Clinical Screening Interview Using a Social Robot for Geriatric Care

Ha Manh Do, *Member, IEEE*, Weihua Sheng, *Senior Member, IEEE*, Erin E. Harrington, and Alex J. Bishop

***Abstract*— Social robots are coming to our homes and have already been used to help humans in a number of ways in geriatric care. This article aims to develop a framework that enables social robots to conduct regular clinical screening interviews in geriatric care, such as cognitive evaluation, falls' risk evaluation, and pain rating. We develop a social robot with essential features to enable clinical screening interviews, including a conversational interface, face tracking, an interaction handler, attention management, robot skills, and cloud service management. Besides, a general clinical screening interview management (GCSIM) model is proposed and implemented. The GCSIM enables social robots to handle various types of clinical questions and answers, evaluate and score responses, engage interviewees during conversations, and generate reports on their well-being. These reports can be used to evaluate the progression of cognitive impairment, risk of falls, pain level, and so on by caregivers or physicians. Such a clinical screening capability allows for early detection and treatment planning in geriatric care. The framework was developed and implemented on our 3-D-printed social robot. It was tested on 30 older adults with different ages, achieved satisfying results, and received their high confidence and trust in the use of this robot for human well-being assessment.**

***Note to Practitioners*—This article is motivated by the goal of using a social robot to perform geriatric well-being assessment through clinical screening interviews. In order to conduct clinical screening interviews, the social robot needs the following essential features: having a verbal conversational interface, adapting to different types of clinical screening interviews, scoring and evaluating answers, having nondirective listening responses, and enabling directive listening responses. The proposed general clinical screening interview management (GCSIM) model demonstrates these capabilities on the social robot. The robot can give structured clinical screening interviews with different question–answer sheets. This will help advance assistive technologies for use by geriatric physicians, nurses, and social service professionals to keep older adults healthy, safe, and independent at home. Robots will become more and more essential in working alongside geriatric practitioners to help monitor older adults at home and to provide early detection and warning of cognitive/mental health problems, falls' risk, and so on. This early detection property can improve quality-of-care and help older adults remain living at home.**

***Index Terms*— Clinical interview, clinical screening, conversation management, elderly care, geriatric care, healthcare, social robot, well-being assessment.**

Manuscript received October 2, 2019; revised December 20, 2019; accepted May 14, 2020. This article was recommended for publication by Associate Editor G. Nejat and Editor Y. Sun upon evaluation of the reviewers' comments. This work was supported in part by the National Science Foundation (NSF) under Grant CISE/IIS 1231671, Grant CISE/IIS 1427345, Grant CISE/IIS 1910993, and Grant EHR/DUE 1928711 and in part by the Department of Education, Title III, Part f, CSU-Pueblo's Communities to Build Active STEM Engagement (C-BASE) Program, under Award P031C160025. *(Corresponding author: Weihua Sheng.)*

Ha Manh Do is with the Communities to Build Active STEM Engagement (CBASE), Colorado State University-Pueblo, Pueblo, CO 81001 USA, and also with the Department of Engineering, Colorado State University-Pueblo, Pueblo, CO 81001 USA (e-mail: ha.do@okstate.edu).

Weihua Sheng is with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: weihua.sheng@okstate.edu).

Erin E. Harrington is with the Department of Psychology, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: erin.harrington@okstate.edu).

Alex J. Bishop is with the Human Development and Family Science Department, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: alex.bishop@okstate.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2020.2999203

I. Introduction

THE older adult population in the United States is increasing. The number of people 65 years of age and older is anticipated to double from almost 43.1 million in 2012 to 83.7 million by 2050 [1]. This trend leads to both economical and sociological challenges in geriatric care. Many of these persons will live to age 85 and older. According to the U.S. census, persons aging 85 and older represent the fastest-growing age demographic in the country. Persons who are 85 and older often require greater health monitoring due to increased vision and hearing impairment, memory deficits, and so on. The increasing older population creates serious social problems on the geriatric care. There are not enough young people to take care of the aging population in the United States [2]. The aging of the American society is resulting in an ever-increasing population of old adults. Many of them age alone with underlying functional and mental health problems. Moreover, despite challenges in daily living performance, most of the older adults prefer to remain living in their homes for as long as possible. This is mainly due to the fact that staying in one's own home is cheaper and commits greater privacy and autonomy [3]. More than a third of older Americans live alone in their own homes [4].

Living alone in old age can prove hazardous to personal safety, emotional health, mental health, and physical health. The feeling of loneliness, negative moods, or depression creates social and mental health problems of older adults. More than five million people in the United States are living with dementia and aging illnesses [5]. It is estimated that one in three older adults will die from complications caused by dementia. In the United States only, every 66 s, one person is diagnosed with Alzheimer's disease [5]. Besides, physical health is another serious issue older adults face in their life.

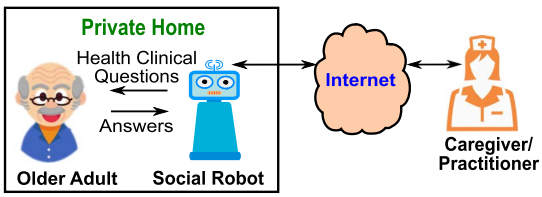

Fig. 1. Overview of the proposed clinical screening interview using a social robot for geriatric care.

For an older adult, the risk of falls increases over time [1]. Falling within one's home can result in injury and contribute to nursing home admission, chronic health complications, social isolation, and loss of self-confidence [6]. An estimated 30%–50% of all persons over 65 years of age fall in their homes annually [7]. Half of them do so repeatedly during the day, making falls the number one cause of consequential disablement, unintentional injury, and accidental death among older adults [8]. Therefore, it is important to keep track of the progression of both mental and physical health in geriatric care.

Social robots have proved their capabilities in geriatric care, such as providing companionship [9], improving the users' mood [10], and conducting cognitive orientation assessment [11], which have a positive influence on the lives of older adults [12], [13]. We believe that social robots can offer a solution to monitoring the mental and physical health of older adults who live independently in their residence.

Clinical assessment is the process of collecting information about a patient for diagnosing disease and planning treatment. There are three common types of clinical assessments: clinical screening interview, neurological and biological testing, and intelligence testing. A clinical screening interview is a procedure that practitioners use to diagnose what is wrong and initiate treatment for a patient. A clinical screening interview, often referred to as a conversation with a purpose, is a dialog in which the practitioner asks specific, open-ended questions in order to assess a client's cognition, behaviors, feelings, and physical well-being. Equipped with verbal conversation capability, the social robot can be used to perform clinical screening interviews. The use of social robots for this task could help detect mental and physical problems of older adults, save patients' time to visit practitioners in person, and decrease their financial burden. Therefore, it is highly beneficial to develop clinical screening interview capability for social robots.

In this article, we propose an approach that uses a social robot for a clinical screening interview. As shown in Fig. 1, a social robot can perform clinical screening interviews through verbal conversation in order to monitor the progression of cognitive impairment, pain, or physical well-being. In addition, the robot is capable of not only conducting clinical screening interviews but also automatically evaluating answers given by the patient during clinical assessment sessions, as well as generating scores and saving them for further analysis by a remote caregiver or practitioner.

The main contribution of this article is that it proposed and implemented a comprehensive clinical screening interview framework that enables social robots to proactively perform geriatric well-being assessment through verbal conversations, which has not been fully developed in existing dialogue or question–answering (QA) systems. We introduced a complete software architecture with a conversational interface (CI) for social robots, which supports not only general conversations as many other social robots do but also clinical screening interviews. We proposed and developed a general clinical screening interview management (GCSIM) model for the human–robot conversation manager. This novel GCSIM model enables a social robot to handle various clinical screening interviews with several essential interview skills to engage interviewees during the conversation. Our work provides a reference design to develop social robots for not only clinical interviews but also other healthcare applications. For example, in pandemics such as the Covid-19, social robots can find themselves invaluable in monitoring the progression of patients' health while they are quarantined at their homes.

The rest of this article is organized as follows. Section II discusses related previous works. Section III describes the design of our social robot platform. Sections IV and V present the development of the CI and the clinical screening interview framework for social robots, respectively. Section VI gives the experimental results. Section VII concludes the article and discusses the future work.

II. Related Works

This section gives a review of the related works in the area of social robots with a focus on geriatric care, healthcare, and dialogue systems.

In recent years, there has been much interest in developing robotic technologies for geriatric care. Several animal-like therapy robots and companion robots have been made commercially available, such as PARO [14] and NeCoRo [15]. In academia, researchers have developed many robots for elderly care in domestic environments, such as the CompanionAble, GiraffPlus, Accompany, Robot-Era, Mobiserv, Hobbit, and SoCoNet projects [16]. Those robots were capable of performing music therapy, teleconferencing, and reminding older adults of taking medicines, eating or drinking, and guiding them through their homes. Several robots have been developed as assistive robots to give baths to in-bed patients [17], help patients dress [18] and eat [19], monitor the older adult's daily activities, and detect falls [20]. Their work showed that older adults, when trained beforehand, are able to effectively take advantage of the assistance of a robot.

Several social robots have been developed for mental health care, such as the Ludwig robot [21] that is capable of analyzing speech patterns to identify early signs of Alzheimer's [22], [23] and ZORA robot that can help older adults with interactive therapeutic and recreational activities [24]. Chang *et al.* [25] found that the PARO robot can act as a stress reliever for older adults suffering from dementia. In [26], the PARO robot was used in experiments to investigate the effect of a social robot intervention on depression, loneliness, and the quality of life of older adults. Their results showed a noticeable improvement in the mental well-being of older adults. Wang *et al.* [27] investigated the

responses of older adults with dementia and their caregivers following direct interaction with a teleoperated assistive robot. In [28], a home service robot was developed to remind the older adults to drink water once fluid intake by the older adults is found to be insufficient. The aforementioned robots show the capabilities of geriatric care and healthcare. However, the abovementioned work failed to develop a general framework for clinical assessment through screening interviews.

Dialogue systems or conversational agents have many applications [29] and are generally classified into three main types: task-oriented dialogue agents, chatbots, and QA systems. Task-oriented dialogue agents are designed for a particular task to get information from the user to help complete the task [30], such as searching for flights, ordering a pizza, and finding restaurants. Voice assistants, such as Apple's Siri, Microsoft's Cortana, Amazon's Alexa, and Google's Assistant, are now popular on smart devices for open conversations that mimic unstructured human–human conversations rather than focus on a particular task [31]. QA systems are developed to allow users to ask questions in natural languages and receive a concise answer [32]. The Web search engines, a type of QA systems, have been developed by Google, Microsoft, Yahoo, and so on for decades. Recently, IBM's Watson question–answer system based on the DeepQA system [33] won the TV game show, Jeopardy. However, these dialogue systems and QA systems are hard to adopt for clinical screening interviews since they are mainly designed to handle the questions from the users. Recently, several studies have used robots or computers with audio systems for interview tasks, such as gathering adults' privacy information [34] and interviewing children about special needs [35] or bullying [36]. However, the audio systems in these studies played limited roles in the interview tasks, which focused on playing a list of prerecorded questions and recording responses without evaluation or scoring. These systems still lack the essential capabilities of a clinical interview system, including handling various types of clinical questions and answers, scoring or evaluating responses, adapting to the users' response, and integrating emotion recognition and attention management to engage users during conversations.

Social robots with conversational capabilities can be the appropriate technology to achieve the goal of caring for the elderly in their own homes [37]. Such robots can provide companionship and assistive services through verbal conversations. Although conversational computers and social robots are getting more popular, these systems mainly deal with general conversations [38]. Therefore, there is a great need to develop a robotic framework for conversations with the purpose of clinical screening interviews for geriatric care.

III. Social Robot Platform for Clinical Screening Interviews

A. Essential Features of Social Robots

Social robots have a wide range of applications, designs, features, and functionalities. This section aims to propose several essential features that a social robot should have to be capable of the clinical screening interview.

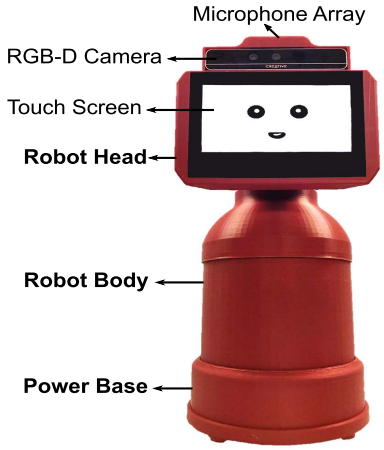

Fig. 2. 3-D-printed social robot used for clinical screening interviews.

Different from a normal conversation, a clinical screening interview is a conversation with a purpose, which is clearly defined and conducted within a certain time frame. The structured clinical screening interview approach is mainly used to gather reliable and valid assessment data. Structured clinical screening interviews involve asking the same questions in the same order as every client. The intake interview and the mental status exam are two of the most common clinical screening interviews. The primary goals of clinical screening interviews are relationship development, assessment, and helping. To achieve these goals, besides honesty and integrity, the ability to remain calm in stressful situations, self-awareness, and observational and assessment skill, a practitioner must meet requirements of basic communication and listening skills [39]. Basic communication and listening skills consist of nondirective listening responses, directive listening responses, and directives and action responses. Nondirective listening responses, including attending behaviors (eye contact, body posture, voice tone, or verbal tracking) and other behaviors (silence, clarifications/verbal prompt, paraphrasing, the reflection of feeling, or summarization) serve to establish a therapeutic alliance. Directive listening responses (feeling validation, interpretive reflection of feeling, interpretation, reframe, or confrontation) help bring the interviewer's perspective into the interview. In addition, they must master technical knowledge associated with clinical screening interviews. This means that they must know different types of clinical screening interviews and the range of available interview responses that likely affect clients. Practitioners are required to maintain these skills as a basic skill set that can be learned and improved with time and practice.

It is challenging for a social robot to possess all the aforementioned skills. However, in order to be used for a clinical screening interview, a social robot should have the essential features as follows: having a verbal CI, adapting to different types of clinical screening interviews, scoring and evaluating answers, having nondirective listening responses, and enabling directive listening responses. In this article, we mainly develop and evaluate the CI, the clinical screening interview skill that enables cognitive assessment, falls' risk evaluation, and pain rating. We also implement an attention management function

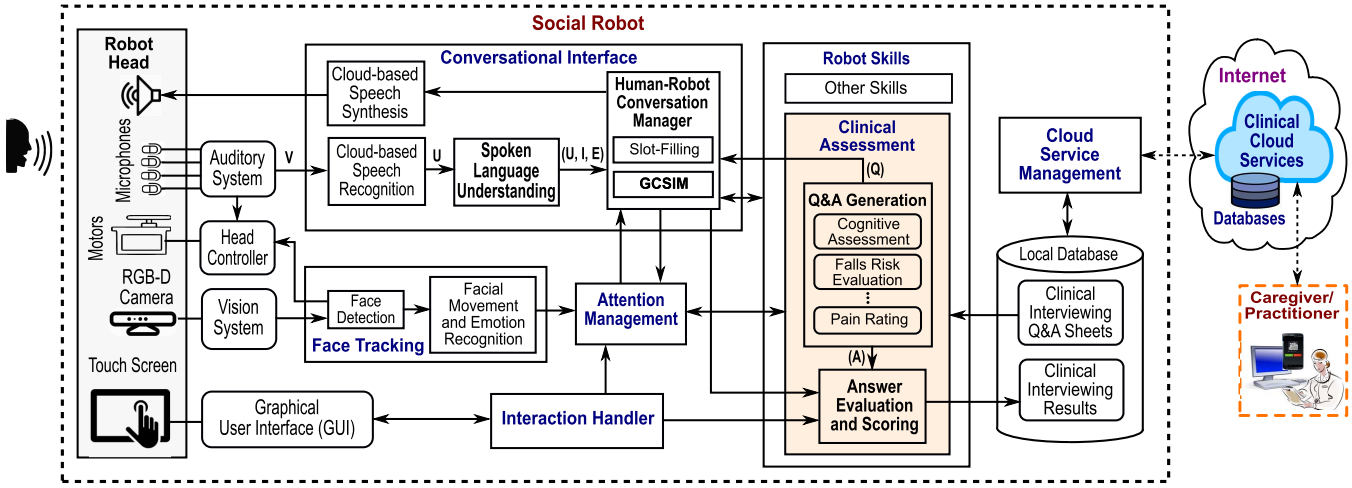

Fig. 3. Software architecture of the social robot.

as a simplified version of nondirective and directive listening responses.

B. Hardware Design

The social robot is a 3-D-printed desktop robot, as shown in Fig. 2. The robot consists of three parts: robot head, body, and base. The robot head is powered by two servo motors and has two degrees of freedom (yaw and pitch). It features a touch screen, a vision system, and an auditory system. Therefore, the robot can track the user's face and turn to the direction where the sound is coming from. The touch screen connected to an Android embedded board is used for video communication and graphic user interface (GUI). The vision system consists of an RGB-D camera allowing face detection. The auditory system is built by extending the built-in microphone array of the PS3eye camera [40]. It consists of four microphones and employs technologies for echo cancellation and background noise suppression. This allows the auditory system to be used for speech recognition, sound localization, and sound separation in noisy environments. The microphone array operates with each channel processing 16-bit samples at a sampling rate of up to 48 kHz per channel and a large dynamic range of signal-to-noise ratio up to 90 dB. The auditory system is based on HARK [41], an open-sourced robot audition software consisting of modules for acoustic signal processing, sound source localization, sound source separation, and automatic speech recognition for various microphone array configurations. Using a microphone array, the robot is able to localize and separate multiple sound sources, which has been realized in our previous works [42], [43]. The robot body has two speakers and a minicomputer inside. The minicomputer is an Intel NUC with a Core i7 processor and runs the robot software. The robot base contains batteries and other power supplies inside.

C. Software Architecture

In order to develop the essential features, as described earlier, the robot software architecture is proposed, as shown in Fig. 3. The architecture features six key components: 1) CI; 2) face tracking (FT); 3) interaction handler; 4) attention management; 5) robot skills; and 6) cloud service management.

The CI combines speech recognition and spoken language understanding (SLU) to enable human–robot verbal conversations.

The FT detects the human face and recognizes facial emotions during the interview for clinical assessment. It consists of face detection and facial movement and emotion recognition. The face detection is implemented using a pretrained facial landmark detector inside the Dlib library [44]. Each detected facial landmark consists of the locations of 68 (x, y) coordinates that map to face regions, including the mouth, right eyebrow, left eyebrow, right eye, left eye, nose, and jaw. The detected landmarks are used to recognize the facial movements using hidden Markov models (HMMs) and estimate the face position in the camera field of view (FoV). We trained seven independent HMMs to detect facial movements, including blink_eyes, smile, turn_head_to_the_left, turn_head_to_the_left, straight_face, speaking, and no_speaking. The face location consists of the X and Y coordinates from the center point of the camera FoV. Besides, the detected faces are cropped and sent to Microsoft cloud servers for facial emotion recognition using Face APIs [45]. This cloud service can return eight facial emotions, including neutral, anger, contempt, disgust, happy, fear, sadness, and surprise. The detected face along with recognized movements and emotions is used for attention management.

The interaction handler sets visual displays for the GUI and handles events generated by the user's interaction on the touch screen. This handler and the GUI are implemented on Android and connected to other modules on the Linux computer using TCP/IP sockets. We call the GUI a robot face.

The attention management determines the user's engagement in an interview by tracking the face, speech, and interaction of the user on the touch screen. It also regulates and maintains the robot face to react to the user's emotions and responses during the interview. During clinical screening interview, if negative verbal intents or negative facial emotions

(anger, contempt, disgust, fear, and sadness) are detected, the robot plays verbal cues (e.g., no worries, you did a good job, fantastic, and nice job) to motivate the user. In addition, if it detects no face, no response, and no interaction on the touch screen, the robot replays the question or prompts the user to the response.

The robot skills enable the robot to assist a human at home by interacting with him or her in a nonintrusive way. The robot is able to provide companionship through socialization skills, such as telling jokes or quotes, playing music, describing the weather, playing the latest news, and playing rock-paper-scissors. In addition, the robot can perform clinical screening interviews for cognitive assessment, mood, and loneliness detection, as well as pain evaluation.

The cloud service management connects the robot to clinical cloud services and databases. The cloud database stores the health clinical question and answer (Q&A) sheets, as well as users' profiles and reports of clinical assessment scores, which can be updated and reviewed by a caregiver. The local database stores the profile, testing schedules, and Q&A sheets customized for the user who is being cared for by the social robot.

IV. Conversational Interface

Enabled by speech synthesis, speech recognition, SLU, and a human–robot conversation manager, as shown in Fig. 3, the robot can interact with the human through verbal conversations.

A. Speech Synthesis

The speech synthesis module is known as text-to-speech (TTS), where questions are verbally communicated to the user. Recently, a number of commercial deep learning-based TTS engines that produce quality sound are available as cloud and web services, such as Amazon IVONA [46], IBM Watson TTS [47], and Google TTS APIs [48]. We implemented online TTS based on Google TTS APIs that can generate human-like sound with a small delay.

B. Speech Recognition

The autonomous speech recognition (ASR) module is known as speech-to-text (STT), where the answers given by the user are converted to text for analysis. Recently, deep neural networks (DDNs) are utilized for academic research in ASR and deployed in most commercial ASR systems that are available for end users to use as cloud services and APIs. In order to achieve the state-of-the-art accuracy of speech recognition, we utilized the Google Speech APIs cloud service [48], one of the best cloud services for speech recognition [49], which can return word sequences in real time.

C. Spoken Language Understanding

SLU takes the outputs from the ASR and produces utterances' meaning representations that are then passed on to the human–robot conversation manager. This module determines what the utterance (U) is about. The meaning representations

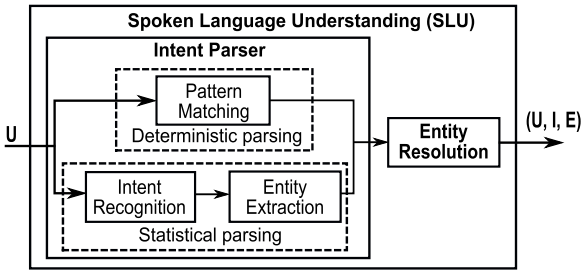

Fig. 4. SLU for social robots.

can be the domain (main topic) of the utterance, user's intent (I) with relevant parameters and entities (E) or tagger slots. An example utterance "Wake me up at 7 am" can be parsed into alarm domain and set.alarm intent with *date-time* entity (*date-time*="*7 am*") that can be resolved into 7:00:00. The alarm is classified by the domain recognition; the set.alarm intent is recognized by the intent recognition; and the date-time entity is extracted by the entity extraction. These modules also estimate the confidence score of each component of the semantic representation. In this article, all intents are implemented in one domain only: the clinical_interview domain. Therefore, the developed SLU consists of an intent parser and entity resolution. Various approaches and techniques have been applied for the intent parser, such as deterministic parsing [e.g., regular expressions (REs)], statistical parsing (e.g., hidden vector state model, stochastic finite state transducers, dynamic Bayesian networks, support vector machines, conditional random fields (CRFs), and deep learning), dialogue act recognition, user's utterance classification, user's utterance content analysis, semantic interpretation, and syntax-driven semantic analysis [38]. In this article, we implemented the intent parser by a combination of deterministic parsing and statistical parsing, as shown in Fig. 4. The deterministic parsing model can cover all the sample utterances that are used to train the model. This model makes the intent parser predictable and easy to be used for clinical screening interviews, where the answers may be changed and updated at the time of operation but are well-bounded and the possible answers are well defined by the practitioners. This model is implemented based on pattern matching. If the first model fails to find a match, the statistical parsing model is used. This second model is trained beyond the set of sample utterances and, therefore, has generalization capabilities. It can parse utterances even though they are not part of the training examples. This model involves two components: intent recognition and entity extraction. The last step after identifying the intent and the entities is to resolve entity values from raw strings of entities. These components are presented in the following.

1) Pattern Matching: Pattern matching checks and locates the constituents of some patterns among a given sequence of tokens. The pattern can be represented by either tree structures or sequences. Sequence patterns are commonly used for intent parsing and often described using REs. The example utterances of each intent are represented as regular languages [50]. REs are used to denote regular languages. They can represent regular languages and operations on them succinctly. An RE

is a formal mathematical expression using a limited set of operators. It can be used to specify a set of strings. REs can be concatenated to form new REs, contain both special and ordinary characters, and have repetition qualifiers (*, +). The parser checks if an input text matches any given RE that was built given example utterances of intent. We used three main operations to present an RE, including disjunction |, grouping {}, and parentheses (). An example utterance of the weather intent is "(tell me|let's me know|what is) the weather in {location}(Stillwater) {time}(tomorrow)," where {location} and {time} are entities. We implemented an RE pattern matching engine based on nondeterministic finite automate (NFA) [51]. REs are represented as NFA machines; then, these NFA machines are run using an input string. The engine checks whether an RE matches an input string and classifies this input string into the corresponding intent.

2) Intent Recognition and Entity Extraction: There have been several approaches for intent recognition or classification. We can use classical approaches of text mining, such as models on bag-of-words with N-Grams [52] and term frequency-inverse document frequency (TF-IDF) [53]. Recently, neural network methods have made significant progress in natural language processing [54], including speech recognition, speech synthesis, intent classification, and entity extraction. We can use recurrent networks RNNs [e.g., long short-term memory (LSTM) and gated recurrent unit (GRU)] or convolutional neural networks (CNNs) for intent recognition. The recent study [55] shows that simple models (shallow neural networks) can achieve significant performance on data sets where the intent recognition task is based on a key phrase recognition task, such as sentiment detection and intent recognition in SLU. We implemented a statistical parser using compact feed-forward neural networks that require a small training data set, fast training, and efficiency in running on embedded systems in social robots. We used four layers of neurons (three hidden layers with the symmetrical sigmoid activation functions and an output layer with the sigmoid activation function) and a "bag-of-words" approach to organizing our training data.

Named entity recognition (NER) or entity extraction locates and classifies the named entities that present in the text. NER classifies the named entities into predefined categories, such as the names of persons, organizations, locations, quantities, monetary values, specialized terms, product terminology, and expressions of time. We implemented the entity extraction based on the slot filling approach using a linear chain CRF [56] that is specifically trained to extract the slots of the recognized intent.

Besides building example utterances and trained intents for general conversations (e.g., setting alarm, telling jokes, playing music, playing games, weather, and playing news), we created example utterances to train clinical intents for interviews. By observing the sample videos of clinical screening interviews, we identified intents that are relevant to clinical interviews. The most popular intents that we identified are clinic.repeat (the human asks the robot to repeat the question), clinic.unknown (the human does not know the answer), clinic.unclear (the question is unclear), clinic.stop (the human wants to stop the interview), clinic.pause (the human wants to pause the interview), and clinic.change (the human changes the answer). We created 30 query examples for each clinical intent based on our observation of these videos.

3) Entity Resolution: Entity resolution involves the disambiguation of entities. It finds all expressions that refer to the same entity in a text and presents the value of the entity in a predefined format that the robot can recognize. For example, the extracted time entity "the first Monday of June 2018" from the utterance "Remind me to visit the doctor at 9:00 am on the first Monday of June 2018" is resolved into a time value in ISO format "2018-06-04T09:00:00Z." In this article, we implemented the entity resolution for dates, times, and money that are mainly used for cognitive assessment.

D. Human–Robot Conversation Manager

The conversation manager coordinates the activity of all components, controls dialogue flows, and communicates with external applications, devices, services, or resources. Its roles include discourse analysis, knowledge database query, dialog management, and task management. In addition, the conversation manager may contact one or more robot skills that have knowledge of specific task domains, such as playing the latest news, weather, setting alarms, playing music, playing games, and clinical assessment. This module is implemented based on the slot-filling approach for general conversations. This approach proved its significance in a conversation flow for parameter value collection within a single intent. Slot-filling requires all parameter values of each intent to be set. If users omit one or more of the parameters in their response, the robot will generate questions to ask or prompt them to provide values for each missing parameter. Besides, we propose and implement a new human–robot conversation management model for clinical interviews, namely, GCSIM that can handle the required timing for well-being assessment, attention management, listening responses, interaction management, and scoring. This model will be presented in Section V.

V. Framework for Clinical Screening Interview

This section presents our development of a clinical screening interview framework that enables multiple clinical screening interviews and features a verbal CI, a clinical screening interview procedure, answer evaluation and scoring, and attention management.

A. Clinical Screening Interviews

A clinical screening interview is a conversation in which the psychologist asks specific, open-ended questions in order to evaluate cognition, health, behaviors, feelings, or other capacities of a client. The clinical screening interview approach is commonly used for an intake evaluation and cognitive impairment diagnostic in geriatric care. There are various interview tools for screening, diagnostic, and assessment widely used for geriatric care. In this article, we mainly discuss interview tools for cognitive assessment, falls' risk evaluation, and pain rating.

1) Cognitive Assessment: Cognitive assessments are commonly implemented to appraise and monitor age-associated changes in human cognition. A cognitive orientation assessment is a practical method for grading the cognitive state of a person [57]. These assessments can be used to check time and space orientation, as well as short- and long-term recalls of information. Generally, it is employed by therapists and physicians. These assessments can also be used to evaluate the cognitive development of children. However, the focus of this work is solely on older adults who are at high risk of dementia and Alzheimer's diseases. There are different types of cognitive orientation assessments that can be conducted. The minimental status examination (MMSE) and MMSE2 consist of daily life questions, such as the date, month, and state of residence. The MMSE and MMSE2 test the ability of a person to register, calculate, recall, and orient oneself at a given situation [57]. The MMSE2, a standardized version of the MMSE, is used to reduce the interrater variability in scores of the original version [58]. The Montreal cognitive assessment (MoCA) is a 10-min-long cognitive screening tool [59] to evaluate the concentration, quantitative thinking, language, and orientation of a person, and it is able to identify mild cognitive impairment with a higher sensitivity than the MMSE. The Addenbrookes cognitive assessment (ACA) is based on the MMSE, and it was developed as a way to differentiate between Alzheimer's disease and dementia. It can be applied in approximately 15 min and also less relies on verbal abilities and focuses more on the executive abilities of the elderly than the MMSE [60], [61]. We built a new cognitive evaluation interview toolkit that covers different types of memory and concentration tests. The toolkit consists of 11 types of questions: 1) repetition (the human is asked to repeat and remember some information); 2) time (time of the day, year, month, day of the week, and date); 3) location (state, state capital, city/town, building, and floor); 4) recall (the human is asked to recall the information she/he heard before); 5) saying series of numbers in reverse; 6) solving problems with basic operations; 7) describing the images; 8) spelling the word in reverse; 9) command (the human is asked to do commands: smile, turn head to the left or right, and blink eyes); 10) figure selection and drawing; and 11) listening comprehension. Each type of question has multiple questions.

2) Falls' Risk Evaluation: Falls are one of the leading causes of both fatal and nonfatal injuries for older adults. The most popular falls' risk assessment tools used in hospitals are the Morse Falls Scall, STRATIFY Scale, Hendrich II Fall Risk Model, and Johns Hopkins Fall Risk Assessment Tool [62]. These toolkits collect multiple factors (e.g., history of falling, gait, mental status, age, cognition, and mobility) from a patient and use their own scoring systems to assess the falls' risk. The most recent falls' risk assessment tool is STEADI [63]. Using in-depth interviews, STEADI helps identify patients at low, moderate, and high risk for a fall. The abovementioned tools mainly use interviews for falls' risk screening and use gait, strength, and balance tests to evaluate the fall-risk levels. However, these tools mainly collect the history of falling in general. In order to evaluate the falls' risk of older adults at their homes, we built a new interview tool with a series of questions about how confidence an older adult is in doing daily activities without falling. The confidence levels are from 1 (not confident at all) to 10 (very confident).

3) Pain Rating: A pain rating tool is a communication method that allows doctors to track patients' pain and rate their pain and evaluate what they are feeling and how bad their pain is. The main approaches of current pain rating tools are using verbal, visual and numeric self-rating scales, physiological responses, and behavioral observation scales to measure the pain. There are many types of pain scales used for general clinical and research settings, including 1–10 pain scales, faces pain scales, global pain scale, visual analog pain scale, McGill pain scale, Mankoski pain scale, color scales for pain, and so on [64]. The 1–10 pain scales, the Wong–Baker FACES pain rating scales, and the McGill pain questionnaire are most frequently used as self-rating instruments for pain rating. Therefore, we built a pain rating tool by combining all these scales and questionnaires.

B. Modeling for Clinical Screening Interview

In a structured clinical screening interview, a practitioner asks a list of predefined questions to collect information for the purpose of assessment. A general process of each question consists of four phases: 1) providing a setting or context of the question by saying statement or providing a sheet of pictures or descriptions; 2) asking the question; 3) waiting for responses and scoring or evaluating the responses; and 4) saying a closing statement for the question before moving to the next one. Listening responses are mainly applied in the first three phases. After each phase, there is a responding time given for the patient to answer, recognize, or memorize. Clinical screening interview using a social robot is performed by combining the robot's CI and graphical user interface on the touch screen. In this section, we propose a general clinical screening interview model for the structured clinical screening interviews using a social robot, such as cognitive assessment, pain rating, and falls' risk evaluation.

A structured clinical screening interview CLIN includes a list of N questions and expected answers $\mathrm{QA} = \{qa_i,\ i = 1,\ 2,\ \ldots,\ N\}$, which is designed by caregivers or practitioners and adapted to each patient at the interview time by the robot. Each question–answer pair qa features four components: a setting or context, a question, expected answers, and a post-question statement that provides a closing statement after each question. The question–answer pair is defined as follows:

$$\begin{aligned} qa &= \{\text{context},\ \text{question},\ \text{answer},\ \text{post\_question}\} \\ &= \{c,\ q,\ a,\ p\} \\ &= \{[vc, tc, ic],\ [vq, tq, iq],\ [ea, sc],\ [vp, tp, ip]\} \end{aligned}$$

where $[vc, tc, ic]$ are the verbal statement, response time, and GUI status of the context c; $[vq, tq, iq]$ are the verbal statement, responding time, and GUI status or graphical content of the question q; $[vp, tp, ip]$ are the verbal statement, responding time, and GUI status or graphical content of the post_question p; and ea and sc are expected answers and corresponding scores. The GUI status or graphical content can

be a sequence of images or figures displayed on the robot face, such as a speaking face, a normal face with blinked eyes, a smiling face, an image, a figure, or a video clip.

An example of a question is as follows:

```
{
"context": {
      "vc": "Pretend this figure is a clockface",
      "tc": 1,
      "ic": ["display:clockface"]},
"question": {
      "vq": "Please use your finger to draw the hour
         and minute hands at 15 minutes to 5 o'clock",
      "tq": 15,
      "iq": ["display:clockface"]},
"answer": {
      "ea": {"draw_clockface": {"hour_hand": 4,
                                "minute_hand": 45}},
      "sc": {"hour_hand": 1, "minute_hand": 1}},
"post_question": {
      "vp": "Thank you. Move to the next question",
      "tp": 1,
      "ip": ["display:robot_face"]}
}
```

The clinical screening interview Q&A sheets are prepared by remote caregivers or practitioners and sent to the robot through cloud services. Question generation constructs instructive statements, and questions and correct answers of the current Q&A sheet during a clinical screening interview. This module can be implemented in different approaches, such as generative models, retrieval-based models, and pattern-based heuristics. In this work, we implemented it using the heuristics-based approach. The questions and correct answers are defined and stored in the database. The statements and questions are selected from if-else conditional logic using a set of rules. The correct answers to the questions related to dates and time are automatically updated using the current time of a given assessment session. These sheets are converted into the above format. A Q&A sheet is created right before each clinical screening interview. All expected verbal answers of this Q&A sheet are used as example utterances to train a new intent, namely, clinic.answering, together with clinical intents (clinic.repeat, clinic.unknown, clinic.unclear, clinic.stop, clinic.pause, and clinic.change).

The GCSIM model for the robot to conduct an interview question is shown in Fig. 5. As mentioned earlier, the robot conducts an interview question through four phases as follows.

In the first phase starting at time t_0, the robot says an opening statement or a verbal context vc of the question and displays a graphical content ic on the robot face. During the time period of this phase from t_0 to t_1, the human face is tracked by the FT. If no face is detected or the human face is not in the center, the robot says a directive statement dlr_1 to get the human ready for answering the question, and the robot may repeat vc and ic.

In the second phase starting at time t_1, the robot asks a question vq and displays a graphical content iq on the robot face. It also checks human's attention in the same way as in the first phase and says a directive statement dlr_2 in response to the recognized attention.

The human is required to answer in a time period tq from t_2 to t_3 in the third phase using verbal or interaction on the touch screen as follows.

1) Verbal Responses: The robot shows the timer and records the human response by capturing audio phrases one by one. For example, the recorded audio phrases are s_1, s_2,..., s_M. Whenever any audio phrase is recorded, it is accumulatively merged to previous audio phrases, and a nondirective listening response nlr_i (e.g., a smiling face) is shown in the robot face to respond to the human voice. If s_i is recorded, the audio data S_i is sent to the Google speech recognition cloud service, where $S_i = s_1 + s_2 + \cdots + s_i$. The Google cloud service returns a word sequence W_i such that

$$W_i = \underset{W}{\text{argmax}}\, p(S_i|W)p(W) \tag{1}$$

where $p(S_i|W)$ is the acoustic model and $p(W)$ is the language model. The SLU parses W_i to detect clinical intents. If any nonanswering intent (clinic.repeat, clinic.unknown, clinic.unclear, clinic.stop, and clinic.pause) is detected, the robot says a corresponding directive listening response dlr_3 to instruct the human in answering the question; otherwise, the intent clinic.answering or clinic.change with its entity e_i is recognized, where e_i is expected information that may answer the question. If these intents are detected at low confidence, then e_i is set to be equal to W_i. The scoring module compares e_i with expected answers ea_i to evaluate and score the human's answer. If this answer is correct, the robot stops the timer and moves to the next phase (the fourth phase); otherwise, the robot keeps recording the human responses until the time is over. Based on the score and evaluation, the robot plays a short verbal cue to motivate the human before saying a closing statement for the question in the fourth phase.

2) Interaction Responses: If the question is required to be answered by human's interactions on the touch screen, the robot shows the timer and a graphical interface for the human to select answers for the question. If any interaction is detected, it is recorded and converted to a corresponding response that is also scored and evaluated in the same way as a verbal response.

C. Answer Evaluation and Scoring

1) Verbal Responses: As presented in Section V-B, the expected information e_i extracted from the human verbal response is scored by comparing with all expected answers ea_i for the question through syntax analysis. The comparison is performed in two steps. First, both the user response e_i and each expected answer of ea_i are analyzed and parsed using the Natural Language Toolkit (NLTK) [65]. The parsed results are converted into two ASCII trees, namely user response tree and correct answer tree. These trees represent the part of speech (POS) and the relationship between words. Second, a matching tree search is performed to compare these two trees. The answering score is awarded based on the matching between these two trees. For each question, the robot generates

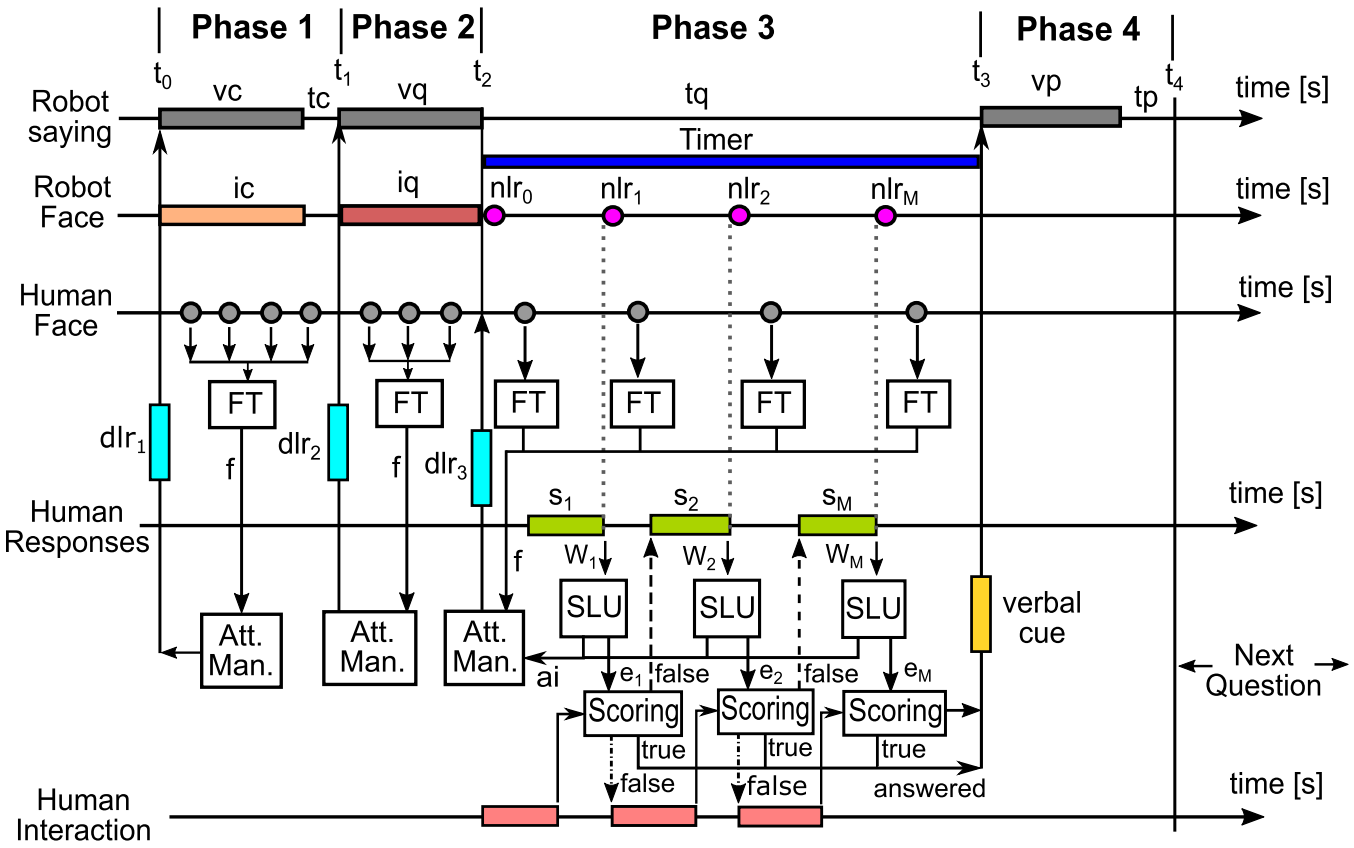

Fig. 5. GCSIM model: FT, SLU, Att. Man. (attention management), dlr (directive listening response), and nlr (nondirective listening response).

a score of 1 if the correct answer tree is a subtree of the user response tree, which means the human answered it correctly or 0 if the correct answer tree is not a subtree of the user response tree, which means that the user response was incorrect. It is also possible to be awarded a partial score if the answer is partially correct by evaluating the matching score of the two trees.

2) Touchscreen Interaction Responses: In the clinical screening interviews, there are three types of touchscreen interactions the human uses to answer questions. The first type is multiple-choice questions (e.g., selecting a shape). The human points to his/her choices on the screen, and they are recorded and scored. The second type is graphical marking questions. The human is asked to mark in an image shown on the screen (e.g., selecting body parts they feel the most pain). The image with marks is captured. The third type is drawing questions. The human is asked to draw a figure (e.g., drawing the hour and minute hands on a presented clockface). The drawn figure is captured and scored by comparing it with the reference figures. The example questions are shown in Fig. 6. This section mainly discusses the evaluation process for a drawing question.

The drawing question aims to check the coordination between hand movement and vision. For this question, the following steps are involved: 1) showing the requirement of the figure to be drawn; 2) giving a place for the figure to be drawn; 3) after the completion of the drawing, check for resemblance. The evaluation of manually drawn shapes requires many factors to be considered, which may be age-related

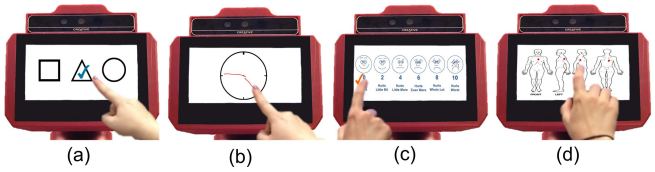

Fig. 6. Interaction responses. (a) Selecting a shape. (b) Drawing hour and minute hands. (c) Selecting the face that best describes the level of pain. (d) Selecting the body part with the most pain.

ailments, such as stiffness to hold and draw, diseases such as Parkinson's disease, steadiness, and vision impairment. If the user commits some minor mistakes, the robot should give positive evaluation results. In addition, drawing on a vertical screen yields a different experience compared with that of drawing on a paper. Considering all these factors, an evaluation module was developed. The process consists of two phases: similarity evaluation and quantitative evaluation. The evaluation methodologies used in these phases depend on each question. Evaluating the drawing of the hour and minute clock hands will be presented to demonstrate this process as follows.

a) Similarity evaluation: This phase estimates the similarity between the drawn figure and the references. For example, the robot asks the human to draw the hour and minute hands at 3:00 on a clockface, as shown in Fig. 7(a). The references of a clockface at 3:00 are shown in Fig. 7(b)–(e). This step includes the following phrases: 1) load the original color image of the drawn figure; 2) convert the image to

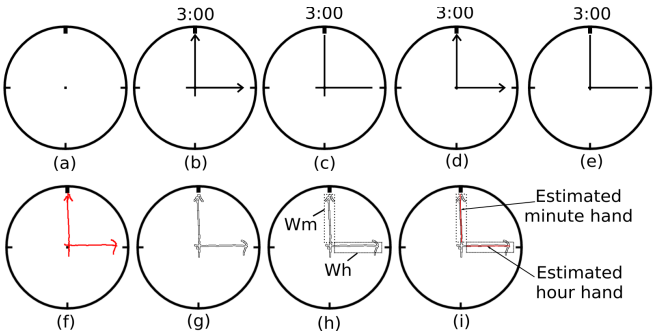

Fig. 7. Clockface drawing evaluation. (a) Clockface. (b)–(e) References of the hour and minute hands at 3:00. (f) Hour and minute hands drawn in the touch screen. (g) Edge of the clock hands. (h) Window of minute hand (Wm) and hour hand (Wh). (i) Estimated minute and hour hands.

gray scale; 3) convert the gray scale image to a binary image; 4) extract the contours from the binary image and the reference figures; and 5) compare the contour of the figure and that of each reference figure and match them. If there is sufficient similarity, then the score of similarity evaluation is returned true, otherwise false. If the similarity score is true, the quantitative evaluation is performed to evaluate the figure in detail.

b) Quantitative evaluation: This phase mainly estimates the size, position, and other features of the drawn figure. For example, this phase estimates the direction and size of the minute and hour hands drawn in the clockface through the following steps: 1) find out the edge points using the Laplacian of Gaussian (LoG) algorithm [66] in the image, as shown in Fig. 7(g); 2) collect two sets of edge points inside the windows of minute hand (Wm) and hour hand (Wh), as shown in Fig. 7(h); and 3) estimate the minute and hour hands from these data sets using the linear regression algorithm. The two data sets collected in the second step are MP and HP, where MP $= [(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)]$ and HP $= [(x_1, y_1), (x_2, y_2), \ldots, (x_H, y_H)]$). They are used to identify the minute and hour hands that can be represented as follows:

$$y = w_0 + w_1 x \tag{2}$$

where w_0 and w_1 are the coefficients of the equation. By using least-squares estimation, the coefficients (W_m and W_h) of the minute and hour hands are computed as follows:

$$W_m = \left[w_0^m \;\; w_1^m\right]^T = \left(X_m^T X_m\right)^{-1} X_m^T Y_m \tag{3}$$

$$W_h = \left[w_0^h \;\; w_1^h\right]^T = \left(X_h^T X_h\right)^{-1} X_h^T Y_h \tag{4}$$

where X_m, Y_m, X_h, and Y_h are given as follows:

$$X_m = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_M \end{bmatrix} Y_m = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_M \end{bmatrix}$$

$$X_h = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_H \end{bmatrix} Y_h = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_H \end{bmatrix}.$$

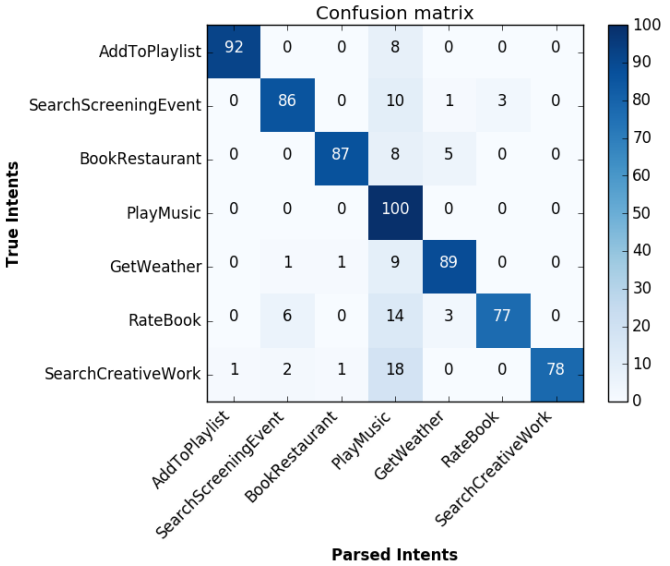

Fig. 8. Confusion matrix of the intent parser evaluated on the SNIPS benchmark.

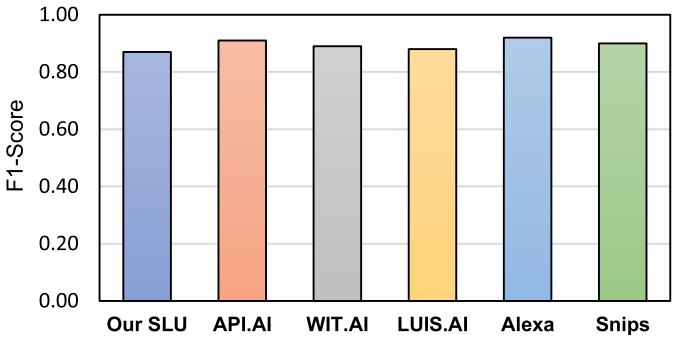

Fig. 9. Average F1-score of our SLU and five main existing NLU systems [67].

The minute and hour hands are estimated as follows:

$$y_i^m = w_0^m + w_1^m x_i, \quad i = [1, 2, \ldots, M] \tag{5}$$

$$y_i^h = w_0^h + w_1^h x_i, \quad i = [1, 2, \ldots, H]. \tag{6}$$

As shown in Fig. 7(i), the direction and size of the estimated hands are used to evaluate and score the response of the human to the drawing question.

VI. Experimental Results

This section presents our experiments and evaluations of the intent parser and the field tests of clinical interviews.

A. Intent Parser

We evaluated the intent parser using the SNIPS benchmark [67], which has been tested on several main existing NLU systems (Google's API.ai, Facebook's Wit, Microsoft's Luis, Amazon's Alexa, and Snips' NLU) for seven chosen intents, including AddToPlaylist, GetWeather, BookRestaurant, SearchCreativeWork, PlayMusic, SearchScreeningEvent, and RateBook. We used 300 sample utterances to train for each intent and 100 utterances for validation. The confusion matrix is shown in Fig. 8. As shown in Fig. 9, the average

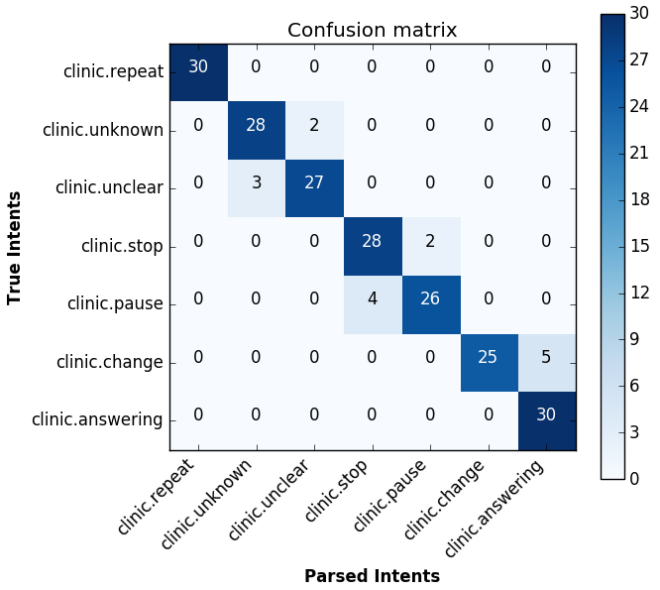

Fig. 10. Confusion matrix of the intent parser for clinical intents. There are a total of 30 validation samples for each intent.

F1-score of our intent parser achieves more than 0.87 and is compatible with the abovementioned NLU systems on the same benchmark.

Besides, we trained and evaluated clinical intents; 30 expected answers for the questions of clinical question–answer sheets are randomly selected to create example utterances for the clinic.answering intent. In addition, there are 30 query examples created for each of the other clinical intents. The intent parser for clinical screening interviews was evaluated using 5-fold cross validation. The confusion matrix of this evaluation is shown in Fig. 10. The average F1-score is more than 91%. For example, the response for the question "What day of the week is it?" is "Today is Monday oh no Tuesday." This response is correctly parsed as follows:

{"query": "Today is Monday oh no Tuesday", "intent": "clinic.change", "entity": {"first-answer": "Monday", "answer": "Tuesday"}, "confidence": 0.86 }.

B. Field Test and Evaluation

Thirty local older adults from 60 to 89 years old completed clinical screening interviews using our social robot during which human-administered evaluation was simultaneously performed. The demographic information about this group is as follows: (10 male, 20 female; Mean age: 73.4; and Standard deviation: 7.90); Races [White/Caucasian: 86.7%, Native Hawaiian/Pacific Islander: 6.7%, American Indian: 3.3%, and Multiracial: 3.3%]; Education [Some college: 30%, High School: 20%, College Degree: 16.6%, Graduate/Ph.D.: 26.6%, and Less than high school: 6.8%]; and Marriage status [Married: 50%, Widowed: 33.3%, Divorced/Separated: 13.3%, and Never Married: 3.4%]). These subjects were predominately in their early 70's, women, white-Caucasian, over 1/3 were college-educated, and most were currently married or widowed. All participants were asked to read and sign a university-approved institutional review board (IRB) consent form before participation. The clinical screening interview lasted about 50 min, which includes cognitive assessment, falls' risk evaluation, pain rating, health screening, mood evaluation, companionship assessment, memory testing, fatigue, and short blessed testing. Survey and debriefing were conducted, which consists of open-ended questions on the subjective experience in using the social robot. The field test scene is shown in Fig. 11. The clinical screening interviews were conducted by the robot, while the answers were evaluated by both the robot and an independent practitioner. Each subject was asked 120 to 128 questions. The answers were evaluated based on the categories of these questions, including Yes/No, Address/Name, Time, Location, Numerical Series, Count_backwards, Spelling, Command, Interaction, Verbal_Choice, Visual_Choice, and Others. The distribution of these categories is shown in Fig. 12. The robot successfully conducted all clinical screening interviews in the field test and received high trust and good feedback from the participants.

Fig. 11. Field test scene: the older adult (right), the local practitioner (middle), and the social robot (left). The clinical screening interviews are conducted by the robot and evaluated by both the robot and the independent practitioner.

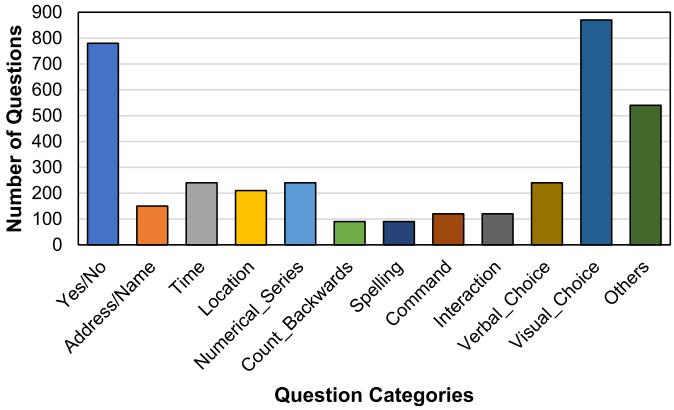

Fig. 12. Question categories of the conducted clinical screening interviews. There are a total of 3690 questions to be given to 30 subjects in the field tests.

We evaluated the scoring performance of the robot in the clinical screening interviews by comparing the points the robot and the practitioner score for each question. The percentage of questions the robot scores the same as the practitioner is shown by the blue bars in Fig. 13. The robot failed most in the spelling responses with a scoring accuracy of about 90% but was able to correctly evaluate and score more than 93% for the other responses, as shown in the blue bars. The incorrect evaluation occurred when there was wrong speech recognition or when the older adults answered in unexpected ways. The robot failed to recognize around 7% of the spelling

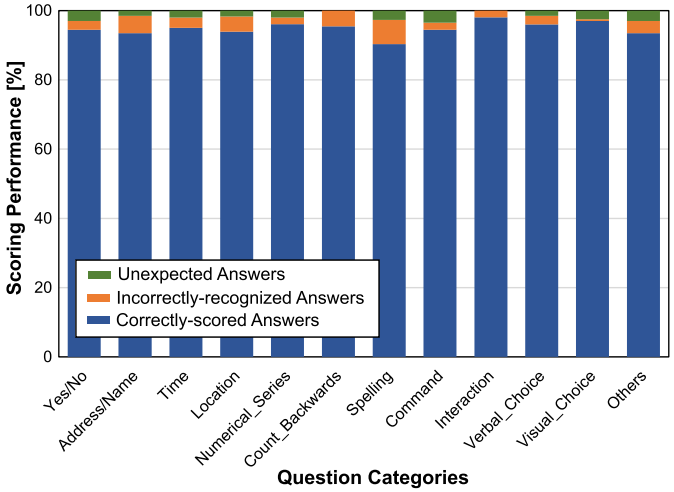

Fig. 13. Scoring performance: the percentage of answers the robot scores the same as the practitioner (blue bars), the rate of wrong speech recognition (orange bars), and the rate of the responses answered in unexpected ways (green bars).

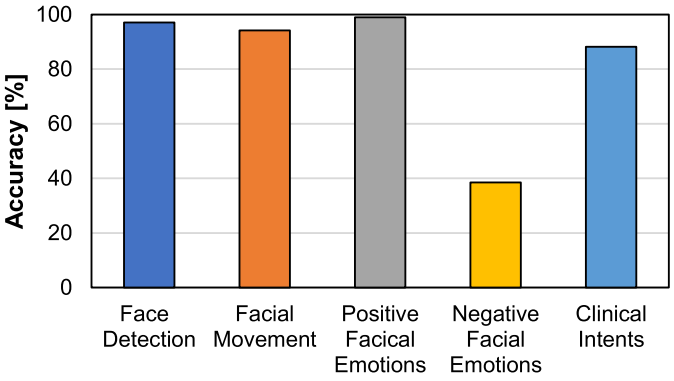

Fig. 14. Accuracy of face detection, facial movement recognition, facial emotion recognition, and clinical intent recognition in the filed test.

responses but less than 3% of almost any other responses, as shown by the orange bars in Fig. 13. There were less than 3% of the responses that were answered in unexpected ways, as shown by the green bars. The robot correctly evaluated all touch screen interaction responses that include the clockface figures drawn by the older adults.

Moreover, in order to evaluate the attention management, we tested the FT, facial movement and emotion recognition, and clinical intent recognition with 30 participants in the field test. As shown in Fig. 14, the accuracy of face detection is more than 97%. The robot mainly failed in detecting the side faces when the participants turned left or right. The accuracy of facial movement recognition is about 94%. The cloud-based facial emotion recognition mainly returned the neutral emotion and had low accuracy of other emotions. Comparing with the evaluation of the practitioner, the accuracy of recognizing negative facial emotions was below 40%. However, the robot can recognize clinical intents of the participants during clinical interviews in the filed test with an accuracy of 88.2%. Overall, attention management played an important role in handling the whole interview process and improving the human–robot communication in interview tasks.

In addition, we conducted post-surveys and debriefing with questions on the subjective experience after each participant completed the robot-administrated clinical interviews. The participants were asked to provide ratings from 0 to 10 on whether: 1) they have confidence that this robot can assess human well-being like a care provider (aide, nurse, and clinician) and 2) they have trust in this robot to assess human well-being. A total of 25 participants voluntarily completed the written debriefing survey. The other five chose not to do after completing the robot interaction. For confidence, there were 11% or 44% of the 25 subjects who indicated low confidence that a robot can assess human well-being as accurate as a care provider, whereas 56% indicated they had a high degree of confidence that this robot can assess human well-being comparable to a human care provider. For the trust, results indicated that 8% or 32% of subjects had a low degree of trust that a social robot could access well-being like a human care provider, whereas 17% or 68% indicated that they trusted this social robot to assess well-being. It is observed that older adults appear to trust the robot somewhat more than they may actually feel the confidence that the social robot can accurately rate well-being. A Chi-Square test was conducted, which results in $\chi^2 = 4.58$, $\rho = 0.03$, or < 0.05, suggesting that the subjects were significantly more trustful of the robot to perform a well-being assessment; then, they were confident that it could access well-being as accurate as a human care provider.

VII. Conclusion

In this article, we proposed and developed a framework that enables a social robot to perform clinical screening interviews for well-being assessment based on verbal communication. The framework is mainly powered by a CI capable of speech recognition, speech synthesis, SLU, and conversation management with the GCSIM model. The whole system was developed and implemented in our social robot. The robot has the capability of handling cognitive assessment, falls' risk evaluation, pain rating, health screening, mood evaluation, companionship assessment, memory testing, fatigue, and short blessed testing. A series of field tests on older adults was conducted, which showed the effectiveness of the social robots in performing a comprehensive geriatric well-being assessment. In future works, we will enhance the facial emotion recognition and integrate the facial emotions with the CI to assess the well-being of older adults. We would like to test the robot with older adults who suffer from memory loss or aging-associated diseases. We will also let the robot autonomously determine when to perform an assessment session. This work has the potential to be used for robot-assisted home geriatric care and healthcare.

References

[1] L. A. West, S. Cole, D. Goodkind, and W. He, "65+ in the United States: 2010," U.S. Census Bur., Suitland, MD, USA, Tech. Rep. P23-212, 2014.

[2] Apple. *The 2030 Problem: Caring for Aging Baby Boomers*. Accessed: Oct. 1, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1464018/

[3] J. Secker, R. Hill, L. Villeneau, and S. Parkman, "Promoting independence: But promoting what and how?" *Ageing Soc.*, vol. 23, no. 3, pp. 375–391, May 2003.

[4] J. Vespa, J. M. Lewis, and R. M. Kreider, "America's families and living arrangements: 2012," *Current Population Rep.*, vol. 20, no. 2013, p. P570, 2013.

[5] Alzheimer's Association, "2016 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 12, no. 4, pp. 1–80, 2016.

[6] R. Stepler, "Smaller share of women ages 65 and older are living alone: More are living with spouse or children," Pew Res. Center, Tech. Rep., 2016.

[7] CDC. *Accidents or Unintentional Injuries*. Accessed: Oct. 1, 2019. [Online]. Available: https://www.cdc.gov/nchs/fastats/accidental-injury.htm

[8] M. E. Tinetti and M. Speechley, "Prevention of falls among the elderly," *New England J. Med.*, vol. 320, no. 16, pp. 1055–1059, 1989.

[9] Key Points. (2005). *Paro Found to Improve Brain Function in Patients With Cognition Disorders*. pp. 1–6. [Online]. Available: http://www.parorobots.com/pdf/pressreleases/Paro found to improve BrainFunction.pdf

[10] S. Sabanovic, C. C. Bennett, W.-L. Chang, and L. Huber, "PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia," in *Proc. IEEE 13th Int. Conf. Rehabil. Robot. (ICORR)*, Jun. 2013, pp. 1–6.

[11] F. Erivaldo Fernandes, H. M. Do, K. Muniraju, W. Sheng, and A. J. Bishop, "Cognitive orientation assessment for older adults using social robots," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2017, pp. 196–201.

[12] M. Niemelä and H. Melkas, "Robots as social and physical assistants in elderly care," in *Human-Centered Digitalization and Services*. Cham, Switzerland: Springer, 2019, pp. 177–197.

[13] E. Harrington, H. Do, A. J. Bishop, C. Reese-Melancon, and A. W. Sheng, "Examining discrepancies in social robot versus human assessments of geriatric well-being," *Innov. Aging*, vol. 3, p. S329, Nov. 2019.

[14] *Paro Therapeutic Robot*. Accessed: Oct. 1, 2019. [Online]. Available: http://www.parorobots.com/

[15] E. Mordoch, A. Osterreicher, L. Guse, K. Roger, and G. Thompson, "Use of social commitment robots in the care of elderly people with dementia: A literature review," *Maturitas*, vol. 74, no. 1, pp. 14–20, Jan. 2013.

[16] D. Portugal, P. Alvito, E. Christodoulou, G. Samaras, and J. Dias, "A study on the deployment of a service robot in an elderly care center," *Int. J. Social Robot.*, vol. 11, no. 2, pp. 317–341, Apr. 2019.

[17] C.-H. King, T. L. Chen, A. Jain, and C. C. Kemp, "Towards an assistive robot that autonomously performs bed baths for patient hygiene," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 319–324.

[18] W. Yu, A. Kapusta, J. Tan, C. C. Kemp, G. Turk, and C. K. Liu, "Haptic simulation for robot-assisted dressing," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 6044–6051.

[19] D. Park, H. Kim, Y. Hoshi, Z. Erickson, A. Kapusta, and C. C. Kemp, "A multimodal execution monitor with anomaly classification for robot-assisted feeding," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017.

[20] H. M. Do, M. Pham, W. Sheng, D. Yang, and M. Liu, "RiSH: A robot-integrated smart home for elderly care," *Robot. Auto. Syst.*, vol. 101, pp. 74–92, Mar. 2018.

[21] F. Rudzicz, S. Raimondo, and C. Pou-Prom, "Ludwig: A conversational robot for people with Alzheimer's," in *Proc. Alzheimer's Assoc. Int. Conf.*, London, U.K., 2017, p. P164.

[22] J. R. Orozco-Arroyave *et al.*, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digit. Signal Process.*, vol. 77, pp. 207–221, Jun. 2018.

[23] F. Rudzicz, S. Raimondo, and C. Pou-Prom, "Ludwig: A conversational robot for people with Alzheimer's," *Alzheimer's Dementia*, vol. 13, no. 7, p. P164, Jul. 2017.

[24] E. Martinez-Martin and A. P. del Pobil, "Personal robot assistants for elderly care: An overview," in *Personal Assistants: Emerging Computational Technologies*. Cham, Switzerland: Springer, 2018, pp. 77–91.

[25] W.-L. Chang, S. Sabanovic, and L. Huber, "Use of seal-like robot PARO in sensory group therapy for older adults with dementia," in *Proc. 8th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2013, pp. 101–102.

[26] S.-C. Chen, W. Moyle, C. Jones, and H. Petsky, "A social robot intervention on depression, loneliness, and quality of life for Taiwanese older adults in long-term care," *Int. Psychogeriatrics*, Apr. 2020.

[27] R. H. Wang, A. Sudhama, M. Begum, R. Huq, and A. Mihailidis, "Robots to assist daily activities: Views of older adults with Alzheimer's disease and their caregivers," *Int. Psychogeriatrics*, vol. 29, no. 1, pp. 67–79, Jan. 2017.

[28] M. Pham, Y. Mengistu, H. Do, and W. Sheng, "Delivering home healthcare through a cloud-based smart home environment (CoSHE)," *Future Gener. Comput. Syst.*, vol. 81, pp. 129–140, Apr. 2018.

[29] M. Burtsev *et al.*, "The first conversational intelligence challenge," in *The NIPS'17 Competition: Building Intelligent Systems*. Cham, Switzerland: Springer, 2018, pp. 25–46.

[30] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–15.

[31] I. V. Serban *et al.*, "A deep reinforcement learning chatbot," 2017, *arXiv:1709.02349*. [Online]. Available: http://arxiv.org/abs/1709.02349

[32] S. Sukhbaatar, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.

[33] D. Ferrucci *et al.*, "Building Watson: An overview of the DeepQA project," *AI Mag.*, vol. 31, no. 3, pp. 59–79, 2010.

[34] A. E. Kurth *et al.*, "A comparison between audio computer-assisted self-interviews and clinician interviews for obtaining the sexual history," *Sexually Transmitted Diseases*, vol. 31, no. 12, pp. 719–726, Dec. 2004.

[35] C. L. Bethel *et al.*, "Using robots to interview children about bullying: Lessons learned from an exploratory study," in *Proc. 25th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2016, pp. 712–717.

[36] L. J. Wood, K. Dautenhahn, H. Lehmann, B. Robins, A. Rainer, and D. S. Syrdal, "Robot-mediated interviews: Do robots possess advantages over human interviewers when talking to children with special needs?" in *Proc. Int. Conf. Social Robot.* Cham, Switzerland: Springer, 2013, pp. 54–63.

[37] C. L. Sidner *et al.*, "Creating new technologies for companionable agents to support isolated older adults," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 3, pp. 1–27, 2018.

[38] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*, 1st ed. Cham, Switzerland: Springer, 2016.

[39] J. Sommers-Flanagan and R. Sommers-Flanagan, *Clinical Interviewing: Intake, Assessment & Therapeutic Alliance*. Mill Valley, CA, USA: Psychotherapy.net, 2014.

[40] *Ps3eye Camera*. Accessed: Oct. 1, 2019. [Online]. Available: http://www.sony.co.in/product/

[41] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system 'HARK'—Open source software for listening to three simultaneous speakers," *Adv. Robot.*, vol. 24, nos. 5–6, pp. 739–761, Jan. 2010.

[42] H. Manh Do, W. Sheng, and M. Liu, "An open platform of auditory perception for home service robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 6161–6166.

[43] H. M. Do, W. Sheng, and M. Liu, "Human-assisted sound event recognition for home service robots," *Robot. Biomimetics*, vol. 3, no. 1, p. 7, Dec. 2016.

[44] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.

[45] Microsoft. *Speech Api*. Accessed: Oct. 1, 2019. [Online]. Available: http://www.bing.com/partners/developers

[46] Amazon. *Ivona Text-to-Speech*. Accessed: Oct. 1, 2019. [Online]. Available: https://www.ivona.com/

[47] IBM. *Text to Speech*. Accessed: Oct. 1, 2019. [Online]. Available: https://www.ibm.com/watson/developercloud/text-to-speech.html

[48] Google. *Powerful Speech Recognition*. Accessed: Oct. 1, 2019. [Online]. Available: https://cloud.google.com/speech/

[49] A. Škraba, V. Stanovov, E. Semenkin, A. Koložvari, and D. Kofjač, "Development of algorithm for combination of cloud services for speech control of cyber-physical systems," *Int. J. Inf. Technol. Secur.*, vol. 10, no. 1, pp. 73–82, 2018.

[50] S. Yu, "Regular languages," in *Handbook of Formal Languages*. Cham, Switzerland: Springer, 1997, pp. 41–110.

[51] K. Thompson, "Programming techniques: Regular expression search algorithm," *Commun. ACM*, vol. 11, no. 6, pp. 419–422, Jun. 1968.

[52] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*. [Online]. Available: http://arxiv.org/abs/1607.01759

[53] A. Aizawa, "An information-theoretic perspective of TF-IDF measures," *Inf. Process. Manage.*, vol. 39, no. 1, pp. 45–65, Jan. 2003.

[54] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[55] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882
[56] M. Firdaus, A. Kumar, A. Ekbal, and P. Bhattacharyya, "A multi-task hierarchical approach for intent detection and slot filling," *Knowl.-Based Syst.*, vol. 183, Nov. 2019, Art. no. 104846.
[57] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "Mini-mental state," *J. Psychiatric Res.*, vol. 12, no. 3, pp. 189–198, Nov. 1975.
[58] Psychological Assessment Resources. 2nd Edition (MMSE-2). *Mini-Mental State Examination*. Accessed: Oct. 1, 2019. [Online]. Available: http://www4.parinc.com/Products/Product.aspx?ProductID=MMSE-2
[59] Z. S. Nasreddine *et al.*, "The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment," *J. Amer. Geriatrics Soc.*, vol. 53, no. 4, pp. 695–699, Apr. 2005.
[60] P. S. Mathuranath, P. J. Nestor, G. E. Berrios, W. Rakowicz, and J. R. Hodges, "A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia," *Neurology*, vol. 55, no. 11, pp. 1613–1620, Dec. 2000.
[61] A. J. Larner, *Dementia in Clinical Practice: A Neurological Perspective*. London, U.K.: Springer, 2012. [Online]. Available: http://link.springer.com/10.1007/978-1-4471-2377-4
[62] K. L. Perell, A. Nelson, R. L. Goldman, S. L. Luther, N. Prieto-Lewis, and L. Z. Rubenstein, "Fall risk assessment measures: An analytic review," *J. Gerontol. A, Biol. Sci. Med. Sci.*, vol. 56, no. 12, pp. M761–M766, Dec. 2001.
[63] J. A. Stevens, "The STEADI tool kit: A fall prevention resource for health care providers," *IHS Primary Care Provider*, vol. 39, no. 9, p. 162, 2013.
[64] S. B. McMahon, M. Koltzenburg, I. Tracey, and D. Turk, *Wall & Melzack's Textbook of Pain*. Amsterdam, The Netherlands: Elsevier, 2013.
[65] S. Bird, E. Klein, and E. Loper. *Natural Language Toolkit*. Accessed: Oct. 1, 2019. [Online]. Available: http://www.nltk.org/
[66] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
[67] Snips. *Natural Language Understanding Benchmark*. Accessed: Oct. 1, 2019. [Online]. Available: https://github.com/snipsco/nlu-benchmark

Ha Manh Do (Member, IEEE) received the B.Sc. degree in electronics and telecommunications from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Oklahoma State University (OSU), Stillwater, OK, USA, in 2015 and 2018, respectively.

He was a Post-Doctoral Fellow with OSU in spring 2019. He is currently a CBASE Postdoctoral Researcher/Professor of the Practice in Engineering with the Department of Engineering, Colorado State University-Pueblo, Pueblo, CO, USA. His research interests include home service robots and smart homes for geriatric care and healthcare, human–robot interaction, robotic perception, artificial intelligence, natural language processing, and deep learning.

Weihua Sheng (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 1994 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from Michigan State University, East Lansing, MI, USA, in May 2002.

He is currently an Associate Professor with the School of Electrical and Computer Engineering, Oklahoma State University (OSU), Stillwater, OK, USA, where he is also the Director of the Laboratory for Advanced Sensing, Computation and Control (ASCC Lab). He is the author of more than 200 peer-reviewed articles in major journals and international conferences. His current research interests include social robots, human–robot interaction, wearable computing, and intelligent transportation systems. His research has been supported by the U.S. National Science Foundation (NSF), the Department of Defense (DoD), the Oklahoma Transportation Center (OTC)/Department of Transportation (DoT), and so on.

Dr. Sheng's eight articles won the best paper or best student paper awards in major international conferences and journals. He has served as an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING from 2010 to 2019.

Erin E. Harrington received the M.S. degree in psychology from Oklahoma State University (OSU), Stillwater, OK, USA, in 2017, where she is currently pursuing the dual-track degree in cognitive and developmental psychology and the Ph.D. degree with the Experimental Psychology Program.

Her primary research interests are in prospective memory, mnemonic strategies, and cognitive aging.

Alex J. Bishop is currently an Associate Professor with the Human Development and Family Science Department and the College of Human Sciences Bryan Close Professor of Adulthood and Aging with Oklahoma State University, Stillwater, OK, USA.

His research is focused on the interplay of biopsychosocial processes that contribute to perceived well-being and healthy longevity. Of particular interest is on understanding how such variables may impact the ability of old and very old adults to access and manage resources that promote successful and long-term aging-in-place.