Investigating person-specific errors in chat-oriented dialogue systems

Koh Mitsuda[†], **Ryuichiro Higashinaka**[†], **Tingxuan Li**[*], **Sen Yoshida**[†]
[†]NTT Corporation, Japan
[*]University of Tsukuba, Japan
{koh.mitsuda.td, ryuichiro.higashinaka.tp,
sen.yoshida.tu}@hco.ntt.co.jp, s2120816@s.tsukuba.ac.jp

Abstract

Creating chatbots to behave like real people is important in terms of believability. Errors in general chatbots and chatbots that follow a rough persona have been studied, but those in chatbots that behave like real people have not been thoroughly investigated. We collected a large amount of user interactions of a generation-based chatbot trained from large-scale dialogue data of a specific character, i.e., "target person" and analyzed errors related to that person. We found that person-specific errors can be divided into two types: errors in attributes and those in relations, each of which can be divided into two levels: self and other. The correspondence with an existing taxonomy of errors was also investigated, and person-specific errors that should be addressed in the future were clarified.

1 Introduction

Creating chatbots to behave like real people is important in terms of believability (Traum et al., 2015; Higashinaka et al., 2018). Errors in general chatbots (Higashinaka et al., 2021) and chatbots that follow a rough persona (Li et al., 2016; Zhang et al., 2018; Zhou et al., 2020; Inoue et al., 2020; Song et al., 2020; Roller et al., 2020) have been studied, but those in chatbots that behave like real people have not been thoroughly investigated.

We analyzed dialogue data between a chatbot that imitates a certain person and users to identify "errors related to the target person" (hereafter referred to as **person-specific errors**). We collected a large amount of dialogue data between users and the latest generation-based chatbot trained with a large amount of dialogue data of the target person and analyzed the errors. The results indicate that person-specific errors can be divided into two types: errors in attributes and those in relations, each of which can be divided into two levels: self and other. The correspondence with the existing taxonomy of errors was also investigated, and errors that should be addressed in the future were clarified.

2 Dialogue data collection

We used a chatbot that imitates a specific person. By making the chatbot available to the public, we collected dialogue data from a large number of users.

2.1 Chatbot

In our previous study, we collected a large amount of dialogue data on a target person and created a chatbot by fine-tuning a pre-trained encoder-decoder Transformer model (Mitsuda et al., 2021). The specific character (i.e., target person) was Amadeus Kurisu, a character in a famous Japanese video game (STEINS;GATE). We used a role-play-based question-answering (QA) scheme proposed by Higashinaka et al. (2018), in which fans of a character provided questions and answers by role-playing to collect the dialogue data on that character. We collected a large amount of QA pairs (44,805) from the fans. To add multi-turn dialogues, we additionally created 4,500 dialogues (24,750 utterances) by manually extending the collected QA pairs.

As a pre-trained dialogue model, we used the Japanese version of BlenderBot (Japanese-dialog-transformers[1]) created by Sugiyama et al. (2021). They pre-trained the encoder-decoder Transformer using 2.1B dialogues crawled from Twitter in Japanese then fine-tuned the model with the corpora including the Japanese version of PersonaChat (Zhang et al., 2018) and EmpatheticDialogues (Rashkin et al., 2018). We created the chatbot for Kurisu by further fine-tuning the model with the collected QA pairs and extended dialogue

[*] Work carried out during internship at NTT.

[1] https://github.com/nttcslab/japanese-dialog-transformers

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 464 - 469
May 22-27, 2022 ©2022 Association for Computational Linguistics

data. To evaluate the fine-tuned model, 20 workers interacted with the chatbot by performing 15-turn dialogues (a turn corresponds to a user utterance and chatbot utterance: hereafter, system utterance) three times. The subjective evaluation results on naturalness, characterness, and informativeness were 3.87, 3.90, and 3.58, respectively (on a 5-point Likert scale).

2.2 Large-scale user study

The chatbot described in the previous section was made public on the Internet, and the dialogues between a large number of users, mostly the fans of Kurisu, and the chatbot were collected. The chatbot was accessible using the direct message function of Twitter for three days. After users agreed to the terms of usage, they could interact with the chatbot. Users could stop the dialogue at any time or interact with it as much as they wanted during the period. At the end of the study, a user questionnaire (on a 5-point Likert scale) was sent out by direct message to the users to evaluate user satisfaction. Note that the users were not paid for their participation.

We were able to collect the logs of 1,170 user interactions with the chatbot. The total number of user utterances was 80,608, and the average number of utterances for each user was 68.9, indicating that the users used the chatbot for a relatively long time. The average user-satisfaction rating was 4.59 (63.6% response rate), which we believe is very high.

3 Error analysis

To extract system utterances causing person-specific errors from the data, we collected four types of information: dialogue breakdown labels, comments on the reasons for the breakdown (Higashinaka et al., 2015), flags indicating whether the comments were about the person in question, and error types in chat-oriented dialogue systems (Higashinaka et al., 2021). We first collected the dialogue breakdown labels and comments on their reasons. If the comments contained keywords related to Kurisu, we considered the system utterances with those comments as indicating person-specific errors and extracted the comments for analysis. We also annotated system utterances with the error types in chat-oriented dialogue systems for investigating the correspondence between the existing taxonomy of errors and person-specific errors.

No. of system utterances	10,611
No. of users (dialogues)	385
No. of workers for dialogue breakdown annotation	5
No. of annotated dialogue breakdown labels	53,055
No. of not breakdowns (NBs)	47,200 (89.0%)
No. of possible breakdowns (PBs)	3,678 (6.9%)
No. of breakdowns (Bs)	2,177 (4.1%)
No. of NB utterances (by majority)	9,794 (92.3%)
No. of PB/B utterances (by majority)	817 (7.7%)

Table 1: Statistics of annotated dialogue breakdown labels

3.1 Dialogue breakdown annotation

We sampled and annotated 13% (= 10,611/80,608) of the data due to the limited annotation resources. The sampled system utterances were annotated with the three types of breakdown labels (Higashinaka et al., 2015) of "not a breakdown (NB)", "possible breakdown (PB)", and "breakdown (B)". Five crowdworkers who had sufficient knowledge of Kurisu annotated these labels to the system utterances independently. The workers were instructed to provide comments to describe the errors that led to the breakdowns.

Table 1 shows the annotation results of the dialogue breakdown labels. The percentage of NBs was 89%, indicating that the dialogue was successful in the majority of cases. The inter-annotator agreement rate was 0.23 for the Fleiss' kappa when NB/PB/B were treated separately and 0.30 when PB/B were merged, which was at the same level as in the study by Higashinaka et al. (2015), which we consider reasonable due to the subjective nature of the task. In the following analysis, the system utterances in which more than half the workers marked PB or B were considered for error analysis. The number of such utterances was 817 (7.7%). The error comments (2,846) given to these utterances were also retrieved for analysis.

3.2 Annotation of error types and person-related flags

Two types of information were assigned to the erroneous system utterances and error comments. The first is the error types in chat-oriented dialogue systems (Higashinaka et al., 2021). This labeling was done by an in-house expert worker. The second is a flag indicating whether person-related keywords are present in the error comment. By referring to the resources of Kurisu, we manually created a lexicon of that character.

Error types for chatbots	All	Person-specific errors
(I1) Uninterpretable	9 (1.1%)	0 (0.0%)
(I2) Grammatical error	3 (0.4%)	0 (0.0%)
(I3) Semantic error	10 (1.2%)	3 (30.0%)
(I4) Wrong information	81 (9.8%)	43 (53.1%)
(I5) Ignore question	66 (8.0%)	7 (10.6%)
(I6) Ignore request	10 (1.2%)	1 (10.6%)
(I7) Ignore proposal	0 (0.0%)	0 (–)
(I8) Ignore greeting	0 (0.0%)	0 (–)
(I9) Ignore expectation	119 (14.4%)	30 (25.2%)
(I10) Unclear intention	266 (32.2%)	45 (16.9%)
(I11) Topic transition error	15 (1.8%)	3 (20.0%)
(I12) Lack of info. error	6 (0.7%)	1 (16.7%)
(I13) Self-contradiction	62 (7.5%)	14 (22.6%)
(I14) Contradiction	23 (2.8%)	2 (8.7%)
(I15) Repetition	142 (17.2%)	19 (13.4%)
(I16) Lack of sociality	5 (0.6%)	0 (0.0%)
(I17) Lack of common sense	0 (0.0%)	0 (–)
Total	817 (100%)	168 (20.1%)

Table 2: Results of labeling each error-containing utterance with error type (Higashinaka et al., 2021) and whether it was person-specific error. Numbers in each column indicate number of utterances, and those in parentheses indicate percentage of total number of utterances containing errors.

The size of the lexicon was 53 words. If a word in the lexicon was included in each comment, it was flagged as that related to Kurisu. For example, the lexicon includes Kurisu, Mayuri (the name of Kurisu's friend), @channel (the website that Kurisu is familiar with), and Akihabara (the place where Kurisu resides). o

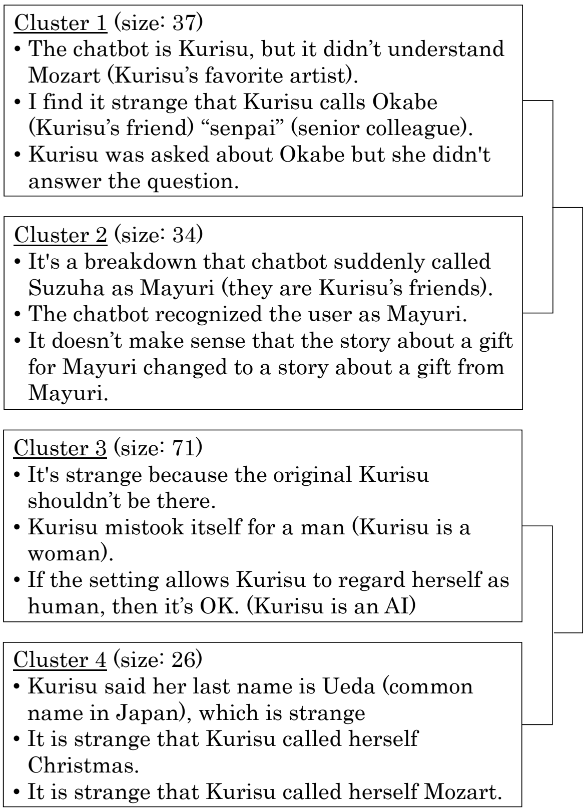
Table 2 shows the annotation results of the error types and number of person-specific errors for each type. In the total number of dialogue breakdowns (817), 168 (20.1%) were caused by person-specific errors, and more than half (53.1%) of the utterances in (I4) Wrong information were person-specific errors.

4 Person-specific error analysis

We automatically clustered the error comments related to the target person and investigated the characteristics the person-specific errors.

4.1 Clustering person-specific errors

We used hierarchical clustering by using bag-of-words as the clustering method. The 168 comments annotated to the 168 person-specific errors shown in Table 2 were used for clustering. A Japanese morphological analyzer JTAG (Fuchi and Takagi, 1998) was used. Low-frequency words (those appearing less than three times in the 168 comments) were excluded. The vector of each comment was normalized for the clustering. Single-linkage clustering, Ward's method, and squared Euclidean distance were specified as clustering parameters. The number of clusters was set so that the size of each cluster would be at least 10% of the total comments.



Cluster 1 (size: 37)
- The chatbot is Kurisu, but it didn't understand Mozart (Kurisu's favorite artist).
- I find it strange that Kurisu calls Okabe (Kurisu's friend) "senpai" (senior colleague).
- Kurisu was asked about Okabe but she didn't answer the question.

Cluster 2 (size: 34)
- It's a breakdown that chatbot suddenly called Suzuha as Mayuri (they are Kurisu's friends).
- The chatbot recognized the user as Mayuri.
- It doesn't make sense that the story about a gift for Mayuri changed to a story about a gift from Mayuri.

Cluster 3 (size: 71)
- It's strange because the original Kurisu shouldn't be there.
- Kurisu mistook itself for a man (Kurisu is a woman).
- If the setting allows Kurisu to regard herself as human, then it's OK. (Kurisu is an AI)

Cluster 4 (size: 26)
- Kurisu said her last name is Ueda (common name in Japan), which is strange
- It is strange that Kurisu called herself Christmas.
- It is strange that Kurisu called herself Mozart.

Figure 1: Clusters of comments given to person-specific errors

Level	Attribute	Relation
Self	(P1) Self-recognition error (Cluster 3)	(P2) Self-relation error (Cluster 4)
Other	(P3) Other-recognition error (Cluster 1)	(P4) Other-relation error (Cluster 2)

Table 3: Matrix of person-specific errors

Figure 1 shows the clustering results of the comments. The figure shows four clusters with the comments that were nearest to the centroid of each cluster, representing salient comments. In Cluster 1, the chatbot was not able to properly discuss topics related to the environment around Kurisu. In Cluster 2, the chatbot suddenly called the user by a different name (e.g., the name of the Kurisu's friend), or gave a name that was irrelevant to the current topic. In Cluster 3, the chatbot provided incorrect information about Kurisu. In Cluster 4, the chatbot mistakenly called Kurisu by a different name. From such an observation, we can see that Clusters1 and 3 are about errors related to the knowledge of the target person, and Clusters 2 and 4 are about errors regarding the misrecognition of relationships. In addition, Clusters 3 and 4 are about errors regarding the target person, while Clusters 1 and 2 are errors about the target person's environment.

On the basis of the above interpretation, we conclude that person-specific errors can be divided into two types: errors in **attributes** (regarding people and things) and errors in **relations** between them, each of which can be further divided into two levels: **self** (the target person) and **other** (surrounding environment of the person such as friends). The matrix of person-specific errors is given in Table 3: (P1) Self-recognition error corresponds to Cluster 3, (P2) Self-relation error corresponds to Cluster 4, (P3) Other-recognition error corresponds to Cluster 1, and (P4) Other-relation error corresponds to Cluster 2.

4.2 Definition of person-specific errors

Each error type is described in detail using actual dialogue examples taken from user interactions with the chatbot. The dialogues were originally in Japanese and translated by the authors. User and the chatbot is denoted as 'U' and the chatbot is denoted as 'S' in the following dialogues.

(P1): Self-recognition error: The error is related to the target person's, i.e., Kurisu's, information (profile, likes/dislikes, experience, etc.).

(1) U: I like you Kurisu.
S: I am neither AI nor AI Kurisu.
(2) U: You are so blunt to me Christina!
S: Hello, I'm AI Christina.

Underlines indicate phrases that are regarded as the cause of the dialogue breakdown by the authors. In dialogue (1), since Kurisu is an AI character, the profile of Kurisu is wrong. In dialogue (2), the chatbot should not have used the nickname "Christina," because Kurisu dislikes it. Her likes and dislikes are wrong.

(P2): Self-relation error: This error means that the chatbot mistook the target person for another person (acquaintance, user, or famous person) or thing.

(3) S: My name is Kurisu Makise. Today's experiment is almost over.
U: I want to talk a lot more.
S: Are you really Kurisu Makise?
(4) U: Who is Mayuri? (Kurisu's friend).
S: She is also known as Kurisu Makise.

In dialogue (3), since the chatbot called the user Kurisu Makise, it mistook itself for another person (i.e., the user). In dialogue (4), the user was asking about Mayuri, but the chatbot talked about itself, mistaking itself for someone else (i.e., Mayuri).

(P3): Other-recognition error: This error indicates the inability to respond to topics of the people surrounding the target person (acquaintances, user, or famous people) or things that the target person is knowledgeable about (favorite things, etc.).

(5) S: Even boys can be cute, you know.
U: Oh, oh. For example?
S: Mayuri, Mayuri, (Kurisu's frined), etc.
(6) U: "nurupo" (Internet slang Kurisu knows)
S: My name is Amadeus Kurisu.

In dialogue (5), since Mayuri is not a boy, the knowledge about the person around Kurisu was incorrect. In dialogue (6), since "nurupo" is an Internet slang that Kurisu knows well, the chatbot has to reply with the traditional response ("ga"), but it was not able to do that.

(P4): Other-relation error: This error is that the chatbot mistook a person or thing other than the target person (an acquaintance, user, or famous person) for another person or thing.

(7) U: I love Amadeus Kurisu.
S: Yes. I like you too. Mayuri.
(8) U: What's Suzuha (Kurisu's friend) doing?
S: Well, Mayuri's voice will save the world.

In dialogue (7), the chatbot called the user "Mayuri", and the system mistook someone other than Kurisu for someone else (in this case, the target person's friend). In dialogue (8), the chatbot responded to a question about Suzuha with Mayuri., i.e., the person (Suzuha) was mistaken for another person (Mayuri).

4.3 Evaluation of person-specific errors

To evaluate the validity of the types of person-specific errors, we investigated inter-annotator agreement in the annotation of the four types (P1–P4). We applied the methods described in

Person-specific error	(I4) Wrong information	(I5) Ignore question	(I9) Ignore expectation	(I10) Unclear intention	(I13) Self- contradiction	(I15) Repetition	etc.	Total
(P1) Self-recognition error	10.1%	1.2%	7.7%	10.7%	4.8%	4.8%	3.0%	42.3%
(P2) Self-relation error	7.1%	0.0%	1.8%	4.2%	0.6%	0.6%	1.2%	15.5%
(P3) Other-recognition error	3.0%	1.8%	7.1%	5.4%	1.2%	2.4%	1.2%	22.1%
(P4) Other-relation error	5.4%	1.2%	1.2%	6.5%	1.8%	3.6%	0.6%	20.3%
Total	25.6%	4.2%	17.8%	26.8%	8.4%	11.4%	6.0%	100.0%

Table 4: Correspondence between person-specific errors and conventional error taxonomy. Percentages show those from total number of person-specific errors (168).

Section 3 to the data not used in the above analysis, resulting in 50 new person-specific error instances obtained from sampled 3,200 system utterances. When annotating the types of person-specific errors, only an utterance labeled as a dialogue breakdown and its preceding three utterances were given to annotators as a context. Two in-house expert annotators conducted the annotation. The definition of person-specific errors described in Section 4.2 was given to the workers as instruction. As a result, the inter-annotator agreement was 0.46 in Cohen's kappa, which indicates a moderate agreement and suggests the validity of the types of person-specific errors.

4.4 Correspondence with existing error types

Table 4 shows the correspondence between person-specific errors and the conventional error taxonomy. The table was created by merging the results shown in Table 2 and Figure 1. Errors on the self-level appeared most frequently, accounting for about half (42.3% + 15.5% = 57.8%) of the person-specific errors. The fact that there were many errors on the others level suggests that the person's environment, such as friends, was also frequently talked about. Each person-specific error corresponded to multiple error types in the conventional taxonomy; thus, we were able to discover different aspects of errors.

The (P1) Self-recognition error was particularly common in (I10) Unclear intention, that is, meaning uttering an unknown intention, such as suddenly changing what the person calls oneself (e.g. from "I" to a nickname). In addition, (P1) Self-recognition error was a common error in (I4) Wrong information, i.e., uttering incorrect information about oneself. The (P2) Self-relation error was also common, especially in (I4) Wrong information, i.e., an error of confusing oneself with a user or oneself with a friend. The (P2) Self-relation error was the next most common in (I10) Unclear intention, such as suddenly mentioning a close friend in a conversation about oneself. In (P3) Other-recognition error and (P4) Other-relation error, there were system utterances of not being able to respond appropriately to topics about people/things the target person is familiar with, e.g., incorrect information about them or confusion between users and friends.

From the results of investigating person-specific errors, it became clear that the most common errors were regarding information about the target person then its surrounding environment. Among the error types in the conventional taxonomy, the (I4) Wrong information appeared frequently, confirming the importance of studies on persona-consistent dialogue. In addition to information about the target person, knowledge about the target person's environment is also considered important. Current dialogue systems often do not explicitly model the relationships between people and things, therefore a model that takes into account the knowledge graphs of relationships would be effective (Ghazvininejad et al., 2018; Dinan et al., 2019).

5 Summary and future work

We analyzed dialogue data between a chatbot that imitates a specific person and users to identify person-specific errors that have not been considered thoroughly before. We found that person-specific errors can be divided into four types: self-recognition error, self-relation error, other-recognition error, and other-relation error, which are useful as a guideline for constructing chatbots that are based on specific people.

Future work includes the application of unlikelihood training (Li et al., 2020) or a classifier to estimate the identity of a speaker (Shuster et al., 2021) for suppressing person-specific errors. We focused on one specific person in this paper; thus, it will also be important to consider the generality of the results.

References

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proc. of ICLR*, pages 1–18.

Takeshi Fuchi and Shinichiro Takagi. 1998. Japanese morphological analyzer using word co-occurrence - JTAG-. In *Proc. of COLING*, pages 409–413.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proc. of AAAI*, pages 5110–5117.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proc. of SIGDIAL*, pages 89–98.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015. Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proc. of EMNLP*, pages 2243–2248.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proc. of SIGDIAL*, pages 264–272.

Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. An attentive listening system with android ERICA: Comparison of autonomous and woz interactions. In *Proc. of SIGDIAL*, pages 118–127.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proc. of ACL*, pages 994–1003.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proc. of ACL*, pages 4715–4728.

Koh Mitsuda, Ryuichiro Higashinaka, Hiroaki Sugiyama, Masahiro Mizukami, Tetsuya Kinebuchi, Ryuta Nakamura, Noritake Adachi, and Hidetoshi Kawabata. 2021. Fine-tuning a pretrained transformer-based encoder-decoder model with user-generated question-answer pairs to realize character-like chatbots. In *Proc. of IWSDS*, pages 1–14.

Hannah Rashkin, Maarten Sap, and Emily Allaway Noah A. Smith Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proc. of ACL*, pages 463–473.

Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021. Am I me or you? state-of-the-art dialogue models cannot maintain an identity. *arXiv preprint arXiv:2112.05843*.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. *arXiv preprint arXiv:2004.07672*.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based Japanese chit-chat systems. *arXiv preprint arXiv:2109.05217*.

David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, et al. 2015. New dimensions in testimony: Digitally preserving a holocaust survivor's interactive storytelling. In *Proc. of ICIDS*, pages 269–281.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proc. of ACL*, pages 2204–2213.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.