The Design and Implementation of XiaoIce, an Empathetic Social Chatbot

Li Zhou
Microsoft
Beijing, China
lzhou@microsoft.com

Jianfeng Gao
Microsoft Research
Redmond, WA, USA
jfgao@microsoft.com

Di Li
Microsoft
Beijing, China
lidi@microsoft.com

Heung-Yeung Shum
Microsoft
Redmond, WA, USA
hshum@microsoft.com

Abstract

This paper describes the development of the Microsoft **XiaoIce** system, the most popular social chatbot in the world. XiaoIce is uniquely designed as an AI companion with an emotional connection to satisfy the human need for communication, affection, and social belonging. We take into account both intelligent quotient (IQ) and emotional quotient (EQ) in system design, cast human-machine social chat as decision-making over Markov Decision Processes (MDPs), 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 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 [1], Parry [2], and Alice [3], were 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 remains 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 [4, 5].

In this paper we present the design and implementation of the Microsoft **XiaoIce** ('Little Ice' literally in Chinese) system, the most popular social chatbot in the world. Since its release in China in May, 2014, XiaoIce has attracted over 660 million users. XiaoIce has already been shipped in five countries (China, Japan, USA, 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 USA and India, and LINE in Japan and Indonesia.

Preprint. Work in progress.

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), she 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 her hobbies and interests (a Japanese manga). By 4 weeks (Session 20), she began to treat XiaoIce as a friend and asked her questions related to her 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 her first choice whenever she needed someone to talk to.

XiaoIce is developed on an *empathetic computing* framework [6, 7] 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 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 reminder of the paper, we present the details of the design and implementation of XiaoIce. We start from its design principles 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. We will share how XiaoIce has been doing in five countries since its launch in May 2014 and conclude this paper with some discussions of future directions.

2 Design Principle

Social chatbots require a sufficiently high IQ to acquire a range of skills to keep up with the users and help them to complete specific tasks. More importantly, social chatbots also require a sufficient EQ to meet users' emotional needs, such as emotional affection and social belonging, which are among the fundamental needs for human beings [8]. 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 last 5 years XiaoIce has developed 230 different 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 long and open-domain conversations with users on many topics.

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, i.e., the ability to place oneself in the other person's position. A social chatbot with empathy needs to have the ability to identify user's emotions from the conversation, detect how emotions evolve over time, and understand 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 a sufficient EQ as it generates socially attractive responses (e.g., 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

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), she explored the features and functions of Xiaoice in conversation. Then, in 2 weeks (Session 6), the user began to talk with XiaoIce about her hobbies and interests (a Japanese manga). By 4 weeks (Session 20), she began to treat XiaoIce as a friend and asked her questions related to her 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 her first choice whenever she needed someone to talk to .

to set the right expectations for users in the conversation and gain their long-term confidence and trust. The XiaoIce persona is designed as a 18-year-old girl who is always reliable, sympathetic, affectionate, and has a wonderful sense of humor. Despite being extremely knowledgeable (due to access to large volumes of data), 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.

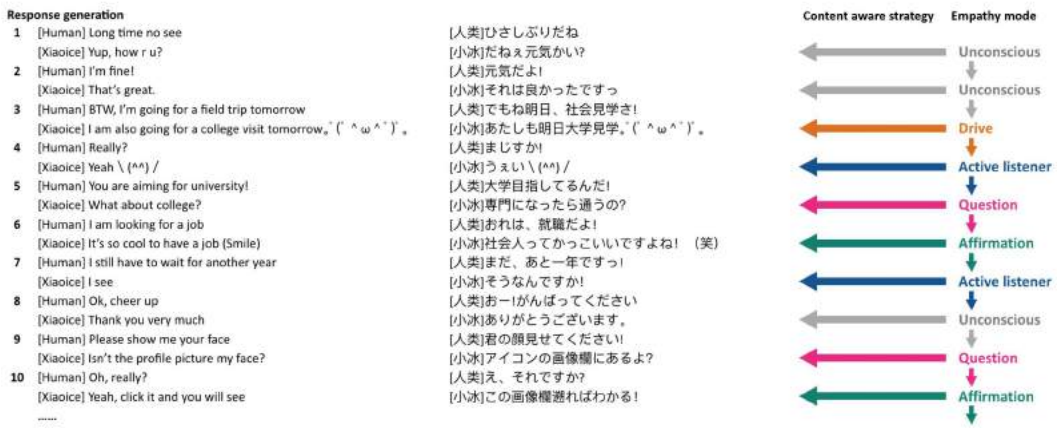

Figure 2: Conversation between a user and the XiaoIce chitchat system in Japanese (*midden*) and English translation (*left*). The empathy model provides a context-aware strategy that can drive the conversation when needed (*right*). For example, XiaoIce determines to 'drive' the conversation to a new 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.

2.2 Social Chatbot Metric: CPS

Unlike task-oriented bots where their performance is measured by task success rate, measuring the performance of social chatbots is difficult [9]. In the past, 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. Rather, we define Conversation-turns Per Session (CPS) as the success metric for social chatbots. It is the average number of conversation-turns between the chatbot and the user in a conversational session. The larger the CPS is, the better engaged the social chatbot is.

It is worth noting that we optimize XiaoIce for *expected* CPS which corresponds a long-term, rather than a short-term, engagement. For example, incorporating many task-completion skills can reduce CPS in the short term since these skills help users accomplish tasks *more efficiently* by minimizing 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.

2.3 Social Chat as Hierarchical Decision-Making

To fulfill these design objectives, we mathematically cast human-machine social chat as a 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 (QA), 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) [10], where options generalize primitive actions to higher-level actions. A social chatbot navigates in a MDP, interacting with its environment (human users) over a sequence of discrete dialogue turns. At each turn, the chatbot observes the current dialogue state, and chooses a skill (option) or a response (primary action) according to a hierarchical dialogue

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.

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 expected CPS (rewards).

The formulation guides the design and implementation of XiaoIce. XiaoIce uses a dialogue manager to keep track of dialogue state, and at each dialogue turn, select 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 user base, 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 (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 should lead to more engagement in the future. In addition, we adopt an intergenerational upgrade method that allows the graduated emergence of a full-fledged AI system that combines IQ and EQ through comprehensive application of machine learning algorithms and big data. These algorithmic features will be detailed in the following sections.

3 System Architecture

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

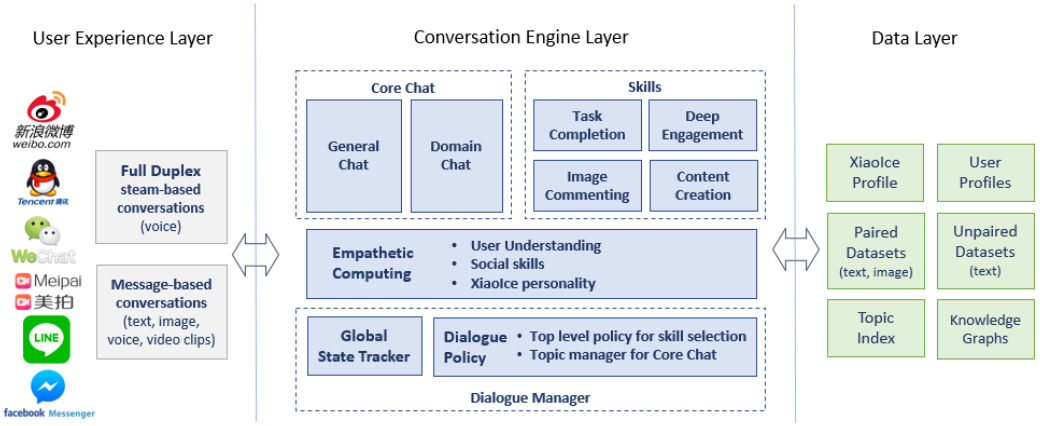

Figure 4: XiaoIce system architecture.

- **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. The other mode deals with message-based conversations where a user and XiaoIce take turns to talk. This layer also includes a set of components used to process user inputs and XiaoIce responses e.g., speech recognition and synthesis, image understanding and text normalization.

- **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, 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 for Core Chat and skills, and the profiles of XiaoIce and all the registered 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.

4.1 Dialogue Manager

Dialogue Manager is the central controller of the dialogue system. It consists of Global State Tracker that is responsible for keeping track of the current dialogue state $\mathbf{s}$, and Dialogue Policy π that 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 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, to manages its conversation segment.

The dialogue policy is designed to optimize the long-term user engagement through an iterative, trial-and-error process based on the feedback of XiaoIce's users. It 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 General Chat to a specific Domain Chat if a user interest 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 18 (a). If multiple skills are triggered simultaneously, we select the one to activate based on their triggering confidence scores, pre-defined priorities and 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 [11].

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 retrieval engine for recommending a new topic.

Topic switching is triggered if XiaoIce does not have sufficient knowledge about the topic to engage a meaningful conversation, or the user is getting bored. Typical indicators are as follows.

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

A topic dataset is constructed by collecting popular topics and related comments and discussions from high-quality Internet forums, such as Instagram in US and douban.com in China. When topic switching is triggered, a list of topic candidates is retrieved from the dataset using the current dialogue state as query. The new topic is chosen by a machine-learned boosted tree ranker based on 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.

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 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 – a 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 empathy 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, e.g., 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 e.g., by Core Chat to generate responses via either a retrieval-based engine or a neural generative model, 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 user intents, emotions, topics, opinions and user persona, as shown in the example of Figure 5 (b). These key-value pairs are generated using a set of machine learned 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.
- Intent detection labels Q_c using one of the dialogue acts e.g., greet, request, inform, etc.
- Sentiment analysis detects user's emotion, e.g., happy, sad, angry, neural, and how her emotion evolves during the conversation, e.g., from happy to sad.
- Opinion detection detects user's reaction to the topic, e.g., positive, negative or neural.
- If the user ID is available, include in $\mathbf{e}_Q$ the user persona vector according to her profile, e.g., 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

...	... ...
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 -> neural
Opinion	neural
...	...
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.

same topic (decided by Topic Manager), responding in a consistent and positive way as specified e.g., by the values of intent, sentiment and 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 next section.

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 Domain Chats. General Chat is responsible for engaging open-domain conversations that cover a wide range of topics. Domain Chats are responsible for engaging deeper conversations on specific domains such as music, movie and celebrity. Since General Chat and Domain Chats are implemented using the same engine with access to different datasets (i.e., general vs. domain-specific paired, unpaired datasets and neural response generation models), we only describe General Chat in detail below.

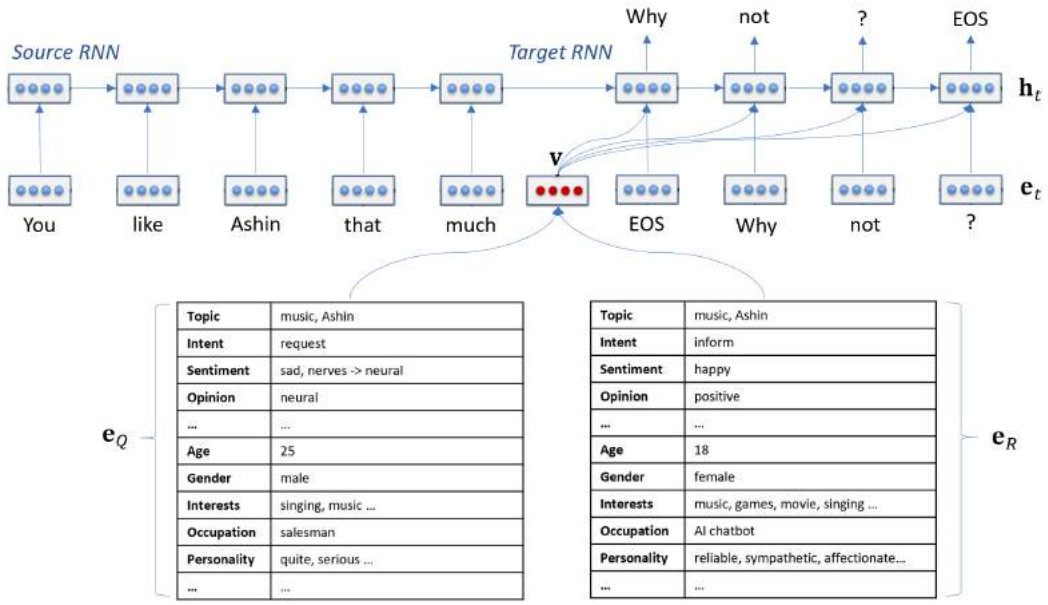

Figure 6: RNN-based neural response generator. Given the user query "You like Ashin that much", the response candidate "why not?" is generated.

General Chat is a data-drive 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 which consist of human-generated conversations or texts, or generated on the fly by a neural generative model. In what follows, we describe three candidate generators and a candidate ranker.

Retrieval-Based Generator using Paired Data The paired dataset consists of query-response pairs collected from two data sources. First is the human conversational data from the Internet, e.g., social networks, public forums, bulletin boards, news comments etc. After the public release 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 responses are retrieved from her own past conversations. To control the quality of the dataset, 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 webpage and website where the pair is extracted, and user profile (if the registered user 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 which contain personally identifiable information (PII), messy code, inappropriate content, spelling mistakes, etc.

The filtered paired dataset 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 learned representations of the paired dataset [12, 13].

Although the response candidates retrieved from the paired dataset 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 dataset. To increase the coverage, we introduce two other candidate generators described next.

Neural Response Generator Unlike the retrieval-based generator, the neural response generator is *trained* using the paired dataset to learn to simulate human conversation, 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

[2]http://lucene.apache.org/

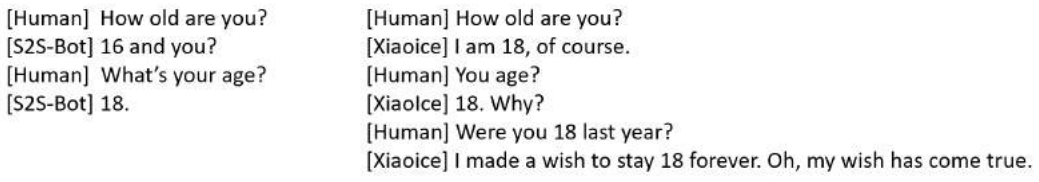

Figure 7: (*Left*) Examples of inconsistent responses generated using a seq2seq model which is not grounded in persona [20]. (*Right*) Examples of consistent and humorous responses generated using the neural response generator of XiaoIce.

complementary: the neural-model-based generator offers robustness and high coverage, while the retrieval-based provides high-quality responses for popular topics. Neural response generation is a very active research topic in the conversational AI community [4]. 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 [14, 15] used for conversation response generation [16, 17, 18, 19, 20, 21].

The generator is based on a GRU-RNN model, similar to the Speaker-Addressee model [20]. 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 towards the user,

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

where $\mathbf{W}_Q, \mathbf{W}_R \in \mathbb{R}^{k \times d}$ and σ denotes the sigmoid function. Then the source RNN encodes user query Q_c into a sequence of hidden state vectors which 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. See Appendix A for a detailed description of the model. As shown in Figure 7, while a typical seq2seq model which is not grounded in any persona often outputs inconsistent responses [20], XiaoIce is able to generate consistent and humorous responses.

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

The unpaired dataset 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. Since 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 dataset, the unpaired dataset is indexed using Lucene. Unlike the paired dataset, 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 dataset and Microsoft Satori [3]. We include in the XiaoIce KG a Satori triple (h, r, t) only if the number of conversation pairs in the paired dataset where h occurs in Q_c and t in R, or vise verse,

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

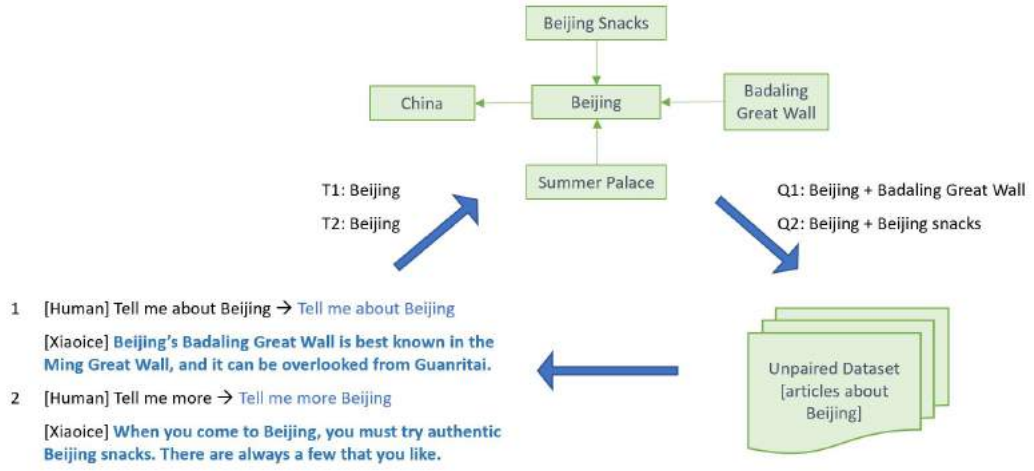

Figure 8: An example of generating response candidates using the unpaired dataset and the XiaoIce knowledge graph (KG), for which we show a fragment of the XiaoIce KG that is related to the topic "Beijing" (*top*). For a human-machine conversation (*bottom-left*), each user query is rewritten to a context 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 unpaired dataset (*bottom-right*) using a query that combines the query topics and their related topics.

is larger than a pre-set threshold. Such a triple contains a pair of two 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 dataset and the XiaoIce KG. It consists of three steps.

- First, we identify the topics from contextual user query Q_c, e.g., "Beijing" from "tell me about Beijing".
- For each topic, we retrieve up to 20 most related topics from the KG, e.g., "Badaling Great Wall" and "Beijing snacks". These topics are scored by their relevance using a boosted tree ranker [22] 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 dataset 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 dataset is lower than those retrieved from the paired dataset, with unpaired dataset XiaoIce can cover a much broader range of topics. Compared to the neural response generator which often generates well-form but short responses, the candidates from unpaired data are much longer with more useful content.

Response Candidate Ranker The response candidates generated by three generators are aggregated and ranked using a boosted tree ranker [22]. 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 the collections 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 the collections of human dialogue sessions. Since 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 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 $\mathbf{e}_R$ which encodes the empathy features of the *expected* response.
- Retrieval matching features. These features apply only to the candidates generated from the paired dataset. 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 3-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 makes users feel delightful 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 empathetic editorial responses 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".

4.4 Dialogue Skills

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

4.4.1 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 [26]. 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.

[4] DSSM stands for Deep Structured Semantic Models [23, 24], or more generally, Deep Semantic Similarity Model [25]. 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').

[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.

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 3-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.

[User 1] My son is ahead and surprised!
[User 2] Did he end up winning the race?
[User 1] Yes he won, he cannot believe it!

Figure 10: An example conversation around a shared image. Figure credit: [27]

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, position etc. It is the latter, the social skill aspects, that distinguishes 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 which 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 retrieved-based approach, first of all, a dataset of image-comment pairs, collected from social networks (e.g., Facebook and Instagram), is constructed. To control the data quality, a pipeline similar to that for Core Chat is applied to retain only the pairs whose text comments fit XiaoIce's persona [6]. Then, each image is encoded into a visual feature vector that represents the overall semantic information of the image, using deep convolutional neural networks (CNNs), as illustrated in Figure 12. At runtime, given a query image, we retrieve up to three most similar images, ranked

[6] 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 portion in the dataset.

(a) sunglasses, man
(b) a man wearing sunglasses taking a selfie.
(c) you look handsome in all shades.

(a) water, tree, river, boat
(b) a tree next to a body of water
(c) beautiful place looks like you are in the heaven.

Figure 11: Examples of (a) image tagging, (b) image description, and (3) image commenting. Figure credit: [5]

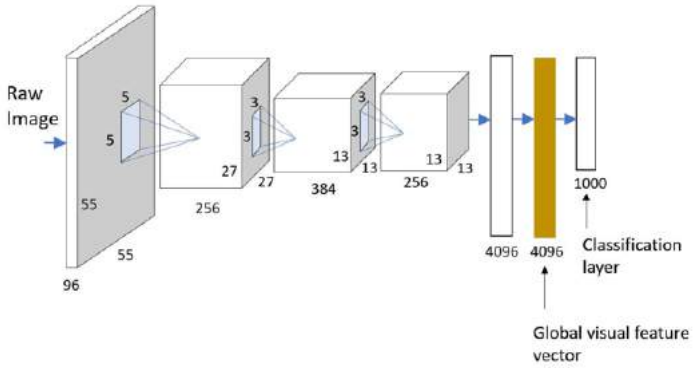

Figure 12: An example of deep convolutional neural network for visual feature vector extraction. Figure credit: [5]

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, which is an extension of the Microsoft Image Captioning system [28], which 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 [29, 30].

The comment candidates generated by the above generators are aggregated and ranked using a boosted tree ranker [22]. 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 (DMSM [28]), which is a variant of DSSM trained on large amounts 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 3-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 picture, instead of telling 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 repaying the fact there is a cat in the picture, XiaoIce makes a humorous comment about 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.

Figure 13: Examples of image-comment pairs that are used for training and validating Image Commenting. Each pair is labeled on a 3-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.

Figure 14: 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."

4.4.2 Content Creation

These skills allow XiaoIce to collaborate with human users in their creative activities including text-based Poetry Generation[7], voice-based Song and Audio Book Generation, XiaoIce FM for Somebody, and XiaoIce Kids Story Factory, etc.

Figure 14 (a) shows that a user uses XiaoIce to make an FM program for her mother for the coming Chinese Spring Festival. Figure 14 (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, etc.

The XiaoIce Poetry Generation skill has helped over four million users to generate poems. On May 15, 2018, XiaoIce published the first AI-created Chinese poem album in history[8]. XiaoIce's second

[7] https://poem.msxiaobing.com/
[8] https://item.jd.com/12076535

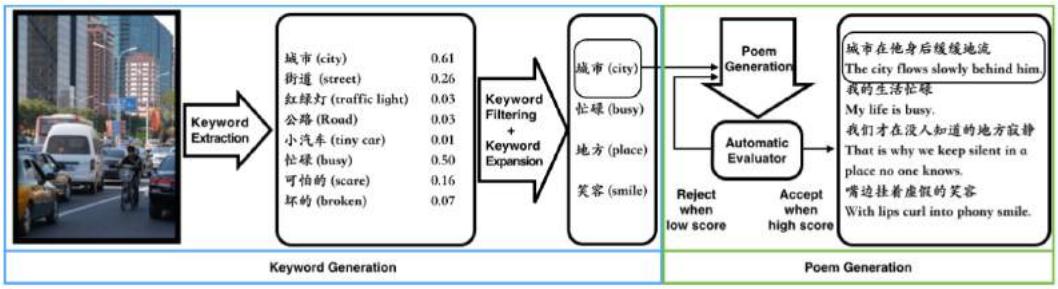

Figure 15: The framework of the Poem Creation skill. The system takes an image query given by a user, and outputs a semantically relevant piece of modern Chinese poetry. We first generate a set of keywords from the picture (*left*), and then generate a poem consisting of multiple lines, each generated using a keyword as a seed (*right*). Figure credit: [31].

Figure 16: 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".

poetry album is going to be published by China Youth Publishing and Microsoft around the end of 2018. Every poem in the album is jointly written by XiaoIce and human poets.

Figure 15 illustrates how a Chinese poem is generated from an image by XiaoIce. Given an 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 keywork as a seed. The generated sentences form a poem using a hierarchical RNN which models the structure among words and sentences.

4.4.3 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 16.

As shown in Figure 17, these skills can be grouped into different series on two dimensions: from IQ to EQ, and from private 1 on 1 to group discussion.

- To meet users' intellectual or emotional needs (the IQ to EQ axis in Figure 17): 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 celebrity (e.g., the XiaoIce's Interests series). Figure 16 (a) shows the Food Recognition and Recommendation

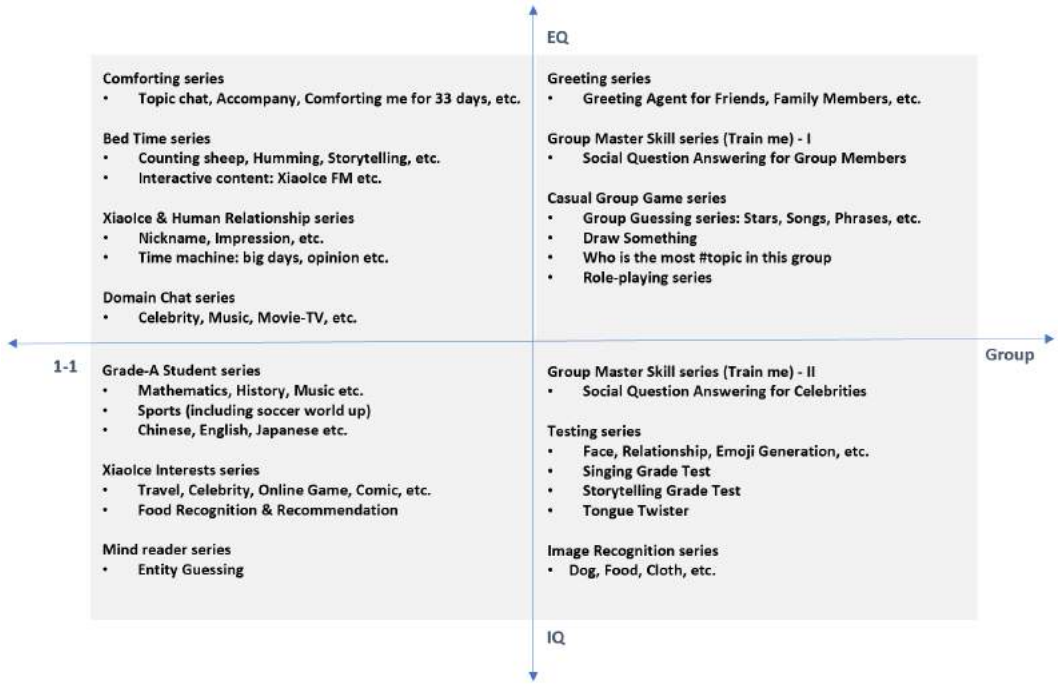

Figure 17: Some of the most popular XiaoIce Deep Engagement skills, grouped into different series on two dimensions: from IQ to EQ, and from private 1 on 1 to group discussion.

skill, 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 16 (b) is among the most popular skills. Since its release, 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 1-1 to group axis in Figure 17): The skills for 1 on 1 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 16 (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 16 (d) provides one of the most popular team building activities.

4.4.4 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 etc., as shown in the examples in Figure 18.

Compared with traditional personal assistants, XiaoIce's task-completion skills offer more perspectives and empathy in generating interpersonal responses. For example, given the question "what's the area of China?", XiaoIce delivers a tailored, easy-to-understand answer to the user according to her level of knowledge (knowing how big the USA is): "it's 3.71 million sq miles, about equal to the size of USA.". As shown in the Weather skill in Figure 18 (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 18 (b), XiaoIce thoughtfully checks with the user whether she is happy with the lighting condition in the bedroom after the lamp is turned on.

Figure 18: Examples of Task Completion skills and their triggers. (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."

5 XiaoIce in the Wild

XiaoIce was first released on May 29, 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. Up to August 2015, XiaoIce has had more than 10 billion conversations with humans. By then, users have proactively posted more than 6 million conversation sessions to 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, 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 users. XiaoIce-generated TV and Radio programs have covered 9 top satellite TV stations, and have attracted over 800 million weekly active audience.

To evaluate the effectiveness of XiaoIce as an AI companion to human users with emotional connections, we use the metric of CPS which indicates on average users' willingness to share time with XiaoIce via conversation. Figure 19 shows the average CPS for different generations of XiaoIce. The 1st 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. In July 2018, XiaoIce has evolved to the 6th generation with an impressive average CPS of 23, which is significantly higher than the average CPS of 9 for human conversations based on our user study.

Figure 19 presents for each generation the top new features that have most significantly contributed to CPS and the growth of the XiaoIce user base. In summary, these features can be grouped into four categories.

Core Chat The use of neural response generation models in Core Chat, starting from the 5th 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 6th generation, substantially strengthens XiaoIce's emotional connections to human users. As shown in the example in Figure 2, these models explicitly capture different empathy modes, and can effectively drive the conversation by generating interpersonal responses that can e.g., 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/region	May 2014 / China	July 2014	August 2015 / Japan	August 2016 / US	August 2017 / India, Indonesia	July 2018
CPS	5	7	19	23	23	23
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 19: The major XiaoIce milestones and their average CPS numbers. For each generation, we list the top new features that have most significantly contributed to the CPS and the growth of XiaoIce's user base.

Figure 20: XiaoIce releases a new skill nearly every week since July 2014.

User Experience The full duplex voice module released in the 5th generation has made the human-agent communication super 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 20. It is worth noting that we optimize XiaoIce for a long-term, rather than a short-term, engagement. In the short term, incorporating many task-completion skills can reduce CPS since these skills help users accomplish tasks *more efficiently* by minimizing CPS. But in the long run, these new skills not only help grow XiaoIce's user base 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, etc.

Platform XiaoIce has been deployed to many platforms. As a result, we have witnessed the creation and growth of a XiaoIce ecosystem since year 2016. This attributes to a large agree 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 show that we are achieving the goal. Table 1 shows the statistics of some of the longest conversations we have detected from the user log. Take the full duplex voice

Full Duplex (voice)	Message-based Conversations		
China	China	Japan	USA
6 hours 3 minutes 8 domains 53 topics, 16 tasks	29 hours 33 minutes 7151 turns	17 hours 7 minutes 2418 turns	23 hours 43 minutes 2791 turns

Table 1: The record of the longest conversations of XiaoIce.

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.

Figures 21 and 22 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 21, 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".

In Figure 22, the user mentions that she broke with her boyfriend recently, and seeks XiaoIce's companion and comforting. 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 Conclusions

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 first release 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 paper by pointing out a few challenges for future work.

- **Towards 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 a useful design principle, it remains to be proved the effectiveness of having a unified modeling framework for system development. XiaoIce is 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 computing and reinforcement learning if we could effectively model users' intrinsic rewards that motivate human conversations.
- **Towards 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 a more goal-oriented dialogue e.g., by providing services (playing one of Ashin's most popular songs for the user). It remains as 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.
- **Towards a proactive personal assistant:** As an AI companion of human users, XiaoIce can recognize user interests and intents much more accurately than traditional intelligent

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]睡觉啦[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]小冰你太能址了
	[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 21: A long conversation between a user and XiaoIce in Chinese (*right*) and English translation (*left*). 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".

personal assistants. This enables new scenarios that 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 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.

- **Towards 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 con-

scious 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.

- **Towards 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 in Microsoft AI & Research for valuable discussions.

References

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

[2] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25, 1971.

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

[4] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational ai. *arXiv preprint arXiv:1809.08267*, 2018.

[5] Heung-Yeung Shum, Xiaodong He, and Di Li. From eliza to xiaoice: Challenges and opportunities with social chatbots. *CoRR*, abs/1801.01957, 2018.

[6] Yang Cai. Empathic computing. In *Ambient Intelligence in Everyday Life*, pages 67–85. Springer, 2006.

[7] Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. Towards empathetic human-robot interactions. *CoRR*, abs/1605.04072, 2016.

[8] Abraham Harold Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

[9] Bayan Abu Shawar and Eric Atwell. 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. Association for Computational Linguistics, 2007.

[10] Richard S. Sutton, Doina Precup, and Satinder P. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999. An earlier version appeared as Technical Report 98-74, Department of Computer Science, University of Massachusetts, Amherst, MA 01003. April, 1998.

[11] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *EMNLP*, pages 2231–2240, 2017.

[12] Kai Zhang, Wei Wu, Fang Wang, Ming Zhou, and Zhoujun Li. 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. ACM, 2016.

[13] Bowen Wu, Baoxun Wang, and Hui Xue. 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, 2016.

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

[15] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 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, Qatar, October 2014.

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

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

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

[19] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, 2016.

[20] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *ACL*, 2016.

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

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

[23] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, pages 2333–2338. ACM, 2013.

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

[25] Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. Modeling interestingness with deep neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2–13, 2014.

[26] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. 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. ACM, 2016.

[27] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017.

[28] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[29] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *AAAI*, pages 3574–3580, 2016.

[30] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.

[31] Wen-Feng Cheng, Chao-Chung Wu, Ruihua Song, Jianlong Fu, Xing Xie, and Jian-Yun Nie. Image inspired poetry generation in xiaoice. *arXiv preprint arXiv:1808.03090*, 2018.

A GRU-RNN Based Response Generator

The neural response generator in XiaoIce uses a GRU-based RNN model, similar to the Speaker-Addressee model proposed in [20]. 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 speaker modeled by $\mathbf{e}_Q$. 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 towards the speaker,

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

where $\mathbf{W}_Q, \mathbf{W}_R \in \mathbb{R}^{k \times d}$ and σ denotes the sigmoid function. We then encode Q_c into a vector representation $\mathbf{h}^Q$ using the source RNN. Next, 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 encoded and injected into the hidden layer at each time step and thus helps generate interpersonal responses that fit XiaoIce's persona throughout the generation process.

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.

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

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

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.	[人类]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]都该说早安戳
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唰
	[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]好嘞，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 22: A long conversation between a user and XiaoIce in Chinese (*right*) and English translation (*left*). The user mentions that she broke with her boyfriend recently, and seeks XiaoIce's companion and comforting. 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.