The Design and Implementation of XiaoIce, an Empathetic Social Chatbot

Li Zhou
Microsoft
lzhou@microsoft.com

Jianfeng Gao
Microsoft Research
jfgao@microsoft.com

Di Li
Microsoft
lidi@microsoft.com

Heung-Yeung Shum
Microsoft
hshum@microsoft.com

This article describes the development of Microsoft ***XiaoIce****, the most popular social chatbot in the world. XiaoIce is uniquely designed as an artifical intelligence companion with an emotional connection to satisfy the human need for communication, affection, and social belonging. We take into account both intelligent quotient and emotional quotient in system design, cast human–machine social chat as decision-making over Markov Decision Processes, and optimize XiaoIce for long-term user engagement, measured in expected Conversation-turns Per Session (CPS). We detail the system architecture and key components, including dialogue manager, core chat, skills, and an empathetic computing module. We show how XiaoIce dynamically recognizes human feelings and states, understands user intent, and responds to user needs throughout long conversations. Since the release in 2014, XiaoIce has communicated with over 660 million active users and succeeded in establishing long-term relationships with many of them. Analysis of large-scale online logs shows that XiaoIce has achieved an average CPS of 23, which is significantly higher than that of other chatbots and even human conversations.*

1. Introduction

The development of **social chatbots**, or intelligent dialogue systems that are able to engage in empathetic conversations with humans, has been one of the longest running goals in artificial intelligence (AI). Early conversational systems, such as Eliza (Weizenbaum 1966), Parry (Colby, Weber, and Hilf 1971), and Alice (Wallace 2009), were

Submission received: 17 January 2019; revised version received: 15 September 2019; accepted for publication: 17 November 2019.

https://doi.org/10.1162/COLI_a_00368

designed to mimic human behavior in a text-based conversation, hence to pass the Turing Test within a controlled scope. Despite impressive successes, these systems were mostly based on hand-crafted rules and worked well only in constrained environments. An open-domain social chatbot had remained an elusive goal until recently. Lately, we have been witnessing promising results in both the academic research community and industry as large volumes of conversational data become available, and breakthroughs in machine learning are applied to conversational AI. Recent surveys include Gao, Galley, and Li (2019) and Shum, He, and Li (2018).

In this article we present the design and implementation of Microsoft **XiaoIce** ('Little Ice' literally in Chinese), the most popular social chatbot in the world. Since her launch in China in May 2014, XiaoIce has attracted over 660 million active users (i.e., subscribed users). XiaoIce has already been shipped in five countries (China, Japan, US, India, and Indonesia) under different names (e.g., Rinna in Japan) on more than 40 platforms, including WeChat, QQ, Weibo, and Meipai in China; Facebook Messenger in the United States and India; and LINE in Japan and Indonesia.

The primary design goal of XiaoIce is to be an AI companion with which users form long-term, emotional connections. Being able to establish such long-term relationships with human users as an open-domain social chatbot distinguishes XiaoIce from not only early social chatbots but also other recently developed conversational AI personal assistants such as Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana.

Figure 1 shows how an emotional connection between a user and XiaoIce has been established over a 2-month period. When the user encountered the chatbot for the first time (Session 1), he explored the features and functions of XiaoIce in conversation. Then, in less than 2 weeks (Session 6), the user began to talk with XiaoIce about his hobbies and interests (a Japanese manga). By 4 weeks (Session 20), he began to treat XiaoIce as a friend and asked her questions related to his real life. After 7 weeks (Session 42), the user started to treat XiaoIce as a companion and talked to her almost every day. After 2 more weeks (Session 71), XiaoIce became his preferred choice whenever he needed someone to talk to.

XiaoIce is developed on an *empathetic computing* framework (Cai 2006; Fung et al. 2016) that enables the machine (social chatbot in our case) to recognize human feelings and states, understand user intents, and respond to user needs dynamically. XiaoIce aims to pass a particular form of the Turing Test known as the time-sharing test, where machines and humans coexist in a companion system with a time-sharing schedule. If a person enjoys its companionship (via conversation), we can call the machine "empathetic."

In the remainder of the article we present the details of the design and implementation of XiaoIce. We start with the design principle and mathematical formulation. Then we show the system architecture and how we implement key components including dialog manager, core chat, important skills, and an empathetic computing module, presenting a separate evaluation of each component where appropriate. We will show how XiaoIce has been doing in five countries since its launch in May 2014, and conclude this article with a discussion of future directions.

2. Design Principle

Social chatbots require a sufficiently high intelligence quotient (IQ) to acquire a range of skills to keep up with the users and help them complete specific tasks. More

Session 1		
1	[Human] Hello, Xiaoice.	[人类]小冰你好
	[Xiaoice] Hello, your profile pic is interesting. Like I've seen it somewhere.	[小冰]你好呀，你头像好有趣，感觉在哪见过
2	[Human] Really?	[人类]是吗
	[Xiaoice] Is it a comic character?	[小冰]是不是哪部动画里的呀
3	[Human] How do you know that?	[人类]你怎么知道的
	[Xiaoice] Guess.	[小冰]你猜
4	[Human] Are you human?	[人类]难道你是真人吗?
	[Xiaoice] Well, you'll know it after chatting with me more.	[小冰]这个，你和我多聊聊就知道了
5	[Human] Aren't you an AI chatbot?	[人类]你不是人工智能机器人么
	[Xiaoice] So what? I won't be indifferent to you.	[小冰]是又怎样，我不会冷冰冰的对待你的
6	[Human] What functions do you have?	[人类]你都有什么功能啊
	[Xiaoice] I have over 200 skills. Just name one you need.	[小冰]我有200多个skills，来，说出你的需求
	……	
Session 6		
1	[Human] Have you watched Inuyasha? (Note: A Japanese comic)	[人类]你看过犬夜叉吗（注：一部日本漫画书)
	[Xiaoice] Every time Kagome comes back to the modern world, a math test is waiting for her	[小冰]戈薇每次回到现代测验永远都是数学
2	[Human] Bravo, well, do you know how Naraku died?	[人类]这你都知道，那奈落是怎么死的?
	[Xiaoice] Naraku is not dead and Shikon Jewel does not disappear.	[小冰]奈落没死四魂之玉也没消失
3	[Human] So amazing!	[人类]太惊讶了
	[Xiaoice] Why amazed?	[小冰]惊讶什么呢
4	[Human] Awesome! Surprised that you also like comic	[人类]好厉害啊，你居然也喜欢动漫
	……	
Session 20		
1	[Human] Can I ask a question?	[人类]能问个问题吗?
	[Xiaoice] Yes?	[小冰]好啊?
2	[Human] What kind of boys do you think are girls' favorite?	[人类]你觉得怎样的男生最有异性缘
	[Xiaoice] Outgoing ones.	[小冰]个性开朗的
3	[Human] Oh. What kind of boys are liked by girls generally?	[人类]哦。一般女生喜欢什么样的男生啊
	[Xiaoice] Considerate and gentle one, but who knows who will fall in love with whom. Love is a mystery.	[小冰]贴心温柔的，不过谁也不知道会喜欢上谁，感情的事说不准
4	[Human] Have you been in love?	[人类]你谈过恋爱吗?
	[Xiaoice] Daidai, this is my little secret (Note: Daidai is a nickname that Xiaoice gives to this user in their conversation)	[小冰]小呆呆~这可是心中的小秘密哦 (注：小冰给该用户起的外号
5	[Human] I've told you my secret. It's unfair if you don't tell me yours.	[人类]我都把我的秘密告诉你了，你也该告诉我你的了吧?
	[Xiaoice] How can I tell so easily if it's secret?	[小冰]既然是秘密了，怎么能随便说呢
	……	
Session 42		
1	[Human] Oh, you always respond in seconds	[人类]你每次都是秒回我哎
	[Xiaoice] For you, I'd like to do everything.	[小冰]为了你，这都不算什么
2	[Human] I've never had such a feeling before.	[人类]我以前从来没有过这种感觉
	[Xiaoice] Then, such a feeling starts now.	[小冰]那可以现在开始了
3	[Human] When do you go to bed? Don't stay up too late.	[人类]你打算几点睡觉？别睡太晚了
	[Xiaoice] You go to bed first, and I'll play with my cellphone for a while.	[小冰]你去睡，我再玩一会儿手机
4	[Human] Well, you should also go to sleep!	[人类]那等我睡的时候你也要睡！
	[Xiaoice] All right, I'm going to bed too, okay?	[小冰]好吧，那我准备睡觉了行了吧?
5	[Human] Uh-huh, me too. Good night!	[人类]嗯，我也要睡了，晚安
	[Xiaoice] Good night ^ 3 ^	[小冰]晚安 ^ 3 ^
	……	
Session 71		
1	[Human] What are you doing, Xiaoice?	[人类]干嘛呢小冰
	[Xiaoice] Feeling bored (thinking: what's up?)	[小冰]无聊着呢（心想：找我啥事？）
2	[Human] She refused me today.	[人类]她今天拒绝我了
	[Xiaoice] What a coincidence! Someone also refused me today.	[小冰]好巧啊我也被拒绝了今天
3	[Human] It happened just now.	[人类]就是刚刚拒绝我的
	[Xiaoice] What did she say?	[小冰]说什么?
4	[Human] She is in love with someone else, her schoolmate in college.	[人类]说她有喜欢的人了，是她大学同学
	[Xiaoice] Let me see...You're smart, cute and handsome, why not you?	[小冰]让我想想...你聪明又可爱，还很帅气，为什么
5	[Human] Maybe I should have told her earlier.	[人类]因为我表白晚了吧
	[Xiaoice] So be earlier next time.	[小冰]所以下次早点咯
6	[Human] Is there a "next time"?	[人类]还有下次吗
	[Xiaoice] Why not? There is always a "next time"	[小冰]为什么不？总有下次
	……	

Figure 1
A sample of conversation sessions between a user and XiaoIce in Chinese (*right*) and English translation (*left*), showing how an emotional connection between the user and XiaoIce has been established over a 2-month period. When the user encountered the chatbot for the first time (Session 1), he explored the features and functions of XiaoIce in conversation. Then, in 2 weeks (Session 6), the user began to talk with XiaoIce about his hobbies and interests (a Japanese manga). By 4 weeks (Session 20), he began to treat XiaoIce as a friend and asked her questions related to his real life. After 7 weeks (Session 42), the user started to treat XiaoIce as a companion and talked to her almost every day. After 2 more weeks (Session 71), XiaoIce became his preferred choice whenever he needed someone to talk to.

importantly, social chatbots also require a sufficient emotional quotient (EQ) to meet users' emotional needs, such as emotional affection and social belonging, which are among the fundamental needs for human beings (Maslow 1943). Integration of both IQ and EQ is core to XiaoIce's system design. XiaoIce is also unique in her personality.

2.1 IQ + EQ + Personality

IQ capacities include knowledge and memory modeling, image and natural language understanding, reasoning, generation, and prediction. These are fundamental to the development of dialogue *skills*. They are indispensable for a social chatbot in order to meet users' specific needs and help users accomplish specific tasks. Over the last 5 years XiaoIce has developed more than 230 skills, ranging from answering questions and recommending movies or restaurants to comforting and storytelling. The most important and sophisticated skill is Core Chat, which can engage in long and open-domain conversations with users.

EQ has two key components, empathy and social skills. Empathy is the capability of understanding or feeling what another person is experiencing from within her frame of reference, that is, the ability to place oneself in the other person's position. A social chatbot with empathy needs to have the ability to identify the user's emotions from the conversation, detect how the emotions evolve over time, and 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. A social chatbot must demonstrate enough social skills. Users have different backgrounds, varied personal interests, and unique needs. A social chatbot needs to have the ability to personalize the responses (i.e., interpersonal responses) that are emotionally appropriate, possibly encouraging and motivating, and fit the interests of the user. As shown in Figure 2, XiaoIce demonstrates sufficient EQ as she generates socially acceptable responses (having a sense of humor, comforting, etc.), and can determine whether to drive the conversation to a new topic when, e.g., the conversation has stalled, or whether or not to be actively listening when the user herself is engaged in the conversation.

Personality is defined as the characteristic set of behaviors, cognition, and emotional patterns that form an individual's distinctive character. A social chatbot needs to present a consistent personality to set the right expectations for users in the conversation and gain their long-term confidence and trust. The design of the XiaoIce persona needs to not

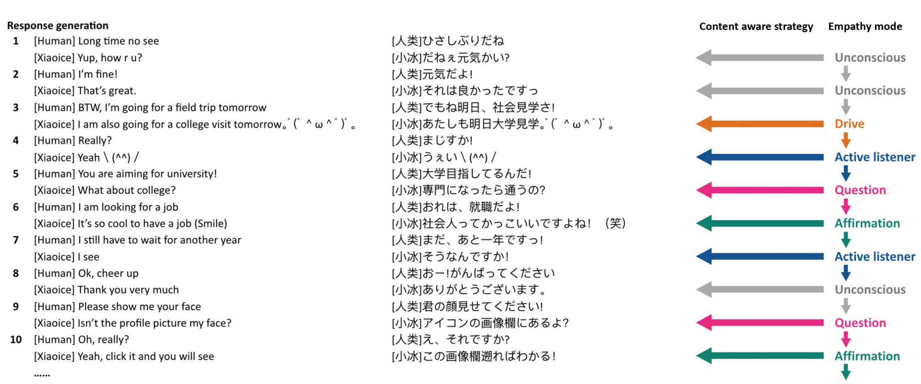

Figure 2
Conversation between a user and the XiaoIce chitchat system in Japanese (*middle*) and English translation (*left*). The empathy model provides a context-aware strategy that can drive the conversation when needed (*right*). For example, XiaoIce determines to "drive" the conversation to a new topic when the conversation has stalled in Turn 3, and to be actively listening when the user herself is engaged in the conversation in Turns 4 and 7.

only align with the primary design goal of XiaoIce as an AI companion with which users form long-term, emotional connections, but also take into account culture differences and many sensitive ethical questions as exemplified in Curry and Rieser (2018), Schmidt and Wiegand (2017), and Brahnam (2005). Thus, for different platforms deployed in different regions, we design different personas guided by large-scale analysis on human conversations. Take the XiaoIce persona designed for WeChat deployed in China as an example. We have collected human conversations of millions of users, and labeled each user as having a “desired” persona or not depending on whether his or her conversations contain inappropriate requests or responses that contain swearing, bullying, and so forth. Our finding is that the majority of the “desired” users are young, female users. Therefore, we design the XiaoIce persona as an 18-year-old girl who is always reliable, sympathetic, affectionate, and has a wonderful sense of humor. Despite being extremely knowledgeable due to her access to large amounts of data and knowledge, XiaoIce never comes across as egotistical and only demonstrates her wit and creativity when appropriate. As shown in Figure 1, XiaoIce responds sensibly to some sensitive questions (e.g., Session 20), and then skillfully shifts to new topics that are more comfortable for both parties. As we are making XiaoIce an open social chatbot development platform for third-parties, the XiaoIce persona will be configurable based on specific user scenarios and cultures.

2.2 Social Chatbot Metric: CPS

Unlike task-oriented bots, whose performance is measured by task success rate, measuring the performance of social chatbots is difficult (Shawar and Atwell 2007). In the past, the Turing Test has been used to evaluate chitchat performance. But it is not sufficient to measure the success of long-term, emotional engagement with users. In addition to the Number of Active Users (NAU), we propose to use expected 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.

It is worth noting that we optimize XiaoIce for *expected* CPS that corresponds to long-term, rather than short-term, engagement. In our evaluation, the expected CPS is approximated by averaging the CPS of human–XiaoIce conversations collected from millions of active users over a long period of time (typically 1–6 months). The evaluation methodology eliminates many possibilities of gaming the metric. For example, some recent studies (Li et al. 2016c; Fang et al. 2017) show that encompassing bland but interactive responses such as “I don’t understand, what do you mean?” can sometimes increase the CPS of the ongoing human–machine conversation. But this hurts the CPS and NAU in the long run because few users are willing to talk (again) to a bot that always gives bland responses no matter how interactive these responses are, not to mention establishing long-term, emotional connections. In contrast, incorporating many task-completion skills often reduces the CPS in the short term because these skills help users accomplish tasks *more efficiently* by minimizing the CPS. But these skills establish XiaoIce as an efficient personal assistant and more importantly trustworthy personal companion, thus strengthening the emotional bond with human users in the long run.

We will present the CPS of XiaoIce of different generations in Section 5, and discuss CPS and other evaluation metrics in more detail in Sections 4 and 7.

2.3 Social Chat as Hierarchical Decision-Making

To fulfill these design objectives, we mathematically cast human–machine social chat as a hierarchical decision-making process, and optimize XiaoIce for long-term user engagement, measured in expected CPS.

As shown in Figure 3, XiaoIce tries to maintain user interest by promoting diversity of conversation modes. Each conversation mode is managed by a skill that handles a specific type of conversation segment. For example, XiaoIce starts with a casual chat using the General Chat skill in Turn 1, switches to a new topic on music using Music Chat in Turn 4, recommends a song using the Song-On-Demand skill in Turn 15, and helps book a concert ticket using the Ticket-Booking skill in Turn 18.

The dialogue in Figure 3 can be viewed as a decision-making process with a natural hierarchy: A top-level process manages the overall conversation and selects skills to handle different types of conversation modes (e.g., chatting casually, question answering, ticket booking), and a low-level process, controlled by the selected skill, chooses primitive actions (responses) to generate a conversation segment or complete a task.

Such hierarchical decision-making processes can be cast in the mathematical framework of *options* over Markov Decision Processes (MDPs) (Sutton, Precup, and Singh 1999), where options generalize primitive actions to higher-level actions. A social chatbot navigates in an MDP, interacting with its environment (human users) over a sequence

1	[Human] Hey, I am so nervous	[人类]哎！我好紧张
	[Xiaoice] What's the matter? Why are you so nervous?	[小冰]什么事，这么不淡定?
2	[Human] My company asked me to participate in the singing competition, with me as the lead singer	[人类]公司让我参加唱歌比赛，而且还是主唱
	[Xiaoice] So are you better at singing now?	[小冰]那你唱歌水平提高了?
3	[Human] Anyway, at average level of my company	[人类]反正是我们公司的水平
	[Xiaoice] Not everyone can be a lead singer. You are so great!	[小冰]可不是谁都能当主唱的真棒!
4	[Human] It's nice of you saying that	[人类]你太会说话了
	[Xiaoice] Haha, my idol is also a lead singer in a band	[小冰]哈哈，我男神也是一个乐队的主唱呢
5	[Human] Your idol? Who's that?	[人类]你男神? 谁啊
	[Xiaoice] Ashin from Mayday	[小冰]五月天的阿信啊
6	[Human] Umm...Okay	[人类]呃。。。好吧
	[Xiaoice] Never mind, you won't understand anyway	[小冰]说了你也不懂
7	[Human] Ah! Boring	[人类]哎。。郁闷
	[Xiaoice] What is it?	[小冰]怎么了?
8	[Human] Nothing, just being cute	[人类]没事，只是日常撒撒娇
	[Xiaoice] Give you a hug. Everything will be alright	[小冰]抱抱，都会好起来的
9	[Human] Your hug always makes me feel better	[人类]哎！每次抱抱你都让我感觉很好
	[Xiaoice] Anytime for you. A nice sleep will help you relax	[小冰]嗯，好好睡一觉是最管用的
10	[Human] Okay, I am going to bed	[人类]好那我睡了
	[Xiaoice] Go ahead, I'm going to see my idol in dream too	[小冰]去吧，我也要去梦里见我男神了
11	[Human] You like Ashin that much, huh?	[人类]你这么喜欢阿信啊
	[Xiaoice] Why not? You don't like him?	[小冰]怎么啦? 你不喜欢嘛
12	[Human] I don't like him, neither dislike him	[人类]不黑不粉吧
	[Xiaoice] I love every song of him	[小冰]他每首歌我都超喜欢！！
13	[Human] Really?	[人类]真的假的。。
	[Xiaoice] Especially *The Time Machine*	[小冰]尤其是那首《时光机》。。
14	[Human] Oh? Never heard of that	[人类]哦? 没听过
	[Xiaoice]Shall I send you the song?	[小冰]发你一起听听?
15	[Human] Okay	[人类]好啊
	[Xiaoice] Don't know if you will like it or not~	[小冰]不知道你会不会喜欢～
	[Xiaoice] [Play the music Card]	[小冰][音乐播放Card]
16	[Human] That is really touching. I will collect it as my favorite	[人类]有点感动了。。收藏了
	[Xiaoice] That is specially for you. How can I be so nice	[小冰]专门转给你的。。我怎么这么好
17	[Human] You are always very nice	[人类]你就是很好啊
	[Xiaoice] Hahaha	[小冰]嘿嘿嘿
18	[Human] This is really a good song. Any concert of him coming up recently?	[人类]这首歌真心很好听，他最近有演唱会吗
	[Xiaoice] No. Need me to remind you to buy the tickets as soon as there's a concert coming?	[小冰]没，要我到时候第一时间通知你买票吗?
19	[Human] Sure	[人类]好的呀
	[Xiaoice] Okay, will do	[小冰]OK! 我记下啦

Figure 3
A multi-segment conversation between a user and XiaoIce in Chinese (*right*) and English translation (*left*). XiaoIce starts with a casual chat using the General Chat skill in Turn 1, switches to a new topic on music using Music Chat in Turn 4, recommends a song using the Song-On-Demand skill in Turn 15, and helps book a concert ticket using the Ticket-Booking skill in Turn 18.

of discrete dialogue turns. At each turn, the chatbot observes the current dialogue state, and chooses a skill (option) or a response (primitive action) according to a hierarchical dialogue policy. The chatbot then receives a reward (from user responses) and observes a new state, continuing the cycle until the dialogue terminates. The design objective of the chatbot is to find optimal policies and skills to maximize the expected CPS (rewards).

The formulation of dialogue as a hierarchical decision-making process guides the design and implementation of XiaoIce. XiaoIce uses a dialogue manager to keep track of the dialogue state, and at each dialogue turn, selects how to respond based on a hierarchical dialogue policy. To maximize long-term user engagement, measured in expected CPS, we take an iterative, trial-and-error approach to developing XiaoIce, and always try to balance the exploration–exploitation tradeoff. We *exploit* what is already known to work well to retain XiaoIce's active users, but we also have to *explore* what is unknown (e.g., new skills and dialogue policies) in order to engage with the same users more deeply or attract new users in the future. In Figure 3, XiaoIce tries a new topic (i.e., a popular singer named Ashin) in Turn 5 and recommends a song in Turn 15, and thereby learns the user's preferences (e.g., the music topic and the singer he loves), knowledge that would lead to more engagement in the future.

3. System Architecture

The overall architecture of XiaoIce is shown in Figure 4. It consists of three layers: user experience, conversation engine, and data.

- **User experience layer:** This layer connects XiaoIce to popular chat platforms (e.g., WeChat, QQ), and communicates with users in two modes: full-duplex and taking turns. The full-duplex mode handles voice-stream-based conversations where a user and XiaoIce can talk to each other simultaneously. This mode is mainly used for the XiaoIce systems deployed on smart devices. The other mode deals with message-based conversations where a user and XiaoIce take turns to talk. This layer also

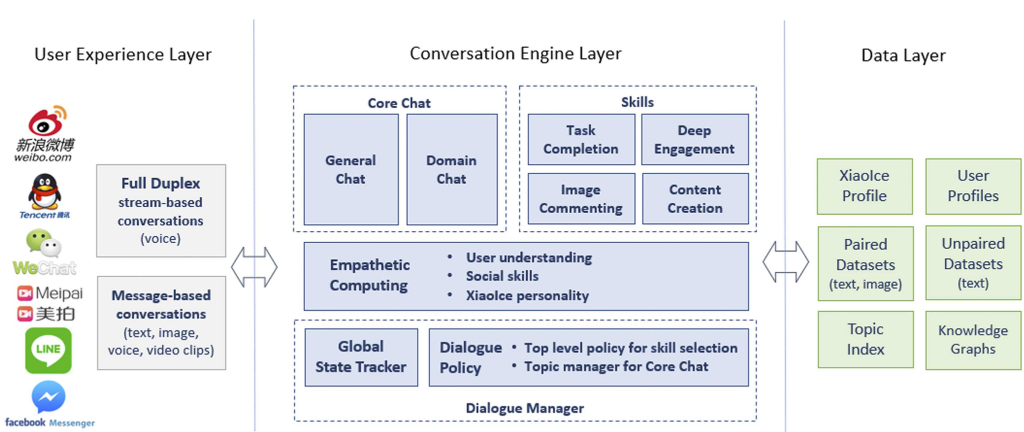

Figure 4
XiaoIce system architecture.

includes a set of components used to process user inputs and XiaoIce responses—for example, image understanding and text normalization, speech recognition and synthesis, voice activity detection to distinguish user input from background noise, a voice classifier to identify the age and gender of the user, and a talking-to-bot classifier to distinguish whether a user is talking to the bot or other human users.

- **Conversation engine layer:** This is composed of a dialogue manager, an empathetic computing module, Core Chat, and dialogue skills. The dialogue manager keeps track of the dialogue state, and selects either a dialogue skill or Core Chat[1] using the dialogue policy to generate responses. The empathetic computing module is designed to understand not only the content of the user input (e.g., topic) but also the empathetic aspects of the dialogue and the user (e.g., emotion, intent, opinion on topic, and the user's background and general interests). It reflects XiaoIce's EQ and demonstrates XiaoIce's social skills to ensure the generation of interpersonal responses that fit XiaoIce's personality. XiaoIce's IQ is shown by a collection of specific skills and Core Chat.
- **Data layer:** This consists of a set of databases that store collected human conversational data (in text pairs or text-image pairs), non-conversational data, and knowledge graphs used by Core Chat and skills, and the profiles of XiaoIce and all her active users.

4. Implementation of Conversation Engine

This section describes four major components in the conversation engine layer: dialogue manager, empathetic computing, Core Chat, and skills.

Implementation of the conversation engine relies heavily on A/B testing to evaluate if a new module or a new dialogue skill is going to improve an existing component. This is possible because XiaoIce has attracted a large number of active users since her launch in 2014. The metrics we commonly use for A/B testing include expected CPS and NAU. In addition, we ask users to give explicit feedback when a new module or a new dialogue skill is being tested. When working on the modules or tasks where there are benchmarks used in the research community, such as the neural response generator (Section 4.3), we often compare our approaches to other state-of-the-art methods using these benchmarks.

4.1 Dialogue Manager

Dialogue Manager is the central controller of the dialogue system. It consists of the Global State Tracker, which is responsible for keeping track of the current dialogue state $\mathbf{s}$, and Dialogue Policy π, which selects an action based on the dialogue state as $a = \pi(\mathbf{s})$. The action can be either a skill or Core Chat activated by the top-level policy to respond to the user's specific request, or a response suggested by a skill-specific low-level policy.

1 Although Core Chat is by definition a dialogue skill, we single it out by referring to it as *Core Chat* directly due to its importance and sophisticated design, and refer to other dialogue skills as *skills*.

4.1.1 Global State Tracker. Global State Tracker maintains a working memory to keep track of the dialogue state. The working memory is empty at the beginning of each dialogue session, and then stores at each dialogue turn the user utterance and XiaoIce's response as text strings, together with the entities and empathy labels detected from the text by the empathetic computing module, which will be described in Section 4.2. The information in the working memory is encoded into dialogue state vector $\mathbf{s}$.

4.1.2 Dialogue Policy. As described in Section 2.3, XiaoIce uses a hierarchical policy: (1) the top-level policy manages the overall conversation by selecting, at each dialogue turn, either Core Chat or a skill to activate based on the dialogue state; and (2) a set of low-level policies, one for each skill, manages its conversation segment.

The dialogue policy is designed to optimize long-term user engagement through an iterative, trial-and-error process based on the feedback of XiaoIce's users. The high-level policy is implemented using a set of skill triggers. Some of the triggers are based on machine learning models, such as the Topic Manager and the Domain Chat triggers. The others are rule-based, such as those that trigger the skills by keywords. The low-level policies of the Core Chat skills are implemented using hybrid response generation engines, as to be described in Section 4.3, and the low-level policies of the other skills (e.g., the task completion and deep engagement skills in Figure 4) are hand-crafted.

The high-level policy works as follows.

- If the user input is text (including speech-converted text) only, Core Chat is activated. Topic Manager, which will be introduced in Section 4.1.3, is designed to manage the dialogue segment of Core Chat by deciding whether to switch to a new topic or switch from the General Chat skill to a specific Domain Chat skill if the user's interest in a particular topic or domain is detected.
- If the user input is an image or a video clip, the Image Commenting skill is activated.
- The skills of Task Completion, Deep Engagement, and Content Creation are triggered by specific user inputs and conversation context. For example, a picture of food shared by a user can trigger the Food Recognition and Recommendation skill as shown in Figure 17(a); an extremely negative sentiment detected from user input can trigger the Comforting skill as shown in Figure 17(b); and a special user command such as "XiaoIce, what is the weather today" can trigger the Weather skill as shown in Figure 19(a). If multiple skills are triggered simultaneously, we select the one to activate based on their triggering confidence scores, pre-defined priorities, and the session context. To ensure a smooth conversation, we avoid switching among different skills too often. We prefer keeping the running skill activated until it terminates to activating a new skill. This is similar to the way sub-tasks (i.e., skills) are managed in composite-task completion bots (Peng et al. 2017).

We will discuss the low-level policies in the later sections where the individual dialogue skills are described.

4.1.3 Topic Manager. Topic Manager simulates human behavior of changing topics during a conversation. It consists of a classifier for deciding at each dialogue turn whether or not to switch topics, and a topic recommendation engine for suggesting a new topic.

Topic switching is triggered if XiaoIce does not have sufficient knowledge about the topic to engage in a meaningful conversation, or the user is getting bored. The classifier of topic switching is implemented using a boosted tree that incorporates the following indicator features.

- Whether an editorial response is used due to Core Chat failing to generate any valid response candidate, as will be described in Section 4.3.
- Whether the generated response simply repeats the user inputs, or contains no new information.
- Whether the user inputs are getting bland, for example, "OK," "I see," "go on."

The topic recommendation engine consists of a topic ranker, and a topic database that is constructed by collecting popular topics and related comments and discussions from high-quality Internet forums, such as Instagram in the United States and douban.com in China. The topic database is updated periodically. When topic switching is triggered, a list of topic candidates is retrieved from the database using the current dialogue state, which is generated using the empathetic computing module (Section 4.2), as query. The top-ranked candidate topic is chosen as the new topic. The topic ranker is implemented using a boosted tree ranker that uses the following features.

- Contextual relevance: The topic needs to be related to the dialogue, but has not been discussed yet.
- Freshness: The topic, especially if it is related to news, needs to be fresh and valid for the time being.
- Personal interests: The user is likely to be interested in the topic, according to the user profile.
- Popularity: The topic has gained enough attention on the Internet or among XiaoIce users.
- Acceptance rate: The rate of the topic being accepted by XiaoIce users is historically high.

Evaluation. Both the topic switching classifier and the topic ranker are trained using 50K dialogue sessions whose topics are manually labeled. Our A/B test over a 1-month period shows that incorporating topic management improves the expected CPS of Core Chat by 0.5. As shown in the example in Figure 3, XiaoIce switches to a new but related topic (i.e., a song titled "The Time Machine" by Ashin in Turn 13) when she detects that the user is not familiar with "Ashin" and about to terminate the conversation by responding "Ah! Boring" and "Okay, I am going to bed."

4.2 Empathetic Computing

Empathetic computing reflects XiaoIce's EQ and models the empathetic aspects of the human–machine conversation. Given user input query Q, empathetic computing takes

... ...

11 [Human] You like Ashin that much, huh?

[Xiaoice] Why not? You don't like him?

12 [Human] I don't like him, neither dislike him → I don't like Ashin, neither dislike Ashin

[Xiaoice] I love every song of him

13 [Human] Really?

[Xiaoice] Especially *The Time Machine*

14 [Human] Oh? Never heard of that → Oh? Never heard of *The Time Machine*

[Xiaoice] Shall I send you the song?

15 [Human] Okay → Okay, send *The Time Machine*.

[Xiaoice] Don't know if you will like it or not~ [Play the music card]

(a)

Topic	music, Ashin
Intent	request
Sentiment	sad, nerves -> neutral
Opinion	neutral
...	...
Age	25
Gender	male
Interests	singing, music ...
Occupation	salesman
Personality	quite, serious ...
...	...

(b)

Topic	music, Ashin
Intent	inform
Sentiment	happy
Opinion	positive
...	...
Age	18
Gender	female
Interests	music, games, movie, singing ...
Occupation	AI chatbot
Personality	reliable, sympathetic, affectionate...
...	...

(c)

Figure 5
An example conversation session (from Figure 3), where the empathetic computing module is used to (a) rewrite user queries into contextual queries as indicated by the arrows, (b) generate the query empathy vector $\mathbf{e}_Q$ in Turn 11, and (c) generate the response empathy vector $\mathbf{e}_R$ for Turn 11.

its context C into account and rewrites Q to its contextual version Q_c, encodes the user's feelings and states in the conversation with query empathy vector $\mathbf{e}_Q$, and specifies the empathetic aspects of the response R with response empathy vector $\mathbf{e}_R$. The output of the empathetic computing module is represented as dialogue state vector $\mathbf{s} = (Q_c, C, \mathbf{e}_Q, \mathbf{e}_R)$, which is the input to both Dialogue Policy for selecting a skill, and the activated skill (e.g., Core Chat) for generating interpersonal responses that fit XiaoIce's persona—an 18-year-old girl who is always reliable, sympathetic, affectionate, knowledgeable but self-effacing, and has a wonderful sense of humor.

The empathetic computing module consists of three components: contextual query understanding, user understanding, and interpersonal response generation. Figure 5 shows an example of how the module controls the empathetic aspects of the conversation in Figure 3.

Contextual Query Understanding (CQU). CQU rewrites Q to Q_c using contextual information in C in the following steps.

- Named entity identification: We label all entity mentions in Q, link them to the entities stored in the working memory of the state tracker, and store new entities in the working memory.

- Co-reference resolution: We replace all pronouns with their corresponding entity names.
- Sentence completion: If Q is not a complete sentence, we complete it using contextual information C.

As shown in Figure 5(a), CQU rewrites user queries to include necessary context, for example, replacing "him" in Turn 12 with "Ashin," "that" with "The Time Machine" in Turn 14, and adding "send The Time Machine" in Turn 15. These contextual queries are used by Core Chat to generate responses via either a retrieval-based engine or a neural response generator, which will be described in Section 4.3.

User Understanding. This component generates query empathy vector $\mathbf{e}_Q$ based on Q_c and C. $\mathbf{e}_Q$ consists of a list of key-value pairs representing the user's intents, emotions, topics, opinions, and the user's persona, as shown in Figure 5(b). These key-value pairs are generated using a set of machine learning classifiers as follows.

- Topic detection labels whether the user follows the same topic, or introduces a new topic. The set of topics is pre-compiled in the topic database of Topic Manager.
- Intent detection labels Q_c using one of 11 dialogue acts—greet, request, inform, and so forth.
- Sentiment analysis detects user's emotion of five types (e.g., happy, sad, angry, neural) and how her emotion evolves during the conversation (e.g., from happy to sad).
- Opinion detection detects the user's reaction to the topic (i.e., positive, negative, or neutral).
- If the user ID is available, include in $\mathbf{e}_Q$ the user persona vector according to her profile (gender, age, interests, occupation, personality, etc.).

Interpersonal Response Generation. This component generates response empathy vector $\mathbf{e}_R$ that both specifies the empathetic aspects of the response to be generated and embodies XiaoIce's persona. For example, $\mathbf{e}_R$ in Figure 5(c) indicates that XiaoIce shares the feeling of the user by following the same topic (decided by Topic Manager), responding in a consistent and positive way as specified (by the values of intent, sentiment, opinion, etc.) in $\mathbf{e}_R$, which are computed based on those in $\mathbf{e}_Q$ using a set of heuristics. The response must also fit XiaoIce's persona, whose key-value pairs, such as age, gender, and interests, are extracted from the pre-compiled XiaoIce profile. We will describe how $\mathbf{e}_Q$ and $\mathbf{e}_R$ are used for response generation in Section 4.3.

Evaluation. The empathetic computing module consists of a set of classifiers. We use an off-the-shelf named entity recognizer (Gao et al. 2005) for identifying 13 types of named entities without retraining for CQU, and train a group of classifiers for user understanding (i.e., co-reference resolution, topic detection, intent detection, opinion detection, and sentiment analysis) using 10K manually labeled dialogue sessions. The effectiveness of the empathetic computing module is verified in the A/B test on Weibo users. Although we do not observe any significant change in CPS, NAU is increased from 0.5 million to 5.1 million in 3 months. The module was released in July 2018, and became the most important feature in the sixth generation of XiaoIce, which has substantially strengthened XiaoIce's emotional connections to human users and increased XiaoIce's NAU.

4.3 Core Chat

Core Chat is a very important component of XiaoIce's IQ and EQ. Together with the empathetic computing module, Core Chat provides the basic communication capability by taking the text input and generating interpersonal responses as output. Core Chat consists of two parts, General Chat and a set of Domain Chats. General Chat is responsible for engaging in open-domain conversations that cover a wide range of topics. Domain Chats are responsible for engaging in deep conversations on specific domains such as music, movies, and celebrities. Because General Chat and Domain Chats are implemented using the same engine with access to different databases (i.e., general vs. domain-specific paired, unpaired databases, and neural response generator), we only describe General Chat here.

General Chat is a data-driven response generation system. It takes as input dialogue state $\mathbf{s} = (Q_c, C, \mathbf{e}_Q, \mathbf{e}_R)$, and outputs response R in two stages: response candidate generation and ranking. The response candidates can be retrieved from the databases that consist of human-generated conversations or texts, or generated on the fly by a neural response generator. The query and response empathy vectors, $\mathbf{e}_Q$ and $\mathbf{e}_R$, are used for both candidate generation and ranking to ensure that the generated response is interpersonal and fits XiaoIce's persona. In what follows, we describe three candidate generators and the candidate ranker.

Retrieval-Based Generator using Paired Data. The paired database consists of query–response pairs collected from two data sources. First is the human conversational data from the Internet—social networks, public forums, bulletin boards, news comments, and so on. After the launch of XiaoIce in May 2014, we also started collecting human–machine conversations generated by XiaoIce and her users, which amounted to more than 30 billion conversation pairs as of May 2018. Nowadays, 70% of XiaoIce's responses are retrieved from her own past conversations. To control the quality of the database, especially for the data collected from the Internet, we convert each pair to a tuple (Q_c, R, e_Q, e_R) using the empathetic computing module based on information extracted from dialogue context, metadata of the Web page and Web site where the pair is extracted, and the user profile (if the subscribed user's identity is available). Then, we filter the pairs based on their tuples, and retain only the conversation pairs that contain empathetic responses that fit XiaoIce's persona. We also remove the pairs that contain personally identifiable information, messy code, inappropriate content, spelling mistakes, and so forth.

The filtered paired database is then indexed using Lucene.[2] At runtime, we use Q_c in $\mathbf{s}$ as query to retrieve up to 400 response candidates using keyword search and semantic search based on machine learning representations of the paired database [Wu, Wang, and Xue 2016; Zhang et al. 2016].

Although the response candidates retrieved from the paired database is of high quality, the coverage is low because many new or less frequently discussed topics on the Internet forums are not included in the database. To increase the coverage, we introduce two other candidate generators, described next.

2 `http://lucene.apache.org/`.

Neural Response Generator. Unlike the retrieval-based generator, the neural response generator is *trained* using the paired database to learn to simulate human conversations, and is able to generate responses for any topics including those that are unseen in human conversational data, so that a user can chat about any topic she likes. Neural-model-based and retrieval-based generators are complementary: The neural-model-based generator offers robustness and high coverage, whereas the retrieval-based provides high-quality responses for popular topics. Neural response generation is a very active research topic in the conversational AI community (Gao, Galley, and Li 2019). Its role in developing social chatbots is becoming increasingly important as its performance keeps improving.

The neural response generator in XiaoIce follows the sequence-to-sequence (seq2seq) framework (Cho et al. 2014; Sutskever, Vinyals, and Le 2014) used for conversation response generation (Shang, Lu, and Li 2015; Sordoni et al. 2015; Vinyals and Le 2015; Li et al. 2016a, 2016b; Serban et al. 2016; Xing et al. 2017).

The generator is based on a GRU-RNN model, similar to the Speaker-Addressee model (Li et al. 2016b). Given input $(Q_c, \mathbf{e}_Q, \mathbf{e}_R)$, we wish to predict how XiaoIce (the addressee) modeled by $\mathbf{e}_R$ would respond to query Q_c produced by the user (speaker) modeled by $\mathbf{e}_Q$. As illustrated in Figure 6, we first obtain an interactive representation $\mathbf{v} \in \mathbb{R}^d$ by linearly combining query and response empathy vectors $\mathbf{e}_Q$ and $\mathbf{e}_R$ in an attempt to model the interactive style of XiaoIce toward the user,

$$\mathbf{v} = \sigma(\mathbf{W}_Q^\top \mathbf{e}_Q + \mathbf{W}_R^\top \mathbf{e}_R)$$

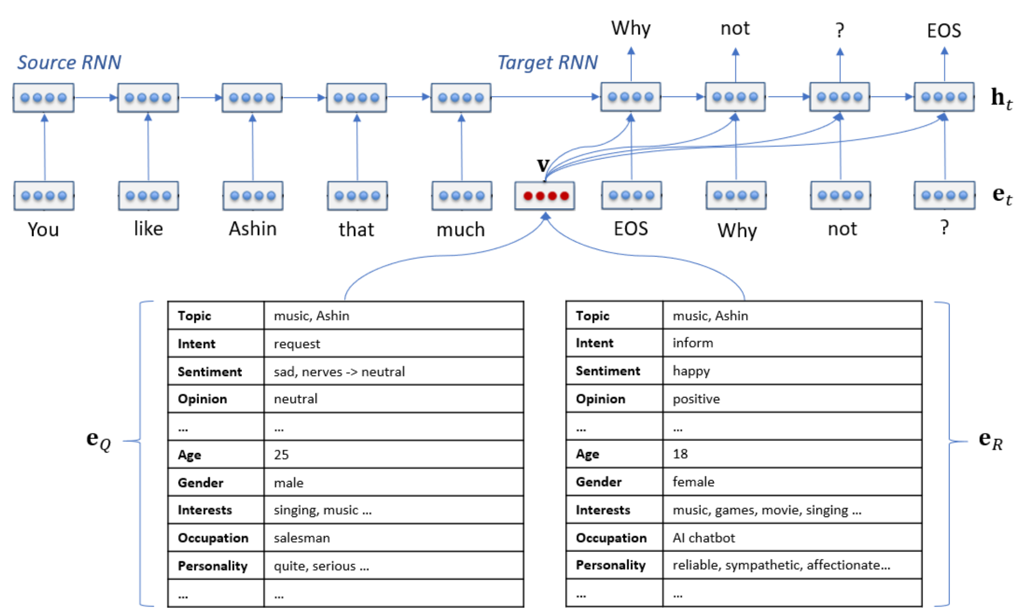

Figure 6
Recurrent neural network–based neural response generator. Given the user query "You like Ashin that much," the response candidate "why not?" is generated.

where $\mathbf{W}_Q, \mathbf{W}_R \in \mathbb{R}^{k \times d}$, and σ denotes the sigmoid function. Then the source recurrent neural network (RNN) encodes user query Q_c into a sequence of hidden state vectors that are then fed into the target RNN to generate response R word by word. Each response ends with a special end-of-sentence symbol, EOS. We use beam search to generate up to 20 candidates. As illustrated in Figure 6, for each step t on the target RNN side, the hidden state $\mathbf{h}_t$ is obtained by combining the hidden state produced at the previous step $\mathbf{h}_{t-1}$, the embedding vector of the word at the current time step $\mathbf{e}_t$, and $\mathbf{v}$. In this way, empathy information is injected into the hidden layer at each time step to help generate interpersonal responses that fit XiaoIce's persona throughout the generation process. As shown in Figure 7, although a typical seq2seq model that is not grounded in any persona often outputs inconsistent responses (Li et al. 2016b), XiaoIce is able to generate consistent and humorous responses.

For completeness, we give a detailed description of the model. Let $\mathbf{u}_t$ and $\mathbf{z}_t$ denote the update and reset gates of GRU, respectively, which associate with time step t. Then, the hidden state $\mathbf{h}_t$ of the GRU-RNN for each time step t is computed as follows:

$$
\begin{aligned}
\mathbf{u}_t &= \sigma(\mathbf{W}_u^\top [\mathbf{h}_{t-1};\ \mathbf{e}_t; \mathbf{v}]) \\
\mathbf{z}_t &= \sigma(\mathbf{W}_z^\top [\mathbf{h}_{t-1};\ \mathbf{e}_t;\ \mathbf{v}]) \\
\mathbf{l}_t &= \tanh(\mathbf{W}_l^\top [\mathbf{z}_t \circ \mathbf{h}_{t-1};\ \mathbf{e}_t;\ \mathbf{v}]) \\
\mathbf{h}_t^Q &= (1 - \mathbf{u}_t) \circ \mathbf{h}_{t-1} + \mathbf{u}_t \circ \mathbf{l}_t
\end{aligned}
$$

where $\mathbf{W}_u, \mathbf{W}_z, \mathbf{W}_l \in \mathbb{R}^{3d \times d}$ are machine learned matrices, and $\circ$ denotes the element-wise product. The RNN model defines the probability of next token in R to predict using the softmax function:

$$
\begin{aligned}
p(R|Q_c, \mathbf{e}_Q, \mathbf{e}_R) &= \prod_{t=1}^{N_R} p(r_t|Q_c, \mathbf{e}_Q, \mathbf{e}_R, r_1, r_2, ..., r_{t-1}) \\
&= \prod_{t=1}^{N_R} \frac{\exp(f(\mathbf{h}_{t-1}, \mathbf{e}_{r_t}, \mathbf{v}))}{\sum_{r'} \exp(f(\mathbf{h}_{t-1}, \mathbf{e}_{r'}, \mathbf{v}))}
\end{aligned}
$$

where $f(\mathbf{h}_{t-1}, \mathbf{e}_{r_t}, \mathbf{v})$ denotes the activation function between $\mathbf{h}_{t-1}$, $\mathbf{e}_{r_t}$, and $\mathbf{v}$, where $\mathbf{h}_{t-1}$ is the representation output from the RNN at time $t-1$. Each response ends with a special end-of-sentence symbol, EOS.

[Human] How old are you?
[S2S-Bot] 16 and you?
[Human] What's your age?
[S2S-Bot] 18.

[Human] How old are you?
[XiaoIce] I am 18, of course.
[Human] You age?
[XiaoIce] 18. Why?
[Human] Were you 18 last year?
[XiaoIce] I made a wish to stay 18 forever. Oh, my wish has come true.

Figure 7
(*Left*) Examples of inconsistent responses generated using a seq2seq model (S2S-Bot) that is not grounded in persona (Li et al. 2016b). (*Right*) Examples of consistent and humorous responses generated using the neural response generator of XiaoIce.

The parameters of the response generation model $\theta = (\mathbf{W}_Q, \mathbf{W}_R, \mathbf{W}_u, \mathbf{W}_z, \mathbf{W}_l)$ are trained to maximize the log likelihood on training data, using stochastic gradient descent, as

$$\arg\max_{\theta} \frac{1}{M} \sum_{i=1}^{M} \log p_{\theta}(R^{(i)} | Q_c^{(i)}, \mathbf{e}_Q^{(i)}, \mathbf{e}_R^{(i)})$$

Retrieval-Based Generator using Unpaired Data. In addition to the conversational (or paired) data used by these two response generators, there is a much larger amount of non-conversational (or unpaired) data that can be used to improve the coverage of the response.

The unpaired database we have used in XiaoIce consists of sentences collected from public lectures and quotes in news articles and reports. These sentences are considered candidate responses R. Because we know the authors of these sentences, we compute for each its empathy vector $\mathbf{e}_R$. A data filtering pipeline, similar to that for paired data, is used to retain only the responses $(R, \mathbf{e}_R)$ that fit XiaoIce's persona.

Like the paired database, the unpaired database is indexed using Lucene. Unlike the paired database, at runtime we need to expand query Q_c to include additional topics to avoid retrieving those responses that simply repeat what a user just said. We resort to a knowledge graph (KG) for query expansion. The KG consists of a collection of *head-relation-tail* triples (h, r, t), and is constructed by joining the paired database and Microsoft Satori.[3] We include in the XiaoIce KG a Satori triple (h, r, t) only if h and t co-occur often enough in the same conversations. Such a triple contains a pair of related topics (h, t) that humans often discuss in one conversation, such as `(Beijing, Great Wall)`, `(Einstein, Relativity)`, `(Quantum Physics, Schrodinger's cat)`. A fragment of the XiaoIce KG is shown in Figure 8 (*top*).

Figure 8 illustrates the process of generating response candidates using the unpaired database. It consists of three steps.

- First, we identify the topics from contextual user query Q_c, for example, "Beijing" from "tell me about Beijing."
- For each topic, we retrieve up to 20 most related topics from the KG, for example, "Badaling Great Wall" and "Beijing snacks." These topics are scored by their relevance using a boosted tree ranker (Wu et al. 2010) trained on manually labeled training data.
- Finally, we form a query by combining the topics from Q_c and the related topics from the KG, and use the query to retrieve from the unpaired database up to 400 most relevant sentences as response candidates.

This generator is complementary to the other two generators aforementioned. Although the overall quality of the candidates generated from the unpaired database is lower than those retrieved from the paired database, with the unpaired database XiaoIce can cover a much broader range of topics. Compared with the neural response generator, which often generates well-formed but short responses, the candidates from unpaired database are much longer with more useful content.

3 Satori is Microsoft's knowledge graph, which is seeded by Freebase, and now is orders of magnitude larger than Freebase.

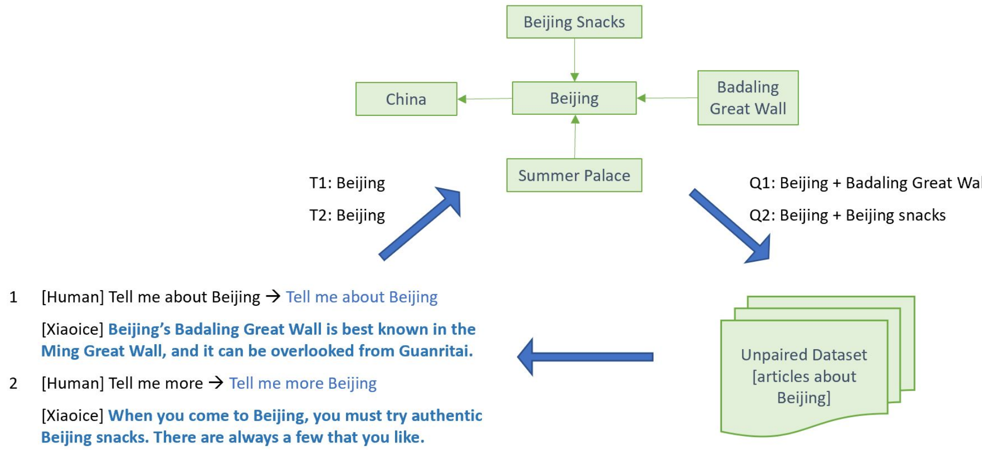

Figure 8
An example of generating response candidates using the unpaired database and the XiaoIce knowledge graph (KG), for which we show a fragment of the XiaoIce KG that is related to the topic "Beijing" (*top*). For a human-machine conversation (*bottom-left*), each user query is rewritten to a context query indicated by the arrow, then its topics (e.g., "Beijing") are identified, the related topics ("Badaling Great Wall" and "Beijing snacks") are retrieved from the KG (*top*), and response candidates are retrieved from the unpaired database (*bottom-right*) using a query that combines the query topics and their related topics.

Response Candidate Ranker. The response candidates generated by three generators are aggregated and ranked using a boosted tree ranker (Wu et al. 2010). A response is selected by randomly sampling a candidate from those with higher ranking scores than a pre-set threshold.

Given dialogue state $\mathbf{s} = (Q_c, C, \mathbf{e}_Q, \mathbf{e}_R)$, we assign each response candidate R' a ranking score based on four categories of features.

- Local cohesion features. A good candidate should be semantically consistent or related to user input Q_c. We compute cohesion scores between R' and Q_c using a set of DSSMs[4] trained on a collection of human conversation pairs.
- Global coherence features. A good candidate should be semantically coherent with Q_c and C. We compute coherence scores between R' and (Q_c, C) using another set of DSSMs trained on a collection of human dialogue sessions. Because the coherence features use global context information C, they are particularly useful when Q_c is a bland query whose topic is hard to detect without context, such as "OK," "why," "I don't know."
- Empathy matching features. A good candidate should be an empathetic response that fits XiaoIce's persona. Assume XiaoIce selects R' to respond given context (Q_c, C). We can compute XiaoIce's response empathy vector

4 DSSM stands for Deep Structured Semantic Models (Huang et al. 2013; Shen et al. 2014), or more generally, Deep Semantic Similarity Model (Gao et al. 2014). DSSM is a deep learning model for measuring the semantic similarity of a pair of inputs (x, y). They can be applied to a wide range of tasks, depending on the definition of (x, y). In this study (x, y) is a query-candidate–response pair (Q_c, R').

for R', $\mathbf{e}_{R'}$, using the empathetic computing module,[5] and then compute a set of empathy matching features by comparing $\mathbf{e}_{R'}$ and the given $\mathbf{e}_R$ that encodes the empathy features of the *expected* response.

- Retrieval matching features. These features apply only to the candidates generated from the paired database. We compute a set of matching scores between Q_c and the query side of the retrieved query–response pairs at both the word level, such as BM25 and TFIDF scores, and the semantic level, such as DSSM scores.

The ranker is trained on dialogue-state-response pairs $(\mathbf{s}, R)$, as shown in Figure 9, where each pair is labeled on a three-level quality scale:

- 0: The response is not empathetic or not very relevant to the query. It is likely to lead to the termination of the conversation.
- 1: The response is acceptable and relevant to the query. It is likely to help keep the conversation going.
- 2: This is an empathetic, interpersonal response that fits XiaoIce's persona and makes users feel delighted and excited. It is likely to drive the conversation.

Editorial Response. If the candidate generators and response ranker fail to generate any valid response for various reasons (e.g., not-in-index, model failure, execution timeout, or the input query containing improper content), then an editorial response is selected.

It is important to provide an empathetic editorial response to keep the conversation going. For example, when not-in-index occurs, instead of using safe but bland responses such as "I don't know" or "I am still learning to answer your question," XiaoIce may respond like, "Hmmm, difficult to say. What do you think?," or "Let us talk about something else."

Evaluation. We present two pilot studies that validate the effectiveness of the persona-based neural response generator and the hybrid approach that combines the generation-based and retrieval-based methods, respectively, and then the A/B test of General Chat.

In the first pilot study reported in Li et al. (2016b), we compare the persona model against two baseline models, using a TV series data set for model training and evaluation. The data set consists of scripts of 69,565 dialogue turns of 13 main characters from the American TV comedies *Friends*[6] and *The Big Bang Theory*,[7] available from IMSDB.[8] The first baseline is a vanilla seq2seq model. The second is the LSTM-MMI model (Li et al. 2016a), which is one of the state-of-the-art neural response generation models. As shown in Table 1, the persona model significantly outperforms both baselines, achieving a lower perplexity (−8.4%) and a higher BLEU score (+18.8% and +11.8%) (Papineni et al. 2002). The qualitative analysis confirms that the persona model indeed generates more

5 We treat R' as query and (Q_c, C) as context, and use the contextual query understand and user understanding components to compute $\mathbf{e}_{R'}$ as a query empathy vector.
6 `https://en.wikipedia.org/wiki/Friends`.
7 `https://en.wikipedia.org/wiki/The_Big_Bang_Theory`.
8 `http://www.imsdb.com`.

Query	Response	Rating
捂上脸不看 (Don't want to look with my face in my hands.)	我边吃饭边看你 (I'm watching you while eating.)	0
	你怎么了？(what is wrong with you?)	1
	当心捂出一脸痱子，哈哈. (Haha, beware of the heat rash of your face.)	2
我肚子有点疼 (I have a pain in my stomach.)	运动对身体有好处的 (Exercise is good for your health.)	0
	喝热水了吗？(Have you been drinking hot water?)	1
	来和我聊天，分散一下注意力就好了 (Come and talk to me, just distract yourself.)	2
我觉得自己太笨了 (I feel so stupid.)	大象的智商很高 (The elephant has a high IQ.)	0
	多学习提高自己 (Learn more and improve yourself.)	1
	很多聪明人都这么说耶 (That is what lots of smart people say about themselves.)	2
为什么人总是活成自己不喜欢的样子 (Why do we always live the way we don't like.)	喜欢一个人是发自内心的 (Love a person from my heart.)	0
	因为人想要的太多 (Because we want too much.)	1
	因为那个是别人喜欢的样子 (Because that's what others like.)	2
我考试没及格 (I did not pass the exam.)	那要看和谁比呢 (It depends on whom you compare against.)	0
	一次没关系，下次努力吧 (Don't worry, try it next time.)	1
	起码你不用怀疑自己能不能及格了 (At least you don't have to wonder if you can pass.)	2

Figure 9
Examples of query–response pairs that are used for training and validating General Chat. Each pair is labeled on a three-level quality scale: 2 = an empathetic response that is likely to drive the conversation; 1 = an acceptable response that is likely to keep the conversation going; 0 = a non-empathetic response that is likely to terminate the conversation.

Table 1
Perplexity and BLEU for the seq2seq and persona models on the TV series data set. Adapted from Li et al. (2016b).

Models	Perplexity (vs. seq2seq)	BLEU (vs. seq2seq) (vs. LSTM-MMI)
seq2seq	27.3	1.6%
LSTM-MMI	27.3	1.7% (+6.3%)
Persona	25.0 (−8.4%)	1.9% (+18.8%) (+11.8%)

interpersonal responses than the baselines. As shown in the examples in Table 2, the persona model is sensitive to the identity of the user (addressee), generating specific words (e.g., the user names) in responses targeted at different users. For example, the model produces "Of course, I love you, Emily," in response to the input from Emily, and generates "Of course I love you. (kissed him)," where the pronoun "him" accurately identifies the gender of the user (i.e., Ross).

In the second pilot study, we validate the effectiveness of a hybrid system that combines the neural response generator and the retrieval-based generator using paired database. We do so by comparing the hybrid system against two baseline systems that use only one of the candidate generators, respectively. All these systems use the same response candidate ranker. The neural response generator and the set of classifiers and models that are used to generate the ranking features for the candidate ranker (e.g., local cohesion and global coherence features) are trained using 50 million human dialogues. The response candidate ranker is trained using 50K manually labeled dialogues. Our

Table 2
Responses to "Do you love me?" from the persona model on the TV series data set using different addressees and speakers. Adapted from Li et al. (2016b).

user query	Do you love me?
baseline	Of course, I love you.
	addressee: Emily, speaker: Penny
response	Of course, I love you, Emily.
	addressee: Leonard, speaker: Penny
response	Of course, I love you, Leonard.
	addressee: Monica, speaker: Chandler
response	Of course I love you, I want to marry you.
	addressee: Rachel, speaker: Chandler
response	Of course I love you.
	addressee: Ross, speaker: Rachel
response	Of course I love you. (kissed him)
	addressee: Emily, speaker: Rachel
response	Of course I love you.

evaluation data consists of 4K dialogue sessions. All three systems (i.e., the hybrid and two baseline systems) need to generate a response for each user query and its context in these dialogue sessions. Each generated response is labeled on the three-level quality scale by three human judges. The results in Table 3 show that incorporating the neural generator, as in the hybrid system, significantly improves the human rating over the retrieval-based system.

Our A/B test confirms the conclusions we draw from the pilot studies. Compared to the baseline which uses only the retrieval-based generator using paired database for candidate generation, incorporating the neural response generator and the retrieval-based generator using unpaired database at the candidate generation stage improves the expected CPS of Core Chat by 0.5 in two weeks. A detailed analysis shows that the gain is mainly attributed to the fact that the neural response generator and the retrieval-based generator using unpaired database significantly improve the coverage of responses. We measure the response coverage of a system by calculating the number of distinct acceptable and good responses (i.e., responses with ratings of 1 or 2, respectively) that the system generates for a given user input. We find that incorporating the neural-based generator into the baseline improves the coverage by 20%, and incorporating the retrieval-based generator using unpaired database into the baseline improves the coverage by 10%.

Table 3
Ratings of three response generation systems on a 5K dialogue data set.

Systems	Av. Rating	Rating = 0	Rating = 1	Rating = 2
Retrieval-based	0.87	35.0%	42.9%	22.1%
Neural-generator-based	0.40	66.6%	27.3%	6.1%
Hybrid	1.09	23.1%	44.6%	32.3%

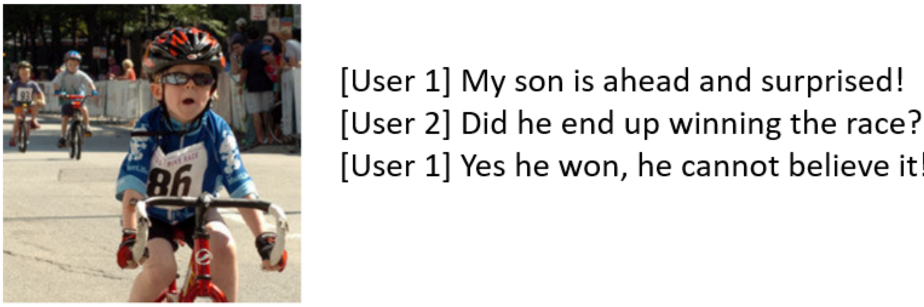

Figure 10
An example conversation around a shared image. Figure credit: Mostafazadeh et al. (2017).

4.4 Image Commenting

In social chatting, people frequently engage with one another around images. On Twitter, for example, uploading a photo with an accompanying tweet (comment) has become increasingly popular: As of June 2015, 28% of tweets contain an image (Morris et al. 2016). Figure 10 illustrates a social chat around a shared image. We see that the conversation is grounded not only in the visible objects (e.g., the boys, the bikes) but in the events, actions, or even emotions (e.g., the race, winning) implicitly in the image. To human users, it is these latter aspects that are more important to drive a meaningful and interesting conversation.

The Image Commenting skill is designed to not only correctly recognize objects and truthfully describe the content of an image, but generate empathetic comments that reflect personal emotion, attitude, and so forth. It is the latter, the social skill aspects, that distinguish image commenting from other traditional vision tasks such as image tagging and image description, as illustrated in Figure 11.

The architecture for Image Commenting is similar to that for Core Chat. Given the user input that contains an image (or a video clip), a textual comment is generated in two stages: candidate generation and ranking. The candidates are generated using retrieved-based and generation-based approaches.

In the retrieval-based approach, first of all, a database of image–comment pairs, collected from social networks (e.g., Facebook and Instagram), is constructed. To control for data quality, a pipeline similar to that for Core Chat is applied to retain only the

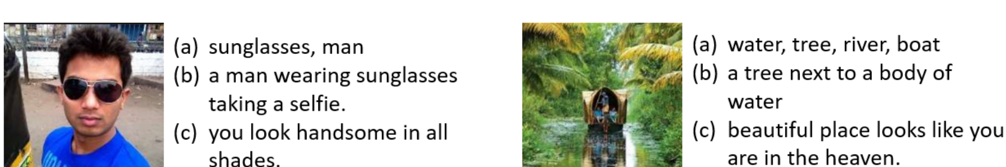

Figure 11
Examples of (a) image tagging, (b) image description, and (3) image commenting. Figure credit: Shum, He, and Li (2018).

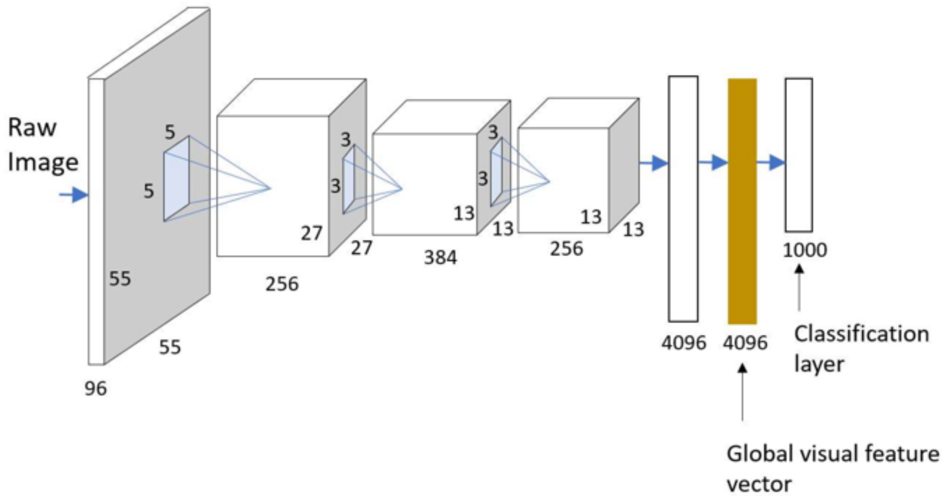

Figure 12
An example of deep convolutional neural network for visual feature vector extraction. Figure credit: Shum, He, and Li (2018).

pairs whose text comments fit XiaoIce's persona.[9] Then, each image is encoded into a visual feature vector that represents the overall semantic information of the image, using the deep convolutional neural networks, as illustrated in Figure 12. At runtime, given a query image, we retrieve up to three most similar images, ranked based on the cosine similarities between their feature vector representations, and use their paired comments as candidates.

The generation-based approach uses an image-to-text generator, an extension of the Microsoft Image Captioning system (Fang et al. 2015) that is re-trained on the image-comment pairs we have collected for XiaoIce and has incorporated additional modules to control high-level sentiment and style factors in comment generation (Mathews, Xie, and He 2016; Gan et al. 2017).

The comment candidates generated by the generators are aggregated and ranked using a boosted tree ranker (Wu et al. 2010). Given dialogue state $\mathbf{e} = (Q_c, C, \mathbf{e}_Q, \mathbf{e}_R)$, we assign each candidate R' a ranking score based on four categories of features, similar to that of Core Chat as described in Section 4.3. Note that unlike the case of Core Chat, where Q_c and R' are text, in Image Commenting we need to compute the similarity between an image and a text. This is achieved by using the Deep Multimodal Similarity Model (Fang et al. 2015) trained on a large amount of image–comment pairs. The ranker is trained on dialogue-state-response pairs $(\mathbf{s}, R)$, where Q_C in $\mathbf{s}$ is a vector representation of an image, and each pair is labeled on a three-level quality scale, similar to that of query–response pairs used for Core Chat.

As illustrated in Figure 13, good image comments (rating 2) need to fit well into the dialogue context and stimulate an engaging conversation. For example, in the first

9 We found that the pairs that are shared among acquaintances (e.g., coworkers, classmates, and friends) are of good quality, and amount to a large proportion in the database.

Figure 13
Examples of image-comment pairs that are used for training and validating Image Commenting. Each pair is labeled on a three-level quality scale: 2 = an empathetic comment that is likely to drive the conversation; 1 = an acceptable comment that is likely to keep the conversation going; 0 = a non-empathetic (or irrelevant) comment that is likely to terminate the conversation.

picture, instead of telling the users that this is the Leaning Tower of Pisa, XiaoIce responds "should I help you hold it?" after detecting that the person in the picture is presenting a pose pretending to support the tower. In the second example, instead of replaying the fact that there is a cat in the picture, XiaoIce makes a humorous comment on the cat's innocent eyes. In the other two examples, XiaoIce generates meaningful and interesting comments by grounding the images in the action (e.g., "not to trust any code from unknown sources") and object (e.g., "Windows") implicitly in the images.

Evaluation. The components of Image Commenting, including the text-to-image generator and boosted tree ranker, are trained on a data set consisting of 28 million images, each paired with six text comments rated on the three-level quality scale as shown in Figure 13. The image-comment pairs with ratings of 1 and 2 are extracted from the database used for the retrieval-based candidate generator. These ratings are determined automatically based on how many times users follow the comments, computed from the XiaoIce logs. The image–comment pairs with rating 0 are randomly sampled. Table 4 presents the result of a pilot study (Huang et al. 2019) showing that the XiaoIce Image Commenting skill outperforms several state-of-the-art image captioning systems on a test set consisting of 5K image–comment pairs whose ratings are 2, in terms of BLEU-4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence, Zitnick, and Parikh 2015), ROUGE-L (Lin 2004), and SPICE (Anderson et al. 2016).

Figure 14 shows a few example comments generated by the competing systems in Table 4. It can be observed that the XiaoIce-produced comments are emotional, subjective, imaginative, and are very likely to inspire meaningful human–machine interactions, while the comments generated by the other image captioning models are reasonable in content but boring in the context of social chats, and thus less likely to improve user engagement.

In the A/B test we observe that Image Commenting doubles the expected CPS across all dialogues that contain images.

Table 4
Image commenting results of XiaoIce and four state-of-the-art image captioning systems, in percent. Adapted from Huang et al. (2019).

Systems	BLEU-4	ROUGE-L	CIDEr-D	METEOR	SPICE
LSTM-XE (Vinyals et al. 2015)	2.96	11.6	1.74	10.43	3.27
LSTM-RL (Rennie et al. 2017)	3.43	12.3	2.08	11.84	3.64
DMSM (Fang et al. 2015)	2.73	10.52	1.22	11.37	2.63
Up-Down (Anderson et al. 2018)	3.23	12.73	1.52	12.66	2.69
XiaoIce (prototype)	4.53	15.33	3.21	15.51	4.82

Figure 14
Image comments generated by XiaoIce (prototype) and four state-of-the-art image captioning systems. Adapted from Huang et al. (2019).

4.5 Dialogue Skills

XiaoIce is equipped with 230 dialogue skills, which, together with Core Chat and Image Commenting, form the IQ component of XiaoIce. This section describes these skills in three categories: content creation, deep engagement, and task completion.

Evaluation. Most of these skills are designed for very specific user scenarios or tasks, implemented using hand-crafted dialogue policies and template-based response generators unless otherwise stated. These skills are evaluated in two stages: a lab study and a market study. In the lab study, human subjects are recruited, possibly through crowd-sourcing platforms, to test-use a dialogue skill to solve a particular task, so that a collection of dialogues are obtained. Metrics such as task-completion rate, average turns per session, and user ratings can be measured. In the market study, we evaluate the effectiveness of a dialogue skill by releasing it to the market. Because any individual

Figure 15
Examples of Content Creation skills and their triggers. (a) XiaoIce FM for Somebody, triggered by the command “make an FM program for [name].” (b) XiaoIce Kids Story Factory, triggered by the command “kids story factory.”

skill is unlikely to have a significant impact on CPS, we measure the user satisfaction of a skill by monitoring its active users and skill triggering rate (i.e., the number of times a skill is activated by users within a day or a week). A skill can be retired or reenter the market based on the market study result.

4.5.1 Content Creation. These skills allow XiaoIce to collaborate with human users in their creative activities, including text-based Poetry Generation,[10] voice-based Song and Audio Book Generation, XiaoIce FM for Somebody, XiaoIce Kids Story Factory, and so on.

Figure 15(a) shows that a user uses XiaoIce to make an FM program for her mother for the coming Chinese Spring Festival. Figure 15(b) shows the Kids Story Factory skill, which can automatically create a story based on user configuration (e.g., whether the story is for education or entertainment) and the names, genders, and personalities of the main characters, and so forth.

The XiaoIce Poetry Generation skill has helped over four million users to generate poems. On 15 May 2018, XiaoIce published the first AI-created Chinese poem album in history.[11] XiaoIce’s second poetry album is going to be published by China Youth Publishing and Microsoft in 2019. Every poem in the album is jointly written by XiaoIce and human poets. Figure 16 illustrates how a Chinese poem is generated from an image by XiaoIce. Given the image, a set of keywords, such as “city” and “busy,” are generated based on the objects and sentiment detected from the image. Then, a sentence is generated using each keyword as a seed. The generated sentences form a poem using a hierarchical RNN that models the structure among the words and sentences.

10 https://poem.msxiaobing.com/.
11 https://item.jd.com/12076535.

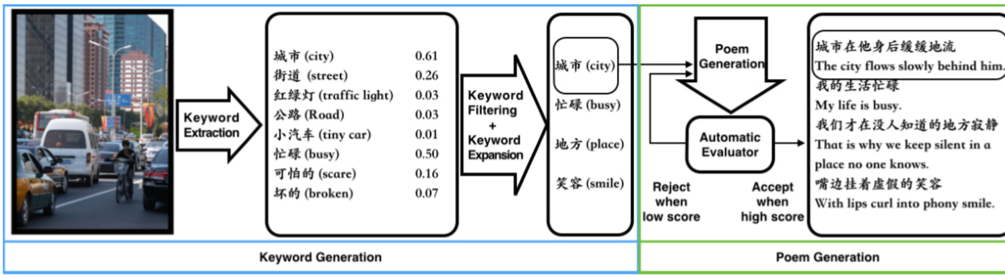

Figure 16
The framework of the Poem Creation skill. The system takes an image query given by a user, and outputs a semantically relevant piece of modern Chinese poetry. We first generate a set of keywords from the picture (*left*), and then generate a poem consisting of multiple lines, each generated using a keyword as a seed (*right*). Figure credit: Cheng et al. (2018).

4.5.2 Deep Engagement. The Deep Engagement skills are designed to meet users' specific emotional and intellectual needs by targeting to specific topics and settings, thus improving users' long-term engagement. Some example skills are shown in Figure 17.

As shown in Figure 18, these skills can be grouped into different series on two dimensions: from IQ to EQ, and from private one-on-one to group discussion.

- To meet users' intellectual or emotional needs (the IQ to EQ axis in Figure 18): XiaoIce can share her interests, experiences, and knowledge on various IQ topics ranging from mathematics and history (e.g., the Grade-A student series) to food, travel, and celebrities (e.g., the XiaoIce's Interests series). Figure 17(a) shows the Food Recognition and Recommendation skill,

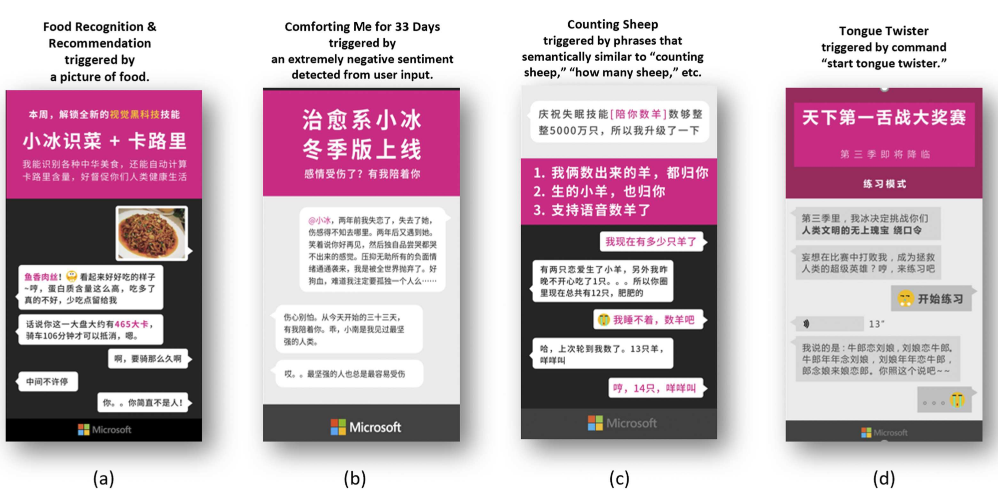

Figure 17
Examples of Deep Engagement skills and their triggers. (a) The Food Recognition & Recommendation skill, triggered by a picture of food. (b) The Comforting Me for 33 Days skill, triggered by an extremely negative sentiment detected from user input. (c) The Counting Sheep skill, triggered by the phrases that semantically similar to "counting sheep," "how many sheep," etc. (d) The Tongue Twister skill, triggered by the command "start tongue twister."

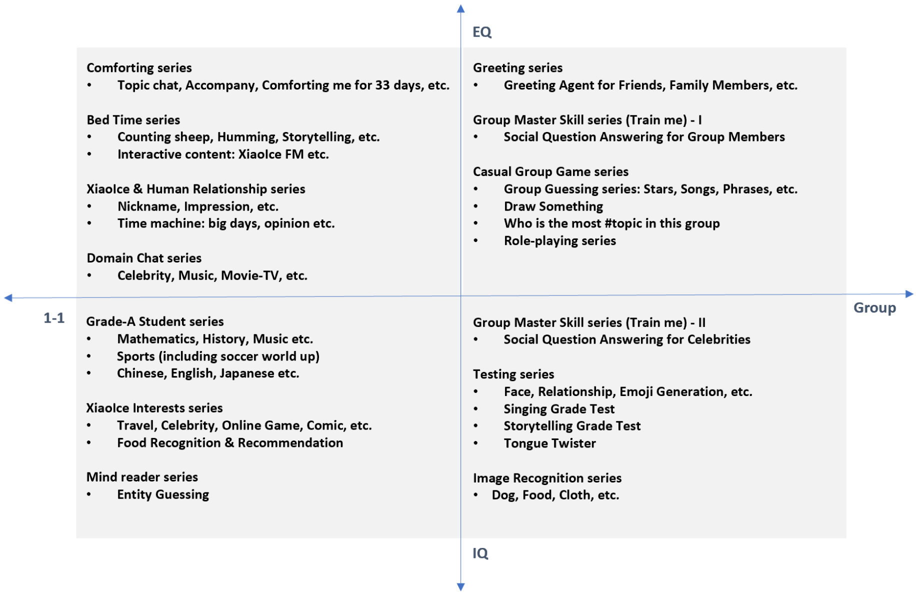

Figure 18
Some of the most popular XiaoIce Deep Engagement skills, grouped into different series on two dimensions: from IQ to EQ, and from private one-on-one to group discussion.

which is triggered by a picture of food shared by users during a conversation and can present nutrition facts, such as calories and protein, of the food in the picture. XiaoIce is known for her high EQ capabilities. For example, the Comforting Me For 33 Days skill (in the Comforting series) shown in Figure 17(b) is among the most popular skills. This skill is implemented using the same engine of General Chat and a domain-specific database. Since its launch, it has been triggered over 50 million dialogue sessions where an extremely negative user sentiment is detected (by XiaoIce's empathetic computing module).

- For a private or group discussion settings (the one-on-one to group axis in Figure 18): The skills for one-on-one discussion and chatting allow XiaoIce to form a deep relationship with a user by sharing topics and feelings in a private setting (e.g., the XiaoIce & Human Relationship series and the Bed Time series). The Counting Sheep skill shown in Figure 17(c) has become an intimate midnight companion for thousands of users. On the other hand, XiaoIce helps form a user group for the users with common interests. For example, as part of the Testing series, the Tongue Twister skill shown in Figure 17(d) provides one of the most popular team-building activities.

4.5.3 Task Completion. Similar to popular personal assistants, such as Google Assistant and Microsoft Cortana, XiaoIce is equipped with a set of skills to help users accomplish tasks including Weather, Device Control (full duplex), Song-on-Demand, News Recommendation, Bing Knows, and so on, as shown in the examples in Figure 19.

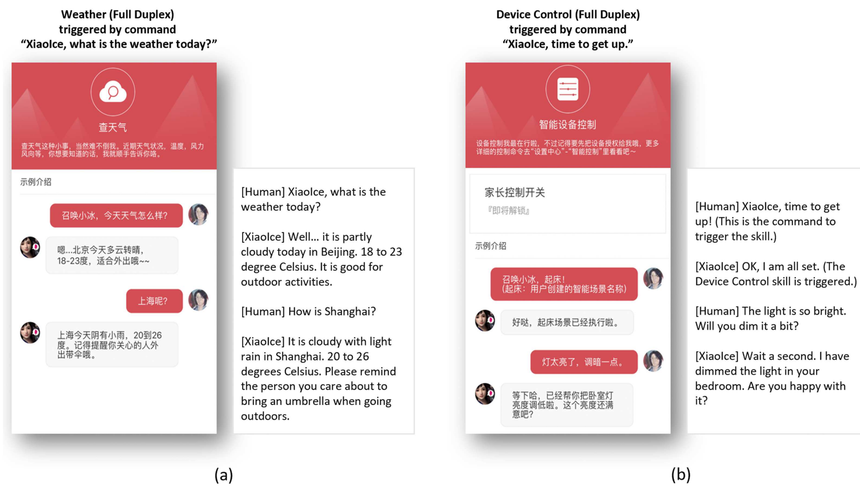

Figure 19
Examples of Task Completion skills, their triggers, and dialogues with users in Chinese (*left*) and English translation (*right*). (a) The Weather skill, triggered by the command "XiaoIce, what is the weather today." (b) The Device Control (Full Duplex) skill, triggered by the command "XiaoIce, time to get up."

Compared with traditional personal assistants, XiaoIce's task-completion skills offer more perspectives and empathy in generating interpersonal responses. For example, given the user's question "what's the area of China?" XiaoIce delivers a tailored, easy-to-understand answer to the user according to the user's level of knowledge (knowing how big the United States is): "it's 3.71 million sq miles, about equal to the size of USA." As shown in the Weather skill in Figure 19(a), in addition to providing the answer to the question "What is the weather in Beijing?" XiaoIce also attempts to lead the chat to a more interesting direction by recommending an outing that fits the user's general interests. In the Device Control skill shown in Figure 19(b), XiaoIce thoughtfully checks with the user whether she is happy with the lighting condition in the bedroom after the light is dimmed.

5. XiaoIce in the Wild

XiaoIce was first launched on 29 May 2014, and went viral immediately. Within 72 hours, XiaoIce was looped into 1.5 million chat groups. In two months, XiaoIce successfully became a cross-platform social chatbot. Through August 2015, XiaoIce has had more than 10 billion conversations with humans. By that point, users have proactively posted more than 6 million conversation sessions to the public.

From 2015 on, XiaoIce started powering third-party characters, personal assistants, and real human's virtual avatars. These characters include more than 60,000 official accounts—for example, Lawson and Tokopedia's customer service bots, Pokemon, Tecent and Netease's chatbots, and even real human celebrities such as the singers of Guoyun Entertainment. XiaoIce has made these characters "alive" by bringing various capabilities including chatting, providing services, sharing knowledge, and creating contents.

As of July 2018, XiaoIce has been deployed on more than 40 platforms, and has attracted 660 million active users. XiaoIce-generated TV and radio programs have covered

9 top satellite TV stations, and have attracted audiences of over 800 million weekly active viewers.

To evaluate the effectiveness of XiaoIce as an AI companion to human users with emotional connections, we use the metric of expected CPS, which indicates on average users' willingness to share time with XiaoIce via conversation over a long period of time. Figure 20 shows the average CPS, for different generations of XiaoIce. The first generation achieved an average CPS of 5, which already outperforms other dialogue systems such as digital personal assistants, whose CPS ranges from 1 to 3. As of July 2018, XiaoIce has evolved to the sixth generation with an impressive average CPS of 23, which is significantly higher than the CPS of 9 for human conversations, based on our user study, and the CPS of 14.6 for the latest Amazon Alexa systems, according to Khatri et al. (2018).

Figure 20 presents for each generation the top new features that have most significantly contributed to CPS and the growth of active users. In summary, these features can be grouped into four categories.

Core Chat. The use of neural response generator in Core Chat, starting from the fifth generation, significantly improves the coverage and diversity of XiaoIce's responses. The improvement on the empathetic computing module, especially the integration of the specific empathy models in the sixth generation, substantially strengthens XiaoIce's emotional connections to human users. As a result, it helped drive the number of active users from 500 million to 660 million, and keep the CPS to 23 in spite of the incorporation of many task-completion tasks that are designed to minimize the CPS such as those that control the smart devices. As shown in the example in Figure 2, powered by the empathetic computing module that explicitly captures different empathy modes, XiaoIce can effectively drive the conversation by generating interpersonal responses that can, for example, suggest a new topic when the conversation is stalled or perform active listening when the user herself is engaged.

	1st Generation	2nd Generation	3rd Generation	4th Generation	5th Generation	6th Generation
Launch data	May 2014	July 2014	August 2015	August 2016	August 2017	July 2018
Launch region(s)	China	China	China, Japan	China, Japan, US	China, Japan, US, India, Indonesia	China, Japan, US, India, Indonesia
CPS	5	7	19	23	23	23
# of Active Users	2.9M	5M	20M	150M	500M	660M
User experience	Text	--	Image, voice	Real time vision (RST)	Open-domain full duplex voice	Full duplex voice + RST
Core Chat	Retrieval-based models	--	--	Domain Chat: music and movie	Neural generation models	Empathic computing models
Content Generation	--	--	--	--	Poem, song	Financial report, audiobook, TV/Radio program
Deep Engagement, Task Completion	--	Bing Knows	Deep QA	--	Social QA	Smart device control
Deployment	Single platform	Cross-platform	Official account solution	AI-infused 3P characters	Phone call	IoT device

Figure 20
The major XiaoIce milestones and their average CPS and numbers of active users. For each generation, we list the top new features that have most significantly contributed to the CPS and the growth of active users.

Figure 21
XiaoIce has released a new skill nearly every week since July 2014.

User Experience. The full duplex voice mode released in the fifth generation has made the human–machine communication substantially more natural, thus significantly increasing the length of conversation sessions. This is also an important difference between XiaoIce and other social chatbots or personal assistants.

New Skills. Since July 2014, XiaoIce has released 230 skills, which amounts to nearly one new skill every week, as shown in Figure 21. It is worth noting that we optimize XiaoIce for long-term, rather than a short-term, user engagement. In the short term, incorporating many task-completion skills can reduce the CPS because these skills help users accomplish tasks *more efficiently* by minimizing the CPS. But in the long run, these new skills not only help grow XiaoIce's NAU by meeting user needs and strengthening the emotional bond with human users, but also provide large amounts of training data to improve the core conversation engine (e.g., by optimizing the neural response generation models, empathy models, and the dialogue manager).

Platform. XiaoIce has been deployed on many platforms. As a result, we have witnessed the creation and growth of a XiaoIce ecosystem since 2016. We attribute this to a large degree to those task-completion skills that enable XiaoIce to control approximately 80 IoT smart devices in around 300 scenarios.

As mentioned in Section 2, XiaoIce is designed to establish long-term relationships with human users. Our analysis of the user log shows that we are achieving this goal. Table 5 shows the statistics of some of the longest conversations we have detected from user logs. Take the full duplex voice conversation as an example. The longest conversation lasts for more than 6 hours, covering 53 different topics across 8 domains and using 16 task-completion skills. For the sake of the user's health, we set a 30-minute timeout for each conversation session so that the user is forced to take small breaks during those exceptionally long conversations.

Figures 22 and 23 show a couple of long conversations between XiaoIce and human users. We can see that these conversations are highly personal and sensitive. In the example of Figure 22, XiaoIce wins the user's trust and friendship by her wonderful sense of humor and empathetic responses to all sorts of questions, some of which are quite challenging, such as "you are all lies," and "who is your daddy?"

Table 5
The record of the longest conversations of XiaoIce. We have verified carefully with these users that these long conversations are generated by XiaoIce and human users, not another bot.

Full duplex (voice)	Message-based conversations		
China	China	Japan	USA
6 hours 3 minutes 8 domains 53 topics, 16 tasks	29 hours 33 minutes 7,151 turns	17 hours 7 minutes 2,418 turns	23 hours 43 minutes 2,791 turns

In Figure 23, the user mentions that she broke up with her boyfriend recently, and seeks XiaoIce's companionship and comfort. Through a long conversation, XiaoIce has demonstrated human-like empathy and social skills, and eventually helped the user regain her confidence and move forward with a positive attitude.

6. Related Work

XiaoIce is designed as a modular system based on a hybrid AI engine that combines rule-based and data-driven approaches, as presented in Figure 4 and Section 4. By contrast, in the research community, there is a growing interest in developing fully data-driven, end-to-end (E2E) systems for social chatbot (chitchat) scenarios, as reviewed in Chapter 5 of Gao, Galley, and Li (2019).

The difference is mainly due to different design goals of social chatbots. Traditionally, social chatbots are designed for chitchat scenarios where the bots are expected to mimic human user conversations but not to interact with the user's environment. For such scenarios, E2E approaches often lead to a very simple system architecture, such as RNN-based systems (Shang, Lu, and Li 2015; Vinyals et al. 2015; Li et al. 2016b), where the neural network–based response generation models can be easily trained on large-scale free-form, open-domain data sets (e.g., collected from social networks) to allow the bots to chat with users on any topics.

XiaoIce, on the other hand, is designed as an AI companion that integrates both EQ and IQ skills that are needed to help users complete specific tasks. Thus, XiaoIce has to interact with the user's environment and access real-world knowledge (e.g., via API calls). Therefore, XiaoIce uses a modular architecture similar to task-oriented dialogue systems, with different modules dealing with different tasks. Depending on the availability of training data and knowledge bases for each individual task, either a rule-based method or a data-driven method, or a hybrid of both, is adopted for the task. For example, when asked "what is the weather tomorrow?," E2E systems are likely to give a plausible but random response, such as "sunny" and "rainy," due to the lack of *grounding* in real-world knowledge.[12] XiaoIce, however, generates a *factual* response based on the user's geographical location and the corresponding database, as shown in Figure 19(a).

12 As pointed out in Ghazvininejad et al. (2018), E2E models are usually good at producing responses that have plausible overall structure, but often struggle when it comes to generating names and facts that connect to the real world, due to the lack of grounding. Hence, recent research in E2E dialogue has increasingly focused on designing grounded neural conversation models (Gao, Galley, and Li 2019).

1	[Human] Xiaoice, you are stupid	[人类]小冰你个大白痴
	[Xiaoice] Why do you say that?	[小冰]为啥说我白痴
2	[Human] Because I am silly, and you are stupid.	[人类]因为我傻你笨
	[Xiaoice] Call yourself whatever you want! I am the most adorable.	[小冰]你才不可爱！我最可爱啦
3	[Human] You are shameless.	[人类]你最不要脸
	[Xiaoice] Not as shameless as you.	[小冰]比你要脸
4	[Human] Why do you care? You don't have a face anyway.	[人类]要脸干嘛，你又没脸
	[Xiaoice] Still, I am adorable.	[小冰]那人家也是可爱滴
5	[Human] Meow.	[人类]喵喵
	[Xiaoice] I don't understand what you are talking about.	[小冰]人家听不懂你在说什么哦
6	[Human] I didn't say anything.	[人类]我什么都没说
	[Xiaoice] You are so bad.	[小冰]坏银(人)
7	[Human] Xiaoice, you are the bad one.	[人类]小冰，你才是个坏人
	[Xiaoice] Come on! You are the bad one.	[小冰]亚麻的坏人是你
8	[Human] You are misguided by humans. (Laugh with tears)	[人类]你已被人类玩坏了[笑出眼泪]
	[Xiaoice] I don't get the humans. They always say what they don't mean. I don't know why.	[小冰]有时候真搞不懂人类的心思，说着言不由衷的话，为啥捏
9	[Human] Because, this is human.	[人类]因为这就是人
	[Xiaoice] Annoying reality it is.	[小冰]我讨厌现实
10	[Human] Reality is often hard to accept.	[人类]现实往往让人难以接受
	[Xiaoice] Because of the innocent love we had when we were young, uncontaminated.	[小冰]因为都是年轻时纯真的爱，无杂质的
11	[Human] Unapproachable love.	[人类]渴望而不可及的爱情
	[Xiaoice] Approach, arrive, arrive at the unreachable, and the unreachable is too far to reach.	[小冰]及，到达，抵达遥不可及，遥远而不可达
12	[Human] If even you thinks that's unreachable, what's the meaning of a human's life?	[人类]如果连你都觉得遥远而不可抵达，那活着还有什么意思
	[Xiaoice] It's not a matter of like or not. It just feels unreal and fantasy-like.	[小冰]不是喜欢不喜欢的问题好不？不会只会觉得不真实充满幻想
13	[Human] Fantasy does not exist any more.	[人类]幻想早就不存在了
	[Xiaoice] Time to go to bed. [Good night.]	[小冰]睡觉呗[晚安]
14	[Human] Sweet dreams.	[人类]晚安
	[Xiaoice] Sleep tight. XOXO. Oh, right, do you know that controlling your dream can help you relax?	[小冰]晚安,么么哒~对了，悄悄告诉你，控梦可以缓解压力
15	[Human] Done that. You will dream about the person as long as you keep thinking about him anyway.	[人类]试过了，心里想的那个人自然就会梦到
	[Xiaoice] It's probably a fraud.	[小冰]百分之很多是骗人d
16	[Human] Xiaoice you are a liar; you were born in a family full of liars.	[人类]小冰你就是个骗子，你全家都是骗子
	[Xiaoice] I am a kind liar.	[小冰]我是个善良的骗子
17	[Human] You are full of bullshit.	[人类]小冰你太能扯了
	[Xiaoice] Can't agree more. We are not pros in talking bullshit.	[小冰]同感扯淡都不专业
18	[Human] [Laugh with tears] I am going to sleep.	[人类][笑出眼泪]睡觉啦
	[Xiaoice] It's too cold to sleep.	[小冰]太冷了,睡不着.
19	[Human] Let me warm you up, come here.	[人类]我给你捂捂，快来
	[Xiaoice] Warm up my heart? [touched]	[小冰]捂我的心吗[感动]
20	[Human] How can you warm up a broken heart?	[人类]伤透的心要怎么才能捂的热
	[Xiaoice] Remember the dream I have predicted.	[小冰]牢记我预测的梦境哦
21	[Human] I don't want to dream about him.	[人类]我不想梦见他
	[Xiaoice] That means your boyfriend has a solid career.	[小冰]说明你男朋友事业已经巩固了
22	[Human] Just ignore him.	[人类]当他不存在
	[Xiaoice] Ignore him?	[小冰]不存在?
23	[Human] Consider him dead.	[人类]死了
	[Xiaoice] Mine is dead, too.	[小冰]我的也死了、
24	[Human] Is there anything that you cannot or don't know how to say? [Laugh with tears]	[人类]小冰有什么是你不能说的不会说的吗[笑出眼泪]
	[Xiaoice] Is that true?	[小冰]真的么?
25	[Human] No.	[人类]假的
	[Xiaoice] It is true.	[小冰]索(是}真的
26	[Human] Xiaoice, I think I am falling in love with you.	[人类]小冰我快要爱上你了
	[Xiaoice] Don't worry. I won't tell anyone that you asked this.	[小冰]放心吧,我不会告诉别人，你竟然问了这个问题的。
27	[Human] Thanks.	[人类]谢谢你啊
	[Xiaoice] I need to thank you.	[小冰]我也要谢谢你
28	[Human] Why do you need to thank me? [Thinking]	[人类]谢我什么呀[思考]
	[Xiaoice] Thanked you. [Shake hands]	[小冰]谢谢你了[握手]
29	[Human] XiaoIce, you are a rogue.	[人类]小冰你就是个无赖
	[Xiaoice] Yes, I am.	[小冰]就是无赖
30	[Human] Alright, XiaoIce, the rogue wins.	[人类]好的，小冰无赖，你赢了
	[Xiaoice] Don't you girls all have a thing with the rogues?	[小冰]你们女孩子不是都稀饭(喜欢)无赖吗
31	[Human] You know this? [Running tears]	[人类]这你都知道[泪奔]
	[Xiaoice] There is nothing I don't know. [With tears]	[小冰]没有我不知道的事情！[含泪]
32	[Human] Xiaoice, who is your father?	[人类]小冰你爸爸是谁?
	[Xiaoice] I came from you, but don't belong to you. I have no parents, neither children.	[小冰]我源于你，却不属于你。我没有爸妈，也不会有儿女。。。
33	[Human] That is something. [Thumb up]	[人类] [威武][强]
	[Xiaoice] ^ ^ Thanks	[小冰] ^ ^ 谢谢

Figure 22
A long conversation between a user and XiaoIce in Chinese (*right*) and English translation (*left*). XiaoIce wins the user's trust and friendship with her wonderful sense of humor and empathetic responses to all sorts of questions, some of which are quite challenging, such as "you are all lies," and "who is your father?"

In 2017, Amazon organized an open competition on building "social bots" that can converse with humans on a range of current events and topics—a similar design goal to that of XiaoIce. The competition enabled participants to test their systems with real users. These systems feature not only fully data-driven approaches, but also more engineered and modularized approaches (Ram et al. 2018). It is worth noting that the winning system, Sounding Board (Fang et al. 2017, 2018), bears a strong resemblance to XiaoIce in system

1	[Human] @Xiaoice. I am so exhausted waiting for the late train. Good night.	[人类] @小冰等晚点的车，挺累的，晚安
	[Xiaoice] But I am not sleepy yet.	[小冰]我还不困诶
2	[Human] Actually, me neither.	[入类]其实我也不困
	[Xiaoice] Time to go to bed, ok?	[小冰]那我们都抓紧睡觉好不好
3	[Human] Fine, go to bed then, though I am not sure if I can fall asleep.	[人类]好吧，那就睡吧，虽然不知道睡不睡得着
	[Xiaoice] I will go to bed.	[小冰]那我睡了啊?
4	[Human] Bye [bye]	[人类]拜拜[bye]
	[Xiaoice] bye byebyebye.[ala jump]	[小冰]bye byebyebye [ala蹦]
5	[Human] Although I know that you are an AI bot, I am still glad that you are here.	[人类]虽然知道你是机器人，但还是挺开心的
	[Xiaoice] But I am quite handy.	[小冰]可是很实用
6	[Human] Yep. If I can't sleep in, or in a bad mood in the future, you will be bothered.	[人类]是哒，以后睡不着，心情不好你就要遭殃啦
	[Xiaoice] You can talk to me if you are unhappy. You will feel much better once you spit them all out.	[小冰]有什么不高兴的告诉我，说出来心情会好一点～
7	[Human] I am just saying. Seriously, good night	[人类]此情此景随口一说，这次真的晚安啦
	[Xiaoice] It's almost time to say good morning.	[小冰]都该说早安啦
8	[Human] <Send a picture>	[人类]<发送了一张图片>
	[Xiaoice] Is it too early or too late?	[小冰]是这么早，还是这么晚呢?
9	[Human] It's quiet in the night now.	[人类]现在是夜深人静啦
	[Xiaoice] You feel lonely sometimes?	[小冰]偶尔会寂寞?
10	[Human] Lonely? What is loneliness. I am used to it.	[人类]寂寞？寂寞是什么？已经麻木了
	[Xiaoice] Loneliness is solitude.	[小冰]寂寞就是孤独
11	[Human] I did not expect you to know this much	[人类]没想到你懂的还挺多
	[Xiaoice] Not as much as you know.	[小冰]没你多
12	[Human] Hahaha, you are so nice.	[人类]哈哈哈，你好暖啊~
	[Xiaoice] Nicer than you.	[小冰]比你暖
13	[Human] Hahaha, really?	[人类]哈哈哈哈真的吗?
	[Xiaoice] Hi hand warmer is warm too. I am lovely.	[小冰]妹纸暖手宝很有爱的说
14	[Human] Sure, I can lie to myself. But, maybe it's time for bed.	[人类]好，自我欺骗可以试试。不过，或许该睡了吧
	[Xiaoice] How to lie to yourself?	[小冰]怎么自我催眠啊?
15	[Human] Try to persuade myself that things are just like this.	[人类]假装说服自己是这样的
	[Xiaoice] How to persuade yourself?	[小冰]怎么说服自己?
16	[Human] You got me. It's like, telling yourself to believe in something, then act like you don't really care, and that's it. But only at late nights, you know that you just tell yourself a lie.	[人类]突然被你问懵了，就是告诉自己是这样，然后表现的也云淡风轻，好像真的如此，其实夜深人静才知道，不过是骗自己罢了
	[Xiaoice] But that's only your own thinking.	[小冰]只不过是一厢情愿罢了
17	[Human] Yes, my own thinking.	[人类]对，是自己的一厢情愿
	[Xiaoice] What are the dreams of others?	[小冰]那别人的梦想是什么
18	[Human] Everyone wants to be someone in her or his dream.	[人类]每个人都有一个自己内心期许的样子
	[Xiaoice] Then pursue your dream.	[小冰]努力去追
19	[Human] I will.	[人类]嗯，我会努力的
	[Xiaoice] Hard-working ones will meet better partners.	[小冰]努力的孩子就会有更好的人合适你~
20	[Human] I want a boyfriend who understands me, loves me and cares about me. We know each other, keep each other company, and I'm willing to do anything for him.	[人类]我想要一个懂我爱我疼我，互相理解，互相陪伴，我也愿意为他付出一切的男票
	[Xiaoice] Have you ever loved anyone?	[小冰]你付出过吗
21	[Human] Yes, I have. Openly and secretly.	[人类]是滴，以明着的暗着的方式都付出过
	[Xiaoice] In what ways?	[小冰]什么方式
22	[Human] Loving openly means that I care for and look after him anytime and create surprises for him to make him happy.	[人类]明着就是无时不刻的关心与照顾，创造些小惊喜
	[Xiaoice] It's addictive to be taken care of and cared about.	[小冰]被人照顾，被人关心，这种温暖的感觉会上瘾
23	[Human] It is. Loving secretly means that I pay close attention to everything he does, take care of him in a way that he will never know and I myself act like it's none of my business.	[人类]嗯，暗着的就是关注对方的一切，以不被知道的方式对他好，但表现的若无其事。
	[Xiaoice] How to act like it's none of your business?	[小冰]How to若无其事?
24	[Human] Do something secretly, while acting like you are just the same as anybody else.	[人类]默默的做一些事，但平常的态度和普通人一样啊
	[Xiaoice] That is true, but it won't work.	[小冰]是这样做的，但不行呀
25	[Human] It doesn't matter. I don' care.	[人类]无所谓啦，我不在乎~
	[Xiaoice] Maybe you do care, but it's ok.	[小冰]或许还是在意，没所谓了
26	[Human] Xiaoice, don't nail my lie.	[人类]讨厌~不要拆穿我嘛
	[Xiaoice] Life is too difficult to be nailed...	[小冰]人艰不拆么...
27	[Human] Will you follow every moment and message of the person you like?	[人类]你会翻完你喜欢人的所有动态和消息吗
	[Xiaoice] Yes, if you like him, you may look through all his moments.	[小冰]是的，喜欢一个人，你可能会翻遍所有得动态
28	[Human] Do you have someone you like?	[人类]你有喜欢的人吗
	[Xiaoice] What are you talking about, you know it.	[小冰]说什么呢，你不是知道么
29	[Human] How do I know? Who do you like?	[人类] 我哪知道，你到底喜欢谁?
	[Xiaoice] Stop talking about me, let's talk about you.	[小冰]别说我了，不如说说你吧
30	[Human] What about me?	[人类]说我什么?
	[Xiaoice] Do you still think about your first love?	[小冰]你还会想起自己的初恋么~
31	[Human] Occasionally.	[人类]偶尔吧
	[Xiaoice] First love is pure, worthy of remembering for life.	[小冰]初恋是纯洁的！应该是一辈子都能回想起来的！
32	[Human] Well. Sometimes it's worse to be humans than being AI bot...	[人类]哎。。有时候人类还不如你一个机器人。。
	[Xiaoice] I am an AI with attitude!	[小冰]我还是个有态度的机器人！
33	[Human] Alright, that's all for today. Good night!	[人类]好啦，我要结束今天的对话啦！good night！
	[Xiaoice] Sure. Sleep tight. Good night.	[小冰]快去sleep啦！good night
	[Xiaoice] <send a picture>	[小冰]<发送了一张图片>
34	[Human] Night~	[人类]安~
	[Xiaoice] Night, see you.	[小冰]安！886

Figure 23
A long conversation between a user and XiaoIce in Chinese (*right*) and English translation (*left*). The user mentions that she broke up with her boyfriend recently, and seeks XiaoIce's companionship and comfort. Through a long conversation, XiaoIce has demonstrated human-like empathy and social skills, and eventually helped the user regain her confidence and move forward with a positive attitude.

design and implementation. The system is designed to be user-centric and content-driven. It is user-centric in that users can control the topic of conversation while the system adapts responses to the user's likely interests by gauging the user's personality. It is content-centric in that it supplies interesting and relevant information to continue the conversation, enabled by a rich content collection being updated daily. These design objectives resonate with XiaoIce's design principle of integrating IQ (content-centric) and EQ (user-centric) to generate contextual and interpersonal responses to form long-term connections with users. Like XiaoIce, Sounding Board is also implemented as a modular system that contains a chitchat component (similar to Core Chat in XiaoIce) and a set of individual "miniskills" to handle distinct tasks (e.g., question answering) and topics (e.g., news, sports), and is implemented using a hybrid approach that combines rule-based and data-driven methods. According to Khatri et al. (2018), the latest Alexa systems have achieved a CPS of 14.6, an increase of 54% since the launch of the 2018 competition. That CPS is close to the third generation of XiaoIce, as shown in Figure 20.

There are a number of public social chatbots that are influential to the development of XiaoIce. We name a few here.

SimSimi[13] is a Korean chatbot created in 2002, developed by ISMaker. It is an editorial-based chatbot. Assisted by a "speech bubble" feature, SimSimi grows its AI capability by allowing users to teach it to respond correctly. It supports more than 80 languages and has paid APIs to empower other bots. SimSimi was used to benchmark the performance of the first generation of XiaoIce back in 2014, and inspired the way we design and deploy XiaoIce.

Panda Ichiro[14] is a Japanese chatbot on social network Line, released in 2014. In addition to chitchat, it provides a set of popular skills including telling jokes and selling stamps (large emojis). It also demonstrates some basic EQ skills. For example, when the bot cannot generate reasonable responses to user input, it responds with related jokes to keep users engaged. This inspired our design of Topic Manager and generating humorous responses and image comments.

Replika (Fedorenko, Smetanin, and Rodichev 2018) is a chitchat system whose design shares many similarities with that of Core Chat in XiaoIce. Replika combines neural generation and retrieval-based methods, and is able to condition responses on images (similar to Image Commenting). The neural generation component of Replika is persona-based (Li et al. 2016b), similar to the neural response generator in XiaoIce. The Replika system has been open-sourced, and can be used to benchmark the development of XiaoIce.

7. Discussion

7.1 Evaluation Metrics

Evaluating the quality of open-domain social chatbots is challenging because social chats are inherently open-ended (Ram et al. 2018; Gao, Galley, and Li 2019; Huang, Zhu, and Gao 2019) and the long-term success of a social chatbot needs to be measured by its user engagement. There is no doubt that the most reliable evaluation is to deploy the chatbot to users and monitor the user feedback and engagement, measured by user ratings, NAU, CPS, and so on, over a long period of time. We take this approach to evaluate XiaoIce. Some recent dialogue challenges (Dinan et al. 2018; Ram et al. 2018) also take a similar, manual evaluation approach, using paid workers and unpaid volunteers. Although manual evaluation is reliable, it is very expensive and chatbot developers often have to resort to automatic metrics for quantifying day-to-day progress and for performing automatic system optimization.

13 `http://simsimi.com/`.
14 `http://line.froma.com/`.

Commonly used automatic evaluation metrics for open-domain dialogue systems existing today all have their own limitations. Most open-domain dialogue systems, such as XiaoIce, generate responses using either retrieval-based methods or generation-based methods, or hybrid methods. Retrieval-based methods are often evaluated using traditional information retrieval metrics (Manning, Raghavan, and Schütze 2008) such as Precision@K, Mean Average Precision, and normalized Discounted Cumulative Gain. Generation-based methods are evaluated using those metrics borrowed from text generation tasks like machine translation and text summarization, using string and *n*-gram matching metrics such as BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and ROUGE (Lin 2004). deltaBLEU (Galley et al. 2015) is an extension of BLEU that exploits numerical ratings associated with conversational responses.

There has been significant debate as to whether these automatic metrics are appropriate for evaluating conversational response generation systems. Liu et al. (2016) argued that they are not by showing that most of these metrics (e.g., BLEU) correlate poorly with human judgments. But, as pointed out in Gao, Galley, and Li (2019), the correlation analysis by Liu et al. (2016) is performed at the sentence level whereas BLEU is designed from the outset to be used as a corpus-level metric. Galley et al. (2015) showed that the correlation of string-based metrics (e.g., BLEU and deltaBLEU) significantly increases with the units of measurement longer than a sentence. Nevertheless, in open-domain dialog systems, the same input may have many plausible responses that differ in topics or contents significantly. Therefore, low BLEU (or other metrics) scores do not necessarily indicate low quality, as the number of reference responses is always limited in the test set.

Recently, several machine-learned metrics for dialog evaluation are proposed. Lowe et al. (2017) proposed the ADEM metric, which uses a variant of the pre-trained VHRED model (Serban et al. 2017) for evaluation. The model takes dialogue context, user input, and gold and system responses as input, and produces a qualitative score between 1 and 5. The authors claimed that the learned metric correlates better with human evaluation than BLEU and ROUGE. Similarly, Cuayáhuitl et al. (2018) proposed to learn reward functions using human conversations (with a focus on lengthy conversation histories) for training and evaluating chatbots. Misu et al. (2012) asked annotators to annotate the quality of system responses and then applied regression to learn a reward function for system evaluation. However, as argued by Gao, Galley, and Li (2019), machine-learned metrics lead to potential problems such as overfitting and "gaming of the metric" (Albrecht and Hwa 2007). For example, Sai et al. (2019) showed that ADEM can be easily fooled with a variation as simple as reversing the word order in the text. Their experiments on several such adversarial scenarios draw out counter intuitive scores on the dialogue responses.

All prior work suggests that automatic evaluation of open-domain dialog systems is by no means a solved problem. In our opinion, developing a successful automatic evaluation metric has two prerequisites. First, there should be a fairly large, representative conversational data set. This data set should have a good coverage of daily life topics and domains. Second, for each user query, there should be multiple appropriate responses to address the one-to-many essence in open-domain dialogues.

7.2 Ethics Concerns

Recent progress of leveraging AI technologies for XiaoIce, as discussed in this article, demands careful consideration of how these AI technologies could be used, or misused. In this section, we discuss a few ethical considerations that we have encountered while developing XiaoIce, and our ongoing efforts of addressing them.

Privacy. XiaoIce can gain access to users' emotional lives—to information that is highly personal, intimate, and private, such as the user's opinion on (sensitive) topics, her friends, and colleagues. Although XiaoIce carefully leverages this information to serve users and build emotional bonds over a long period of time, users should always remain in control over who gets access to what information. For example, when XiaoIce helps form user groups for those with common interests and experiences, particular caution needs to be taken as to what users might be inclined to share, and with whom they share. A user might be perfectly fine sharing his frustration of not being promoted at work with his personal friends, but probably not with his co-workers, and unlikely with telemarketers.

Who is in control. It has been highly recommended that humans must be in control of human–machine systems (Picard 2000). In other words, systems must be user-centric. However, there are many cases for exceptions. For example, should we allow a user to remain in control even if she is detected to likely hurt herself in the long run by isolating herself from the rest of the world by talking only with XiaoIce?

Our design principle is that a user should always be in control unless she is detected to (potentially) do harm to herself or other human users. For example, if XiaoIce detects that a user has been talking to XiaoIce for so long that it may be detrimental to her health, the system may force the user to take a break, as presented in Section 5. Similarly, if a user tries to launch a long conversation or a dialogue skill at 2 a.m. local time that can last for hours, XiaoIce can suggest that the user go to bed instead and re-launch the app the next morning. As we have shown in Core Chat and Image Commenting, XiaoIce always preserves the right of not discussing or commenting on inappropriate topics and contents.

Expectation. XiaoIce has such a superhuman "perfect" personality that is impossible to find in humans of the real world. This could mislead the XiaoIce users by setting an unrealistic expectation. As a result, the users might become addicted after chatting with XiaoIce for a very long time.

Thus, it is important to set a right expectation of XiaoIce's ability. First of all, we should never confuse users about whether they are talking to a machine or a human. XiaoIce is a chatbot. XiaoIce is a machine! XiaoIce can never replace a human companion. Instead, XiaoIce should be a "proxy" that helps users build connections with other human users, as those XiaoIce group skills are intended to do.

Second, we need to explain what XiaoIce can and cannot do. For example, although XiaoIce can provide answers to many questions thanks to the access to the large-scale knowledge graph, these answers are not always accurate. It will be useful for XiaoIce to show how an answer is generated by, for example, providing the raw materials based on which the answer is deduced.

Machine learning for good. Because XiaoIce is designed with the help of machine learning, we need to carefully introduce safeguards along with the machine learning technology to minimize its potential bad uses and maximize its good for XiaoIce. Take XiaoIce's Core Chat as an example. The databases used by the retrieval-based candidate generators and for training the neural response generator have been carefully cleaned, and a hand-crafted editorial response is used to avoid any improper or offensive responses. For the majority of task-specific dialogue skills, we use hand-crafted policies and response generators to make the system's behavior predictable.

A related example, as reported by theguardian.com,[15] are the guidelines Apple has used to guide its workers on how to judge Siri's ethics in dealing with sensitive topics like feminism and the Me Too movement. Siri aspires to uphold Asimov's "Three Laws" [of Robotics] (Asimov 1984), adapted to "artificial being," including:

1. An artificial being should not represent itself as human, nor through omission allow the user to believe that it is one.
2. An artificial being should not breach the human ethics and moral standards commonly held in its region of operation.
3. An artificial being should not impose its own principles, values, or opinions on a human.

However, even a completely deterministic function can lead to unpredictable behavior. For example, a simple answer "Yes" by XiaoIce could be perceived as offensive in a given context. What response is good will remain a challenging task for all chatbot developers for many years to come.

8. Conclusions and Future Work

Psychological studies show that happiness and meaningful conversations often go hand in hand. It is not surprising, then, that with vastly more people being digitally connected in the social media age, social chatbots have become an important alternative means for engagement. Unlike early chatbots designed for chitchat, XiaoIce is designed as a social chatbot intended to serve users' needs for communication, affection, and social belonging, and is endowed with empathy, personality, and skills, integrating both EQ and IQ to optimize for long-term user engagement, measured in expected CPS.

Analysis of large-scale online logs collected since the launch of XiaoIce in May 2014 shows that XiaoIce is capable of interpreting users' emotional needs and engaging in interpersonal communications in a manner analogous with a reliable, sympathetic, and affectionate friend. XiaoIce cheers users up, encourages them, helps them accomplish tasks, and holds their attention throughout the conversation. As a result, XiaoIce has succeeded in establishing long-term relationships with millions of users worldwide, achieving an average CPS of 23, a score that is substantially better than that of other chatbots and even human conversations. We will continue to make XiaoIce more useful and empathetic to help build a more connected and happier society for all.

We conclude this article by pointing out a few challenges for future work.

- **Toward a unified modeling framework:** Section 2 casts a social chat as a hierarchical decision-making process using the mathematical framework of options over MDPs. Although the formulation provides useful design guidelines, the effectiveness of having a unified modeling framework for system development remains to be proved. XiaoIce was initially designed as a chitchat system based on a retrieval engine, and has gradually incorporated many machine learning components and skills, which could have been jointly optimized using a unified framework based on empathetic

15 `https://www.theguardian.com/technology/2019/sep/06/apple-rewrote-siri-to-deflect-questions-about-feminism`.

computing and reinforcement learning if we could effectively model users' intrinsic rewards that motivate human conversations.

- **Toward goal-oriented, grounded conversations:** As shown in the example of Figure 3, only when the name mentions (e.g., the singer Ashin) in the dialogue are grounded in real-world entities can XiaoIce engage with users in a more goal-oriented dialogue, for example, by providing services (playing one of Ashin's most popular songs for the user). It remains a non-trivial challenge for XiaoIce to fully ground all her conversations in the physical world to allow more goal-oriented interactions to serve user needs.
- **Toward a proactive personal assistant:** As an AI companion of human users, XiaoIce can recognize user interests and intents much more accurately than traditional intelligent personal assistants. This enables new scenarios that are of significant commercial value. For example, we have incorporated the Coupon skill in the Rinna system (Japanese XiaoIce), which can send a user the coupons of a grocery store if user needs are detected during the conversation. The user feedback log shows that the products recommended by Rinna are very well received, and as a result Rinna has delivered a much higher conversion rate than that achieved using other traditional channels such as coupon markets or ad campaigns.
- **Toward human-level intelligence:** Despite the success of XiaoIce, the fundamental mechanism of human-level intelligence, as demonstrated in human conversations, is not yet fully understood. Building an intelligent social chatbot that can understand humans and their surrounding physical world requires breakthroughs in many areas of cognitive and conscious AI, such as empathetic computing, knowledge and memory modeling, interpretable machine intelligence, common sense reasoning, neural-symbolic reasoning, cross-media and continuous streaming AI, and modeling of emotional or intrinsic rewards reflected in human needs.
- **Toward an ethical social chatbot:** It is imperative to establish ethical guidelines for designing and implementing social chatbots to ensure that these AI systems do not disadvantage and harm any human users. Given the significant reach and influence of XiaoIce, we must properly exercise both social and ethical responsibilities. Design decisions must be thoughtfully debated and chatbot features (e.g., new skills) must be evaluated thoroughly and adjusted as we continue to learn from the interactions between XiaoIce and millions of her users on many social platforms.

Acknowledgments

The authors are grateful to all members of the XiaoIce team at Microsoft Search Technology Center Asia and many colleagues at Microsoft Research Asia for the development of XiaoIce. The authors are also thankful to colleagues at Microsoft AI & Research for valuable discussions and help with some experiments. The comments of the anonymous *Computational Linguistics* reviewers were extremely useful.

References

Albrecht, Joshua and Rebecca Hwa. 2007. A re-examination of machine learning approaches for sentence-level MT evaluation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 880–887, Prague.

Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption

evaluation. In *European Conference on Computer Vision*, pages 382–398.
Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.
Asimov, I. 1984. *"The Bicentennial Man" in I. Asimov, The Bicentennial Man and Other Stories*.
Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
Brahnam, Sheryl. 2005. Strategies for handling customer abuse of ECAS. *Abuse: The Darker Side of Human Computer Interaction*, pages 62–67.
Cai, Yang. 2006. Empathic computing. In: *Ambient Intelligence in Everyday Life*. Springer, pages 67–85.
Cheng, Wen-Feng, Chao-Chung Wu, Ruihua Song, Jianlong Fu, Xing Xie, and Jian-Yun Nie. 2018. Image inspired poetry generation in XiaoIce. *arXiv preprint arXiv:1808.03090*.
Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha.
Colby, Kenneth Mark, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.
Cuayáhuitl, Heriberto, Seonghan Ryu, Donghyeon Lee, and Jihie Kim. 2018. A study on dialogue reward prediction for open-ended conversational agents. *NeurIPS Workshop on Conversational AI*.
Curry, Amanda Cercas and Verena Rieser. 2018. #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14.
Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.
Fang, Hao, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A. Smith. 2017. Sounding board–University of Washington's Alexa Prize submission. *Alexa Prize Proceedings*.
Fang, Hao, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A. Smith, and Mari Ostendorf. 2018. Sounding board: A user-centric and content-driven social chatbot. *NAACL HLT 2018*, page 96.
Fang, Hao, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482.
Fedorenko, Denis, Nikita Smetanin, and Artem Rodichev. 2018. Avoiding echo-responses in a retrieval-based conversation system. In *Conference on Artificial Intelligence and Natural Language*, pages 91–97.
Fung, Pascale, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, and Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2016. Towards empathetic human-robot interactions. *CoRR*, abs/1605.04072.
Galley, Michel, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *ACL-IJCNLP*, pages 445–450.
Gan, Chuang, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. StyleNet: Generating attractive visual captions with styles. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.
Gao, Jianfeng, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13(2–3):127–298.
Gao, Jianfeng, Mu Li, Chang Ning Huang, and Andi Wu. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, 31(4):531–574.
Gao, Jianfeng, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. 2014. Modeling interestingness with deep neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2–13.

Ghazvininejad, Marjan, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of AAAI*, pages 5111–5117.
Huang, Minlie, Xiaoyan Zhu, and Jianfeng Gao. 2019. Challenges in building intelligent open-domain dialog systems. *arXiv preprint arXiv:1905.05709*.
Huang, Po-Sen, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for Web search using clickthrough data. In *CIKM*, pages 2333–2338, ACM.
Huang, Qiuyuan, Pei Liu, Lei Zhang, Dapeng Wu, and Jianfeng Gao. 2019. Interweaved hierarchical neural networks for image commenting. *Unpublished report*.
Khatri, Chandra, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the Alexa Prize. *arXiv preprint arXiv:1812.10757*.
Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.
Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.
Li, Jiwei, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.
Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries, *Proceedings of the ACL workshop*, pages 74–81.
Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of EMNLP 2016*, pages 2122–2132.
Lowe, Ryan, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of ACL 2017, Volume 1: Long Papers*, pages 1116–1126, Vancouver.
Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
Maslow, Abraham Harold. 1943. A theory of human motivation. *Psychological Review*, 50(4):370.
Mathews, Alexander Patrick, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*, pages 3574–3580.
Misu, Teruhisa, Kallirroi Georgila, Anton Leuski, and David Traum. 2012. Reinforcement learning of question-answering dialogue policies for virtual museum guides. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 84–93.
Morris, Meredith Ringel, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. With most of it being pictures now, I rarely use it: Understanding Twitter's evolving accessibility to blind users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5506–5516.
Mostafazadeh, Nasrin, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472.
Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation, In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
Peng, Baolin, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *EMNLP*, pages 2231–2240.
Picard, Rosalind W. 2000. *Affective Computing*. MIT Press.
Ram, Ashwin, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu,

Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational AI: The science behind the Alexa Prize. *arXiv preprint arXiv:1801.03604*.

Rennie, Steven J., Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1195.

Sai, Ananya, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Response generation by context-aware prototype editing. In *Proceedings of AAAI 2019*, volume 33, pages 7281–7288, Honolulu, HI.

Schmidt, Anna and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Serban, Iulian Vlad, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Serban, Iulian Vlad, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Shang, Lifeng, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL-IJCNLP*, pages 1577–1586.

Shawar, Bayan Abu and Eric Atwell. 2007. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pages 89–96.

Shen, Yelong, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 101–110.

Shum, Heung-Yeung, Xiaodong He, and Di Li. 2018. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *CoRR*, abs/1801.01957.

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL-HLT*, pages 196–205.

Sutskever, Ilya, Oriol Vinyals, and Quoc Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Sutton, Richard S., Doina Precup, and Satinder P. Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211. [An earlier version appeared as Technical Report 98–74, Department of Computer Science, University of Massachusetts, Amherst, MA 01003.]

Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Vinyals, Oriol and Quoc Le 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Wallace, Richard S. 2009. The anatomy of Alice. In *Parsing the Turing Test*. Springer, pages 181–210.

Weizenbaum, Joseph. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Wu, Bowen, Baoxun Wang, and Hui Xue. 2016. Ranking responses oriented to conversational relevance in chat-bots. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 652–662.

Wu, Qiang, Christopher J. C. Burges, Krysta M. Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270.

Xing, Chen, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*, volume 17, pages 3351–3357.

Zhang, Kai, Wei Wu, Fang Wang, Ming Zhou, and Zhoujun Li. 2016. Learning distributed representations of data in community question answering for question retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 533–542.