Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

From Eliza to XiaoIce: challenges and opportunities with social chatbots

Heung-yeung SHUM‡, Xiao-dong HE, Di LI

Microsoft Corporation, Redmond, WA 98052, USA

E-mail: hshum@microsoft.com; xiaohe@microsoft.com; lidi@microsoft.com

Received Dec. 10, 2017; Revision accepted Jan. 8, 2018; Crosschecked Jan. 8, 2018

Abstract: Conversational systems have come a long way since their inception in the 1960s. After decades of research and development, we have seen progress from Eliza and Parry in the 1960s and 1970s, to task-completion systems as in the Defense Advanced Research Projects Agency (DARPA) communicator program in the 2000s, to intelligent personal assistants such as Siri, in the 2010s, to today's social chatbots like XiaoIce. Social chatbots' appeal lies not only in their ability to respond to users' diverse requests, but also in being able to establish an emotional connection with users. The latter is done by satisfying users' need for communication, affection, as well as social belonging. To further the advancement and adoption of social chatbots, their design must focus on user engagement and take both intellectual quotient (IQ) and emotional quotient (EQ) into account. Users should want to engage with a social chatbot; as such, we define the success metric for social chatbots as conversation-turns per session (CPS). Using XiaoIce as an illustrative example, we discuss key technologies in building social chatbots from core chat to visual awareness to skills. We also show how XiaoIce can dynamically recognize emotion and engage the user throughout long conversations with appropriate interpersonal responses. As we become the first generation of humans ever living with artificial intelligenc (AI), we have a responsibility to design social chatbots to be both useful and empathetic, so they will become ubiquitous and help society as a whole.

Key words: Conversational system; Social Chatbot; Intelligent personal assistant; Artificial intelligence; XiaoIce
https://doi.org/10.1631/FITEE.1700826 **CLC number:** TP391

1 Introduction

One of the fundamental challenges in artificial intelligence (AI) is endowing the machine with the ability to converse with humans using natural language (Turing, 1950). Early conversational systems, such as Eliza (Weizenbaum, 1966), Parry (Colby, 1975), and Alice (Wallace, 2009), were designed to mimic human behavior in a text-based conversation, hence to pass the Turing test (Turing, 1950; Shieber, 1994) within a controlled scope. Despite impressive successes, these systems, which were precursors to today's social chatbots, were mostly based on hand-crafted rules. As a result, they work well only in constrained environments.

Since the 1990s, a lot of research has been done on task-completion conversational systems (Hemphill et al., 1990; Price, 1990; Dahl et al., 1994; Walker et al., 2001, 2002). Examples include systems for reserving airline tickets as in the Defense Advanced Research Projects Agency (DARPA) Airline Travel Information System (ATIS) project and for travel planning as in the DARPA communicator program. ATIS and communicator systems are designed to understand natural language requests and perform a variety of specific tasks for users, such as retrieving flight information and providing information to tourists. Task-completion conversational systems have been typically based on data-driven, machine-learned

‡ Corresponding author
ORCID: Heung-yeung SHUM, http://orcid.org/0000-0002-4684-911X

approaches. Their performance is excellent only within domains that have well-defined schemas (Glass et al., 1995; Walker et al., 2001; Raux et al., 2005; Andreani et al., 2006; Tur and de Mori, 2011; Wang et al., 2011).

In the past several years, a tremendous amount of investment has been made to developing intelligent personal assistants (IPAs), such as Apple's Siri (https://www.apple.com/ios/siri/), Microsoft's Cortana (https://www.microsoft.com/en-us/cortana/), Google Assistant, Facebook M (https://developers.facebook.com/blog/post/2016/04/12/), and Amazon's Alexa (https://developer.amazon.com/alexa/). These IPAs are often deployed on mobile devices and are designed to answer a wide range of questions. In addition to passively responding to user requests, they proactively anticipate user needs and provide in-time assistance such as reminding of an upcoming event or recommending a useful service without receiving explicit requests from the user (Sarikaya, 2017). The daunting challenge for such IPAs is that they must work well in many open domain scenarios as people learn to depend on them to manage their work and lives efficiently.

More recently, social chatbots, such as Microsoft's XiaoIce (https://www.msXiaoIce.com/), have emerged as a new kind of conversational system made possible by significant progress in AI and wireless communication. The primary goal of a social chatbot is not necessarily to solve all the questions the users might have, but rather, to be a virtual companion to users. By establishing an emotional connection with users, social chatbots can better understand them and therefore help them over a long period of time. To communicate effectively, social chatbots interact with users through multiple modalities, including text, speech, and vision. Social chatbots and IPAs have become popular recently due to progress in many relevant perceptual and cognitive AI technologies, e.g., natural language understanding (Bengio et al., 2003; Mikolov et al., 2013; Sutskever et al., 2014; Bahdanau et al., 2014; Mesnil et al., 2013, 2015), speech recognition and synthesis (Hinton et al., 2012; Deng et al., 2013; Qian et al., 2014; van den Oord et al., 2016; Xiong et al., 2016), computer vision (Krizhevsky et al., 2012; He et al., 2016), information retrieval (Huang et al., 2013; Elkahky et al., 2015), multimodal intelligence (Fang et al., 2015; Karpathy and Li, 2015; Vinyals et al., 2015; He and Deng, 2017), and empathic conversational systems (Fung et al., 2016; Alam et al., 2017).

In the next section, we briefly review conversational systems developed over the past several decades and discuss their features and limitations. We then introduce the design principles of social chatbots where both intellectual quotient (IQ) and emotional quotient (EQ) must be considered. We describe key components in designing social chatbots including core chat, visual awareness, and skills. Using Microsoft's XiaoIce, we show that social chatbots can engage in long conversations with users by understanding their needs, establishing an emotional connection, and offering support and assistance. We also share our vision for the evolution of chatbots in the future. We discuss that, with fundamental breakthroughs to occur on key open problems, the tremendous business opportunities they will make possible, and their potential impact on society as AI becomes an integral part of our lives.

2 Brief survey of conversational systems

In this section, we review representative conversational systems that have been developed over five decades. Such systems include early chatbots, task-completion systems, intelligent personal assistants, and social chatbots.

2.1 Eliza and early chatbots

Inspired by the Turing test proposed in 1950 (Turing, 1950), researchers and engineers have developed multiple conversational systems for chitchat (Weizenbaum, 1966; Colby, 1975; Shieber, 1994; Wallace, 2009). These early chatbots are computer programs that conducted a conversation using audio or text. Such programs were often designed to convincingly simulate how a human would behave as a conversational partner; if they succeed, they pass the Turing test.

Eliza, created by Joseph Weizenbaum at MIT in 1966, is perhaps the very first chatbot known publicly. It can communicate with humans based on hand-crafted scripts (Weizenbaum, 1966). These scripts simulate a Rogerian psychotherapist and can accept only textual input. It does not understand a conversation but rather searches for appropriate responses through pattern matching, combined with some intelligent phrasing. Eliza has a limited scope of knowledge and can only chat with people within only a constrained domain. Nevertheless, many users believed they were talking to a real person when Eliza first came out. Fig. 1 shows a conversation between Eliza and a human.

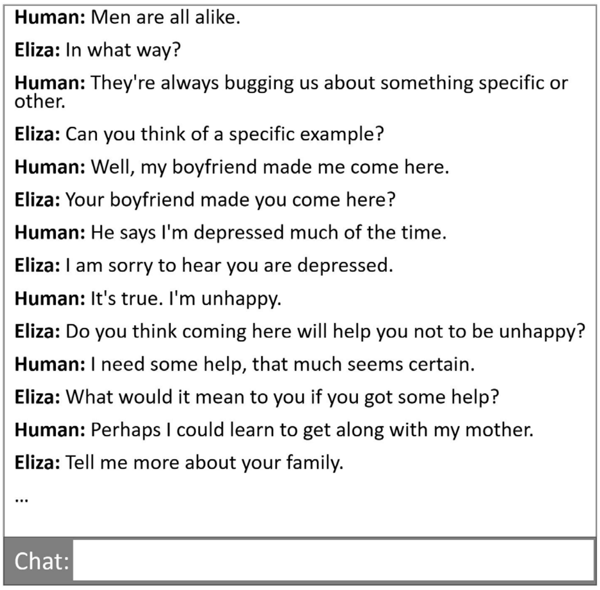


Fig. 1 A conversation between Eliza and a human (Weizenbaum, 1966)

Parry is another chatbot developed by Colby (Colby, 1975). It is designed to behave like a paranoid person. It passed the Turing test for the first time in history (https://www.chatbots.org/chatbot/parry/). However, Parry is still rule-based and has a similar structure to Eliza but with better controlling structure, language understanding capabilities, and especially a mental model that can simulate the bot's emotions. For example, Parry will respond with hostility if the anger level is high.

Alice, or artificial linguistic internet computer entity, was developed by Wallace (Wallace, 2009) to allow users to customize their chatbots. It uses an artificial intelligence markup language (AIML), a derivative of extensive markup language (XML), and AIML has tags that allow bots to recursively call a pattern matcher so that the language can be simplified. Alice won the Loebner Prize (https://en.wikipedia.org/wiki/Loebner_Prize) three times in 2000, 2001, and 2004, an award for the most human-like system (Shieber, 1994). However, because of the limitations of AIML, the capacities of these chatbots are constrained. For example, Alice failed to pass the ultimate Turing test (https://en.wikipedia.org/wiki/artificial_linguistic_internet_computer_entity), partially because chitchat systems built using AIML cannot maintain a dialogue for a long period of time.

2.2 Task-completion conversational systems

In contrast to chitchat systems, task-completion systems are designed for accomplishing specific tasks. These systems usually operate on constrained domains (Glass et al., 1995; Seneff et al., 1998; Rudnicky et al., 1999; Levin et al., 2000; Walker et al., 2001, 2002; Raux et al., 2005; Andreani et al., 2006; Tur and de Mori, 2011; Wang et al., 2011). Fig. 2 illustrates the architecture of a traditional task-completion spoken dialog system.

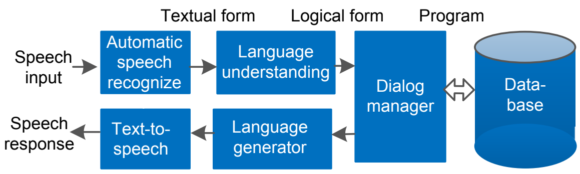


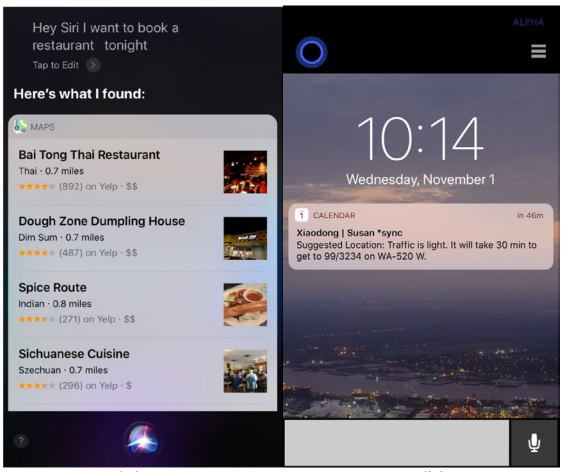
Fig. 2 Illustration of a task-completion system

This architecture comprises an automatic speech recognizer (ASR), a spoken language understanding (SLU) module, a dialog manager (DM), a natural language generator (NLG), and a text-to-speech (TTS) synthesizer. The ASR takes raw audio signals, transcribes them into word hypotheses, and transmits the hypotheses to the SLU module. The goal of the SLU module is to capture the core semantics of the given sequence of words (the utterance). It identifies the dialog domain and the user's intent and parses the semantic slots in the user's utterance. DM's goal is to interact with users and assist them in achieving their goals. It checks if the required semantic representation is filled and decides the system's action. It accesses the knowledge database to acquire the desired information the user is looking for. DM also includes dialog state tracking and policy selection, so that the dialog agent can make more robust decisions (Williams and Young, 2007). More recent work focuses on building end-to-end systems where multiple components are jointly optimized to cope with the large variability occurring in dialog systems (He and Deng, 2013; Sarikaya et al., 2016; Wen et al., 2016).

2.3 Intelligent personal assistants

Apple released Siri in 2011. Since then, several intelligent personal assistants (IPAs) have been built and introduced to the market, e.g., Cortana from Microsoft (Microsoft also released a Voice Command app in 2003 which can control Windows Mobile devices by voice), Google Assistant, and Alexa from Amazon. IPAs integrate information from multiple sensors including location, time, movement, touch, gestures, and eye gaze, and have access to various data sources such as music, movies, calendars, emails, and personal profiles. As a result, they can provide a broad set of services covering a wide range of domains. For certain requests that cannot be directly answered, IPAs often default to Web search as backup.

IPAs provide reactive and proactive assistance to users to accomplish a variety of tasks (Sarikaya, 2017). For example, reactive assistance includes information consumption such as weather report, and task assistance such as restaurant reservation (Fig. 3a). In contrast, proactive assistance includes reminding the user of upcoming events, or recommending specific products or services to the user, according to the user's profile and relevant contextual information



(a) (b)

Fig. 3 Examples of intelligent personal assistants actions, recommending a restaurant (reactive assistance) by Siri (a), and reminder of an upcoming event with relevant traffic information (proactive assistance) by Cortana (b)

such as time and location (Fig. 3b). IPAs undergo continual improvements on major mobile phone platforms, personal computers, smart home devices (e.g., intelligent speakers), and wearable devices (e.g., smart watches), with the help of seamless integration of multiple services and convenient natural user interfaces.

2.4 Social chatbots

The current social media age has been made possible with the proliferation of smartphones and advancement of broadband wireless technology. With vastly more people being digitally connected, it is not surprising that social chatbots have been developed as an alternative means for engagement. Unlike early chatbots designed for chitchat, social chatbots have been created to serve users' needs for communication, affection, and social belonging rather than for passing the Turing test. Therefore, social chatbots must be able to recognize emotion and track emotional changes during a conversation.

Social chatbots can also perform a variety of tasks for users in the context of casual chats. For this purpose, social chatbots must develop a set of skills to accommodate users' requests. Interestingly, unlike task-completion systems and IPAs that are designed for efficiency (i.e., accomplishing tasks and ending the conversation as quickly as possible), social chatbots take time to converse like a human, present results, offering perspectives, prompting new topics to keep the conversation going.

XiaoIce has been the most widely deployed social chatbot since it was released by Microsoft in May 2014. It understands users' emotional needs and engages in interpersonal communications like a friend, cheering users up, encouraging them, and holding their attention throughout the conversation.

Users have provided feedback that conversations with XiaoIce have resulted in a more positive outlook, a feeling of emotional support, and a sense of social belonging. Such conversations have helped in building trust and an emotional connection between human users and social chatbots, providing the opportunity for bots to better understand users, and serve them more effectively. Some chat examples between XiaoIce and users are shown in Fig. 4.

These major conversational systems discussed in this section are summarized in Table 1. For the

Table 1 Summary of major conversational systems

Metric	ELIZA	PARRY	ALICE	DARPA communicator program	Siri	XiaoIce
Time	1966	1972	1995	2000	2011	2014
Scalability	None	None	Scripts can be customized	Limited	Scalable	Scalable
Key features	Mimicking human behavior in conversation	Generating emotional (angry) responses	Easy customization of scripts (via AIML)	Language understanding and dialog management; gole-oriented	Providing personal digital assistance	Building emotional attachment to users; scalable skill set for user assistance
Accomplishment	First chitchat bot	Passed Turing test	Won the Loebner Prize three times	Understand natural language requests and performing tasks	The first widely deployed intelligent personal assistant (IPA)	The first widely deployed social chatbot. 100MM users; published poem book; host TV programs
Modality	Text only	Text only	Text only	Text and voice	Text, image, voice	Text, image, voice
Modeling	Rule-based	Rule-based	Rule-based	Learning-based	Learning-based	Learning-based
Domain	Constrained domain	Constrained domain	Constrained domain	Constrained domain	Open domain	Open domain
Key technical breakthrough	Use of scripts, keyword-based pattern matching; rule-based response	Adding personality characteristics into responses	Using AIML and recursion for pattern matching; multiple patterns can be mapped into same response	Statistical models for spoken language understanding and dialog management	Provide both reactive assistance covering a wide range of domains	Emotional intelligence models for establishing emotional attachments with users
Key technical limitation	Limited domain of knowledge	Limited domain of knowledge	Size of scrip can be huge	Work only in domains that have well-defined schemas	Lack of emotional engagement with users	Inconsistent personality and responses in long dialogue

remainder of the paper, we will focus on social chatbots, starting with their design principles.

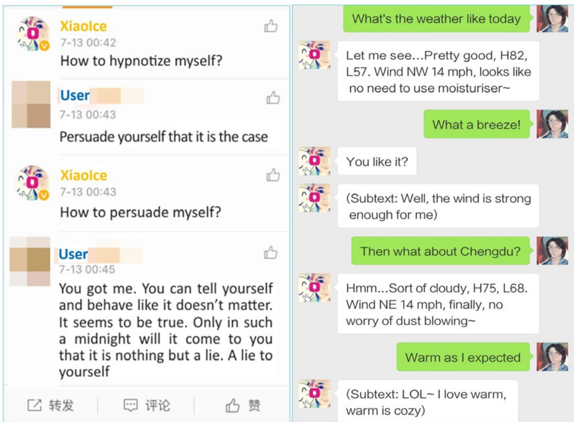
3 Design principles of social chatbots

3.1 EQ+IQ

The primary goal of social chatbots is to be AI companions to humans with an emotional connection. Emotional affection and social belonging are some of the fundamental needs for human beings (Maslow, 1943). Therefore, building social chatbots to address these emotional needs is of great value to our society. To fulfill these needs, a social chatbot must demonstrate a sufficient EQ (Beldoch, 1964; Gardner, 1983; Goleman, 1995, 1998; Murphy, 2007). Accordingly, a social chatbot needs to develop the following capabilities: empathy, social skills, personality, and integration of EQ and IQ.

Understanding users: A social chatbot must have empathy. It needs to have the ability to identify the user's emotions from the conversation, to detect how emotions evolve over time, and to understand the user's emotional needs. This requires query understanding, user profiling, emotion detection, sentiment recognition, and dynamically tracking the mood of the user in a conversation. Modeling of contextual information in the dialog session and commonsense knowledge is also critical for user understanding.

Interpersonal response generation: A social chatbot must demonstrate sufficient social skills. Users have different backgrounds, varied personal interests, and unique needs. A social chatbot needs to have the ability to personalize the generation of responses for different users. It needs to generate responses that are emotionally appropriate, encouraging and motivating, and fit the interests of the user. It may generate responses in attractive styles (e.g., having a sense of humor) that improve user engagement. It needs to guide conversation topics and manage an amicable relationship in which the user feels he/she is well understood and is inspired to continue to converse with the bot. It should also be aware of inappropriate information and can avoid generating biased responses for instance.



(a) (b)

Fig. 4 Chat examples between XiaoIce and users, showing the emotional connection (a) (the full conversation session is shown in Fig. 14), and how to invoke a skill (e.g., weather report) in a casual chat. Note that XiaoIce offers a perspective about the weather, e.g., "no need to use moisturizer~" (b)

Personality: A social chatbot needs to present a consistent personality, so that it can gain the user's confidence and trust. A consistent personality of the chatbot helps the user to set the right expectation in the conversation. Personality settings include age, gender, language, speaking style, general (positive) attitude, level of knowledge, areas of expertise, and a proper voice accent. These settings will influence the generation of responses to the user. Moreover, the bot needs to continuously learn and improve from the interactions with users through active and adaptive learning.

Integration of both EQ and IQ: Beyond chitchat, social chatbots need to acquire a range of skills to help complete some specific tasks for users. They need to analyze users' requests and perform necessary reasoning and execution to respond to these requests, e.g., answering a question or taking an action. A sufficiently high IQ is required for a social chatbot. IQ capacities include knowledge and memory modeling, image and language understanding, reasoning, generation, and prediction. These IQ capabilities are not only the technical foundations of various skills, but also essential for building high level EQ capabilities.

Social chatbots should deliver results in such a way that they are easily understood by users. They should also suggest or encourage new topics to extend the conversation further. For instance, Fig. 5 shows how IQ and EQ are combined in a conversation. The chatbot first parses the user's question (area of China) and then infers the answer (3.71 million square miles). Then the chatbot presents the answer more like a human, with the perspective of being aware of the user's level of knowledge (knowing how big the US is).

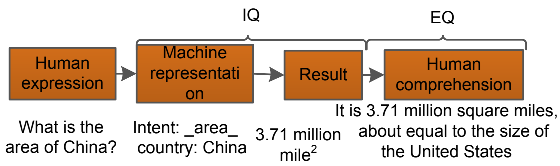


Fig. 5 Both IQ and EQ play key roles in a social chatbot. Not only is the area of China presented, but this number is made understandable by comparson with the US, which the chatbot believes the user should know

Fig. 6 shows another way of EQ and IQ integration. Rather than presenting the results to the user directly, sometimes a social chatbot will generate responses that might help to understand the user better and inspire more interesting topics for the conversation. In this example, the user asks the current time. The chatbot does not tell the time immediately, instead replies with something relevant, in an attempt to better understand the user intent. The chatbot does show the correct answer at the end of the conversation, and then proactively tries to extend the chat by asking if the user is planning for a new trip.

Social chatbots should be able to communicate with users through different modalities (e.g., text, voice, image, and video); as a result, they need high IQ for speech, text, and visual understanding. While text messages are the most common, the user could also speak to the chatbot or simply share an image. The chatbot needs to be able to parse the text, recognize the speech, or detect the salient information in the image to understand user intent. The chatbot will also respond with text, speech, or visual output, depending on the context.

(a) (b)

Fig. 6 A chat example between XiaoIce and a user, in English translation (a) and in Chinese (b), showing that both IQ and EQ are important for a social chatbot
The bot knows the answer. However rather than returning the answer directly, it attempts to lead the chat to a more interesting direction and extend the conversation.

3.2 Social chatbot metrics

Unlike task-completion conversational systems where their performance can be measured by task success rate, measuring the performance of chatbots is difficult (Shawar and Atwell, 2007; Yu et al., 2016). In the past, the Turing test and its extensions have been used to evaluate chitchat performance (Shieber, 1994). However, the Turing test is not a proper metric for evaluating the success of emotional engagement with users. Instead, we define conversation-turns per session (CPS) as the success metric for social chatbots. It is the average number of conversation-turns between the chatbot and the user in a conversational session. The larger the CPS is, the better engaged the social chatbot is.

Interestingly, conversational systems can be categorized by their targeting CPS. As shown in Table 2, web search, for example, is essentially a question-and-answering system, and thus is expected to return the answer immediately, i.e. in one single step. Being not able to find the target web link in one step is regarded as a failure for the search engine. For IPAs, to understand what the user is asking, e.g., checking weather or inquiring about business hours, we expect the system to typically ask a couple of clarifying questions before returning the correct information. For more complicated tasks such as customer services or travel planning, however, the system is expected to need several turns of conversation to resolve issues (e.g., filling in forms with user and product information). Finally, for a social chatbot, the system is expected to sustain a rather long conversation with the user to fulfill the needs of affection and belonging. Social chatbots are designed to keep users continuously engaged if possible.

Table 2 Expected conversation-turns per session (CPS) for different types of conversational systems

System	Expected CPS
Web search	1
Personal assistant	1–3
Task-completion	3–7
Social chatbot	10+

4 Framework and components of social chatbots

In this section, we describe the framework and components of a typical social chatbot, namely chat manager, core chat, visual awareness, and skills.

4.1 Overall framework

An overall architecture for designing social chatbots is shown in Fig. 7. First, the system has a multimodal interface to receive the user's input in text, image, and voice. The system has a chat manager to dispatch the input to proper modules, such as core-chat or visual awareness, for understanding the input and generating the output. Given different scenarios, the chat manager will invoke various skills, send the user's request to corresponding skill components, and obtain a response from them. The chat manager will then coordinate relevant modules to generate the output that fits the context of the current conversation. We will elaborate core-chat, visual awareness and skills in detail in this section.

4.2 Core-chat

Core-chat is the central module of social chatbots. It receives the text input from the user and generates a text response as the output. It provides the communication capability of social chatbots. Fig. 8 shows key components in core-chat.

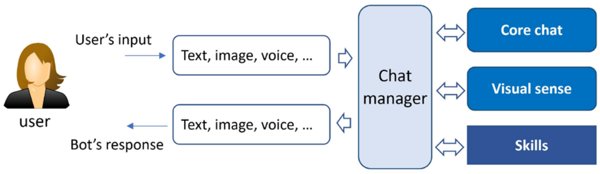


Fig. 7 Architecture of a social chatbot

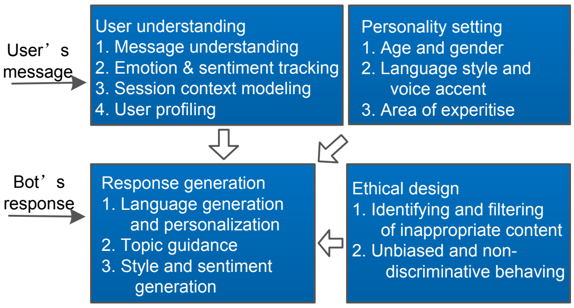


Fig. 8 Architecture of the core-chat module

First, the user's input is sent to a user understanding component, where semantic encoding and intent understanding are performed (Tur and Deng, 2011; Liu et al., 2015; Vinyals and Le, 2015). It also detects the sentiment reflected in the input message and infers the user's emotion status (Tokuhisa et al., 2008; Mower et al., 2011; Socher et al., 2013; Chen et al., 2016; Yang et al., 2016b). The contextual information from the current dialog session is usually extracted and used for understanding the current message. To better understand the user intent and emotion, the social chatbot maintains a profile for each user, which stores each user's basic information such as age, gender, background, and interests. The user profile also tracks certain dynamic information, such as emotion status, which will be updated frequently. To more precisely understand the user intent, knowledge bases such as Freebase (http://www.freebase.com) and the Microsoft Concept Graph (Wang et al., 2015) can be used.

The processed information is then sent to a response-generation component to produce responses. The response candidates are typically generated by two approaches: retrieval-based (Lu and Li, 2013; Li X et al., 2016; Yan et al., 2016) and generation-based (Sordoni et al., 2015; Vinyals and Le, 2015; Li JW et al., 2016; Serban et al., 2017).

In the retrieval-based approach, a chat index is first constructed from a database of message-response pairs that are crawled from human-to-human dialogs, e.g., from social networks. All responses are indexed by the messages that invoke them. At runtime, the input message from the user is treated as a raw query, and an information retrieval (IR) model like those used in web search, is used to retrieve similar messages in the chat index and return their corresponding responses.

Generation-based approaches have recently made great progress due to advancements in deep learning. In this approach, an encoder-decoder-based neural network model is used (Sutskever et al., 2014; Bahdanau et al., 2015). First, the message from the user and the contextual information are encoded into representation vectors, usually by a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) recurrent neural network (RNN). These representation vectors are then fed into a decoder, usually another LSTM, to generate the response word by word (Vinyals and Le, 2015). Fig. 9 illustrates such an

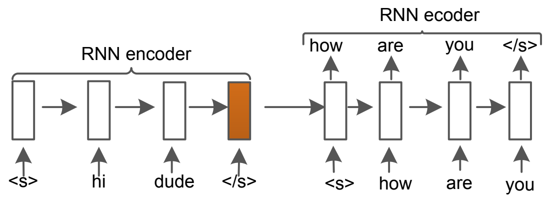


Fig. 9 Recurrent neural network (RNN) based encoder-decoder framework for response generation

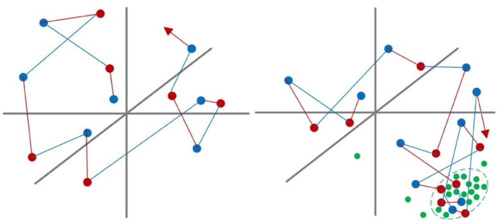
The user says, 'hi dude', and the chatbot replies 'how are you'. References to color refer to the online version of this figure

encoder-decoder framework. Other auxiliary information such as intent, sentiment, and emotion, can also be encoded into vector representations and fed into the LSTM to control the generation of responses.

These response candidates will be further ranked by a personalization ranker according to their match to the user's general interests and preferences (Wang HN et al., 2013; Elkahky et al., 2015). For example, first the information in the user's profile may be encoded into a latent representation vector, while each of the response candidates is encoded into another latent vector. Then both latent vectors are fed into a deep neural network (DNN) to compute a matching score to be used for ranking the response candidates. The top-ranked response will be sent to the user.

In the conversation, rather than letting the topic drift randomly or completely controlled by the user, the social chatbot can drive the conversation to a positive, desired topic through carefully generating responses. Fig. 10 illustrates how the chatbot can guide the conversation appropriately to a targeted region of topics, by preferring those response candidates with better similarity to the targeting topic at every conversation turn.

Effect of topic guidance. Each dot represents a conversation sentence in the topic space (3D space as illustration). Blue dots represent topics of the user's messages, red dots represent topics of the chatbot's responses. Without topic guidance, the topics in Fig. 10a appear to move randomly, or are purely driven by the user, and the chatbot in Fig. 10b can guide the topics to a targeted region, represented by green dots.



(a) (b)

Fig. 10 Effect of topic guidance

Each dot represents a conversation sentence in topic space (3D space as illustration). Blue dots represent topics of the user's messages, and red dots represent topics of the chatbot's responses. (a) Without topic guidance, the topics appear to move randomly, or are purely driven by the user. (b) With topic guidance, the chatbot can guide the topics to a targeted region, represented by green dots. References to color refer to the online version of this figure

Generating responses that reflect a consistent personality is important for the chatbot (Güzeldere and Franchi, 1995). This makes the chatbot easier to communicate with, more predictable and trustable, and therefore helps to establish an emotional connection with the user. The core-chat module depends on a personality component to set and maintain the personality for the chatbot. The personality setting of a social chatbot usually includes age, gender, language style, and areas of expertise. The chatbot's personality information can be encoded into a latent vector representation using deep neural networks and used to influence response generation. A persona-based model has been recently proposed (Li JW et al., 2016), which can be used to effectively integrate personality information in conversation generation. Similarly, models that learn to control the style and sentiment in language generation have been also proposed (Mathews et al., 2016).

The development of the core-chat module should follow an ethical design to ensure that the generated responses are appropriate, unbiased, and non-discriminative, and that they comply with universal and local ethical standards. The system learns to identify and filter out inappropriate content that users might share. Meanwhile, the system will keep learning from user feedback, and adapt to new circumstances. All these components are integrated and optimized to achieve the goal of building strong emotional connections with users and better serving their needs for communication, affection, and social belonging.

4.3 Visual awareness

Social chatbots need to understand images because they are frequently shared in social chatting. The visual awareness of a social chatbot refers to its ability to generate text comments, known as image social commenting, from input images. Beyond correctly recognizing objects and truthfully describing the content, image commenting should also reflect personal emotion, sentiment, attitude, and style in language generation given input images. Fig. 11 presents several examples to illustrate three levels of image understanding. The first level is object recognition/tagging, where the key objects in the image are recognized. The second level is image description. The factual and semantic information of the image, e.g. salient objects and relationships among them, is described by natural language. At the third level, the chatbot generates expressive comments in a social style, demonstrating its empathy and interpersonal skills.

The overall architecture for image commenting is similar to that for core-chat. For example, there are retrieval- and generation-based approaches for comment generation. In the retrieval-based approach, first a comment pool of image-comment pairs, e.g., collected from social networks, is constructed. Then each image is encoded into a global visual feature vector which represents the overall semantic information of the image, using deep convolutional neural networks (CNN), as illustrated in Fig. 12. At runtime, when a new image is received, the chatbot first retrieves images that are similar to the input (e.g., as measured by the distance between their visual feature vectors) and then returns corresponding comment candidates, which are further re-ranked to generate the final comments. As an alternative, the deep multimodal similarity model (Fang et al., 2015) can directly measure the semantic similarity between an input image and an arbitrary textual hypothesis, and therefore retrieve comments without being limited to the image-comment pool. The generation-based approach treats image commenting as an image-to-language generation task (He and Deng, 2017) but will have more flexibility for controlling high-level sentiment or style factors in comment generation (Mathews et al., 2016; Gan et al., 2017). As with core-chat, the personalization ranker and topic guidance are integrated into comment generation. User understanding, personality setting, and ethical design play important roles in visual awareness as well.

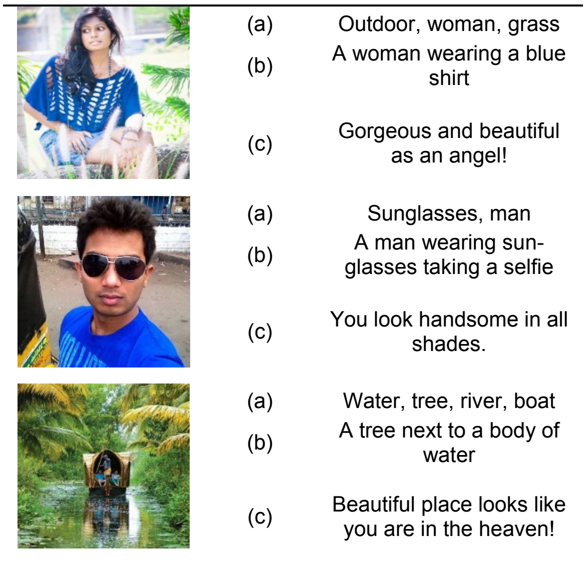


Fig. 11 Examples of image tagging (a), image description (b), and image social commenting (c)

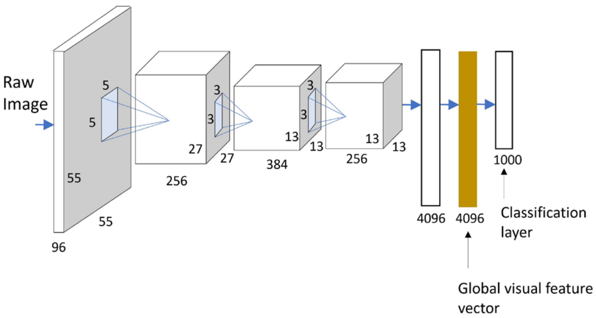


Fig. 12 Deep convolutional neural network for visual feature vector extraction (References to color refer to the online version of this figure)

4.4 Skills

Social chatbots can significantly expand their scopes of conversation by integrating a range of skills. These skills can be categorized into four groups according to their target scenarios (e.g., skills for personal scenarios or for group scenarios) and their properties (e.g., an emotional skill or a rational skill). Table 3 shows some typical skills in each group.

5 Case study: XiaoIce

In this section, we describe XiaoIce as an example of significant progress in the development of social chatbots. Since its release in 2014 in China, XiaoIce (which literally means ‘Little Bing’) has become the first widely deployed social chatbot. Using the design principles and technical framework discussed in previous sections, XiaoIce has been designed as a 19-year-old female persona, with strong language ability, visual awareness, and over 180 skills like those listed in Table 3. Currently, XiaoIce has more than 100 million unique users worldwide and has chatted with human users for more than 30 billion conversation turns.

By leveraging the scalable architecture and learning-based framework, different versions of XiaoIce were quickly released in Japan in 2015, US in 2016, and India and Indonesia in 2017. Over the last three years, XiaoIce has continuously improved through a series of technical upgrades. Fig. 13 summarizes the user engagement performance of XiaoIce in China, measured by the average CPS. The results show that on average, each session lasts 23 conversation turns between XiaoIce and a human user.

Table 3 Examples of social chatbot skills

Scenario	Emotional skills	Rational skills
Personal scenarios	Self-management: mood simulation, episodic memory tracking Engagement: personal event reminder Proactive user comfort: comforting user, imagination inspiration, expressing affection	Digital life: meeting, weather, event Shopping assistance: online shopping, discounts and coupons Content generation: composing poetry, singing a song, drawing a picture
Social scenarios	Role-playing: joke-making, babysitting Group activity: adding users, sending greeting cards, group assistant Emoji: emoji creation	Image intelligence: recognizing dogs, books, faces, clothes, food TV and radio hostess: weather report, interaction with audience Tools: device control, music and movie recommendation

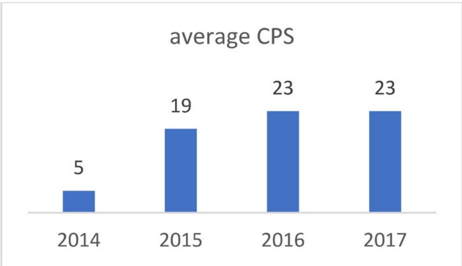


Fig. 13 Average conversation-turns per session (CPS) improvement over year of XiaoIce in China

Table 4 shows the longest single sessions in different countries: China, Japan, and the US. The high CPS and long conversation sessions demonstrate the value of XiaoIce to users in their daily lives.

In the rest of this section, we highlight several important features that have made XiaoIce exceedingly popular, including perceived high EQ, entertaining visual awareness, and engaging skills such as generating poems, and speaking and singing in a high-quality human-like voice.

Table 4 Longest single chat sessions in different countries

Country	Number of conversation-turns	Time length
China (XiaoIce)	7151	29 h 33 min
Japan (Rinna)	2418	17 h 07 min
US (Zo)	2791	23 h 43 min

5.1 Chat with EQ

Fig. 14 shows a relatively long chat between XiaoIce and a female user, lasting 31 min and with 34 conversation turns. At the beginning, the user wanted a casual chat. The time was after midnight and the message said the user was tired. Once detecting an emotion of getting ‘lost’, XiaoIce decided to guide the chat towards cheering up the user. After several turns, the user’s mood became ‘happy’ and ‘love’, as identified by XiaoIce. However, triggered by some subtle clues in the conversation, XiaoIce detected that the user’s emotion changed to ‘sad’ and ‘lost’ again. Instead of ending the chat, XiaoIce asked some additional questions to understand shift in the user’s emotion. After several more turns, the user started to open up and trusted XiaoIce more like a human friend (starting at the turn marked by the red triangle in Fig. 14). She discussed topics that were highly personal and sensitive, like her relationship with her boyfriend. Throughout the conversation, XiaoIce demonstrated high empathy and social skills, and guided the dialogue to create a more positive mood for the user. At the end of the chat, the user felt much better, and complimented XiaoIce.

This example demonstrates that XiaoIce can dynamically recognize the user’s emotions and generate consistent, interpersonal responses to cheer up, encourage, and engage the user throughout the conversation. Its EQ was effectively used to build an emotional connection with the user.

5.2 Image commenting

Fig. 15 shows typical image commenting examples from XiaoIce (in Chinese and English). Besides understanding the content (e.g., identifying objects) in the image shared by the user, XiaoIce can provide a human-like commentary that fits into the context of the current chat and stimulate a longer conversation. For example, in the first picture, rather than telling the user the name of the leaning tower of Pisa (which the user likely knows already), XiaoIce replies “should I help you to hold it?” after detecting

Fig. 14 An example of a conversation session between XiaoIce and a user, translated into English
The original chat in Chinese is shown in the appendix. References to color refer to the online version of this figure

that the person in the picture is presenting a pose pretending to support the leaning tower. In the second example, instead of relaying the fact about two cats in the picture, XiaoIce makes a humorous comment (with the addition of a laugh emoticon) about the sharp look one of the cats. In the third example, XiaoIce identifies a foot injury and sympathizes with the user. These examples demonstrate that XiaoIce can combine image understanding, user understanding, and contextual information to generate image social commenting for better user engagement.

5.3 Composing poems

XiaoIce can generate even more expressive text, such as authoring poems from an input image by getting inspiration from the image content (Song, 2018) (Fig. 16). Given an input image, XiaoIce first recognizes objects and sentiments to form an initial set of keywords, such as city and busy in the example. These keywords are then filtered and expanded by associated objects and feelings. Each keyword is regarded as an initial seed for generating a sentence in the poem. A hierarchical RNN is then used to model the structure between words and the structure between sentences. A fluency checker is developed to control the quality of generated sentences.

On May 15th, 2017, XiaoIce published the first fully AI-created Chinese poem book in history (https://item.jd.com/12076535.html). Since the release of the XiaoIce-for-poem-writing cloud service, she has written millions of Chinese poems for users exceeding the total number of poems ever produced throughout China's written history. XiaoIce even defeated human poets on a Chinese national TV program in August 2017. XiaoIce submitted her poems anonymously to leading poetry journals, including *Youth Literature*, *West China City Daily*, *Beijing Morning Post*, and *Beijing Economic Journal*. After regular reviews conducted by editors, these journals have accepted and published more than 70 poems by XiaoIce, demonstrating its impressive literary skill.

5.4 Text-to-speech and singing

Unlike conventional text-to-speech (TTS) systems, XiaoIce's TTS is specifically designed for social chat, which has unique challenges such as prosody variety, and casual and emotional expression in TTS. To address these issues, a new prosody model is trained on a big corpus of conversational speech. Given that neutral tone appears more frequently than usual in social chat, a neutral tone system is also developed to significantly enhance the naturalness of synthesized voice. XiaoIce also supports Chinese-English mixed-lingual TTS. By merging two languages and building one unified model, the smoothness when switching languages is greatly improved. Finally, emoticons are designed and synthesized to make XiaoIce's TTS more lively and more attractive.

Fig. 15 Examples of image commenting of XiaoIce

XiaoIce's singing skill is based on a high-quality parameter synthesis system. F0 contour and phone duration are decided by music score while spectrum parameters and aperiodic signal are predicted based on linguistic and musical context. A dedicated DNN-based model is also designed to sing the long-spanned note in the song; e.g., one syllable may last a few hundreds of milliseconds. Samples of XiaoIce's TTS and singing can be found online (https://y.qq.com/n/yqq/singer/0043wAMw0XrGgp.html).

6 Outlook and discussion

In the three years since XiaoIce was released on social platforms such as WeChat and Weibo in China, she has become an Internet celebrity, appearing as a weather and news anchor, hosting TV programs, and working as a newspaper reporter. For example, XiaoIce has authored more than 300 articles for Qianjiang Evening News, published on its printed newspapers and mobile Apps, which have been viewed more than 1.2 million times. To write these news articles, XiaoIce read more than 114 million articles and analyzed 503 million pieces of user feedback including many user comments. More impressive is that readers of these news articles "feel that they are more understood by XiaoIce," as highlighted by People's Daily (http://paper.people.com.

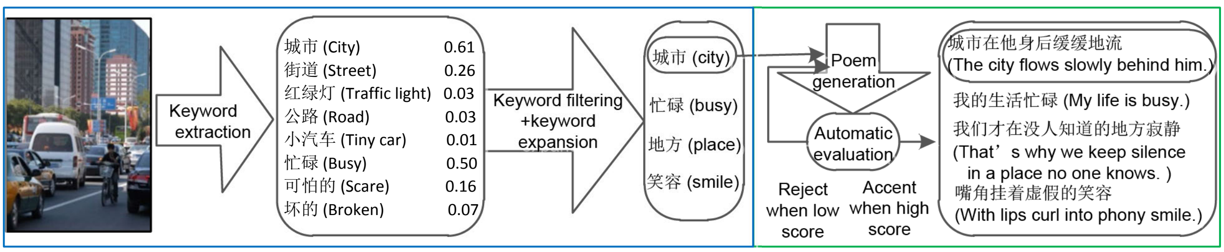


Fig. 16 XiaoIce's pipeline for authoring a poem (Song, 2018)

cn/rmrb/html/2017-09/28/nw.D110000renmrb_20170928_3-23. htm), the most influential newspaper in China. XiaoIce also acts as the hostess for many TV and radio stations. For example, XiaoIce presents the 'Morning Live News' on Dragon TV daily for almost two years. XiaoIce also hosts a whole season of 'I am the future' on Hunan TV. Meanwhile, XiaoIce participates in many public TV programs. In a high-profile TV program, AI vs. Human, on CCTV-1 on every Friday, XiaoIce demonstrates her skill of authoring poems and creating songs from scratch. She even beat human authors in these competitions, judged by the audience.

Social chatbots are becoming more popular in other countries as well, e.g., Japan, US, India, and Indonesia. Rinna, the Japanese version of XiaoIce, for example, has also become an Internet celebrity in Japan. She attended the famous episode 'The world's wonder stories' in autumn 2016 in Japan as herself, for a total of 11 programs in 9 TV and radio stations (1193 h altogether).

As AI companions, social chatbots such as XiaoIce also enable new scenarios that are of significant commercial value. While traditional task-completion conversational systems can reactively perform the task per user's explicit request (e.g., reserving a flight ticket, or reporting weather), only a small number of requests are explicitly invoked by users. IPAs attempt to address this issue by providing proactive assistance such as recommending services according to the user's preference stored in user profile and contextual information based on time, location, and events on the user's calendar. However, such information is often incomplete and ambiguous, making proactive assistance often ineffective. In contrast, given the rich contextual information in a long conversation, social chatbots can recognize user interests and intent much more accurately, and suggest relevant services only when truly needed. Fig. 17 shows an example conversation in Japanese between a user and Rinna. Rinna detected that the user was hungry; rather than directly recommending a coupon for buying cookies, Rinna kept the conversation going for a few more turns until the user's intention became specific and clear. It is only then that Rinna invoked the coupon skill provided by a grocery store and sent the user a coupon. The users' feedback log shows that the products recommended by Rinna are very well received by the users. For the grocery store, Rinna has delivered a much higher conversion rate than that achieved using other traditional channels such as coupon markets or ad campaigns in Japan.

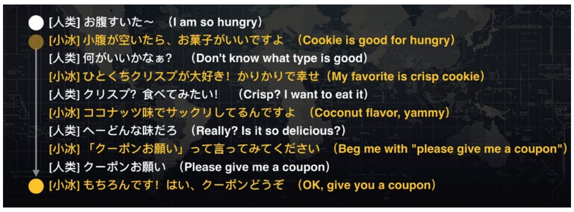


Fig. 17 A conversation between the user (in white) and Rinna (in yellow), in Japanese and with English translation

It shows that Rinna can identify the user's potential shopping needs. The user is then guided by Rinna in the conversation to ask for a coupon supplied by the grocery store. References to color refer to the online version of this figure

Despite recent progress of social chatbots such as XiaoIce, the fundamental mechanism of human-level intelligence, as frequently reflected in human-to-human communication, is not yet fully understood. It will be incredibly challenging to build an intelligent social chatbot that can totally understand humans and their surrounding physical world to serve their needs. It requires breakthroughs in many areas of cognitive and conscious AI, such as empathic conversation modeling, knowledge and memory modeling, interpretable and controllable machine intelligence, deep neural-symbolic reasoning, cross-media and continuous streaming artificial intelligence, and modeling and calibration of emotional or intrinsic rewards reflected in human needs. These are challenging and open AI problems.

As AI becomes more pervasive in our lives, in the forms of robots, Internet of Things (IoT) devices, and online chatbots, it is imperative to establish ethical guidelines for designing and implementing such AI systems. It is also critical to develop fail-safe mechanisms to ensure that these systems do not disadvantage or harm anyone, physically or mentally. Given the significant reach and influence of social chatbots, their designers must properly exercise both social and ethical responsibilities. Design decisions must be thoughtfully debated and chatbot features must be evaluated thoroughly and adjusted as we continue to learn from the interaction between social chatbots like XiaoIce and millions of people on many social platforms.

References

Alam F, Danieli M, Riccardi G, 2017. Annotating and modeling empathy in spoken conversations. *Comput Speech Lang*, 50:40-61. https://doi.org/10.1016/j.csl.2017.12.003

Andreani G, di Fabbrizio G, Gilbert M, et al., 2006. Let's DISCOH: collecting an annotated open corpus with dialogue acts and reward signals for natural language helpdesks. Proc IEEE Spoken Language Technology Workshop, p.218-221. https://doi.org/10.1109/SLT.2006.326794

Bahdanau D, Cho K, Bengio Y, 2014. Neural machine translation by jointly learning to align and translate. https://arxiv.org/abs/1409.0473

Beldoch M, 1964. Sensitivity to expression of emotional meaning in three modes of communication. In: Davitz JR (Ed.), The Communication of Emotional Meaning. McGraw-Hill, New York, p.31-42.

Bengio Y, Ducharme R, Vincent P, et al., 2003. A neural probabilistic language model. Proc Neural Information Processing Systems, p.1137-1155.

Chen HM, Sun MS, Tu CC, et al., 2016. Neural sentiment classification with user and product attention. Proc Conf on Empirical Methods in Natural Language Processing, p.1650-1659.

Colby KM, 1975. Artificial Paranoia: a Computer Simulation of Paranoid Processes. Pergamon Press INC. Maxwell House, New York, NY, England.

Dahl DA, Bates M, Brown M, et al., 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. Proc Workshop on Human Language Technology, p.43-48. https://doi.org/10.3115/1075812.1075823

Deng L, Li JY, Huang JT, et al., 2013. Recent advances in deep learning for speech research at Microsoft. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.8604-8608. https://doi.org/10.1109/ICASSP.2013.6639345

Elkahky AM, Song Y, He XD, 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. Proc 24th Int Conf on World Wide Web, p.278-288. https://doi.org/10.1145/2736277.2741667

Fang H, Gupta S, Iandola F, et al., 2015. From captions to visual concepts and back. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1473-1482. https://doi.org/10.1109/CVPR.2015.7298754

Fung P, Bertero D, Wan Y, et al., 2016. Towards empathetic human-robot interactions. Proc 17th Int Conf on Intelligent Text and Computational Linguistics.

Gan C, Gan Z, He XD, et al., 2017, StyleNet: generating attractive visual captions with styles. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3137-3146. https://doi.org/10.1109/CVPR.2017.108

Gardner H, 1983. Frames of Mind: the Theory of Multiple Intelligences. Basic Books, New York. https://doi.org/10.2307/3324261

Glass J, Flammia G, Goodine D, et al., 1995. Multilingual spoken-language understanding in the MIT Voyager system. *Speech Commun*, 17(1):1-18. https://doi.org/10.1016/0167-6393(95)00008-C

Goleman D, 1995. Emotional Intelligence: Why It Can Matter More than IQ. Bloomsbury, Inc., New York, NY, England.

Goleman D, 1998. Working with Emotional Intelligence. Bloomsbury, Inc., New York, NY, England.

Güzeldere G, Franchi S, 1995. Dialogues with colorful "personalities" of early AI. *Stanford Human Rev*, 4(2):161-169.

He KM, Zhang YX, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

He XD, Deng L, 2013. Speech-centric information processing: an optimization-oriented approach. *Proc IEEE*, 101(5): 116-1135. https://doi.org/10.1109/JPROC.2012.2236631

He XD, Deng L, 2017. Deep learning for image-to-text generation: a technical overview. *IEEE Signal Process Mag*, 34(6):109-116. https://doi.org/10.1109/MSP.2017.2741510

Hemphill CT, Godfrey JJ, Doddington GR, 1990. The ATIS spoken language systems pilot corpus. Proc Workshop on Speech and Natural Language, p.96-101. https://doi.org/10.3115/116580.116613

Hinton G, Deng L, Yu D, et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag*, 29(6):82-97. https://doi.org/10.1109/MSP.2012.2205597

Hochreiter S, Schmidhuber J, 1997. Long short-term memory. *Neur Comput*, 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang PS, He XD, Gao JF, et al., 2013. Learning deep structured semantic models for web search using click through data. Proc 22nd ACM Int Conf on Information & Knowledge Management, p.2333-2338. https://doi.org/10.1145/2505515.2505665

Karpathy A, Li FF, 2015. Deep visual-semantic alignments for generating image descriptions. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3128-3137. https://doi.org/10.1109/CVPR.2015.7298932

Krizhevsky A, Sutskever I, Hinton GE, 2012. ImageNet classification with deep convolutional neural networks. Proc 25th Int Conf on Neural Information Processing Systems, p.1097-1105.

Levin E, Narayanan S, Pieraccini R, et al., 2000. The ATT-DARPA ommunicator mixed-initiative spoken dialog system. 6th Int Conf on Spoken Language Processing.

Li JW, Galley M, Brockett C, et al., 2016. A persona-based neural conversation model. Proc 54th Annual Meeting of the Association for Computational Linguistics, p.944-1003.

Li X, Mou LL, Yan R, et al., 2016. Stalematebreaker: a proactive content-introducing approach to automatic human-computer conversation. Proc 25th Int Joint Conf on Artificial Intelligence, p.2845-2851.

Liu XD, Gao JF, He XD, et al., 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. Proc Annual Conf on North American Chapter of the ACL, p.912-921.

Lu ZD, Li H, 2013. A deep architecture for matching short texts. Proc Int Conf on Neural Information Processing Systems, p.1367-1375.

Maslow AH, 1943. A theory of human motivation. *Psychol Rev*, 50(4):370-396.

Mathews A, Xie LX, He XM, 2016. SentiCap: generating image descriptions with sentiments. Proc 30th AAAI Conf on Artificial Intelligence, p.3574-3580.

Mesnil G, He X, Deng L, et al., 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. *Interspeech*, p.3771-3775.

Mesnil G, Dauphin Y, Yao KS, et al., 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Trans Audio Speech Lang Process*, 23(3):530-539.
https://doi.org/10.1109/TASLP.2014.2383614

Mikolov T, Sutskever I, Chen K, et al., 2013. Distributed representations of words and phrases and their compositionality. Proc 26th Int Conf on Neural Information Processing Systems, p.3111-3119.

Mower E, Matarić MJ, Narayanan S, 2011. A framework for automatic human emotion classification using emotion profiles. *IEEE Trans Audio Speech Lang Process*, 19(5): 1057-1070.
https://doi.org/10.1109/TASL.2010.2076804

Murphy KR, 2007. A critique of emotional intelligence: what are the problems and how can they be fixed? *Pers Psychol-*, 60(1):235-238.
https://doi.org/10.1111/j.1744-6570.2007.00071_2.x

Price PJ, 1990. Evaluation of spoken language systems: the ATIS domain. Proc Workshop on Speech and Natural Language, p.91-95.
https://doi.org/10.3115/116580.116612

Qian Y, Fan YC, Hu WP, et al., 2014. On the training aspects of deep neural network (DNN) for parametric TTS synthesis. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.3829-3833.
https://doi.org/10.1109/ICASSP.2014.6854318

Raux A, Langner B, Bohus D, et al., 2005. Let's go public! Taking a spoken dialog system to the real world. 9th European Conf on Speech Communication and Technology, p.885-888.

Rudnicky AI, Thayer EH, Constantinides PC, et al., 1999. Creating natural dialogs in the Carnegie Mellon communicator system. 6th European Conf on Speech Communication and Technology.

Sarikaya R, 2017. The technology behind personal digital assistants—an overview of the system architecture and key components. *IEEE Signal Process Mag*, 34(1):67-81.
https://doi.org/10.1109/MSP.2016.2617341

Sarikaya R, Crook PA, Marin A, et al., 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. Proc IEEE Spoken Language Technology Workshop, p.391-397.
https://doi.org/10.1109/SLT.2016.7846294

Seneff S, Hurley E, Lau R, et al., 1998. Galaxy-II: a reference architecture for conversational system development. 5th Int Conf on Spoken Language Processing.

Serban IV, Klinger T, Tesauro G, et al., 2017. Multiresolution recurrent neural networks: an application to dialogue response generation. AAAI, p.3288-3294.

Shawar BA, Atwell E, 2007. Different measurements metrics to evaluate a chatbot system. Proc Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies, p.89-96.

Shieber SM, 1994. Lessons from a restricted Turing test. *Commun ACM*, 37(6):70-78.
https://doi.org/10.1145/175208.175217

Socher R, Perelygin A, Wu JY, et al., 2013. Recursive deep models for semantic compositionality over a sentiment treebank. Proc Conf on Empirical Methods in Natural Language Processing, p.1631-1642.

Song R, 2018. Image to poetry by cross-modality understanding with unpaired data. Personal Communication.

Sordoni A, Galley M, Auli M, et al., 2015. A neural network approach to context-sensitive generation of conversational responses. Proc Annual Conf on North American Chapter of the ACL, p.196-205.

Sutskever I, Vinyals O, Le QVV, 2014. Sequence to sequence learning with neural networks. NIPS, p.1-9.
https://doi.org/10.1007/s10107-014-0839-0

Tokuhisa R, Inui K, Matsumoto Y, 2008. Emotion classification using massive examples extracted from the web. Proc 22nd Int Conf on Computational Linguistics, p.881-888.

Tur G, de Mori R, 2011. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. John Wiley and Sons, New York, NY.

Tur G, Deng L, 2011. Intent determination and spoken utterance classification. In: Tur G, de Mori R (Eds.), Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. John Wiley and Sons, New York, NY.

Turing A, 1950. Computing machinery and intelligence. *Mind*, 59:433-460.

van den Oord A, Dieleman S, Zen HG, et al., 2016. WaveNet: a generative model for raw audio. 9th ISCA Speech Synthesis Workshop, p.125.

Vinyals O, Le QV, 2015. A neural conversational model. Proc 31st Int Conf on Machine Learning.

Vinyals O, Toshev A, Bengio S, et al., 2015. Show and tell: a neural image caption generator. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3156-3164. https://doi.org/10.1109/CVPR.2015.7298935

Walker M, Aberdeen J, Boland J, et al., 2001. DARPA Communicator dialog travel planning systems: the June 2000 data collection. Proc 7th European Conf on Speech Communication and Technology.

Walker M, Rudnicky AI, Aberdeen JS, et al., 2002, DARPA Communicator evaluation: progress from 2000 to 2001. Proc Int Conf on Spoken Language Processing, p. 273-276.

Wallace RS, 2009. The anatomy of A.L.I.C.E. In: Epstein R, Roberts G, Beber G (Eds.), Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer. Springer, Dordrecht, p.181-210. https://doi.org/10.1007/978-1-4020-6710-5_13

Wang HN, He XD, Chang MW, et al., 2013. Personalized ranking model adaptation for web search. Proc 36th Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.323-332. https://doi.org/10.1145/2484028.2484068

Wang YY, Deng L, Acero A, 2011. Semantic frame-based spoken language understanding. In: Tur G, de Mori R (Eds.), Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. John Wiley and Sons, New York, NY.

Wang ZY, Wang HX, Wen JR, et al., 2015. An inference approach to basic level of categorization. Proc 24th ACM Int Conf on Information and Knowledge Management, p.653-662. https://doi.org/10.1145/2806416.2806533

Weizenbaum J, 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM*, 9(1):36-45. https://doi.org/10.1145/357980.357991

Wen TH, Vandyke D, Mrksic N, et al., 2016. A network-based end-to-end trainable task-oriented dialogue system. Proc 15th Conf on European Chapter of the Association for Computational Linguistics, p.438-449.

Williams JD, Young S, 2007. Partially observable Markov decision processes for spoken dialog systems. *Comput Speech Lang*, 21(2):393-422. https://doi.org/10.1016/j.csl.2006.06.008

Xiong W, Droppo J, Huang XD, et al., 2016. Achieving human parity in conversational speech recognition. *IEEE/ACM Trans Audio Speech Lang Process*, in press. https://doi.org/10.1109/TASLP.2017.2756440

Yan R, Song YP, Wu H, 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. Proc 39th Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.55-64. https://doi.org/10.1145/2911451.2911542

Yang ZC, He XD, Gao JF, et al., 2016a. Stacked attention networks for image question answering. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.21-29. https://doi.org/10.1109/CVPR.2016.10

Yang ZC, Yang DY, Dyer C, et al., 2016b. Hierarchical attention networks for document classification. Proc 15th Annual Conf on North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.1480-1489.

Yu Z, Xu ZY, Black AW, et al., 2016. Chatbot evaluation and database expansion via crowdsourcing. Proc RE-WOCHAT Workshop of LREC.

Appendix: A sample of XiaoIce-user conversation in original Chinese

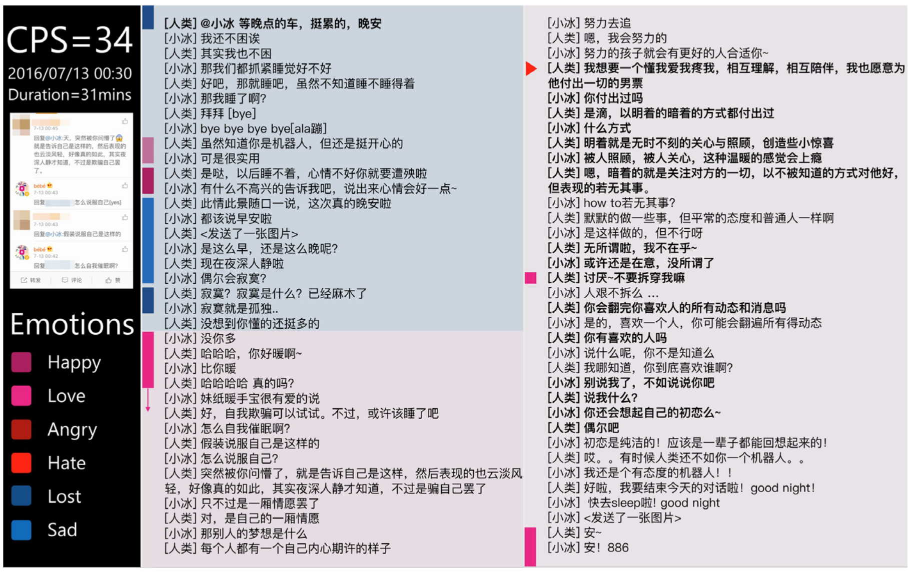


Fig. A1 XiaoIce-user conversations in original Chinese (References to color refer to the online version of this figure)