabilistic-based responses of chatbots.

Ramesh et al. [30] describe various approaches to design and build chatbots. Ahmad et al. [31] provide some examples of chatbots, describe their design, and provide a description of the most-popular techniques used by chatbot developers. Diederich et al [32] analyze 51 CA platforms to develop a taxonomy that would allow the identification of platform archetypes in CA design. The taxonomy consists of eleven dimensions and three archetypes, which can be used by practitioners in the design stages of CA. Lokman and Ameedeen [33] categorize modern chatbot design into the following elements: domain knowledge, response generation (retrieval or generative), text processing (vector embedding or Latin alphabet), and machine learning (ML) (mostly using neural networks). The various components described in this section enable the creation of CAs that are able to communicate with humans through an appropriate textual interface. In the next section, these technologies are also used for other types of CAs, such as voice-based CAs.

3.2. Voice-Related Components

Voice-based virtual agents are CAs that communicate with humans using speech. The process used by CAs usually includes: translating the sound waves into text, understanding the text, producing a text response for the user, and translating the text response to the sound produced by the computer or by the robot. The steps of understanding the text and producing an answer usually rely on the text-related components described above, but there are additional components, such as voice-based virtual agents related to audio analysis and audio production. A voice-based virtual agent may extract additional non-verbal information from the user audio, such as the user's emotional state, e.g., whether the user is being sarcastic, dramatic, decisive, or trying to deceive the system. Some works have also used non-verbal cues to detect whether a user is trying to correct previously made statements [34]. The components responsible for additional voice-based capabilities include:

- An automatic-speech-recognition (ASR) component (speech to text): converts the audio stream to a text representation.
- Non-verbal-information-extraction component: extracts relevant non-verbal information from the audio, such as observing the user's emotional state or understanding the urgency.
- Text-to-speech component: synthesizes the output waveform that is sent to the speakers.

The main components of the audio-process components are described in Figure 5.

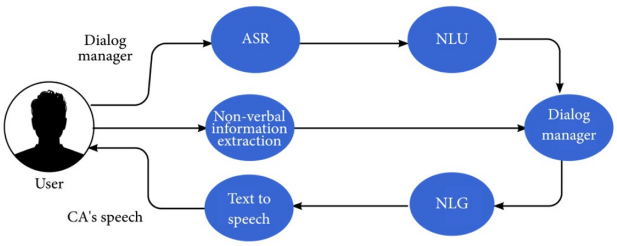

Figure 5. The main voice-based components of CAs.

Additional information on the capabilities and components of speech-based CAs is described by Saund [35]. Benzeguiba et al. [36] review ASR challenges and technologies, and Yu and Deng [37] provide a complete overview on modern ASR technologies with an emphasis on the deep-learning methods adopted in ASR.

3.3. Physical-Related Components

Physical embedded CAs, which obtain visual input from the user, benefit from the ability to understand physical-related gestures, such as body language and facial expressions. In addition, embodied CAs (ECAs) can use facial expressions and body gestures in their reactions.

Sign languages are complete languages that use only physical gestures to communicate. These languages may be used by CAs designed to communicate and/or tutor deaf users. Next, the main components in building an agent with these capabilities are described while referring the reader to articles reviewing this field.

Sadeghipour and Kopp [38] describe an overall model for cognitive processes of embodied perception and generation. According to them, the main components for physical agent–human communication are as follows:

- Perception component: receives visual movements and preprocesses them. The preprocessing pipeline consists of four submodules: (1) The body correspondence solver is responsible for performing required operations (such as rotation and scaling) on the observations. (2) The sensory memory receives the transformed positions and buffers them in chronological order. (3) The working memory holds a continuous trajectory for each hand through agent-centric space. (4) The segmenter submodule decomposes the received trajectory into movement segments called guiding strokes.
- The shared-knowledge component is responsible for the representation of motor knowledge. This component consists of a hierarchical structure, starting with the form of single-gesture performances in terms of movement trajectories and leading into less-contextualized motor levels and then toward more context. The motor-representation hierarchy consists of three levels: motor commands, motor programs, and motor schemas.
- The gesture-generator component is invoked by a prior decision to express an intention through a gesture. This component may also be used by a virtual agent that is built on a motor-control engine.

The main components of the physical-based, embodied CA are described in Figure 6. Krishnaswamy et al. [39]. provide a review on sign languages and gesture interpretation and generation. Homburg et al. [40] describe the process of sign-language (SL) translation, including SL recognition and SL generation. Singh et al. [41] detail the process of recogniz-

ing and interpreting the Indian sign language. Finally, Beck et al. [42] study the generation of emotional body language to be displayed by humanoid robots.

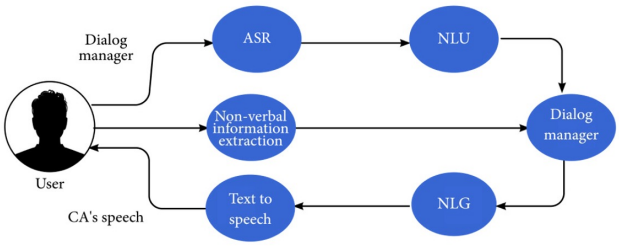

Figure 6. The main components of a physical-based embodied CA.

3.4. Task-Related Components

Goal-oriented CAs assist users in completing tasks requiring multiple steps and decisions, such as CAs booking vacations and planning trips. Goal-oriented CAs may use the text-related and voice-related components described above, in addition to task-related components. Task-related components are special components that handle task-related planning and learn challenges for the successful execution of the required goal. Previous studies on goal-oriented CAs [43,44] describe the processes followed by a conventional goal-oriented CA. This process includes the phases of text understanding, state estimation, dialogue policy, and text generation. The additional task-related components are defined as follows:

- State tracker: estimates the state of the user's goal by tracking the information across all turns of the dialogue.
- Policy manager: determines the next set of actions to help reach that goal. The policy manager uses the goal-related information from the state tracker and may communicate with the dialogue manager.
- Action manager: performs the required cyber actions (e.g., hotel reservations, food ordering, and flight booking) and/or the required physical actions to successfully fulfill the user requests.

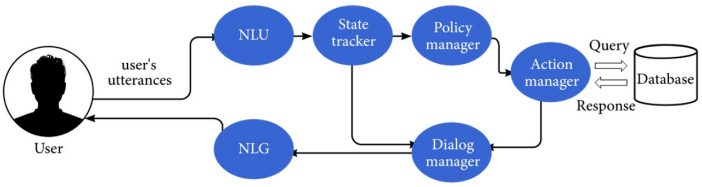

Figure 7. The main components of a goal-oriented CA.

The schematic description of the task-related components is provided in Figure 7, and an overview of the technologies behind goal oriented CAs is provided in Section 4.5.

4. Technologies Behind CA Components

In this section, the technologies behind the CA components presented in Section 3 are described in further detail, detailed examples are provided for the physical components, and the implementation of the technologies in recent CA systems are discussed.

4.1. Natural Language Understanding

Natural language understanding (NLU) typically refers to extracting structured semantic knowledge from text. NLU tasks mainly include tokenizing the text, normalizing it, recognizing the text entities, and performing dependency or constituency parsing. The traditional NLU stack is based on the following five components: phonology, morphology, syntax, semantics, and reasoning [45].

In particular, morphological analysis or parsing can be viewed as resolving natural-language ambiguity at different levels by mapping a natural language sentence to a series of human-defined, unambiguous, symbolic representations, such as part-of-speech (POS) tags, context-free grammar, and first-order predicate calculus. NLU includes the following sub areas: resolution, discourse analysis, machine translation, morphological segmentation, named-entity recognition, POS tagging, and more [27]. For a review on natural language understanding, the reader is referred to the survey of Navigli [46], in which several NLU approaches and modes are reviewed, including explicit versus implicit learning, representation of words and semantics, and a vision on what machines are expected to understand.

In the remainder of this section, the focus is on studies that use NLU for CA development. Initially, CAs using classical NLU technologies are described. Next, CAs using a parser as their NLU component are described. To conclude, recent CAs that use advanced technologies for NLU are described.

A classical approach for designing chatbots is the pattern-matching approach, in which the CA matches the user input with a pattern and chooses the most-suitable response stored in its predefined text corpus. One example of a CA that is based solely on simple pattern matching is ELIZA [19]. Over the years, several studies have developed additional rules and corpora to develop more-adaptive and advanced CAs. Inui et al. [47] use a linguistic corpus to design a CA interface. The dialogue corpus is based on a series of dialogues, and NLU is achieved by adopting corpus-based methods like the stochastic model, the n-gram model, keyword matching, and structural matching.

ALICE [48] is a chatbot based on AIML [49], an XML-based language designed to create chatbots based on pattern matching. ALICE won the Loebner Prize as "the most human computer" at the annual Turing Test contests of 2000, 2001, and 2004. ALICE answers the user's query by using its pattern-matching engine, which searches for a lexical correspondence between the user's query and the chatbot's patterns.

Agostaro et al. [50] outline the limitations of the pattern-matching approach. Pattern matching may fail to answer the user query when the query is composed of words that do not match any pattern. Therefore, when the query is grammatically incorrect, the pattern-matching mechanism will fail. To overcome these limitations, Agostaro et al. developed LSA-bot [50], which is a chatbot based on latent semantic analysis (LSA). LSA applies statistical computations to a large corpus of text to extract and represent the meaning of words. LSA-bot uses LSA to map its knowledge base into a conceptual space. The user input is mapped into the same conceptual space, allowing LSA-bot to find an appropriate response.

The informal response interactive system (IRIS) chatbot, developed by Banchs and Li [51], uses a large database of dialogues to provide candidate responses to a given user utterance. The IRIS response-selection process chooses the candidate utterances using two scores. The first score is determined by the cosine similarities between the current user input vector and all single utterances stored in the database. The second score is determined by the cosine similarity between the current vector dialogue and the dialogue

history of the user. The two scores are combined using a log-linear scheme. The IRIS randomly selects one of the top-ranked utterances as its response.

A context-free-grammar (CFG) parser [52] is often used by CAs for NLU. A CFG parser builds a constituency parse tree from the given user utterance based on a grammar, which is composed of parsing rules. A more generalized CFG, which is more suitable for solving ambiguity, is the probabilistic CFG (PCFG) [53,54]. In a PCFG parser, each rule in the grammar is associated with some probability. A PCFG parser outputs the parse tree with the highest probability.

Azaria et al. [55] present LIA, an agent that uses a combinatory categorial grammar (CCG) parser as its NLU component. The parser maps the commands, which are given in natural language, to logical forms, which contain functions and concepts that can later be executed by the dialogue manager. CCGs benefit from being more expressive than CFGs as they can represent the long-range dependencies appearing in some sentences (e.g., relative clauses), which cannot be expressed using CFGs. Recent ML methods and word-embedding methods are widely adapted to achieve NLU components with higher performance. Rasa NLU and Rasa Core [56] are open-source Python libraries for building conversational software. Rasa NLU allows the use of a predefined pipeline for the NLU process.

Recent ML methods and word embedding methods are widely adapted for achieving NLU components with higher performance. Rasa NLU and Rasa Core [56] are open-source Python libraries for building conversational software.

Rasa NLU allows the use of a predefined pipline for the NLU process. Their recommended pipeline process starts by tokenizing the user input, followed by the conversion of each token to a GloVe embedding vector [57]. Then, a multiclass support vector machine (SVM) [58] is used for deciding which action to take. Custom entities are recognized using a conditional random field [59].

ConvLab-2 [24], which is an open-source toolkit for building goal oriented CAs, provides three NLU models: a semantic tuple classifier, a multi-intent language understanding model [60], and a fine-tuned BERT- [61] based NLU model with the ability of intent classification and slot tagging.

4.2. The Dialogue Manager

Given the input text, the next step in the CA's pipeline is to manage the dialogue with the user. The *dialogue-manager* component is responsible for two main tasks: *Dialogue modeling*: keeps track of the state of the dialogue and *Dialogue control*: decides on the next system action [62].

Harms et al. [63] review the state-of-the-art commercial and research tools available for CA dialogue management. They divide the management approaches into two types: handcrafted-rule-based approaches and probabilistic (data-driven) approaches. The handcrafted dialogue manager defines the state and the control of the system by a set of rules that are defined by developers and experts, while the probabilistic dialogue manager learns the rules from actual conversations.

The studies described next concentrate on dialogue managers, including handcraft-rule-based systems and probabilistic-based systems. Handcraft rule-based management systems may be based on a planning algorithm or a pattern-matching based approach. Nguyen and Wobcke [64] propose a planning-based approach for developing a personal-assistant CA. In their approach, the dialogue manager has a set of plans, which can be divided into four groups: conversational-act determination and domain-task classification, intention identification, task processing, and response generation.

CommandTalk is a spoken-language interface for a battlefield military simulator [65,66]. It manages the representation of linguistic context, interprets user utterances within that context, and plans system responses. The CommandTalk dialogue manager uses a dialogue stack, a recovery mechanism for the stack, reference mechanisms, as well as finite state machines.

The MindMeld Conversational AI platform [67] is a platform designed for building conversational assistants. It uses pattern-matching rules to determine the dialogue state, and, based on this state and the predefined business logic, the CA performs the required task (or response) related to this state.

The Bottery CA creation platform [68] consists of four components: a set of states, a blackboard-style memory, an optional set of global transitions to allow the agent to switch from state to state, and an optional grammar used by the agent to generate the final outputs of the CAs. The Bottery syntax can be simply expressed by using structured JSON and can be extended by using imperative JavaScript code. The Bottery conversation management is performed by a finite state machine, which is displayed as a graph.

We proceed by describing probabilistic-based dialogue-management schemes. Google DialogFlow [69] is a framework for composing CAs. The Google dialogue manager considers the intent or motivation extracted from the user conversation to determine the appropriate action. Another commercial CA framework is Microsoft LUIS [70], a cloud-based conversational AI service that uses ML to understand the conversation to extract relevant information. LUIS can assist developers, who are unfamiliar with ML methods, to create their own cloud-based ML models specific to the application domain. Henderson et al. [71] present a word-based approach to dialogue state tracking using recurrent neural networks (RNNs). The model is capable of generalizing to unseen dialogue states' hypotheses. For long-term effects of the conversation, dialogue managers consider the conversation as a Markov decision process (MDP) and choose their responses by using RL methods. Singh et al. [72] suggest using RL for goal-oriented dialogue management.

Li et al. [73] suggest applying DRL to model future rewards in CAs. The agent's reward is determined according to three useful properties: informativity (non-repetitive turns), coherence, and ease of answering. The dialogue manager of the ensemble-based CA developed by Serban et al. [74] for the Amazon Alexa Prize competition utilizes an ensemble of NLG and retrieval models, including template-based models, bag-of-words models, sequence-to-sequence (seq2seq) neural networks, and latent-variable neural=network models. Their dialogue manager is trained to select an appropriate response by applying RL. The training was carried out on crowdsourced data as well as on real-world-user-interactions data.

4.3. Natural Language Generation

The NLG component translates the CA's representation of the response to natural language. NLG is defined by Reiter and Dale [75] as a subfield of AI and computational linguistics that is concerned with producing understandable texts in some human language from some underlying non-linguistic representation of information. Gatt and Krahmer [76] provide a recent survey on state-of-the-art NLG research, focusing on data-to-text generation. They discuss NLG architectures and approaches and highlight several new developments. In addition, they review the challenges of NLG evaluation and show the relationships between different evaluation methods.

NLG can be performed by template-based systems, which map the non-linguistic input directly to the linguistic surface structure without intermediate representations. Van Dimter et al. [77] describe several template-based systems and compare them to other NLG systems in terms of their potential for performing NLG tasks. They claim that template-based systems can, in principle, perform all NLG tasks in a linguistically well-founded way.

Several recent CAs use deep neural networks (DNNs) to perform the natural language-generation task. Wen et al. [78] present a statistical language generator based on a semantically controlled long-short-term-memory (LSTM) structure. The LSTM generator is trained on unaligned data by jointly optimizing sentence planning and surface realization. Variations in natural-language output are obtained by randomly sampling the network output.

Tran et al. [79] present a semantic component, called an aggregator, which can be integrated into an existing RNN encoder–decoder architecture, to improve NLG performance. The proposed component consists of an aligner and a refiner. The aligner is a component that computes the attention over the encoded input information, while the refiner is a gating mechanism stacked over the attentive aligner to further select and aggregate the semantic elements.

Jeraska et al. [80] focus on language-generation models with inputs structured for meaning representation to describe a single dialogue act with a list of key concepts that need to be conveyed to the user. They present a neural ensemble encoder–decoder model for generating natural utterances from the meaning representations.

Dusek et al. [81] assess the capabilities of recent seq2seq data-driven NLG systems, which can be trained on pairs of sequences, without the need for fine-grained semantic alignments. These pairs of sequences are composed of meaning representations, which are the output of the dialogue manager and the corresponding natural-language texts. They find that seq2seq NLG systems generally score high in terms of word-overlap metrics and human evaluations of naturalness but often fail to correctly express a given meaning or representation if they lack a strong semantic-control mechanism during decoding. Moreover, they can be outperformed by hand-engineered systems in terms of the quality, complexity, and diversity of outputs.

4.4. End to End Models

A popular end-to-end technique used by CAs is based on sequence-to-sequence learning models. These models convert sequences from one domain into sequences in another domain. Sequence-to-sequence models are widely used in different domains, such as machine translation, text summarization, speech to text conversion, image-caption generation, and automated answer generation.

Sordoni et al. [82] present a sequence-to-sequence-based chatbot trained end-to-end on large quantities of unstructured Twitter conversations. A neural-network architecture was used to address sparsity issues that arise when integrating contextual information with classic statistical models, allowing the system to take into account previous dialogue utterances. They extended the recurrent-neural-network language model [83] and proposed a set of conditional language models in which past utterances are encoded in a continuous context vector to help generate the response.

Li et al. [84] propose a method for defining the sequence-to-sequence objective function. They proposed using MMI, a measurement of the mutual dependence between inputs and outputs, as the objective function for the generated conversational responses. They also present practical strategies for neural generation models that use MMI as the objective function. The experimental results demonstrate that the proposed MMI models produce more diverse, interesting, and appropriate responses, yielding substantial gains in BLEU scores and in human evaluations.

Serban et al. [85] investigate the task of building open-domain CAs based on large dialogue corpora using generative models. Generative models produce responses that are generated word-by-word, opening the possibility for realistic, flexible interactions. In their model, a dialogue is considered as a sequence of utterances that, in turn, are sequences of tokens. They extend the hierarchical recurrent encoder–decoder (HRED) neural network to the dialogue domain. Their experiments demonstrate that the hierarchical recurrent-neural-network generative model outperforms both n-gram-based models and baseline neural-network models in the task of modeling utterances and speech acts. In addition, they show that the performance of their system can be improved by bootstrapping the learning from a larger question–answer pair corpus and from pretrained word embeddings.

Some studies concentrate on seq2seq learning for question-answering chatbots. He et al. [86] suggest a model based on sequence-to-sequence learning for a question-answering chatbot, which can answer complex questions in a natural manner. The model incorporates copying and retrieving mechanisms in a bi-directional RNN. The semantic units in the

answers are dynamically predicted from the vocabulary, copied from the given question, and/or retrieved from the corresponding knowledge base.

Qiu et al. [87] present a hybrid open-domain question-and-answer chatbot that combines information retrieval and seq2seq models. Information retrieval methods are used to retrieve a set of question/answer pairs based on a chat log of an online customer service. Then, the seq2seq model is used to rank the candidate answers. If the score of the top candidate answer is above a predefined threshold, it is considered to be the answer; otherwise, the answer is generated by the seq2seq model. Similarly, Ghazvininejad at el. [88] present a general data-driven and knowledge-grounded CA. They condition the CA responses not only on the conversation history but also on external facts through multi-task learning. This makes the CA versatile and applicable to an open-domain setting.

End-to-end models can also be useful in goal-oriented CA developments. Ham et al. [89] describe the use of end-to-end models for goal-oriented CAs, which need to integrate external systems to provide an explanation for the particular responses. They present an end-to-end monolithic neural model that learns to follow the core steps in the dialogue-management pipeline. The model outputs all the intermediate results in the dialogue-management pipeline to enable integration with the external system and to interpret why the system generates a particular response.

Kim [90] presents an end-to-end document-grounded, goal-oriented CA that utilizes a pretrained language model with an encoder–decoder structure. The encoder solves both the knowledge-seeking turn-detection task and the knowledge-selection task; the decoder solves the response-generation task.

Das et al. [91] suggest using DRL to learn the policies of goal-oriented CAs to answer visual questions. They pose a cooperative dialogue between two CAs communicating by natural language. The dialogue involves two collaborative CAs; one CA sees the image; and the second CA asks the first one questions about the image. DRL is used for learning the policies of these agents during the multi-round dialogue. As a result, the two trained CAs invent their own communication protocol without any human supervision.

4.5. Technologies Specific to Goal-Oriented CAs

In the development of goal-oriented CAs, there are additional challenges due to the need to combine both the dialogue handling and the task-performance management. Several ML-based technologies are commonly used to handle these challenges.

Zhang et al. [92] review the recent advances in goal-oriented CAs and discuss three critical topics: data efficiency, multi-turn dynamics, and knowledge integration. They also review the recent progress on task-oriented dialogue evaluation and widely used corpora, and they conclude by discussing some future trends for task-oriented CAs.

Zhao and Eskenazi [43] discuss the limitations of the conventional goal-oriented CA pipeline and suggest an alternative end-to-end task-oriented dialogue-management framework. In their framework, the state tracker is an LSTM-based classifier that inputs a dialogue history and predicts the slot-value of the latest question. The policy manager is implemented by a deep recurrent Q-network (DRQN) that controls the next verbal action. This framework enables the creation of a CA, which can interface with a relational database and learn policies for both language understanding and dialogue strategies.

Noroozi et al. [44] present a fast-schema-guided tracker (FastSGT), which is a BERT-based model for state tracking in goal-oriented CAs. FastSGT enables switching between services and accepting the values offered by the system during the dialogue. Finally, an attention-based projection is suggested to better model the encoded utterances.

Kim et al. [93] propose a two-step ANN-based dialogue-state tracker, which is composed of an informativeness classifier and a neural tracker. The informative CNN-based classifier filters out non-informative utterances, and the neural tracker estimates dialogue states from the remaining informative utterances.

Mrksic et al. [94] consider the issue of developing a state tracker for goal-oriented CAs. They consider the difficulty of scaling the state tracker to large and complex dialogue

domains because of the dependency on large training sets. They propose a neural-belief-tracking (NBT) framework that uses pretrained word embeddings to learn the distribution of user contexts.

Su et al. [95] estimate the task success by inspecting the dialogue as it evolves, by utilizing RNNs and CNNs. Their experiments demonstrate that both RNNs and CNNs can accurately estimate when substantial training data are available, though RNNs are more robust when training data are limited. Many goal-oriented CAs are trained on available goal-oriented datasets (see Section 8.3 for more details on such datasets). Other goal-oriented CAs are trained on human users. While such training may yield richer dialogues, it is more expensive.

Liu and Lane [96] address the challenges of building a reliable user simulator to train a goal-oriented CA by simulating the dialogues between two agents. Initially, a basic conversational agent and a basic-user simulator are trained on dialogue corpora through supervised learning, and then their abilities are improved by allowing them to conduct task-oriented dialogues while iteratively improving the policies using DRL.

5. Human-Related Issues

In addition to the technical issues of natural language understanding and generation, good conversational agents should be aware of human characteristics, observe user emotions, provide empathy in their responses, and engage the user.

According to Clark et al. [97], humans perceive the communication with CA as a means to achieve functional goals. In their study, Clark et al. present the results of semi-structured interviews on how people view the conversation between humans and CAs. They found that several social features reported as crucial in human–human conversation, such as understanding and common ground, trust, active listenership, and humor, are not listed as required for human–CA conversations. CA conversations are described almost exclusively by transactional and utilitarian terms. However, this view of CAs is not satisfactory in domains that require the user to engage and form an emotional bond with the CA.

Yand et al. [98] argue that understanding users' affective experience is crucial to the design of compelling CAs. To elaborate on this claim, they surveyed 171 CA users of Google assistant and examined the affective responses in four major usage scenarios. In addition, they observed the factors that influence affective responses. They found that the overall experience of the user was positive, with the most salient emotion being interest.

Both pragmatic and hedonic qualities influence affective experience. The factors underlying the pragmatic quality are helpfulness, proactivity, fluidity, seamlessness, and responsiveness. The factors underlying the hedonic quality are comfort in human–machine conversation, the pride of using cutting-edge technology, fun during use, the perception of having a human-like assistant, a concern about privacy, and the fear of causing distraction. In the remainder of this section, several issues are discussed that can assist in establishing a deeper connection between the user and the CA during conversations. The focus is on the following aspects: emotional issues, CA personality, and adaptation to the taste and needs of the user.

5.1. Emotional Aspect of Conversations

Emotional understanding and empathy are important abilities for CAs acting in several social domains including healthcare, education, and customer support; however, these abilities are also useful to CAs, in general. Combining emotional awareness with technologies and methods for CAs requires multi-domain knowledge in psychology, artificial intelligence, sociology, and education research.

The challenge in enabling empathy and emotionally adjusted responses is twofold: first, the agent must be able to detect the emotional state of the human; second, it must be able to provide the proper emotional response.

The agent may be able to detect user emotions based on user utterances as well as voice and body language. Emotion detection (ED) is an important branch of sentiment analysis and deals with the extraction and analysis of emotions from text and from audio. Acheampong et al. [99] surveyed models, concepts, and approaches for text-based ED and listed the important datasets available for text-based ED. In addition, they discuss recent ED studies, their results, and their limitations. Allouch et al. [100] concentrate on the problem of emotionally insulting sentences recognized by a CA designed to assist the special needs children with their social interactions. They generated a dataset consisting of insulting and non-insulting sentences and compared the ability of different ML methods in detecting the insulting content. In a related study, Schlesinger et al. [101] focus on race-talk and hate speech. They describe technologies, theories, and experiences that enable the CA to handle race-talk and examine the generative connections between race, technology, conversation, and CAs. Drawing together technological-social interactions involved in race-talk and hate speech, they point out the need of developing generative solutions focusing on this issue.

The challenge of listening to the user and understanding the user's emotional feelings is considered in Sarder's [102] thesis work, which studies the issue of conversational-agent development for mental-health intervention. Sarder built an embodied conversational agent with three different levels of backchannel strategies and ran a within-subject study with a convenience sample of 24 participants. He showed that the emotional content recognized in the words of the user increases as the CA listening capabilities increase.

As stated above, the second challenge for a CA with emotional abilities is to provide the appropriate response given the user's emotional state. The ability to recognize the emotions and feelings of others and replying accordingly is known as empathy, which is a crucial socio-emotional behavior for smooth interpersonal interactions. Therefore, the second emotional challenge is to assimilate empathy into CAs.

Empathy can be verbal and non-verbal. Yalcin [103] suggests that embodied CAs should be equipped with real-time multimodal empathic-interaction capabilities. The empathic framework leverages three hierarchical levels of capabilities to model empathy for CAs. Following the theoretical background on empathic behavior in humans, the embodied CA can express empathy by using facial expressions; gaze, head, and body gestures; as well as verbal responses.

Tellols et al. [104] propose equipping the CA with sentient capacities, using ML technologies. They illustrate their proposal by embedding a virtual tutor in an educational application for children. Their CA has a unique personality, emotional understanding, and needs that the user has to meet. The CA's needs can be expressed by Maslow's hierarchy of needs [105]. Tellols et al. tested the two CA versions with 10–12 year-old students and found that the second version, equipped with ML capabilities, displays higher understanding capacity and yields a nearly 100% user satisfaction rate. Emotional effects, as well as properties of the speaking style, can be added to the CA to generate speech that is closer to human dialogue.

Chen et al. [106] proposed a conditional text-generative adversarial network (CTGAN), in which an emotion label is adopted as an input channel to specify the output text. To match the generated text data to the real scene, they designed an automated word-level replacement strategy such that after generating initial texts by CTGAN, they extract keywords from the training texts and replace them in the generated texts.

XiaoIce is a popular social CA, developed in 2014 by Microsoft. Zhou et al. [107] describe the design of XiaoIce as an AI companion with an emotional connection. The XiaoIce design includes the intelligence quotient (IQ), the emotional quotient (EQ), and a culturally sensitive personality. The IQ capacity is achieved by knowledge and memory modeling. The EQ capacity includes two key components: empathy and social skills. Both IQ and EQ are combined in a unique personality. The CA personality is defined as the characteristic set of behaviors, cognition, and emotional patterns that form an individual's distinctive character. XiaoIce's developers have designed different personas for XiaoIce to

suit the preferences and desires of users in different cultures and regions. By analyzing the XiaoIce online logs, Zhou et al. show that XiaoIce understands user intent, recognizes human feelings, generates appropriate responses, and is capable of establishing a long-term relationship.

Asghar et al. [108] propose three methods to incorporate emotional aspects into encoder–decoder neural-conversation models: affective word embeddings, augmenting affective objectives in the loss function, and incorporating a search for affective responses during text decoding. Affective word embedding, in 3D space, can be performed using a cognitive-engineering affective dictionary. Affective objectives can be augmented in the cross-entropy loss function to generate additional emotional responses. Finally, the CA can be guided to search for effective responses during decoding. Asghar et al. show that incorporating these emotional aspects improves the quality of the CA responses in terms of syntactic coherence, naturalness, and emotional appropriateness.

Zhou et al.[109] explain the range of challenges that exist in addressing the emotion factor in large-scale conversation generation. These include: (i) the difficulty of obtaining high-quality emotion-labeled data since emotion annotation is a subjective task, (ii) the need to balance grammar and emotion in expressions, and (iii) the challenge of embedding emotion information. To express emotion naturally and coherently in a sentence, they designed a seq2seq generation model equipped with new mechanisms for emotion-expression generation.

To summarize, considering that the user's emotional experience and engagement are of great importance in various social and health domains, several studies suggest methods to recognize user's emotional state to provide an appropriate empathic response. The emotional awareness of CAs can make the user more satisfied and can yield longer and meaningful human–CA conversations.

5.2. The Effect of CA Personality

Recent studies have observed that adding personality aspects and human-like characteristics to the conversation may strengthen the connection of the user with the CA. In particular, in the mental-health-care domain, such CAs can elicit higher engagement from humans during the therapeutic process.

Chavesa and Gerosa [110] surveyed 56 studies from various domains to understand how social characteristics in CAs benefit human–CA interactions. They defined eleven social characteristics: proactivity, conscientiousness, communicability, damage control, thoroughness, manners, moral agency, emotional intelligence, personalization, identity, and personality, further grouping them into three social categories: conversational intelligence, social intelligence, and personification. They showed that certain characteristics, such as moral agency and communicability are influenced by the domain, while others, such as manners and damage control, are more generally applicable. They further point out that social-science theories, such as the cooperative principle and mind-perception theories, can contribute to the design of CAs with social characteristics.

Zhang et al. [111] proposed endowing CAs with a profile of a configurable, yet persistent, persona to make them more engaging. This profile is encoded by multiple sentences of textual description. To train the CAs on personal topics, they present a new dialogue dataset consisting of 164,356 utterances between crowd workers who were asked to chat naturally to get to know each other during the conversation.

Inspired by the vision of human-like interactions of conversational agents, Volkel et al. [112] examine the important features of a CA's personality. They used various sources to examine the main adjectives used by CAs, including an online survey, an interaction task in the lab, and a text analysis of 30,000 online reviews of CAs. They aggregated the results into a set of 349 adjectives, which were rated by 744 people in an online survey. A factor analysis revealed that the commonly used big-five model for human personality [113] does not adequately describe the CA personality. As an initial step in developing a personality

model, Vokel et al. proposed an alternative set of main features to be applied to the design of CA personalities.

Feine et al. [114] observed the process of how a social cue evolves into a social signal and subsequently triggers a social reaction. Using the theory of interpersonal communication [115], they identified a taxonomy of social cues of ECAs and classified the social cues into four major categories and ten sub-categories. The four major categories were: verbal, visual, auditory, and invisible. They evaluated the mapping between the identified social cues and the categories using a card-sorting approach.

The effect of ECA personas and cues on user engagement was studied by Liao and He [116]. In their experiment, participants were randomly assigned to racial-mirroring ECAs, non-mirroring ECAs, or control groups. After interacting with the ECA, participants completed a survey assessing their perception and evaluation of the agent. Liao and He demonstrated that racial mirroring has a positive influence on the user's perceived interpersonal closeness with the agent; the participants interacting with mirroring ECAs reported a higher level of satisfaction, a higher desire to continue interacting with the agent, and predicted a closer future relationship. In addition, people were significantly more likely to select same-race agent personas when they were given an opportunity to customize the ECA.

Go and Sundar [117] tested the distinct and combined effects of three types of cues that potentially enhance the humanness of chat agents: human-like visual cues, the use of human names or identities, and the use of human language. For these three factors, the authors examined how interactions among these cues influence psychological, attitudinal, and behavioral outcomes. Their experimental results indicate that CA interactivity is an important factor in determining psychological, attitudinal, and behavioral outcomes, while the identity cue turns out to be a key factor in eliciting certain expectations regarding CA's performance in conversation. However, message interactivity can compensate for the impersonal CA nature.

A good open-domain CA should be able to seamlessly blend all its skills, including the ability to be engaging, knowledgeable, and empathetic into one conversational flow. Smith et al. [118] present a method for training a CA with blended skills and testing it. They show that existing single-skill tasks can effectively be combined to obtain a model that blends all skills into a single CA. To preclude unwanted biases when selecting the skill, fine-tuning was done on the blended data.

5.3. Personalized CAs and their Effect on Human Engagements

In addition to possessing empathy, persona, and knowledge, the ability of the CA to adapt itself to the user's taste and needs is also important in engaging the user.

The studies described in this section are related to personalized CAs that adapt themselves to particular users to increase user satisfaction. However, adaptation may come at the cost of a loss in user privacy, which, if observed by the user, may limit the user's spontaneity in conversation. The effect of users limiting their conversation, upon detecting that the CA is collecting private information to adapt, was reported by [119].

A psycholinguistic characteristic of young adults interacting with a CA is to discuss daily-scheduling concerns and stress levels. Ferland and Koutstaal performed a linguistic analysis that presents the slightly paradoxical effect of reduced user engagement when a conversational agent explicitly discloses information on its user model to the user. They conclude that overt user models may discourage users from self-disclosure and participation in an information-rich spontaneous conversation.

Nevertheless, in task-oriented domains as well as educational domains, adaptation to the user's abilities and skills may assist the CA to be more effective and may result in higher user satisfaction. Carfora et al. [120] envisage goal-oriented agents whose policies take into consideration the psychological features of the user to deliver personalized and more effective messages. They built a probabilistic predictor based on the theory of planned

behavior [121] and a psycho-social model of reference and implemented it by a dynamic Bayesian network.

The smart-learning environment may involve task assignments adapted to the learner's abilities [122], smart hints and feedbacks [123], smart guidance during the learning process [124], and personalized conversational agents who assist in the learning process [125].

In the healthcare domain, Mandy [126], a primary-care CA created to assist healthcare staff by automating the patient-intake process, provides personalized intake service to patients by understanding their symptom descriptions and generating corresponding questions during the intake interview.

Schuetzler et al. [127] focused on the effect of improving the social presence of CAs by enhancing their responsiveness and embodiment. Responsiveness is the ability of the agent to provide responses contingent on user messages, and embodiment is the visual representation of the agent. In particular, they examined the influence of CA responsiveness and embodiment on the answers people give in response to sensitive and non-sensitive questions. They found that CA responsiveness increases socially desirable responses to sensitive questions.

Figure 8 presents an overview of the human-related issues discussed in this section. Each challenge is associated with the appropriate CA component expected to assume the most responsibility for that challenge. Understanding the user's emotional state is mostly a challenge of the ASR, NLU, and perception components; the dialogue manager decides on how to provide an appropriate empathic response; the NLG, the gesture generator, and the text-to-speech components are responsible for generating empathy in verbal and non-verbal responses; the personality of the CA is expressed by the response generators including the text-generator, the speech-generator, and the gesture-generator components; and adaptation of the CA to the user's taste and needs is the responsibility of the dialogue manager.

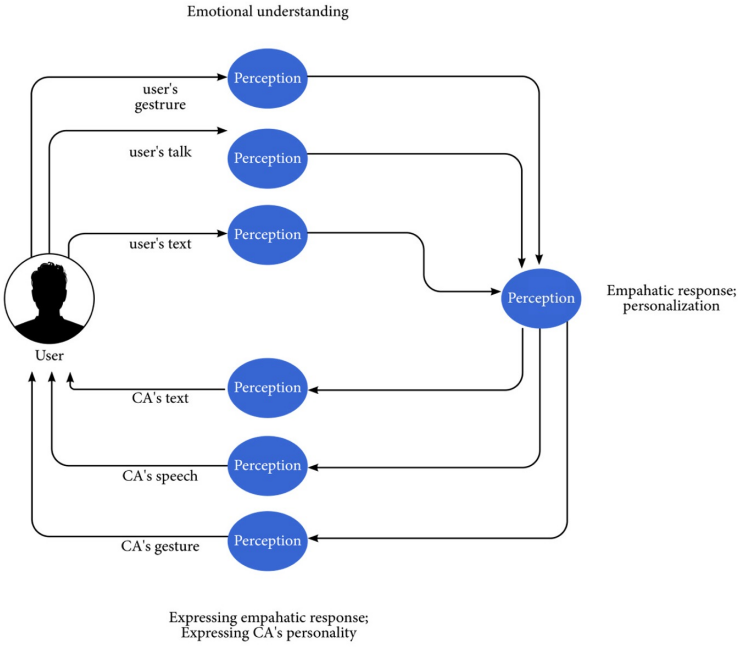

Figure 8. Human-related aspects of the CA: emotion sensitivity, personality expression, and adaptation to the user's taste and needs.

6. Goals and Applications of Conversational Agents

6.1. Personal Assistants and Open-Domain Conversational Agents

The first CA was developed in 1964 by Weizenbaum [19]. It was named ELIZA, and it simulated conversations by using a pattern-matching approach. ELIZA was designed to serve as a psychologist and mimicked certain kinds of natural-language conversation between humans and computers. People mistakenly believed ELIZA to be intelligent enough to comprehend a conversation, and some even became emotionally close to it. In 1972, the psychiatrist Kenneth Colby developed PARRY [128], which is a natural-language program that simulates the thinking of a paranoid individual. PARRY was developed to train users to detect people at psychological risk.

DeepProbe [129], RubyStar [130], and Meena [2] are recently developed open-domain chatbots. DeepProbe uses a sequence-to-sequence mechanism to satisfy user queries. RubyStar combines ML models and template- and rule-based responses; it uses topic detection, engagement monitoring, and context tracking. Meena CA is trained end-to-end on data mined and filtered from conversations on social media.

Currently, mobile devices and smart speakers are equipped with powerful agents such as Siri, Cortana, Alexa, and Google Assistant, offering support for a variety of tasks such as question answering, information retrieval, scheduling meetings, sending messages, and controlling smart home devices [10,131]. These assistants constantly listen to hear a wake-up keyword, for example, "Okay Google", "Alexa," etc. Once a wake-up keyword is said, the assistant records the user's command and sends it to a server. The server translates the voice command to text by using an ASR component that parses the text using a parser and uses a natural-language-understanding component to determine the appropriate response or action to be taken by the assistant. For example, a simple query "How are you today?" may be followed by an answer "I'm fine; thank you." A more-sophisticated question, such as "How many types of mammals are there?" may invoke a web-search that results in an answer such as "There are 6,000 different species of mammals." Commands requesting turning on the lights, setting the temperature of an air conditioner, playing a specific song, or ordering a product are executed accordingly.

Current virtual assistants have several drawbacks. First, they require a steady internet connection. Second, while they usually support multiple languages, they are far from supporting all languages used world-wide. In addition, virtual assistants that order products or book hotels and flights may cause unintentional expenses, e.g., when the user is a child. Misinterpretation may cause the virtual assistant to send an unwanted message. This may be harmful if the wrong message is sent to the wrong person or if a conversation is unintentionally recorded and sent to the wrong person. A virtual assistant may also enable the installation of malware. Misinterpretations may also cause the accidental turning off of the heating in a house with a baby, which may have devastating consequences. Finally, the use of virtual assistants may raise serious privacy concerns, as the user audio is recorded and sent to a server for processing. This challenge is further discussed in Section 9. Virtual assistants usually collect user information during their operation.

Some virtual assistants give programmers the ability to extend their abilities. For example, Alexa allows programmers to extend her abilities using the Alexa Skill Kit (ASK). Participants in the Alexa Prize challenge developed social chatting skills for Alexa. There are few open-domain CAs that enable a lay user, rather than a programmer, to teach the agent to perform new action sequences or new responses. A learning-by-instruction agent (LIA) [132] uses a combinatory categorial grammar (CCG) semantic parser to transform the semantics of each command to a few terms of primitive executable procedures that define the sensors and effectors of the agent. If the user gives the LIA a natural language command and if the LIA does not know how to execute the command, it will ask the user to explain how to realize the command through a sequence of natural-language steps. Once explained, the LIA can execute the command in the future.

SUGILITE [133] is a programming-by-demonstration (PBD) system that uses the Android's accessibility API to enable users to create automation on smartphones. In case

the user specifies commands that SUGILITE does not know how to execute, it prompts the user to demonstrate the command, records the user's explanation, and automatically generates a script. Thus, SUGLITE can learn to execute an unrecognized command from a single demonstration.

Safebot is a collaborative chatbot that allows users to teach the agent new responses [134]. Safebot allows the users to identify inappropriate responses, which are then removed from Safebot's database such that future users are not allowed to teach Safebot responses similar to the ones previously tagged as inappropriate.

KBot [135] is a comprehensive open-access CA that exploits the potential of semantic web technologies, federated databases, and NLU. KBot contributes to a better understanding of user queries in the context of linked data by being able to answer different user queries. It can handle tasks such as conversations in English, social-network conversations, FAQs, and mathematical tasks, using information gathered from multiple sources such as DBpedia, Wikidata, and MyPersonality (http://mypersonality.org) datasets.

Finally, MILABOT [74] is a DRL-based CA, developed for the Amazon Alexa Prize competition. MILABOT is capable of chatting with humans through speech or text. It was trained on crowdsource data and real-world-user interactions.

6.2. Educational Applications

Online learning has shown significant growth over recent years, in particular, during the COVID-19 outbreak. Unfortunately, in online learning, teachers and students are distant from each other, and therefore, the connection and interaction between them may be insufficient. This may cause online learning to be less effective.

There have been multiple attempts to enhance online learning by using intelligent tutoring systems (ITS) [136], which are customized, computer-based instruction and feedback methods without human intervention. Many include conversational agents, which can interact with the students in natural language during the learning process.

Paschoal et al. [137] surveyed 101 pedagogical conversational agents. They identified the different educational areas for which conversational agents have been developed, discussed common development techniques for pedagogical CAs, and also surveyed the communication strategies used by pedagogical CAs to interact with students. Some successful CAs that are recently used in the education domain are next described. Sara is a CA to assist students with learning [138]. Sara shows online video lectures and asks questions to ensure that the student has understood the lecture. It offers additional information and explanations if the student's responses are inaccurate. Sara interacts by voice and text when needed and has a voice-based input mode. It was demonstrated to improve learning in a programming task. A similar CA was developed by Paschoal et al. [139] to support software testing. AutoTutor [140] is a computer tutor that simulates the dialogues and strategies of a human tutor. It presents questions and problems from a curriculum script and, according to the learner's input, decides which action to perform next (e.g., providing a hint or moving on to the next problem). AutoTutor segments the input from the learner into a sequence of words, to assign alternative syntactic tags to words and the correct syntactic class to a word.

MSRBot is a question-answering CA dedicated to software-related issues [141]. It uses a neural network to classify each speech act into one of five speech-act categories: assertion, wh-question, yes/no question, directive, and response. It extracts useful information from software repositories to answer several common software development/maintenance questions.

Hobert [142] presents the design and evaluation of a chatbot-based tutor to help teach beginner programmers to code in university courses. Hobert's coding tutor is based on teaching-assistant requirements that appear in the scientific literature. Hobert claims that his chatbot tutor is suited to take over the tasks of teaching assistants when there is no human teaching assistant available.

Similarly, Kloos et al. and Aguirre et al. [143,144] introduced the design and features of a CA for Google Assistant [145] to complement a massive open online course (MOOC) for learning Java. Both studies run several experiments and report that users find the conversational agents to be very useful.

Lin et al. [146] developed Zhorai, a CA that enables children to explore AI algorithms and machine learning. Lin et al. showed that by training an agent, observing its mistakes, and retraining the agent, children were able to understand the agent's ability to learn, as well as obtaining some level of understanding of the learning algorithms used by it.

Cai et al. [147] introduced MathBot, a rule-based chatbot that explains math concepts, provides practice questions, solves problems, and offers tailored feedback. Using mTurk workers, Mathbot was compared to other baseline methods, such as video tutorials and written material. It was found that students prefer MathBot over other options.

CAs can also be useful in foreign-language learning. Indeed, there have been several recent attempts to develop CAs for that purpose. Duolingo's chatbot with Mondly as well as Andy are some examples of chatbot applications for language learning [148]. Some virtual assistants, such as Alexa, include extensions that enable the learning of foreign languages [149]. Alexa has the skills to assist in building a vocabulary and handling a conversation in a foreign language. Pham et al. [150] developed English Practice, which is a mobile chatbot application to assist a user in learning new vocabulary and to carry on a conversation. Another CA dedicated to language learning is Lucy [151], an embodied virtual agent, designed to help users to learn vocabulary and grammar and to carry on a conversation.

CAs can also be used to support the administration in educational systems. For example, Hien et al. [152] present FIT-EBot, a chatbot that responds to student questions related to services provided by the education system on behalf of the academic staff. Similarly, Ranoliya et al. [153] introduced a chatbot designed to answer visitor questions at Manipal University. It provides an answer based on a dataset of frequently asked questions (FAQ) using AIML. When a user asks a query, the chatbot searches for a similar question and provides the answer to that question. Another chatbot was developed by Keeheon et al. [154] to provide information in educational systems by answering frequently asked questions The chatbot was successfully used by students and department offices in Underwood International College, Korea.

The authors reported that the use of the chatbot had a positive influence on administrative work in reducing workload.

Discussion-bot [155], developed by Feng et al., provides answers to students' discussion-board questions using natural language. Given a question, it mines suitable answers from an annotated corpus of archived discussions and course documents and chooses an appropriate response.

6.2.1. Special-Needs Education and Assistance

In recent years, researchers have expressed a growing interest in using CAs as well as social robots as a positive intervention for children with special needs [156].

PunkBuddy is a tool that includes a chatbot that helps dyslexic students learn through interaction. The chatbot can advise students on the rules of using punctuation, utilizing the benefits of explicit instruction [157].

Park et al. [158] developed a voice-based virtual agent for children with ADHD to help them in their daily tasks. The agent provides vocal feedback to the child and encourages the child to complete the task (on time). The child reports back to the agent about her/his progress.

Xuan et al. [156] developed a chatbot dedicated to children with autistic spectrum disorder (ASD) to improve their conversation abilities. Their chatbot is intended to arouse the curiosity of children and assist them in understanding the conversation better. The chatbot uses a large question-and-answer corpus. Social-assistance CAs are commonly used to assist children and adults with special needs, and especially children with ASD.

Indeed, several studies have shown that social robots can help improve the social skills of children with ASD [159], and some have indicated that a child with ASD might find it easier to interact with a social robot than with a human teacher [160].

Scassellati et al. [161] developed a social robot to increase the social-communication skills of children with ASD. The robot can move or talk according to a selected task defined by the caregiver. For example, the robot can present a social situation and ask the child what the story character is feeling. They reported that after a one-month deployment, the children with ASD improved their behavior and gained their independence.

Costa et al. [162] introduced QTrobot, a social robot developed to assist children with ASD to focus their attention, imitate positive behavior, and reduce repetitive and stereotyped behaviors. QTrobot converses with the child and plays imitation games with the child. Costa et al. showed that children pay more attention to QTrobot than to a person, imitate the robot as if it is a person, and practice fewer repetitive and stereotyped behaviors with the robot than with the person.

Vanderborght et al. [163] developed Probo, which is a social story-telling robot capable of expressing emotions via facial expressions and gaze. Probo uses stories to teach children with ASD how to react in different situations, such as saying "hello" or "thank you." Probo also teaches children to share their toys. Vanderborght et al. showed that there are situations where the social performance of autistic children improves when using Probo.

Another known robot developed in the same project is Nao. [164], an embedded CA that has been tested and deployed in several healthcare scenarios, including care homes and schools.

6.3. Healthcare Conversational Agents

CAs can potentially play an important role in healthcare. There have been several recent reviews on CAs in this field (see [165–168]). Each points to challenges in the healthcare area pertaining to efficiency, security, and privacy.

CoachAI is a system that includes a chatbot and a machine-learning model to support a patient's health activities [169]. The chatbot collects data, sends reminders, and converses with users through text-based, simple, graphical elements to guide the user in health-related issues. The model is based on real-world data provided by a health clinic. The application provides the caregivers with insights on the users and assists with the tracking of user activities and their health conditions.

Daily healthcare can be overwhelming for people with a chronic disease. Neerincx et al. [170] developed a social robot that helps children with diabetes. The robot supports the daily diabetes-management processes, namely, taking pills, shots, and body measurements by conversing with the child.

The Watson assistant for health (Watson Health) is an extension of IBM Watson [171] to the healthcare domain. Watson was originally developed for the Jeopardy challenge. Watson Health [172] is a CA for health support. It uses a text-based natural-language interface. It receives a collection of patient symptoms and produces a list of possible diagnoses. The assistant provides detailed annotation as well as links to supporting medical literature. However, a study conducted by Ross and Swetlitz [173] indicates that, in some cancer cases, Watson Health provided unsafe and incorrect recommendations.

Xu et al. [174] introduced KR-DS, a chatbot for the healthcare domain. KR-DS obtains a set of symptoms from the user, recognizes the bio tags of each word using Bi-LSTM, classifies the intent of each sentence, and finally, provides a diagnosis to the user, in natural language, using a medical-knowledge graph. Experiments show that KR-DS outperforms other state-of-the-art methods in diagnosis accuracy.

Fitzpatrick et al. [175] developed Woebot, a medical voice-based CA for cognitive-behavioral therapy dedicated to nonclinical cases addressing low mood and anxiety. Woebot provides mental-health information, recommends activities for specific mood problems, and handles emergency-support services. The users reported an improvement in their mood after using Woebot.

Edwards et al. [176] introduced Tanya, a graphically embodied female agent that supports breastfeeding. Tanya was deployed in a hospital and was accessible to women after birth. Edwards et al. show that women that interacted with Tanya increased their chance of successful breastfeeding for the first six months.

During the COVID-19 outbreak, people require medical information with respect to the outbreak but cannot obtain the information from medical teams, which are overwhelmed. Yang et al. [177] developed a medical chatbot that can be consulted for COVID19-related issues. The chatbot is trained on two datasets, in English and Chinese, containing conversations between doctors and patients on COVID-19.

Despite all the CAs developed in the field of healthcare, the reception of CAs in this field has not been as positive as expected. Palanica et al. [178] examined the perspectives of practicing medical physicians on the use of healthcare CAs for patients. Their results indicate that many physicians believe that CAs would be most beneficial for scheduling doctor appointments, locating health clinics, and providing medication information. However, most of the physicians believe that CAs cannot effectively take care of patients' needs or provide detailed diagnosis and treatment. Nadarzynski et al. [179] studied the acceptability of CAs in healthcare from the perspective of the general public. While the participants in the study recognized the potential of CAs in healthcare, they stated that their experience is not satisfactory enough and that they are concerned about security issues. Scholten et al. [180] surveyed several CAs in the field of healthcare. They concluded that while CAs can increase the motivation of patients and promote behavioral change, user needs are many times implicit, and these needs cannot be addressed by CAs.

6.4. CAs in the Business Domain

Conversational agents are becoming more and more prominent in a diverse range of applications in the business area. According to Dhanda [181], CAs have reduced costs in organizations by approximately USD 48.3 million in 2018 and are expected to reduce costs by USD 11.5 billion by 2023. See Bavarescoa et al. [182] for a literature review on CAs in the business domain with a focus on machine learning. CAs can be used as customer-service assistants, providing answers to frequently asked questions (FAQs), which is a common task that can be handled by CAs.

The Thomas question-answering chatbot [183] uses artificial-intelligence markup language (AIML) for template-based questions like greetings and general questions and latent semantic analysis (LSA) [183] to answer other related questions. If the chatbot cannot find a relevant answer, it asks the user for a clarification.

Another chatbot in the customer service area is SuperAgent [184], which leverages large-scale and publicly available ecommerce data. Given a user request for information about a specific product, SuperAgent provides relevant information from in-page product descriptions and from ecommerce websites. SuperAgent is provided as an add-on extension to the Microsoft Edge and Google Chrome browsers.

Xu et al. [185] created a chatbot to serve users' requests on social media (Twitter). The chatbot encourages interaction between users and businesses on social media. The chatbot was trained on nearly one million Twitter conversations between users and agents. Their analysis indicates that over 40% of user requests are emotional and do not intend to seek specific information. They showed that their chatbot, which is based on deep learning, yields a higher BLEU score [186] than that of an information-retrieval-based system.

Yan et al. [187] introduce a chatbot, dedicated to online shopping. The goal is to assist online customers in purchase-related tasks by answering specific questions and searching for a product. They integrate this system into a mobile online shopping application with millions of consumers.

Another chatbot is SamBot [188], which is integrated into Samsung's website to answer user questions. Its knowledge base includes: Samsung promotion, Samsung product FAQs, and general information related to Samsung (e.g., open hours and branch locations). If a proper answer cannot be found, SamBot generates a random answer. It can also recommend

users questions to ask. They show that SamBot is capable of handling Samsung-related questions very well.

Kaghyan et al. [189] reviewed the aspects of business-to-business (B2B) tools including the use of CAs. In their article, they describe several methods and platforms for creating Facebook chatbots that support a business. Detailed descriptions are provided for three chatbot-creation platforms: Chatfuel, ManyChat, and "It's Alive!," and a comparison was performed with respect to capabilities, strengths, and limitations.

Another use of CAs in the business domain is for negotiation. Lewis et al. [190] demonstrate that it is possible to train end-to-end CAs for negotiation, which is simultaneously a linguistic and a reasoning problem. To achieve this goal, their CAs contain adversarial elements as well as cooperative elements, and the CAs are required to understand, plan, and generate utterances. They collected a dataset of natural-language negotiations between two people to show that their end-to-end neural models successfully imitate human behavior in this domain.

Luo et al. [191] collaborated with a large financial-services company to design a randomized field experiment on the consequences of chatbots hiding or revealing that they are indeed chatbots. They concluded that when the true identity of chatbots is not disclosed, CAs are as effective as proficient workers and four times more effective than inexperienced workers in increasing customer purchases. However, when chatbots disclose their identity before conversation, the purchase rates are reduced by more than 79.7%, and the conversation becomes shorter. Unfortunately, users do not always trust that CAs can provide the required support.

Følstad et al. [192] present an interview study of thirteen users who interact with chatbots in customer support regarding their experience and the factors affecting their trust. The users' trust was found to be affected by different attributes such as the quality of the CA's interpretation of the requests and whether the generated text seemed human-like.

Chihsun et al. [193] investigated how users cope with conversations with chatbots that do not make any progress in the field of customer support. They analyzed a three-month conversation log with a chatbot, which was taken by one of the top digital-banking institutions in Taiwan. They found 12 types of conversational non-progress and 10 types of coping strategies on the part of the user.

Abdellatif et al. used Google's Dialogflow engine [69] to extract the user intent and the entities mentioned in the user input. Their initial training set was collected from a group of software developers and consisted of different ways developers pose similar questions. Additional training data were collected from developers using the initial CA version during a test period.

6.5. Influence and Malicious CAs in Social Networks

Several conversational agents are developed for deployment in social networks. These CAs attempt to influence public opinion by persuading specific surfers to take certain actions, consume certain products, or influence political views.

Few internet tutorials [194,195] have been written to guide users in the process of Twitter chatbot development. Adams [196] gives an overview of influence-impersonating CAs, which impersonate a human to influence users on social media. They also state that most impersonator chatbots are very simple and therefore,cannot deceive serious interrogators.

The study of Assenmacher et al. [197] provides insights into markets of influence and malicious chatbots as well as an analysis of freely available software tools, which are used to create them. Similar to Adams, they conclude that current influence chatbots are very simple and, despite the major advances in the literature on CAs, still use very simple automation methods.

Another study in the social chatbot area is that of Kollany [198]. According to Kollany, there is an exponential growth in the number of influence chatbots on Twitter. Kollany

gathered data from GitHub on the ways developers collaborate with each other and check social aspects of programming on that platform.

While influence CAs are usually intended only to influence a person's opinion, some malicious CAs utilize a social network to steal personal and private, information including credit-card and bank-account details, or to spread false information in an attempt to manipulate the stock market [199].

Several studies focus on influence and malicious chatbots acting in social media. Varol et al. [200] used a publicly available dataset of Twitter accounts and manually labeled all users either as humans or influence chatbots. They estimated that 9–15% of active Twitter accounts exhibit influence chatbot behavior. They present a machine learning model to detect influence chatbots on Twitter based on features extracted from the dataset, such as user followers and tweet content and sentiment.

DARPA held a four-week competition in 2015 in which multiple teams competed to detect influence chatbots on Twitter [201]. Out of 7038 Twitter accounts, 39 were labeled by DARPA as influence chatbots. The leading group detected all influence chatbots, using a combination of machine learning techniques along with a user support system.

Lee et al. [202] deployed honeypots in the Twitter social network to identify and analyze content polluters. They investigated the attributes of Twitter users, including user behavior over time, user followers, and user following. They also enumerate features that may assist in identifying content polluters automatically, and they present a classification model. Finally, they show that their model successfully identifies content polluters.

To summarize this section, Figure 9 refers to the CA definitions (provided in Figure 1) and, for each type of CA, details the domain of applicability.

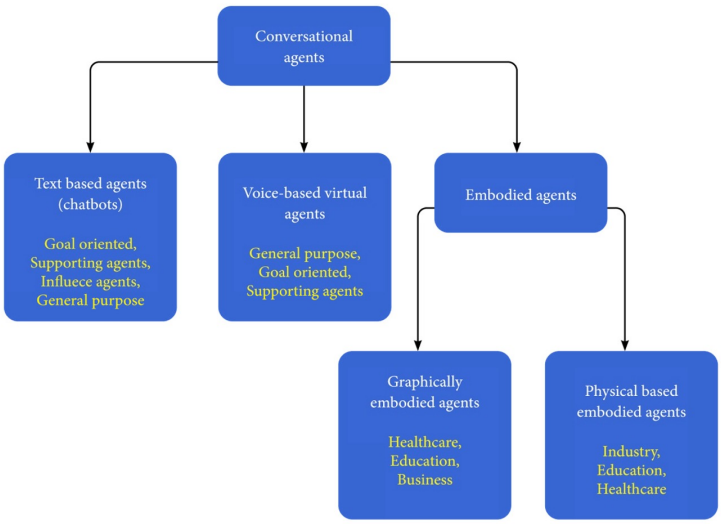

Figure 9. Conversational-agent applications.

7. Evaluation Metrics

Three main approaches are used in the literature for evaluating the quality of a conversation agent: human-based evaluation procedures, machine evaluation metrics based on language characteristics, and an ML approach trained on a dataset consisting of human evaluations. The advantages of human evaluation are clear, as humans can evaluate whether the CA responses seem appropriate and resemble responses. However, since human evaluation procedures are expensive, several automatic metrics have been

proposed for the evaluation process. Unfortunately, due to the linguistic richness of natural languages and the wide variety of reasonable response options, it is still challenging to achieve accurate and meaningful evaluation when using automatic tools. Therefore, the ML approach tries to benefit from both approaches; on the one side, it is based on human evaluation, and, on the other side, it does not require new implicit costly evaluation methods for each new dialogue situation.

Radziwill and Benton [14] present a literature review of quality issues related to CA development and implementation, focusing on two topics: quality-attributes and quality-assessment approaches. Deriu et al. [203] surveyed the main concepts and methods of CA evaluation. For each type of CA, task-oriented, conversational, and question-answering dialogue systems, they defined the main technologies and the evaluation methods that are appropriate for that type. The requirements of the evaluation methods are stated with respect to automated or partially automated evaluation, repeatability of the results, correlation with human judgment, ability to focus on CA features, and explainability. Finally, Masche and Le [16] divide the different evaluation methods into four classes: qualitative analysis, quantitative analysis, pre/post-test, and CA competition.

In this section, the evaluation methods are divided into three classes, according to the way they are obtained, namely, human-based evaluation, machine-based evaluation, and the ML approach, and some popular evaluation methods are further described for each of these three classes.

7.1. Human-Based Evaluation Procedures

As mentioned above, the most accurate method to assess the dialogue quality of a CA is through the score and the qualitative description obtained from humans interacting with the CA. Deriu et al. [203] describe various approaches of human evaluation consisting of lab experiments with users invited to interact with a CA and subsequently asked to fill out a questionnaire; in-field experiments with feedback collected from real users of the CA; and crowdsourcing with crowd workers, either asked to talk to the CA and then rate it or asked to read a produced dialogue and then rate it. The CA rating is based on quality, fluency, appropriateness, and sensibleness.

Venkatesh et al. [18] describe the following metrics to evaluate an open-domain CA: user experience, coherence, engagement, domain coverage, topical depth, and topical diversity. In addition, they propose a unified evaluation strategy, which combines the above metrics into a new evaluation model that correlates well with human judgment. Their unified evaluation strategy was applied throughout the Alexa Prize competition to select the top-performing CAs.

Griol et al. [204] defined a set of specific measures to evaluate the quality of a medically oriented CA. The proposed measures are divided into high-level dialogue features, dialogue style, and cooperativeness. High-level dialogue features evaluate how long the dialogue lasts, how much information is transmitted in individual turns, and how active the dialogue participants are, while dialogue style and cooperativeness features analyze the contents of different speech actions.

To summarize, there are generally three main sources of human-based evaluation: lab sources, real CA users, and crowdsourcing. The information obtained from humans can include: qualitative and quantitative questionnaires, real CA user feedbacks, and dialogue features.

7.2. Machine-Evaluation Metrics

Since a high cost is associated with human evaluation, machine-based evaluation or hybrid human-machine-based evaluation are widely used to examine the quality of CAs. Machine-based CA evaluation is challenging due to the lack of an explicit objective for conversation performance measurement. Several studies utilize machine translation-based metrics for CA quality evaluation.

One such metric is the BLEU score [205], a text summarization metric developed for automatic evaluation of machine translation. BLEU takes the geometric mean of the test corpus modified precision scores and multiplies it by an exponential brevity penalty factor. The main component of BLEU is the n-gram precision, which is the proportion of the matched n-grams out of the total number of n-grams in the evaluated translation.

Recall-oriented understudy for gisting evaluation (ROUGE) [206], originally developed for automatic summarization, is also adapted to CA evaluation. Similar to BLEU, ROUGE counts the number of language units, such as n-grams, that appear both in the evaluated summary and in the ideal human-generated summary.

Another popular evaluation metric for machine translation that is applied to CA evaluation is METEOR [207]. METEOR evaluates a translation by counting word-to-word matches between a translation and the reference sentence. If more than one reference is available, the given translation is scored against each reference independently, and the best score is reported.

Liu et al. [208] investigated the usage of the above translation and summarization evaluation metrics for CA. They note that available machine translation metrics assume that valid responses should have significant word overlap with the ground-truth responses. This is a strong assumption for CAs, which exhibit a significant diversity in the space of valid responses. They show that many commonly used metrics for CA evaluation do not correlate strongly with human judgment, and they conclude that there is a need for a new metric that correlates more strongly with human judgment.

7.3. Machine-Learning-Based Evaluation

A third approach of CA evaluation is to use ML to predict the human rating of CAs' dialogues. Lowe et al. [209] present a dialogue-evaluation model called ADEM that learns to predict human-like scores for CA responses, using a dataset of human scores of responses. The human scores were collected using crowd workers that were shown a dialogue context and a candidate response and asked to rate the responses. ADEM is trained by an RNN and, given a response, can successfully predict the appropriateness rating of the response as if it is a human.

Tao et al. [210] propose a routine for evaluating system responses called RUBER. RUBER consists of a Siamese neural network, trained to predict if a pair of context and response are relevant. RUBER is trained using two metrics: a referenced metric measures the similarity between the generated response and the ground-truth response, and an unreferenced metric measures the relatedness between the generated response and the original query. The referenced and unreferenced metrics are combined with heuristic strategies (e.g., averaging) to further improve RUBER's performance.

Guo et al. [211] propose a topic-based evaluation method on topic breadth, which checks the ability of the CA to talk about a large variety of topics, and topic depth, which checks the ability of the CA to handle a long and cohesive conversation about one topic. A deep average network (DAN) was used to train the topic classifier on a variety of questions and query data, categorized into multiple topics. To summarize, the ML approach of evaluation can be helpful to a wide range of CA researchers and developers as it combines the advantage of human judgment with the advantage of resource saving to rate an unlimited number of CAs and dialogues, utilizing the trained evaluation model.

Tables 1 provide the technologies and the evaluation method(s) behind each of the main CAs described in Section 6.

Table 1: Technologies and evaluation methods for main CA applications: part A

Personal Assistants and Open-Domain CAs			
CA	**Short Description**	**Main Technology**	**Evaluation Method**
ALICE [48]	a general-purpose chatbot	AIML, pattern matching	the most human computer winner, 2000, 2001, 2004
LSA-bot [50]	ad-hoc implementation of the LSA framework	Latent Semantic Analysis (LSA)	-
IRIS [51]	example-based chatbot	vector space model cosine similarity metric	success and failure examples
DeepProbe [129]	an open-domain chatbot chatbot	seq-2-seq	AUC scores
RubyStar [130]	an open-domain chatbot	seq-2-seq, topic detection, engagement monitoring, context tracking	human evaluation by the Alexa Prize evaluation
Siri [1]	Apple's virtual assistant	CNN, LSTM	commercial application
Cortana [3]	voice-controlled assistant for Microsoft windows	NLP, Tellme Networks, Semantic search database	commercial application
Alexa [23]	Amazon voice assistant	NLP, LSTM	commercial application
KBot [135]	knowledge chatbot	SVM+analytical queries engine	F-score, precision, recall, intent classification
MILABOT [74]	speech/text CA	DRL	Amazon Alexa Prize competition
Discussion-Bot [155]	question-answering chatbot	semantically related matching, TF-IDF metric	human judges classified the answers quality
Goal-Oriented CAs			
CA	**Short Description**	**Main Technology**	**Evaluation Method**
SUGILITE [133]	Programming-by-demonstration system	frame-based dialogue management	a lab study: task completion time
Safebot [134]	collaborative chatbot	parser+Word2Vec	users' engagement
LIA [55]	learning by instructions agent	uses combinatory categorial grammar (CCG) parser	speed of task completeness

Table 1: *Cont*.

CAs for Social Support			
CA	**Short Description**	**Main Technology**	**Evaluation Method**
ELIZA [19]	the first CA: emulates a psychologist	pattern matching	people experience
XiaoIce [107]	a popular social CA	IQ + EQ + Personality	human rating
Meena [2]	a sensible chatbot	generative chatbot trained end-to-end on social media conversations	human evaluation metric called Sensibleness and Specificity Average (SSA)

Table 2: Technologies and evaluation methods for main CA applications: part B

Educational CAs			
CA	**Short Description**	**Main Technology**	**Evaluation Method**
Sara [138]	student's assistant	scaffolding strategy	pretest and posttest scores of learners pro-survey and post-survey
AutoTutor [140]	computer tutor	LSA, pattern-matching speech act classification	learning gain
MSRbot [141]	sofware related Q&A	Dialogflow	effectiveness, efficience
Zhorai [146]	CA for children to explore ML concepts	NLTK package Website visualizer	accuracy, child's level of engagement
MathBot [147]	math teaching chatbot	rule based	crowd worker preferences
English Practice [150]	Personal Assistant for Mobile Language Learning	Dialogflow platform	statistics about real users
Lucy [151]	embodied on-line virtual agent for language learning	ALICE offshoot	demonstrative examples
FIT-EBot [152]	administrative chatbot	DialogFlow	students reports
QTrobot [162]	social robot to assist children with ASD	bodied humanoid robot	interviews with the users
Probo [163]	social robot for children with ASD	compliant actuation systems	children performance
Healthcare CAs			
CA	**Short Description**	**Main Technology**	**Evaluation Method**
CoachAI [169]	patient's support chatbot	task-oriented finite state machine (FSM) architecture	user's engagement, system accaptance and rating.
Woebot [175]	therapist CA	AI,NLP,empathy engine	users' reports
Mandy [126]	a primary care CA	NLU, NLG, word2vec	accuracy
Tanya [176]	graphically embodied female agent that supports breastfeeding		increased breastfeeding success
KR-DS [174]	diagnosis chatbot	Bi-LSTM, Deep Q-network	diagnosis accuracy
Commercial CAs			
CA	**Short Description**	**Main Technology**	**Evaluation Method**
SuperAgent [184]	customer-service chatbot	AIML+LSA	2 customer reviews
SamBot [188]	question-answering CA	AIML	Loebner Prize Competition + user interaction

Finally, Figure 10 illustrates the various evaluation methods and their relation to each of the relevant components.

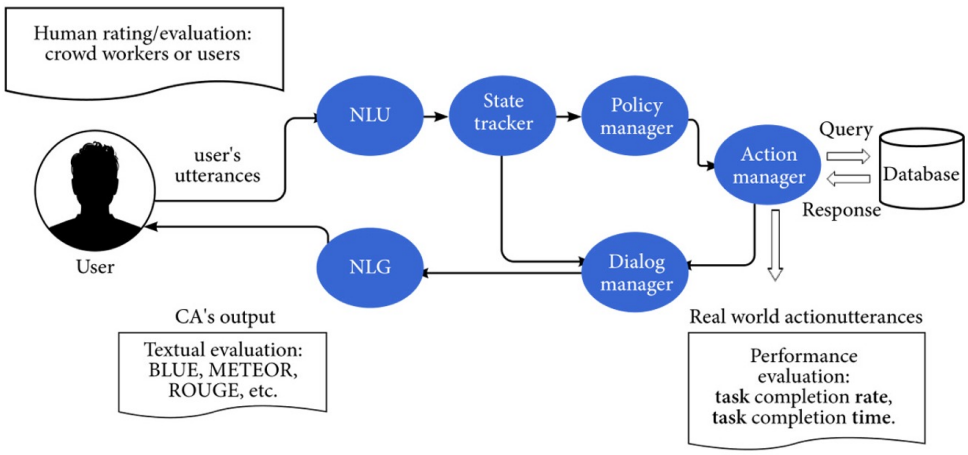

Figure 10. A diagram illustrating the various CA evaluation methods.

8. Publicly Available Conversation Datasets

Conversation datasets are used to train machine learning CA models and to test the quality of the CA. In this section some of the existing datasets used in the literature for CA development and CA evaluation are described. Some recent reviews focusing on available conversation datasets are presented next.

Serban et al. [212] review different types of conversations datasets for CAs and categorize them according to the type (text or speech), topics, length (number of dialogs, average number of turns, and number of words), and description.

Keneshloo et al. [213] provide a list of conversational datasets that can be used for sequence-to-sequence models. Some of the databases provided can be helpful for the dialogues generated by conversational agents, and others are related to other domains, such as image and video captioning, computer vision, speech recognition, and synthesis.

Deriu et al. [203] provide another list of available conversation corpora focusing on task related conversations in several domains, such as the restaurant domain and the tourist information domain. They note that question answering dialogue systems can be extracted either from chat logs or from several available literature sources, news, scientific resources, Wikipedia articles, FAQ sites, and even cooking domains.

In the remainder of this section, some of the most useful corpora for conversation understanding, generation, and evaluation are described and classified according to their applications, using the terms defined in Section 2.

8.1. Datasets for General Purpose CAs

There are various sources of datasets used for general-purpose dialogues. DailyDialog (http://yanran.li/dailydialog) [214] is a dataset consisting of handwritten texts, manually labeled with communication intention and emotion information. DailyDialog contains multi-turn dialogues, reflecting daily communication on various aspects of daily life. The

dialogues in the dataset conform to various common dialogue flows, such as question and answer, bi-turn flows, and multi-turn dialogue-flow patterns reflecting realistic dialogues.

Large amounts of available data on movie reports may also be utilized to build dialogue corpora. The SubTle corpus [215] is designed for general-purpose interaction generation. It is composed of interaction–response pairs, extracted from the OpenSubtitles (http://opus.nlpl.eu) [216,217] movie corpus, which is a multi-language conversation corpus based on movie subtitles. Additional datasets based on movie dialogs are the Movie dialogue dataset (https://www.kaggle.com/abhishek/the-movie-dialog-dataset) [218] and Cornell movie dialogues corpus (https://www.cs.cornell.edu// cristian/Cornell_Movie-Dialogs_Corpus.html) [219].

Serban et al. [212] consider the advantages and disadvantages of training and evaluating CAs based on artificial datasets, such as datasets extracted from movie manuscripts and audio subtitles. The advantages are as follows: (a) the dialogues resemble human spontaneous language; (b) the dialogues are easy to follow and contain less garbling and repetition; (c) there is a diversity of dialogues, topics, environments, actors, and relationships. This enables creating a more flexible CA, which may talk with various users in different situations while using various interaction patterns. However, since CAs must consider the context to provide accurate responses, Serban et al. state that artificial datasets may have a caveat as they do not provide this context. It should be noted that since dialogues from movies can be too extreme and not reflect real-life dialogues, training and evaluating CAs based on them may lead to undesired behavior on the part of the CAs.

Another source of datasets, for the training and evaluation of CAs, is social media. Many datasets are composed of texts extracted from popular conversation websites and applications, such as Reddit (https://www.reddit.com) and Twitter (https://twitter.com).

Dialogue corpora based on Twitter conversations are developed and used by Li et al. [220], Sordoni et al. [82], Xu et al. [185], and Ritter et al. [221]. Dialogue corpora based on Reddit forums have been developed by several other studies, including the study of Dodge et al. [218], Serban et al. [74], Schrading et al. [222], and recently by Zhang et al. [223]. The dialogue-generation model of PLATO [224] is pretrained on both Twitter and Reddit. The Ubuntu dialogue corpus [225] is based on the Ubuntu chat logs.

Serban et al. [212] note that datasets based on conversations extracted from social media have some significant limitations. Generally, they are noisy, and they may include texts generated by non-human CAs, such as influence agents. Another limitation of Twitter-based datasets is the maximum length of 140 characters per Twitter message. As a result, the Twitter corpus has an enormous number of typos, slang, and abbreviations as well as Twitter-specific structures, such as hashtags. Similar to the issue with artificial datasets, Serben et al. note that dialogues extracted from social media may be missing context. In addition, as stated by Kourosh [226], the use of auto-correction by users of social media may cause an additional layer of complication.

8.2. Datasets for Question Answering

Question-answering conversational agents can be trained using publicly available question-and-answer web pages. Zeng et al. [227] surveyed machine-reading-comprehension evaluation and benchmark datasets. They note that the most popular datasets in this category are the Stanford question answering dataset (Squad) versions 1.1 [228] and 2 [229], the CNN/Daily Kail dataset [230], the natural-questions dataset [231], and TriviaQA [232].

The Squad datasets are designed for machine-reading-comprehension training. They consist of more than 100K questions and answers posed by crowd workers in Wikipedia articles; the answers are citations within Wikipedia articles. The CNN/Daily Mail dataset contains question/answer pairs generated from CNN and Daily Mail articles, published during 2007–2015 for CNN and during 2010–2015 for the Daily Mail.

The natural-questions dataset [231] contains real user questions posted on Google search and answers found on Wikipedia by crowd workers. Each real question may have

three types of answers: an associated long answer, which is based on text from a Wikipedia article, a list of short answers, and a yes–no-answer.

Finally, the TriviaQA [232] dataset, designed for machine-reading-comprehension challenges, contains triplets of question–answer-evidence; the evidence aims to ease the answering process. TriviaQA contains relatively complex and challenging questions with syntactic and lexical variability, requiring cross-sentence reasoning in answering TriviaQA questions.

8.3. Datasets for Goal-Oriented CAs

The challenge of designing a goal-oriented CA is twofold: the CA should be both effective in NLU and NLG and efficient in helping to solve the common task. Consequently, the task-oriented conversation should take into consideration both aspects. A useful source for obtaining goal-oriented datasets is the dialogue-system-technology challenge (DSTC) [71], which is a yearly challenge started in 2013. Various well-known datasets have been produced and released for every DSTC edition.

The schema-guided-dialogue (SGD) dataset [233], released for DSTC8, contains approximately 23K annotated multi-domain (bank, media, calendar, travel, and weather), task-oriented dialogues between a human and a virtual assistant. SGD can test state tracking as well as intent prediction, slot filling, and language generation.

MultiWOZ [234] is a tourist-dialogue dataset, annotated with dialogue belief states and dialogue actions. The dialogues in MultiWoz cover seven touristic domains: attractions, hospitals, police, hotels, restaurants, taxis, and trains. Each dialogue in MultiWoz can cover more than one domain.

Taskmaster-1 [235] includes dialogues of the following task-oriented domains: ordering pizza, setting auto-repair appointments, arranging taxi services, ordering movie tickets, ordering coffee drinks, and making restaurant reservations. More than half of the dialogues were created manually, using crowd-workers to compose entire dialogues.

Finally, MultiDoGo [236] is a public human-generated multi-domain dialogue dataset, composed of dialogues created by crowd workers and trained annotators, with a total of over 81K dialogues across six domains. Over 54K of these conversations are annotated for intent classes and slot labels.

For a list of task-related datasets, including DTSC challenges datasets, see Deriu et al. [203].

8.4. Datasets for Social Assistance

Social-assistance CAs aim to provide medical, healthcare, mental, or other educational assistance. In these domains, there may exist a privacy issue: information in medical, mental, or educational dialogues is sensitive, and therefore, it is difficult to publish dialogues in a way that would honor the privacy of the participants. Here are some repositories found in these areas.

The first attempt to create a large medical corpus is MedDialog, developed by Zeng et al. [237]. MedDialog is a medical-dialogue dataset that consists of 3.4M conversations between patients and doctors in Chinese, covering 172 specialties of diseases, and 260K conversations in English, covering 96 specialties of diseases. Each consultation consists of a description of the patient's medical condition, followed by a conversation between the patient and the doctor. The data are gathered from Iclinic (iclinic.com) and HealthcareMagic (caremagic.com), which are online healthcare service platforms.

Another health-related dataset was constructed by Yang et al. [177]. Their dataset consists of a collection of conversations in English and Chinese between doctors and patients about COVID-19. The English dataset contains 603 consultations, and the Chinese dataset contains 1088 consultations.

Sharma et al. [238] introduced the task of transforming low-empathy conversational posts into higher-empathy posts. They focus on mental health-related conversations filtered from posts of TalkLife (talklife.co), which is the largest online peer-to-peer support platform

for mental-health support. The dataset contains 3.33M interactions from 1.48M users posts. The interactions were labeled with empathy measurements using a framework, consisting of three empathy-communication mechanisms: emotional reactions (expressing emotions such as warmth and compassion), interpretations (communicating an understanding, feelings, and experiences), and explorations (improving understanding of the users by exploring feelings and experiences).

Another dataset that can be used for empathic user responses is EmpatheticDialogues (https://github.com/facebookresearch/EmpatheticDialogues) [239]. This dataset consists of 25K conversations grounded in emotional situations, divided into 32 different emotion categories. The conversations are open-domain and handled between two users, with one responding empathetically to the other. Next, some datasets are described that may be helpful in recognizing emotion, detecting abuse, and generating empathic responses, which are all qualities expected from a CA used for mental and psychological assistance. The emotionally recorded corpus SEMAINE, developed by McKeown et al. [240], is based on recorded dialogues of users talking with an operator who tries to evoke emotional reactions. The corpus includes 20 participants and 100 conversations, all recorded with high-resolution cameras and microphones.

Schrading et al. [222] built a text dataset of domestic abuse, extracted from Reddit. The dataset includes abuse and non-abuse texts. Allouch et al. [241] developed a sentence-level dataset based on 13K sentences related to interactions with children having special needs. The sentences are categorized into four classes: normal sentences, insulting sentences, negative sentences about a different person, or sentences that may indicate a dangerous situation. Chai et al. [242] developed an offensive-response dataset, which consists of 110K input–response chat records in which the response is either appropriate or offensive. These databases can assist in training CAs, allowing the CAs to identify different sensitive situations to respond accordingly.

8.5. Educational Datasets

Here, educational datasets that can be helpful for educational CA development are provided.

The BURCHAK dataset [243] is a human–human dialogue dataset for interactive learning of visually grounded word meanings in a foreign language. A learner needs to learn invented words for visual objects (for example, the word ”burchak”for a square) from a tutor. The text-based interactions resemble face-to-face conversations and thus contain many of the linguistic phenomena encountered in spontaneous dialogues. The corpus contains 177 conversations and includes 2454 turns in total.

Wolska et al. [244] annotated a corpus of tutorial dialogues on mathematical-theorem proving. To collect the data, they designed and performed an experiment with a simulated tutorial dialogue system to teach mathematical-theorem proofs. The total corpus comprises 66 sets of dialogue-session logs with 12 turns, on average. There are 1115 sentences in total, of which 393 are student sentences.

Hutzler et al. [245] prepared a bank of questions designed to train high-school students on reading-comprehension skills. The questions were rated by a panel of experts using a set of criteria based on Bloom’s cognitive taxonomy [246].

The CIMA collection [247] includes tutoring dialogues between crowd workers playing the role of students and tutors. The tutoring utterances include educational strategies, such as hint provision and questions asked to check the student’s understanding.

MyPersonality (http://mypersonality.org) is a knowledge base composed of information collected from over six million volunteers on Facebook using a personality questionnaire. MyPersonality is used by KBot [135], a social-media-trained chatbot, to find answers to some questions that cannot be found in other knowledge bases, especially in the psychological and social-science domains.

Tables 3 and 4 describe the list of datasets available online, which are reviewed in this section. For each dataset, a short description is provided along with some important

attributes and the type of conversational agent that uses it, referring to the usage described in Figure 3.

Table 3: Main available datasets for conversational agents—part A.

Dataset	**Source**	**Des**
General-Purpose Datasets		
DailyDialog (http://yanran.li/dailydialog) [214]	hand written, manualy labeled	dail
[217]	subtitles	inte pair
Movie dialogue dataset (https://www.kaggle.com/abhishek/the-movie-dialog-dataset) [218]	movie metadata as knowledge triples	OM and
Cornell Movie Dialogues Corpus ($https://www.cs.cornell.edu//cristian/Cornell_Movie-Dialogs_Corpus.html$) [219]	Short conversations from film scripts	mov
Ubuntu dialogue corpus (https://github.com/rkadlec/ubuntu-ranking-dataset-creator [225]	Ubuntu chat stream	hum
Question-Answering Datasets		
Squad Version 1.1 (https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/) [228]	questions and answers on Wikipedia articles	100 on
Squad Version 2 (https://rajpurkar.github.io/SQuAD-explorer/) [229]	questions and answers and additional questions with no answers	Squ 50k with
CNN/Daily Mail (https://github.com/deepmind/rc-data) comprehension [230]	queries from the CNN and Daily Mail websites	cont tripl
Natural Questions (https://github.com/google-research-datasets/natural-questions) dataset [231]	Google search queries+ Wikipedia answers by crowd workers	Goo long shor
TriviaQA (http://nlp.cs.washington.edu/triviaqa/) [232]	crowdworkers questions	ques evid

Table 4: Main available datasets for conversational agents—part B.

	Datasets for Goal Oriented CAs
Schema Guided (https://github.com/google-research-datasets/dstc8-schema-guided-dialogue) Dialogue [233]	dialogue simulator+ paid crowd-workers
MultiWOZ (https://github.com/budzianowski/multiwoz) [234]	turkers working
Taskmaster-1 (https://g.co/dataset/taskmaster-1) [235]	crowd workers users and center operators
MultiDoGo (https://github.com/awslabs/multi-domain-goal-oriented-dialogues-dataset) [236]	crowd workers paired with trained annotators
	Datasts for Supporting CAs
COVID-19 dialogue (https://github.com/UCSD-AI4H/COVID-Dialogue dataset[177]	online healthcare platform (haodf.com
MedDialog (https://github.com/UCSD-AI4H/Medical-Dialogue-System [237]	medical dialogue platform (haodf.com
SEMAINE (https://semaine-db.eu/ [240]	human–human conversation experiment
EmpatheticDialogues (https://github.com/facebookresearch/EmpatheticDialogues [239]	810 crowd workers select an emotion and talk about it
Offensive response (https://github.com/chaiyixuan/Offensive-Responses-Dataset dataset [242]	input–response records from SimSimi (ww offensivity annotated by crowd workers
BURCHAK dataset (https://sites.google.com/site/hwinteractionlab/babble [243]	dialogues of pairs of participants, discussing visual attributes of 9 objects
The CIMA collection (https://github.com/kstats/CIMA [247]	conversations between crowd workers playing as students and tutors.

9. Conclusions and Open Issues

In this study, the extensive development of CAs in recent years was reviewed. The leap in the progression of CA development is mostly due to recent advances in deep-learning and big-data technologies. These technologies have led to developments in several domains, such as ASR, NLU, NLG, and emotion-recognition given text, voice, or images, which, combined, allow the creation of a new generation of CAs, with human-like dialogue capabilities. The focus has been on describing the current state-of-the-art technologies developed for conversational agents and various practical applications in which these agents are in use. The survey includes several innovative uses of CAs in various practical areas, including general assistance, task performance, assistance in various social areas, and influence agents, designed to impact the business and public sectors. Figure 11 summarizes the information provided by the different illustration diagrams, which appear in this survey, categorized according to their aims.

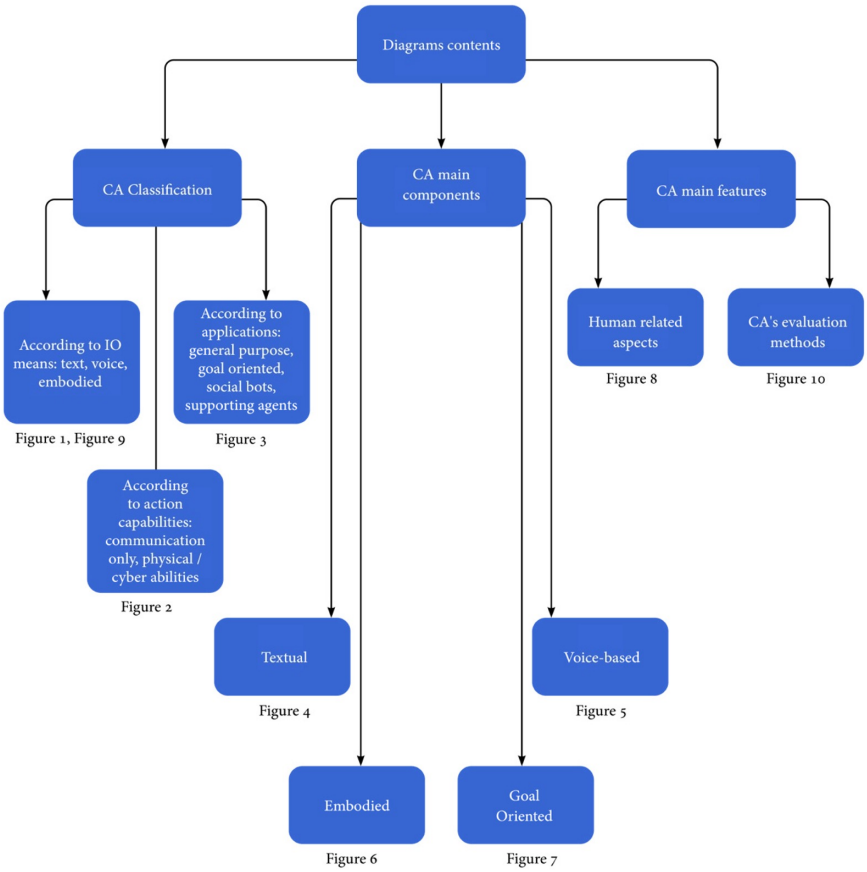

Figure 11. A summary of all diagrams.

There are, however, various additional situations where CAs can be utilized to assist and support people. With state-of-the-art CAs, the most advanced improve themselves based on new data. There are very few CAs, however, that allow humans to teach them additional knowledge and new capabilities or to provide them with the ability to direct their learning process. One of the few systems that can learn directly from humans is commonsense reasoning by instruction (CORGI) [248]. CORGI performs the commonsense reasoning required in applying if-then rules, by initiating a conversation with the user. Another example is Safebot [249], which is taught new responses by the user to avoid learning inappropriate responses. Finally, the learning-by-instruction agent (LIA) [250] asks the user to explain how to execute a new command and associates a sequence of natural-language steps with it. Such systems enable users to fine-tune CAs to adapt them to personal needs and preferences. To further enhance such systems, additional appropriate protocols, algorithms, and rules should be developed and examined.

Another domain where CAs may be useful is in explanatory interactive systems [251,252], which aim to explain to humans the reasons behind decisions made by an automated system. Such explanations are necessary to strengthen the trust between agents and people. CAs may be used to make machine explanations understandable to the human user.

Another area in which CAs are expected to be more prominent is related to consulting a person during his/her conversations. Such a consulting agent would be expected to

support people in their daily interactions with other people. The agent is required to model all participants of the conversation to identify their needs in complex social situations to be able to advise them on how to act, talk, or respond in complex social interactions. In our ongoing study [100,241], technology is being developed to assist children with special needs in their daily interaction while monitoring the environment for them.

It should also be emphasized that as CAs become ubiquitous and their ability to provide human-like responses improves, a significant moral question arises: Is there a need to declare the identity of the service or the technical-support representative? Do CAs acting as support or sales agents have the obligation to share their nature with the clients? While studies have revealed that people feel more engaged when conversing with other humans [97], it remains questionable whether maintaining the obscurity of the agent is right, fair, or justified [253].

Another related moral issue arises when considering influential agents. Considering the current state of the technology, any company, party, or ideological movement may develop a CA as a representative to describe its agenda and influence public opinion to garner support for its position. To what extent is such a practice considered moral? Situations where the CA identity is known or hidden should be distinguished, and situations where the company or party is represented by a single CA or by several, hundreds, or even thousands, to create a representation of mass support should be carefully considered and clarified. Surely, using a mass of CAs to influence public opinion seems to be dishonest and unfair, but where is the moral limit?

In addition, given the possibility of such an unfair usage of influence agents, technology should be developed to be able to detect such unfair influence. In Section 6.5, some studies are described that deal with detecting malicious "influence bots." As the technological ability of such influence bots increases, detecting them becomes more challenging. However, such detection may be crucial, especially when considering extreme groups that may have incentives to utilize such agents for negative purposes.

Several issues arise by the use of assistant agents related to the challenges of protecting user privacy. Mainly, assistant-agent developers must prevent the use of information acquired by the assistance agent by other parties, such as, commercial companies and adversaries. Information-security technologies should be employed to avoid such situations.

To summarize, the rise of CAs and their applications can have a significant influence on our future life. Some of these applications are positive and even crucial, such as health support or social support; others can be beneficial to business and companies; and others should be monitored or even avoided for moral reasons. The limits of fair use of CAs and the technological tools to enforce these limits should be discussed and developed in future research.

Abbreviations

The following abbreviations are used in this manuscript:

AGATA	Automatic generation of IAML from text acquisition
ASD	Autistic spectrum disorder
ASK	Alexa Skills Kit
AI	Artificial intelligence
AIML	Artificial-intelligence Markup Language
ASR	Automatic speech recognition
ASRU	Automatic speech recognition
B2B	Business to business
CA	Conversational agents
CCG	Combinatory categorial grammar
CFG	Context-free grammar
CORGI	Commonsense reasoning by instruction
CTGAN	Conditional text generative adversarial network
DAN	Deep average network
DBN	Dynamic Bayesian network
DNN	Deep neural network
DSTC	Dialogue-state-tracking Challenge
DOAJ	Directory of open-access journals
DRL	Deep reinforcement learning
DRQN	Deep recurrent QNetwork
DSTC	Dialogue system technology challenge
ECA	Embodied conversational agent
ED	Emotion detection
EQ	Emotional quotient
FAQ	Frequently asked questions
GAN	Generative adversarial network
HQ	Hedonic quality
HRED	Hierarchical recurrent encoder–decoder
IoT	Internet of Things
IQ	Intelligence quotient
IR	Information retrieval
IRIS	Informal response interactive system
IS	Information systems
ITS	Intelligent tutoring systems
IVR	Interactive voice response
JA	Joint attention
LD	Linear dichroism
LIA	Learning by instruction agent
LSA	Latent semantic analysis
LSTM	Long short-term memory

MDP	Markov decision process
MDPI	Multidisciplinary Digital Publishing Institute
ML	Machine learning
MMI	Maximum mutual information
MOOC	Massive open online course
MT	Machine translation
NBT	Neural belief tracking
NLG	Natural-language generation
NLP	Natural-language processing
NLU	Natural-language understanding
PCFG	Probabilistic context-free grammar
POS	Part-of-speech
PBD	Programming-by-demonstration
RNN	Recurrent neural network
ROUGE	Recall-oriented understudy for gisting evaluation
SAR	Socially assistive robotics
SCE	Socio-cognitive engineering
SGD	Schema-guided dialogue
SL	Sign language
SQUAD	Stanford question-answering dataset
SSA	Sensibleness and specificity average
SVM	Support vector machine
TF-IDF	Term frequency inverse document frequency
TLA	Three-letter acronym
UX	User experience

Author Contributions: Conceptualization, writing - original draft preparation, A.A., M.A. and R.A.; writing - review and editing, A.A.; visualization, R.A.; supervision, A.A. and R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Ministry of Science, Technology & Space, Israel.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bosker, B. Siri rising: The inside story of siri's origins–and why she could overshadow the iphone. *Huffington Post* https://www.huffpost.com/entry/siri-do-engine-apple-iphone_n_2499165 (accessed on 14 12 2021)
2. Adiwardana, D.; Luong, M.T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; others. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* **2020**.
3. Bhat, H.R.; Lone, T.A.; Paul, Z.M. Cortana-intelligent personal digital assistant: a review. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 55–57.
4. Adamopoulou, E.; Moussiades, L. Chatbots: History, Technology, and Applications. *Machine Learning with Applications* **2020**, *2*, 100006.
5. Adamopoulou, E.; Moussiades, L. An overview of chatbot technology. In Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations, Neos Marmaras, Greece, 5-7 June 2020; Springer Nature, Switzerland AG 2020, pp. 373–383.
6. Nuruzzaman, M.; Hussain, O.K. A survey on chatbot implementation in customer service industry through deep neural networks. In Proceedings of the 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), Xi'an, China, 2–14 October 2018, IEEE: Manhattan, New York, USA, 2018, pp. 54–61.
7. Borah, B.; Pathak, D.; Sarmah, P.; Som, B.; Nandi, S. Survey of Textbased Chatbot in Perspective of Recent Technologies. In Proceedings of the International Conference on Computational Intelligence, Communications, and Business Analytics, Kalyani, India, 27–28, July 2018, Springer, Switzerland AG, 2018, pp. 84–96.
8. Chen, H.; Liu, X.; Yin, D.; Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *ACM Sigkdd Explorations Newsletter* **2017**, *19*, 25–35.
9. Jianfeng Gao, M.G.; Li, L. Neural Approaches to Conversational AI. *arXiv preprint. arXiv* **2019**, *arXiv:1809.08267*.
10. Diederich, S.; Brendel, A.B.; Kolbe, L.M. On Conversational Agents in Information Systems Research: Analyzing the Past to Guide Future Work. In Proceedings of the 14th International Conference on Wirtschaftsinformatiks, 24-27 February, 2019, Siegen, Germany.

11. Meyer von Wolff, R.; Hobert, S.; Schumann, M. How may i help you?–state of the art and open research questions for chatbots at the digital workplace. In Proceedings of the 52nd Hawaii international conference on system sciences, Honolulu, Hawaii, USA, 8-11 January 2019.
12. Vishnoi, L. Conversational Agent: A More Assertive Form of Chatbots, 2020. Available online: https://towardsdatascience.com/conversatio agent-a-more-assertive-form-of-chatbots-de6f1c8da8dd (accessed on 14 12 2021).
13. Nuseibeh, R. What is a Chatbot?, 2018. https://medium.com/rajai_nuseibeh/what-is-a-chatbot-402427354f44 (accessed on 14 12 2021).
14. Radziwill, N.; Benton, M. Evaluating Quality of Chatbots and Intelligent Conversational Agents. *Software Quality Professional* **2017**, *19*, 25.
15. Hussain, S.; Sianaki, O.A.; Ababneh, N. A survey on conversational agents/chatbots classification and design techniques. In Proceedings of the Workshops of the International Conference on Advanced Information Networking and Applications, Matsue, Japan, 27-29 March 2019; Springer Nature, Switzerland AG, 2019, pp. 946–956.
16. Masche, J.; Le, N.T. A review of technologies for conversational systems. In Proceedings of the International conference on computer science, applied mathematics and applications Berlin, Germany, June 30 - July 1, 2017; Springer Nature, Switzerland AG, 2017, pp. 212–225.
17. Nimavat, K.; Champaneria, T. Chatbots: An overview types, architecture, tools and future possibilities. *Int. J. Sci. Res. Dev* **2017**, *5*, 1019–1024.
18. Venkatesh, A.; Khatri, C.; Ram, A.; Guo, F.; Gabriel, R.; Nagar, A.; Prasad, R.; Cheng, M.; Hedayatnia, B.; Metallinou, A.; Goel, R.; Yang, S.; Raju1, A. On Evaluating and Comparing Conversational Agents. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), long Beach, CA, 4-9 Dec 2017.
19. Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9.1*, 36–45.
20. Breazeal, C. Social robots: from research to commercialization. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Vienna, Austria, 6-9 March 2017; p. 1.
21. Gehl, R.W. Teaching to the Turing Test with Cleverbot. *The Journal of Inclusive Scholarship and Pedagogy* **2014**, *24*, 56–66.
22. Hill, J.; Randolph Ford, W.; Farreras, I.G. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* **2015**, *49*, 245–250.
23. Lopatovska, I.; Rink, K.; Knight, I.; Raines, K.; Cosenza, K.; Williams, H.; Sorsche, P.; Hirsch, D.; Li, Q.; Martinez, A. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* **2019**, *51*, 984–997.
24. Zhu, Q.; Zhang, Z.; Fang, Y.; Li, X.; Takanobu, R.; Li, J.; Peng, B.; Gao, J.; Zhu, X.; Huang, M. Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. *arXiv preprint* **2020**, *arXiv:2002.04793* .
25. Taskbot, A.P. Alexa Prize Taskbot, 2021. https://developer.amazon.com/alexaprize (accessed on 14 12 2021).
26. Fernandes, A. NLP, NLU, NLG and how Chatbots work. https://chatbotslife.com/nlp-nlu-nlg-and-how-chatbots-work-dd7861dfc9df (accessed on 14 12 2021).
27. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *arXiv preprint* **2017**, arXiv:1708.05148 .
28. Stoner, D.J.; Ford, L.; Ricci, M. Simulating military radio communications using speech recognition and chat-bot technology. *The Titan Corporation: Orlando, USA* **2004**.
29. Abdul-Kader, S.A.; Woods, J. Survey on Chatbot Design Techniques in Speech Conversation Systems. Int. J. of Adv. Comput. Sci. and App., 6(7), 2015, Science and Information Organization, UK.
30. Ramesh, K.; Ravishankaran, S.; Joshi, A.; Chandrasekaran, K. A Survey of Design Techniques for Conversational Agents. In Proceedings of the 2017 ICICCT Information, Communication and Computing Technology,New Delhi, India, May 13, 2017; pp. 336–350.
31. Ahmad, N.A.; Hamid, M.H.C.; Zainal, A.; Rauf, M.F.A.; Adnan, Z. Review of Chatbots Design Techniques. *Int. J. Comput. Appl.* **2018**, *181*, 56–67.
32. Diederich, S.; Brendel, A.B.; Kolbe, L.M. Towards a taxonomy of platforms for conversational agent design. WI 2019, 2019. https://aisel.aisnet.org/wi2019/track10/papers/1/ (accessed on 14 12 2021
33. A.S., L.; M.A., A. Modern Chatbot Systems: A Technical Review. In Proceedings of the Future Technologies Conference (FTC), San Francisco, USA, 25-26 October 2019; Springer Nature, Switzerland AG, 2019; Vol. 881, pp. 1012–1023.
34. Azaria, A.; Nivasch, K. SAIF: A Correction-Detection Deep-Learning Architecture for Personal Assistants. *Sensors* **2020**, *20*, 5577.
35. Saund, E. How Do Conversational Agents Answer Questions? https://towardsdatascience.com/how-do-conversational-agents-answer-questions-d504d37ef1cc (accessed on 14 12 2021).
36. Benzeghiba, M.; De Mori, R.; Deroo, O.; Dupont, S.; Erbes, T.; Jouvet, D.; Fissore, L.; Laface, P.; Mertins, A.; Ris, C.; others. Automatic speech recognition and speech variability: A review. *Speech communication* **2007**, *49*, 763–786.
37. Yu, D.; Deng, L. *Automatic Speech Recognition*; Springer Nature, Switzerland AG, 2016.
38. Sadeghipour, A.; Kopp, S. Embodied gesture processing: Motor-based integration of perception and action in social artificial agents. *Cognitive computation* **2011**, *3*, 419–435.

39. Krishnaswamy, N.; Narayana, P.; Wang, I.; Rim, K.; Bangar, R.; Patil, D.; Mulay, G.; Beveridge, R.; Ruiz, J.; Draper, B.; others. Communicating and acting: Understanding gesture in simulation semantics. In Proceedings of the 12th International Conference on Computational Semantics (IWCS), Montpellier, France, 19-22 September 2017.
40. Homburg, D.; Thieme, M.S.; Völker, J.; Stock, R. RoboTalk-Prototyping a Humanoid Robot as Speech-to-Sign Language Translator. In Proceedings of the 52nd Hawaii International Conference on System Sciences, Honolulu, Hawaii, USA, 8-11 January 2019.
41. Singh, S.; Jain, A.; Kumar, D. Recognizing and interpreting sign language gesture for human robot interaction. *Int. J. Comput. Appl.* **2012**, *52*.
42. Beck, A.; Stevens, B.; Bard, K.A.; Cañamero, L. Emotional body language displayed by artificial agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* **2012**, *2*, 1–29.
43. Zhao, T.; Eskenazi, M. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560* **2016**.
44. Noroozi, V.; Zhang, Y.; Bakhturina, E.; Kornuta, T. A Fast and Robust BERT-based Dialogue State Tracker for Schema-Guided Dialogue Dataset. *CoRR* **2020**, *abs/2008.12335*, [2008.12335].
45. Bird, S.; Klein, E.; Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*; " O'Reilly Media, Inc.": Sebastopol, California 2009.
46. Navigli, R. Natural Language Understanding: Instructions for (Present and Future) Use. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13-19 July 2018, p. 5697–5702.
47. Inui, N.; Koiso, T.; Nakamura, J.; Kotani, Y. Fully corpus-based natural language dialogue system. In Proceedings of the Natural Language Generation in Spoken and Written Dialogue, AAAI Spring Symposium, Palo Alto, California, 24–26 March 2003.
48. Wallace, R.S. The anatomy of ALICE. In *Parsing the Turing Test*; Springer Nature, Switzerland AG, 2009; pp. 181–210.
49. Marietto, M.d.G.B.; de Aguiar, R.V.; Barbosa, G.d.O.; Botelho, W.T.; Pimentel, E.; França, R.d.S.; da Silva, V.L. Artificial intelligence markup language: a brief tutorial. *arXiv preprint arXiv:1307.3091* **2013**.
50. Agostaro, F.; Augello, A.; Pilato, G.; Vassallo, G.; Gaglio, S. A conversational agent based on a conceptual interpretation of a data driven semantic space. In Proceedings of the Congress of the Italian Association for Artificial Intelligence, Milan, Italy, 21-23 September 2005, Springer Nature, Switzerland AG, 2005; pp. 381–392.
51. Banchs, R.E.; Li, H. IRIS: a chat-oriented dialogue system based on the vector space model. In Proceedings of the ACL 2012 System Demonstrations, Jeju, Republic of Korea, 8-14 July 2012; pp. 37–42.
52. Nijholt, A. *Context-free grammars: covers, normal forms, and parsing (Lecture Notes in Computer Science, 93).* Springer Science and Business Media, Springer-Verlag Berlin Heidelberg 1980.
53. Resnik, P. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, 23-28 August 1992, Volume 1, 1992.
54. Gandhe, A.; Rastrow, A.; Hoffmeister, B. Scalable language model adaptation for spoken dialogue systems. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18-21 December 2018; IEEE: Manhattan, New York, USA, 2018, pp. 907–912.
55. Azaria, A.; Srivastava, S.; Krishnamurthy, J.; Labutov, I.; Mitchell, T.M. An agent for learning new natural language commands. *Autonomous Agents and Multi-Agent Systems* **2020**, *34*, 1–27.
56. Bocklisch, T.; Faulkner, J.; Pawlowski, N.; Nichol, A. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181* **2017**.
57. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25-29 October 2014, pp. 1532–1543.
58. Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.
59. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning (ICML 2001), Williamstown, MA, USA. pp. 282-289, 28 June 2001- 1 July 2001, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
60. Lee, S.; Zhu, Q.; Takanobu, R.; Zhang, Z.; Zhang, Y.; Li, X.; Li, J.; Peng, B.; Li, X.; Huang, M.; Gao, J. ConvLab: Multi-Domain End-to-End Dialog System Platform. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; Association for Computational Linguistics, Florence, Italy, July 28 - August 2 2019; pp. 64–69. doi:10.18653/v1/P19-3011.
61. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
62. McTear, M. The Role of Spoken Dialogue in User–Environment Interaction. *Human-Centric Interfaces for Ambient Intelligence* **2010**, pp. 225–254.
63. Harms, J.G.; Kucherbaev, P.; Bozzon, A.; Houben, G.J. Approaches for dialog management in conversational agents. *IEEE Internet Computing* **2018**, *23*, 13–22, IEEE: Manhattan, New York, USA.
64. Nguyen, A.; Wobcke, W. An agent-based approach to dialogue management in personal assistants. In Proceedings of the 10th international conference on Intelligent user interfaces, San Diego, California, USA, 10-13 January 2005; pp. 137–144.
65. Moore, R.C.; Dowding, J.; Bratt, H.; Gawron, J.M.; Gorfu, Y.; Cheyer, A. CommandTalk: A spoken-language interface for battlefield simulations. In Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, USA, March 31 - April 3, 1997 1997; pp. 1–7.

66. Stent, A.; Dowding, J.; Gawron, J.M.; Bratt, E.O.; Moore, R.C. The CommandTalk spoken dialogue system. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, 20-26 June 1999, pp. 183–190.
67. MindMeld. Introducing MindMeld. https://www.mindmeld.com/docs/intro/introducing_mindmeld.html (accessed on 14 12 2021).
68. Klopfenstein, L.C.; Delpriori, S.; Ricci, A. Adapting a conversational text generator for online chatbot messaging. In Proceedings of the International Conference on Internet Science, St. Petersburg, Russia, October 24-26, 2018; Springer Nature, Switzerland AG, 2019, pp. 87–99.
,
69. Building and deploying a chatbot by using Dialogflow (overview). https://cloud.google.com/solutions/building-and-deploying-chatbot-dialogflow (accessed on 14 12 2021).
70. Williams, J.D.; Kamal, E.; Ashour, M.; Amr, H.; Miller, J.; Zweig, G. Fast and easy language understanding for dialog systems with Microsoft Language Understanding Intelligent Service (LUIS). In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czech Republic, 2-4 September 2015, pp. 159–161.
71. Henderson, M.; Thomson, B.; Young, S. Word-based dialog state tracking with recurrent neural networks. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Philadelphia, PA, USA, 18-20 June 2014, pp. 292-299.
72. Singh, S.P.; Kearns, M.J.; Litman, D.J.; Walker, M.A. Reinforcement learning for spoken dialogue systems. Advances in neural information processing systems, 12, 956-962.
73. Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; Jurafsky, D. Deep Reinforcement Learning for Dialogue Generation. *arXiv preprint arXiv:1606.01541* **2016**.
74. Serban, I.V.; Sankar, C.; Germain, M.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Pieper, M.; Chandar, S.; Ke, N.R.; others. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349* **2017**.
75. Reiter, E.; Dale, R. *Building Applied Natural Language Generation Systems*; Natural Language Engineering, 3(1), 1997, 57-87, Cambridge University Press, Cambridge, UK.
76. Gatt, A.; Krahmer, E. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61, 2018, 65-170, AAAI Press, Palo Alto, California, U.S.
77. van Deemter, K.; Krahmer, E.; Theune, M. Squibs and Discussions: Real versus Template-Based Natural Language Generation: A False Opposition? *Computational Linguistics* **2005**, *31*, 15–24, MIT Press, Cambridge, Massachusetts, USA.
78. Wen, T.H.; Gašić, M.; Mrkšić, N.; Su, P.H.; Vandyke, D.; Young, S. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics, Lisbon, Portugal, 17-21 September 2015; pp. 1711–1721. doi:10.18653/v1/D15-1199.
79. Tran, V.K.; Nguyen, L.M.; Tojo, S. Neural-based Natural Language Generation in Dialogue using RNN Encoder-Decoder with Semantic Aggregation. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue; Association for Computational Linguistics: Saarbrücken, Germany, Saarbruecken, Germany, 15-17 August 2017 2017; pp. 231–240. doi:10.18653/v1/W17-5528.
80. Juraska, J.; Karagiannis, P.; Bowden, K.; Walker, M. A Deep Ensemble Model with Slot Alignment for Sequence-to-Sequence Natural Language Generation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA, 1-6 June 2018; Association for Computational Linguistics: New Orleans, Louisiana, USA, 2018; pp. 152–162. doi:10.18653/v1/N18-1014.
81. Dušek, O.; Novikova, J.; Rieser, V. Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge. *Computer Speech and Language* **2020**, *59*, 123–156, Elsevier, Amsterdam, Netherlands. doi:https://doi.org/10.1016/j.csl.2019.06.009.
82. Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.Y.; Gao, J.; Dolan, B. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 – June 5, 2015, Association for Computational Linguistics: Denver, Colorado, 2015; pp. 196–205. https://www.aclweb.org/anthology/N15-1020, doi:10.3115/v1/N15-1020.
83. Mikolov, T.; Zweig, G. Context dependent recurrent neural network language model. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, 2-5 December 2012; IEEE: Manhattan, New York, USA, 2012, pp. 234–239.
84. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, B. A Diversity-Promoting Objective Function for Neural Conversation Models. *arXiv preprint* **2015**, *arXiv:1510.03055*.
85. Serban, I.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, 12–17 February 2016, Vol. 30.
86. He, S.; Liu, C.; Liu, K.; Zhao, J. Generating Natural Answers by Incorporating Copying and Retrieving Mechanisms in Sequence-to-Sequence Learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, July 30 - August 4 2017; pp. 199–208.

87. Qiu, M.; Li, F.L.; Wang, S.; Gao, X.; Chen, Y.; Zhao, W.; Chen, H.; Huang, J.; Chu, W. Alime chat: A sequence to sequence and rerank based chatbot engine. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, Canada, July 30 - August 4 2017, pp. 498–503.
88. Ghazvininejad, M.; Brockett, C.; Chang, M.W.; Dolan, B.; Gao, J.; tau Yih, W.; Galley, M. A Knowledge-Grounded Neural Conversation Model. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, 2-7 February 2018.
89. Ham, D.; Lee, J.G.; Jang, Y.; Kim, K.E. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5-10 July 2020; Association for Computational Linguistics, pp. 583–592. doi:10.18653/v1/2020.acl-main.54.
90. Kim, J.; Ham, D.; Lee, J.G.; Kim, K.E. End-to-End Document-Grounded Conversation with Encoder-Decoder Pre-Trained Language Model. In Proceedings of the DSTC9 Workshop, Online, 8-9 February 2021.
91. Das, A.; Kottur, S.; Moura, J.M.; Lee, S.; Batra, D. Learning cooperative visual dialog agents with deep reinforcement learning. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22-29 Oct 2017, pp. 2951–2960, IEEE: Manhattan, New York, USA.
92. Zhang, Z.; Takanobu, R.; Huang, M.; Zhu, X. Recent Advances and Challenges in Task-oriented Dialog System. *CoRR* **2020**, *abs/2003.07490*, [2003.07490].
93. Kim, A.; Song, H.J.; Park, S.B.; others. A two-step neural dialog state tracker for task-oriented dialog processing. *Computational Intelligence and Neuroscience* **2018**, *2018*.
94. Mrksic, N.; Seaghdha, D.O.; Wen, T.H.; Thomson, B.; Young, S.J. Neural Belief Tracker: Data-Driven Dialogue State Tracking. ACL, Vancouver, Canada, July 30 - August 4 2017, pp. 1777–1788. https://doi.org/10.18653/v1/P17-1163.
95. Su, P.H.; Vandyke, D.; Gasic, M.; Kim, D.; Mrksic, N.; Wen, T.H.; Young, S. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. *arXiv preprint arXiv:1508.03386* **2015**.
96. Liu, B.; Lane, I. Iterative policy learning in end-to-end trainable task-oriented neural dialog models. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16-20 December 2017. IEEE: Manhattan, New York, U.S., 2017, pp. 482–489.
97. Clark, L.M.H.; Pantidi, N.; Cooney, O.; Garaialde, P.R.D.D.; Edwards, J.; Spillane, B.; Gilmartin, E.; Murad, C.; Munteanu, C. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In Proceedings of the 2019 CHI Conference, Glasgow, Scotland, Uk, 4-9 May 2019.
98. Yang, X.; Aurisicchio, M.; Baxter, W. Understanding Affective Experiences With Conversational Agents. In Proceedings of the 2019 CHI Conference, Glasgow, Scotland, Uk, 4-9 May 2019.
99. Acheampong, F.A.; Wenyu, C.; Nunoo-Mensah, H. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports* **2020**, *2*, e12189, Wiley Online Library, NY, USA.
100. Allouch, M.; Azaria, A.; Azoulay, R.; Ben-Izchak, E.; Zwilling, M.; Zachor, D.A. Automatic detection of insulting sentences in conversation. In Proceedings of the 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), Eilat, Israel, 12-14 December 2018. IEEE: Manhattan, New York, USA, 2018, pp. 1–4.
101. Schlesinger, A.; O'Hara, K.P.; Taylor, A.S. Let's talk about race: Identity, chatbots, and AI. In Proceedings of the 2018 CHI conference on human factors in computing systems, Montreal, QC, Canada, 21-26 April 2018, pp. 1–14.
102. Sarder, M.A. ECActive Embodied Conversational Agent for Mental Health Intervention. Master's Thesis, Delft University of Technology, Delft, Netherlands, 2018.
103. Yalçın, Ö.N. Empathy framework for embodied conversational agents. *Cognitive Systems Research* **2020**, *59*, 123–132, Elsevier, Amsterdam, Netherlands.
104. Puig, D.T.M.L.S.I.R.P.A.A. Enhancing sentient embodied conversational agents with machine learning. *Pattern Recognition Letters* **2020**, *129*, 317–323, Elsevier, Amsterdam, Netherlands.
105. McLeod, S. Maslow's hierarchy of needs. *Simply psychology* **2007**, *1*. https://www.simplypsychology.org/maslow.html (accessed on 14 12 2021)
106. Chen, J.; Wu, Y.; Jia, C.; Zheng, H.; Huang, G. Customizable text generation via conditional text generative adversarial network. *Neurocomputing* **2020**, *416*, 125–135. doi:https://doi.org/10.1016/j.neucom.2018.12.092.
107. Zhou, L.; Gao, J.; Li, D.; Shum, H.Y. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* **2020**, *46*, 53–93, MIT Press, Cambridge, Massachusetts, USA
108. Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; Mou, L. Affective neural response generation. In Proceedings of the European Conference on Information Retrieval. Grenoble, France, 26-29 March 2018, Springer Nature, Switzerland AG, pp. 154–166.
109. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018 2018, Vol. 32.
110. Chaves, A.P.; Gerosa, M.A. How should my chatbot interact? A survey on human-chatbot interaction design, 2020. arXiv preprint https://arxiv.org/abs/1904.02743. **2020**.
111. Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. Personalizing Dialogue Agents: I have a dog, do you have pets too?, 2018. arXiv preprint https://arxiv.org/abs/1709.02349. **2018**

112. Völkel, S.T.; Schödel, R.; Buschek, D.; Stachl, C.; Winterhalter, V.; Bühner, M.; Hussmann, H. Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach. CHI 2020, 2020, pp. 1–14.
113. Roccas, S.; Sagiv, L.; Schwartz, S.H.; Knafo, A. The Big Five Personality Factors and Personal Values. *Personality and Social Psychology Bulletin* **2002**, *28*, 789–801. doi:10.1177/0146167202289008.
114. Feine, J.; Gnewuch, U.; Morana, S.; Maedche, A. A Taxonomy of Social Cues for Conversational Agents. *Int. J. Human-Comput. Stud.* **2019**, *132*, 138–161. https://www.sciencedirect.com/science/article/pii/S1071581918305238, doi:https://doi.org/10.1016/j.ijhcs.2019.07.009.
115. Burgoon, J.; Guerrero, L.; Manusov, V., Nonverbal signals. In *The SAGE Handbook of Interpersonal Communication*; SAGE Publications: Thousand Oaks, CA, USA, 2011; pp. 239–282.
116. Liao, Y.; He, J. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. IUI 20, 2020, pp. 430–442.
117. Go, E.; Sundar, S.S. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* **2019**, *97*, 304–316. doi:https://doi.org/10.1016/j.chb.2019.01.020.
118. Smith, E.M.; Williamson, M.; Shuster, K.; Weston, J.; Boureau, Y.L. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. *arXiv preprint arXiv:2004.08449* **2020**.
119. Ferland, L.; Koutstaal, W. How's Your Day Look? The (Un)Expected Sociolinguistic Effects of User Modeling in a Conversational Agent. CHI 2020), 2020, pp. 482–489. doi:10.1109/ASRU.2017.8268975.
120. Carfora, V.; Massimo, F.D.; Rastelli, R.; Catellani, P.; Piastra, M. Dialogue management in conversational agents through psychology of persuasion and machine learning. *Multimedia Tools and Applications volume* **2020**, *79*, 35949–35971.
121. Ajzen, I. The theory of planned behavior. *Organizational behavior and human decision processes* **1991**, *50*, 179–211.
122. Rina Azoulay, Esther David, M.A.; Hutzler, D., Adaptive Task Selection in Automated Educational Software: A Comparative Study. In *Intelligent Systems and Learning Data Analytics in Online Education*; Elsevier, Amsterdam, Netherlands 2021; chapter 7.
123. Azevedo, R.; Landis, R.S.; Feyzi-Behnagh, R.; Duffy, M.; Trevors, G.; Harley, J.M.; Bouchet, F.; Burlison, J.; Taub, M.; Pacampara, N.; others. The effectiveness of pedagogical agents' prompting and feedback in facilitating co-adapted learning with MetaTutor. In Proceedings of the International conference on intelligent tutoring systems, Chania, Crete, Greece, 14-18 June 2012. Springer Nature, Switzerland AG, 2012, pp. 212–221.
124. Ueno, M.; Miyazawa, Y. IRT-based adaptive hints to scaffold learning in programming. *IEEE Transactions on Learning Technologies* **2017**, *11*, 415–428, IEEE: Manhattan, New York, USA.
125. Winkler, R.; Hobert, S.; Salovaara, A.; Söllner, M.; Leimeister, J.M. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In Proceedings of the 2020 CHI conference on human factors in computing systems, Honolulu, HI, 26 April 2020, pp. 1–14.
126. Ni, L.; Lu, C.; Liu, N.; Liu, J. Mandy: Towards a smart primary care chatbot application. In Proceedings of the International symposium on knowledge and systems sciences, Bangkok, Thailand, 17-19 November 2017. Springer Nature, Switzerland AG, pp. 38–52.
127. Schuetzler, R.M.; Grimes, G.M.; Giboney, J.S.; Nunamaker Jr, J.F. The influence of conversational agents on socially desirable responding. In Proceedings of the 51st Hawaii International Conference on System Sciences, Hawaii, USA, 3-6 January 2018, p. 283.
128. Colby, K.M. Ten criticisms of parry. *ACM SIGART Bulletin* **1974**, *48*, 5–9.
129. Yin, Z.; Chang, K.h.; Zhang, R. Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13-17 August 2017, pp. 2131–2139.
130. Liu, H.; Lin, T.; Sun, H.; Lin, W.; Chang, C.W.; Zhong, T.; Rudnicky, A. Rubystar: A non-task-oriented mixture model dialog system. *arXiv preprint arXiv:1711.02781* **2017**.
131. Hoy, M.B. Human-Aided Bots. *Medical Reference Services Quarterly* **2018**, *37*, 81–88. doi:10.1080/02763869.2018.1404391.
132. Azaria, A.; Krishnamurthy, J.; Mitchell, T. Instructable intelligent personal agent. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, 12-17 February 2016, Vol. 30.
133. Li, T.J.J.; Azaria, A.; Myers, B.A. SUGILITE: creating multimodal smartphone automation by demonstration. Proceedings of the 2017 CHI conference on human factors in computing systems, Denver, CO, USA, 06-11 May 2017, pp. 6038–6049.
134. Chkroun, M.; Azaria, A. Safebot: A safe collaborative chatbot. In Proceedings of the AAAI Workshops, New Orleans, Louisiana, USA, 2-7 February 2018.
135. Ait-Mlouk, A.; Jiang, L. KBot: a Knowledge graph based chatBot for natural language understanding over linked data. *IEEE Access* **2020**, *8*, 149220–149230.
136. Paladines, J.; Ramirez, J. A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access* **2020**, *8*, 164246–164267.
137. Paschoal, L.N.; Krassmann, A.L.; Nunes, F.B.; de Oliveira, M.M.; Bercht, M.; Barbosa, E.F.; de Souza, S.d.R.S. A Systematic Identification of Pedagogical Conversational Agents. In Proceedings of the 2020 IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 21-24 October 2020. IEEE: Manhattan, New York, U.S., 2020, pp. 1–9.
138. Rainer, W.; Sebastian, H.; Salovaara, A.; Matthias, S.; Marco, L.J., Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2020; p. 1–14. https://doi.org/10.1145/3313831.3376781.

139. Paschoal, L.N.; Turci, L.F.; Conte, T.U.; Souza, S.R. Towards a conversational agent to support the software testing education. In Proceedings of the 33th Brazilian Symposium on Software Engineering, Salvador, Brazil, 23-27 September 2019, pp. 57–66.
140. Graesser, A.C.; Wiemer-Hastings, K.; Wiemer-Hastings, P.; Kreuz, R.; Group, T.R.; others. AutoTutor: A simulation of a human tutor. *Cognitive Systems Research* **1999**, *1*, 35–51.
141. Abdellatif, A.; Badran, K.; Shihab, E. MSRBot: Using bots to answer questions from software repositories. *Empirical Software Engineering* **2020**, *25*, 1834–1863.
142. Hobert, S. Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program. *Digital Learning Environment and Future IS Curriculum* **2019**.
143. Kloos, C.D.; Catálan, C.; Muñoz-Merino, P.J.; Alario-Hoyos, C. Design of a conversational agent as an educational tool. In Proceedings of the 2018 Learning With MOOCS (LWMOOCS), Madrid, Spain, 26-28 September 2018. IEEE: Manhattan, New York, USA, pp. 27–30.
144. Aguirre, C.C.; Kloos, C.D.; Alario-Hoyos, C.; Muñoz-Merino, P.J. Supporting a MOOC through a conversational agent. Design of a first prototype. In Proceedings of the 2018 International Symposium on Computers in Education (SIIE) Cadiz, Spain, 19-21 Sept 2018. IEEE: Manhattan, New York, U.S., 2018, pp. 1–6.
145. Assistant, G. Google Assistant, your own personal Google. https://assistant.google.com/ (accessed on 14 12 2021).
146. Lin, P.; Van Brummelen, J.; Lukin, G.; Williams, R.; Breazeal, C. Zhorai: Designing a Conversational Agent for Children to Explore Machine Learning Concepts. In Proceedings of the AAAI Conference on Artificial Intelligence, NY, USA, 7-12 February 2020; Vol. 34, pp. 13381–13388.
147. Cai, Grossman, Lin, Sheng, Wei, Williams, and Goel] Cai, W., Grossman, J., Lin, Z.J., Sheng, H., Wei, J.T.Z., Williams, J.J., and Goel, S. Bandit algorithms to personalize educational chatbots. Machine Learning, 110(2), 2021, 1-30. Springer Nature, Switzerland AG
148. Kim, N.Y.; Cha, Y.; Kim, H.S. Future English learning: Chatbots and artificial intelligence. *Multimedia-Assisted Language Learning* **2019**, *22*, 32–53.
149. MARIA, A. Got an Alexa? You've Got a Polyglot Tutor That Can Teach You a Language. https://www.fluentu.com/blog/can-alexa-teach-languages/ (accessed on 14 12 2021).
150. Pham, X.L.; Pham, T.; Nguyen, Q.M.; Nguyen, T.H.; Cao, T.T.H. Chatbot as an intelligent personal assistant for mobile language learning. In Proceedings of the 2018 2nd International Conference on Education and E-Learning, 2018, pp. 16–21.
151. Fei, W.Y.; Petrina, S. Using learning analytics to understand the design of an intelligent language tutor–Chatbot lucy. *Editorial Preface* **2013**, *4*, 124–131.
152. Hien, H.T.; Pham-Nguyen, C.; Nam, L.N.H.; Dinh, T.L. Intelligent Assistants in Higher-Education Environments: The FIT-EBot, a Chatbot for Administrative and Learning Support. In Proceedings of the ninth international symposium on information and communication technology, Danang City, Viet Nam, 6-7 December 2018, Association for Computing Machinery, NY, USA, pp. 69-76.
153. Ranoliya, B.R.; Raghuwanshi, N.; Singh, S. Chatbot for university related FAQs. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Manipal, India, 13-16 Sep 2017. IEEE: Manhattan, New York, U.S., 2017, pp. 1525–1530.
154. Lee, K.; Jo, J.; Kim, J.; Kang, Y. Can Chatbots Help Reduce the Workload of Administrative Officers?-Implementing and Deploying FAQ Chatbot Service in a University. In Proceedings of the International Conference on Human-Computer Interaction, Orlando, Florida, USA, 26-31 July 2019. Springer Nature, Switzerland AG, 2019, pp. 348–354.
155. Feng, D.; Shaw, E.; Kim, J.; Hovy, E. An intelligent discussion-bot for answering student queries in threaded discussions. In Proceedings of the 11th international conference on Intelligent user interfaces, Sydney, Australia, January 29 - February 1, 2006, pp. 171–177.
156. LI, X.; Zhong, H.; Zhang, B.; Zhang, J. A General Chinese Chatbot based on Deep Learning and Its' Application for Children with ASD. *Int. J. Mach. Learn. and Comput.* (IJMLC) **2020**, pp. 1–10.
157. Triantafyllidou, C. Assistive Technologies for Dyslexia: Punctuation and its Interfaces with Speech. Master's Thesis, University of Central Florida, city, country, 2020.
158. Park, D.E.; Shin, Y.J.; Park, E.; Choi, I.A.; Song, W.Y.; Kim, J. Designing a Voice-Bot to Promote Better Mental Health: UX Design for Digital Therapeutics on ADHD Patients. In Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25-30 April 2020, pp. 1–8.
159. Valadão, C.T.; Goulart, C.; Rivera, H.; Caldeira, E.; Bastos Filho, T.F.; Frizera-Neto, A.; Carelli, R. Analysis of the use of a robot to improve social skills in children with autism spectrum disorder. *Research on Biomedical Engineering* **2016**, *32*, 161–175.
160. Boucenna, S.; Narzisi, A.; Tilmont, E.; Muratori, F.; Pioggia, G.; Cohen, D.; Chetouani, M. Interactive technologies for autistic children: A review. *Cognitive Computation* **2014**, *6*, 722–740.
161. Scassellati, B.; Boccanfuso, L.; Huang, C.M.; Mademtzi, M.; Qin, M.; Salomons, N.; Ventola, P.; Shic, F. Improving social skills in children with ASD using a long-term, in-home social robot. *Science Robotics* **2018**, *3*.
162. Costa, A.P.; Charpiot, L.; Lera, F.R.; Ziafati, P.; Nazarikhorram, A.; Van Der Torre, L.; Steffgen, G. More attention and less repetitive and stereotyped behaviors using a robot with children with autism. In Proceedings of the 27th IEEE Int. Symp. Robot Human Interactive Commun, Nanjing, China, 27-31 August 2018 2018.

163. Vanderborght, B.; Simut, R.; Saldien, J.; Pop, C.; Rusu, A.S.; Pintea, S.; Lefeber, D.; David, D.O. Using the social robot probo as a social story telling agent for children with ASD. *Interaction Studies* **2012**, *13*, 348–372.
164. Peca, A.; Tapus, A.; Aly, A.; Pop, C.; Jisa, L.; Pintea, S.; Rusu, A.; David, D. Exploratory study: Children's with autism awareness of being imitated by NAO Robot. *arXiv preprint arXiv:2003.03528* **2020**.
165. Laranjo, L.; Dunn, A.G.; Tong, H.L.; Kocaballi, A.B.; Chen, J.; Bashir, R.; Surian, D.; Gallego, B.; Magrabi, F.; Lau, A.Y.; others. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* **2018**, *25*, 1248–1258.
166. Car, L.T.; Dhinagaran, D.A.; Kyaw, B.M.; Kowatsch, T.; Rayhan, J.S.; Theng, Y.L.; Atun, R. Conversational agents in health care: Scoping review and conceptual analysis. *Journal of medical Internet research* **2020**, *22*, e17158.
167. Theresa Schachner, Roman Keller, F.v.W. Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review. *Journal of Medical Internet Research* **2020**, *22*.
168. Montenegro, J.L.Z.; da Costa, C.A.; da Rosa Righi, R. Survey of conversational agents in health. *Expert Systems with Applications* **2019**, *129*, 56–67.
169. Fadhil, A.; Wang, Y.; Reiterer, H. Assistive conversational agent for health coaching: a validation study. *Methods of information in medicine* **2019**, *58*, 009–023.
170. Neerincx, M.A.; van Vught, W.; Blanson Henkemans, O.; Oleari, E.; Broekens, J.; Peters, R.; Kaptein, F.; Demiris, Y.; Kiefer, B.; Fumagalli, D.; others. Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management. *Frontiers in Robotics and AI* **2019**, *6*, 118.
171. High, R. The era of cognitive systems: An inside look at IBM Watson and how it works. *IBM Corporation, Redbooks* **2012**, *1*, 16.
172. Strickland, E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* **2019**, *56*, 24–31.
173. Ross, C.; Swetlitz, I. IBM's Watson supercomputer recommended 'unsafe and incorrect'cancer treatments, internal documents show. *Stat* **2018**, *25*.
174. Xu, L.; Zhou, Q.; Gong, K.; Liang, X.; Tang, J.; .; Lin, L. End-to-End Knowledge-routed relational dialogue system for automatic diagnosis. In Proceedings of the AAAI, Online, 2-9 February 2019, Vol. 33, p. 7346–7353.
175. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* **2017**, *4*, e19.
176. Edwards, R.A.; Bickmore, T.; Jenkins, L.; Foley, M.; Manjourides, J. Use of an interactive computer agent to support breastfeeding. *Maternal and child health journal* **2013**, *17*, 1961–1968.
177. Yang, W.; Zeng, G.; Tan, B.; Ju, Z.; Chakravorty, S.; He, X.; Chen, S.; Yang, X.; Wu, Q.; et al., Z.Y. On the generation of medical dialogues for covid-19. arXiv preprint https://arxiv.org/abs/2005.05442 **2020**
178. Palanica, A.; Flaschner, P.; Thommandram, A.; Li, M.; Fossat, Y. Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *Journal of medical Internet research* **2019**, *21*, e12887.
179. Nadarzynski, T.; Miles, O.; Cowie, A.; Ridge, D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health* **2019**, *5*.
180. Scholten MR, Kelders SM, V.G.P.J. Self-Guided Web-Based Interventions: Scoping Review on User Needs and the Potential of Embodied Conversational Agents to Address Them. *Journal of medical Internet research* **2017**, *19*. doi:10.2196/jmir.7351.
181. Dhanda, S. How chatbots will transform the retail industry. *Juniper Research* **2018**, Juniper Research Ltd, Basingstoke, Hampshire, UK https://www.brand-news.it/wp-content/uploads/2018/07/How-Chatbots-Will-Transform-The-Retail-Industry-whitepaper.pdf (accessed on 14 12 2021).
182. Bavaresco, R.; Silveira, D.; Reis, E.; Barbosa, J.; Righi, R.; Costa, C.; Antunes, R.; Gomes, M.; Gatti, C.; Vanzin, M.; others. Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review* **2020**, *36*, 100239.
183. Thomas, N. An e-business chatbot using AIML and LSA. In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21-24 September 2016; IEEE: Manhattan, New York, U.S., 2016, pp. 2740–2742.
184. Cui, L.; Huang, S.; Wei, F.; Tan, C.; Duan, C.; Zhou, M. Superagent: A customer service chatbot for e-commerce websites. In Proceedings of the ACL 2017, System Demonstrations, Vancouver, Canada, July 30 - August 4 2017, pp. 97–102.
185. Xu, A.; Liu, Z.; Guo, Y.; Sinha, V.; Akkiraju, R. A new chatbot for customer service on social media. In Proceedings of the 2017 CHI conference on human factors in computing systems, Denver, CO, USA, 06-11 May, 2017, pp. 3506–3510.
186. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6-12 July 2002, pp. 311–318.
187. Yan, Z.; Duan, N.; Chen, P.; Zhou, M.; Zhou, J.; Li, Z. Building task-oriented dialogue systems for online shopping. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 4-9 February 2017, Vol. 31.
188. Pradana, A.; Sing, G.O.; Kumar, Y. Sambot-intelligent conversational bot for interactive marketing with consumer-centric approach. *Int. J. of Comput. Infor. Sys. and Indus. Manage. Appl.* **2017**, *6*, 265–275.
189. Kaghyan, S.; Sarpal, S.; Zorilescu, A.; Akopian, D. Review of Interactive Communication Systems for Business-to-Business (B2B) Services. *Electronic Imaging* **2018**, *2018*, 117–1.
190. Lewis, M.; Yarats, D.; Dauphin, Y.N.; Parikh, D.; Batra, D. Deal or No Deal? End-to-End Learning for Negotiation Dialogues, 2017. https://arxiv.org/abs/1706.05125.

191. Luo, X.; Tong, S.; Fang, Z.; Qu, Z. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* **2019**, *38*, 937–947.
192. Følstad, A.; Nordheim, C.B.; Bjørkli, C.A. What makes users trust a chatbot for customer service? An exploratory interview study. In Proceedings of the International conference on internet science, St. Petersburg, Russia, 24-26 October 2018, Springer Nature, Switzerland AG, 2018, pp. 194–208.
193. Li, C.H.; Yeh, S.F.; Chang, T.J.; Tsai, M.H.; Chen, K.; Chang, Y.J. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, 26 April 2020, pp. 1–12.
194. Agarwal, A. How to Write a Twitter Bot in 5 Minutes. https://www.labnol.org/internet/write-twitter-bot/27902/ (accessed on 14 12 2021).
195. daniel Peterschmidt. How to Make a Twitter Bot in Under an Hour Even if you don't code that often. https://medium.com/science-friday-footnotes/how-to-make-a-twitter-bot-in-under-an-hour-259597558acf (accessed on 14 12 2021).
196. Adams, T. AI-Powered Social Bots. arXiv preprint https://arxiv.org/abs/1706.05143 **2017**.
197. Assenmacher, D.; Clever, L.; Frischlichy, L. Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Social Media + Society* **2020**, pp. 1–14. doi:10.1177/2056305120939264.
198. Kollanyi, B. Automation, Algorithms, and Politics | Where Do Bots Come From? An Analysis of Bot Codes Shared on GitHub. *Int. J. Commun.* **2016**, *10*, 20.
199. Ferrara, E.; Varol, Q.; Davis, C.; Menczer, F.; Flammini, A. The rise of social bots. *Publication:Communications of the ACM* **2016**, *37*, 81–88. doi:10.1145/2818717.
200. Varol, O.; Ferrara, E.; Davis, C.; Menczer, F.; Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. In Proceedings of the International AAAI Conference on Web and Social Media, Montréal, Québec, Canada, 15-18 May 2017, Vol. 11, pp. 280-289.
201. Subrahmanian, V.S.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; Menczer, F. The DARPA Twitter bot challenge. *Computer* **2016**, *49*, 38–46.
202. Lee, K.; Eoff, B.; Caverlee, J. Seven months with the devils: A long-term study of content polluters on twitter. In Proceedings of the International AAAI Conference on Web and Social Media, Cambridge, Massachusetts, USA, 8-11 July 2013. 2011, Vol. 5.
203. Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; Cieliebak, M. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* **2021**, *54*.
204. Griol, D.; Carbó, J.; Molina, J.M. AN AUTOMATIC DIALOG SIMULATION TECHNIQUE TO DEVELOP AND EVALUATE INTERACTIVE CONVERSATIONAL AGENTS. *Applied Artificial Intelligence* **2013**, *27*, 759–780. doi:10.1080/08839514.2013.835230.
205. Papineni, K.A.; Roukos, S.; Ward, T.; Zhu, W. Understanding Affective Experiences With BLEU: a method for automatic evaluation of machine translation. In Proceedings of the ACL 2002, Philadelphia, PA, July 6-12 2002 2002.
206. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. Text summarization branches out, https://aclanthology.org/W04-1013.pdf accessed on 14 12 2021, 2004, pp. 74–81.
207. Banerjee, S.; Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the ACLworkshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Ann Arbor, Michigan, 29 June 2005, Vol. 29, pp. 65–72.
208. Liu, C.W.; Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. arXiv preprint https://arxiv.org/abs/1603.08023 **2016**.
209. Lowe, R.; Noseworthy, M.; Serban, I.V.; Angelard-Gontier, N.; Bengio, Y.; Pineau, J. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149* **2017**.
210. Tao, C.; Mou, L.; Zhao, D.; Yan, R. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, 2-7 February 2018.
211. Guo, F.; Metallinou, A.; Khatri, C.; Raju, A.; Venkatesh, A.; Ram, A. Topic-based evaluation for conversational bots. *arXiv preprint arXiv:1801.03622* **2018**.
212. Serban, I.V.; Lowe, R.; Henderson, P.; Charlin, L.; Pineau, J. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *arXiv preprint arXiv:1512.05742* **2017**.
213. Keneshloo, Y.; Shi, T.; Ramakrishnan, N.; Reddy, C.K. Deep Reinforcement Learning For Sequence to Sequence Models. *arXiv preprint arXiv:1805.09461* **2018**.
214. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, November 27 - December 1, 2017 2017.
215. Ameixa, D.; Coheur, L.; Redol, R.A. From subtitles to human interactions: introducing the subtle corpus. Technical report; https://www.inesc-id.pt/ficheiros/publicacoes/10062.pdf accessed on 14 12 2021
216. Lison, P.; Tiedemann, J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the LREC 2016, Portorož, Slovenia, 23-28 May 2016.
217. Tiedemann, J. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. Recent advances in natural language processing, 2009, Vol. 5, pp. 237–248.

218. Dodge, J.; Gane, A.; Zhang, X.; Bordes, A.; Chopra, S.; Miller, A.H.; Szlam, A.; Weston, J. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. In Proceedings of the ICLR, San Juan, Puerto Rico, 2-4 May 2016.
219. Danescu-Niculescu-Mizil, C.; Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077* **2011**.
220. Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.P.; Gao, J.; Dolan, B. A Persona-Based Neural Conversation Model. *arXiv preprint* **2016**, *arXiv:1603.06155*.
221. Ritter, A.; Cherry, C.; Dolan, B. Unsupervised Modeling of Twitter Conversations. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California, 2-4 June 2010; Association for Computational Linguistics: 2010; pp. 172–180. https://www.aclweb.org/anthology/N10-1020.
222. Schrading, N.; Ovesdotter Alm, C.; Ptucha, R.; Homan, C. An Analysis of Domestic Abuse Discourse on Reddit. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17-21 September; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 2577–2583. https://www.aclweb.org/anthology/D15-1309, doi:10.18653/v1/D15-1309.
223. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. arXiv preprint https://arxiv.org/abs/1911.00536 **2019**.
224. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In Proceedings of the ACL, Online, 5-10 July 2020, p. 85–96.
225. Lowe, R.; Pow, N.; Serban, I.; Pineau, J. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. Proc. SIGDIAL 16, 2017, pp. 285–294. arXiv preprint https://arxiv.org/abs/1506.08909.
226. Alizadeh, K. Limitations of Twitter Data Issues to be aware of when using Twitter text data. https://towardsdatascience.com/limitations-of-twitter-data-94954850cacf (accessed on 14 12 2021).
227. Zeng, C.; Li, S.; Li, Q.; Hu, J.; Hu, J. A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets. *Applied Sciences* **2020**, *10*. doi:10.3390/app10217640.
228. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* **2016**.
229. Rajpurkar, P.; Jia, R.; Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 15-20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 784–789. doi:10.18653/v1/P18-2124.
230. Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching Machines to Read and Comprehend. Advances in Neural Information Processing Systems; Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; Garnett, R., Eds. Curran Associates, Inc., 2015, Vol. 28.
231. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; others. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **2019**, *7*, 453–466.
232. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, July 30 - August 4; Association for Computational Linguistics: Vancouver, Canada, 2017; pp. 1601–1611. doi:10.18653/v1/P17-1147.
233. Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; Khaitan, P. Towards Scalable Multidomain Conversational Agents: The Schema-Guided Dialogue Dataset. arXiv preprint https://arxiv.org/abs/1909.05855 **2020**.
234. Budzianowski, P.; Wen, T.H.; Tseng, B.H.; Casanueva, I.; Ultes, S.; Ramadan, O.; Gašić, M. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In Proceedings of the EMNLP, Brussels, Belgium, October 31 - November 4, 2018.
235. Byrne, B.; Krishnamoorthi, K.; Sankar, C.; Neelakantan, A.; Duckworth, D.; Yavuz, S.; Goodrich, B.; Dubey, A.; Cedilnik, A.; Kim, K.Y. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In Proceedings of the EMNLP-IJCNLP, Hong Kong, China, 3-7 November 2019.
236. Peskov, D.; Clarke, N.; Krone, J.; Fodor, B. Multi-Domain Goal-Oriented Dialogues (MultiDoGO): Strategies toward Curating and Annotating Large Scale Dialogue Data. In Proceedings of the EMNLP-IJCNLP, Hong Kong, China, 3-7 November 2019.
237. Zeng, G.; Yang, W.; Ju, Z.; Yang, Y.; Wang, S.; Zhang, R.; Zhou, M.; Zeng, J.; Dong, X.; Zhang, R.; Fang, H.; Zhu, P.; Chen, S.; Xie, P. MedDialog: Large-scale Medical Dialogue Datasets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16-20 November 2020; Association for Computational Linguistics, pp. 9241–9250. doi:10.18653/v1/2020.emnlp-main.743.
238. Sharma, A.; Lin, I.W.; Miner, A.S.; Atkins, D.C.; Althoff, T. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *arXiv preprint arXiv:2101.07714* **2021**.
239. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. arXiv preprint https://arxiv.org/abs/1811.00207 **2019**.

240. McKeown, G.; Valstar, M.F.; Cowie, R.; Pantic, M. The SEMAINE corpus of emotionally coloured character interactions. In Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, ICME, Singapore, 19-23 July 2010 2010, pp. 1–4. doi:10.1109/ICME.2010.5583006.
241. Allouch, M.; Azaria, A.; Azoulay, R. Detecting sentences that may be harmful to children with special needs. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, November 4-6, 2019. IEEE, 2019, pp. 1209–1213.
242. Chai, Y.; Liu, G.; Jin, Z.; Sun, D. How to Keep an Online Learning Chatbot From Being Corrupted. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, United Kingdom, 19-24 July 2020, pp. 1–8.
243. Yu, Y.; Eshghi, A.; Mills, G.; Lemon, O. The BURCHAK corpus: a challenge data set for interactive learning of visually grounded word meanings. In Proceedings of the Sixth Workshop on Vision and Language). Association for Computational Linguistics, Valencia, Spain, 4 April 2017 2017, pp. 1–10.
244. Wolska, M.; Vo, Q.B.; Tsovaltzi, D.; Kruijff-Korbayová, I.; Karagjosova, E.; Horacek, H.; Fiedler, A.; Benzmüller, C. An Annotated Corpus of Tutorial Dialogs on Mathematical Theorem Proving. In Proceedings of the LREC, Lisbon, Portugal, 26-28 May 2004 2004.
245. Hutzler, D.; David, E.; Avigal, M.; Azoulay, R. Learning methods for rating the difficulty of reading comprehension questions. In Proceedings of the 2014 IEEE International Conference on Software Science, Technology and Engineering, Ramat Gan, Israel, June 11 - 12, 2014.
246. Bloom, B.S.; Engelhart, M.D.; Furst, E.J.; Hill, W.H.; Krathwohl, D.R.; others. Taxonomy of educational objetives: the classification of educational goals: handbook I: cognitive domain. Technical report, Longmans, Green and Company New York, NY, USA: D. Mckay, 1956.
247. Stasaski, K.; Kao, K.; Hearst, M.A. CIMA: A Large Open Access Dialogue Dataset for Tutoring. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, Online, July 10, 2020, pp. 52–64. doi:10.18653/v1/2020.bea-1.5.
248. Arabshahi, F.; Lee, J.; Gawarecki, M.; Mazaitis, K.; Azaria, A.; Mitchell, T. Conversational neuro-symbolic commonsense reasoning. *AAAI'21* **2021**.
249. Chkroun, M.; Azaria, A. A Safe Collaborative Chatbot for Smart Home Assistants. *Sensors* **2021**, *21*, 6641.
250. Chkroun, M.; Azaria, A. Lia: A virtual assistant that can be taught new commands by speech. *Int. J. Human-Comp. Intera.* **2019**, *35*, 1596–1607.
251. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), Opatija, Croatia, May 21-25, 2018. IEEE, 2018, pp. 0210–0215.
252. Rosenfeld, A.; Richardson, A. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* **2019**, *33*, 673–705.
253. Bird, E.; Fox-Skelly, J.; Jenner, N.; Larbey, R.; Weitkamp, E.; Winfield, A. The ethics of artificial intelligence: Issues and initiatives. *European Parliamentary Research Service, Technical Report PE* , https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS accessed on 15 12 2021, European Parliamentary Research Service, **2020**.