Comparison of the Intimacy Process between Real and Acting-based Long-term Text Chats

Tsunehiro Arimoto, Hiroaki Sugiyama, Hiromi Narimatsu, Masahiro Mizukami
NTT Communication Science Laboratories, Kyoto 619-0237, Japan
{tsunehiro.arimoto, hiroaki.sugiyama, hiromi.narimatsu, masahiro.mizukami}@ntt.com

Abstract

Long-term chatbots are expected to develop relationships with users. The major trend in this field's recent long-term chatbot studies is to train systems with virtual long-term chat data called Multi-Session Chat (MSC), which collects text chat from multiple sessions of crowd workers playing the roles of speakers with defined personas. However, no investigation has attempted to determine whether such virtual long-term chat can successfully simulate relationship-building between speakers. To clarify the difference between an actual long-term intimacy process and an MSC intimacy process, this study collects real long-term chat and MSC in Japanese and compares them in terms of speech form and dialogue acts. The results of analyzing these factors suggest that MSC have an unnatural tendency to behave as if they have a close relationship with non-polite speech levels compared to actual long-term chats, but also as if they have a shallow relationship with more questions than real long-term chats.

Keywords: corpus, chatbots, text analytics

1. Introduction

Recent open-domain chatbots can generate natural responses over multiple turns using large-scale language models (Smith et al., 2020; Adiwardana et al., 2020; Sugiyama et al., 2023), but they do not address the speaker intimacy process (Altman and Taylor, 1973) and thus cannot sustain natural dialogue over multiple days and weeks. The ability to generate intimacy-based dialog is one of the challenges facing long-term dialog agents (Coon et al., 2013; Kageyama et al., 2018; Ouchi et al., 2019). As users chat with a chatbot multiple times, if the dialog agent does not change its behavior, users get the impression that it is difficult to develop a relationship with the chatbot (Croes and Antheunis, 2021).

In order to train a response generation model that can incorporate the intimacy process, dialogue data that capture behavioral changes are required. However, while dialogue data exist for specific intimacies (e.g., between friends (Yamazaki et al., 2020)), dialogue data with changing values of intimacy are nearly non-existent. Recently, text chat data called Multi-Session Chat (MSC) has been proposed to simulate long-term (multi-session) dialogues by giving the speaker a virtual persona and a timeline (Xu et al., 2021). Such data are often used to train response generation models that consider long-term contexts (Zhang et al., 2022; Bae et al., 2022). However, since they are only simulated data collected by acting, the dialogue trends that appear in the intimacy process may differ from the natural changes that occur between real speakers.

Therefore, this study represents the first effort to collect actual long-term text chats by first-time speakers with the aim of determining the difference between the actual long-term intimacy process and the MSC intimacy process. The data were collected in Japanese from 60 pairs over a period of eight weeks; MSCs in Japanese were also collected for comparison[1]. Our results suggest that actual long-term chats involve more relationship considerations than does the MSC.

2. Related work

Conventional long-term dialog agents consider intimacy with the user in a rule-based approach. For example, based on the knowledge that topics become deeper as speakers become closer (Altman and Taylor, 1973), a rule-based system has been proposed that defines the intimacy of topics and gradually selects topics with a high level of intimacy (Coon et al., 2013). Furthermore, based on the finding that the use of honorifics decreases as the relationship becomes closer (Brown et al., 1987), some dialog agents gradually reduce their use of honorifics (Kageyama et al., 2018; Ouchi et al., 2019).

However, it is not easy to control the dialogue behavior of the intimacy process by simple rules, and the inability to consider user intimacy in detail poses a major problem. Therefore, statistical approaches that can flexibly respond to the user are expected. For example, a tone-changing neural model has been proposed to allow a system to adapt to the user's use of honorifics (Niu and

[1] We release part of our dataset at https://github.com/nttcslab/japanese-long-term-chat.

LREC-COLING 2024, pages 3639–3644
20-25 May, 2024.

Bansal, 2018). Estimating a speaker's intimacy by using multi-modal features has also been studied (Chiba et al., 2021). Recently, moreover, long-term chat data, MSC, have been published to develop long-term open-domain chatbots (Xu et al., 2021). However, since MSC does not contain chat data from real speakers, it is unclear whether we should regard it as capable of producing a natural intimacy process.

Accordingly, this study collects actual long-term chat data to identify the natural tendencies of the intimacy process that occur in actual text chats and to evaluate the naturalness of the MSC intimacy process.

3. Methods

3.1. Data collection

3.1.1. Long-term Asynchronous Chat

We collected actual long-term chats, here called Long-term Asynchronous Chat (LAC), using the collection procedure described below.

Twenty Japanese men and twenty Japanese women in their 20s participated in the data collection from their own PCs or smartphones via the Slack application[2] as paid workers. They were not only students but also working adults. They were asked to participate in chats at least five days a week for a maximum of 56 days of interaction. They were also asked to post their utterances at least five times a day in each chat room. Each worker participated in three chat rooms. Their interlocutors were of the same gender, and they met each other for the first time. To reduce the chat burden, the workers were allowed to post consecutively if their interlocutor was slow in replying. Since one worker withdrew in the final week, three pairs ended their participation on the 49th day.

We instructed the workers to talk about chat topics that both speakers would enjoy, such as common interests and daily events. For privacy reasons, they were not allowed to mention identifiable information (names, addresses, etc.).

Participants were surveyed in three different stages: before, during, and after the data collection period. The pre-collection survey assessed participants' personality characteristics. An in-work questionnaire was administered weekly, in which workers reported their impressions of the interaction and their interlocutors. In the questionnaire, workers responded with a score of 0 representing a first meeting and 100 representing a best friend in terms of how close they felt to the other person. This was used as the study's intimacy score. In the post-program questionnaire, workers responded with their overall impression while looking back on the data-collection period. A dialog example is displayed in Table 1. The total number of utterances in the collected data was 71,244 (1,301,611 characters).

Table 1: Example of LAC. Polite expressions "*desu/masu (*です/ます*)*" are underlined.

Spk	Utterance
A	カラオケと歌が上手いは比例しないので大丈夫<u>です</u>よ!人が聞いて心地よい歌い方ができてればそれでいいと思い<u>ます</u>! (Karaoke has nothing to do with being a good singer. If you can sing in a way that can entertain someone, that's all that matters!)
B	そう<u>です</u>よね!魂で歌い<u>ます</u>!愚問かもしれない<u>です</u>が、ちなみにカラオケで何点<u>です</u>か? (That's right! I sing with my soul! This may be a silly question, but what is your karaoke score?)
A	魂!笑笑 曲にもより<u>ます</u>けれどいい時は 98 点とか<u>です</u>かね… (Soul! LOL, it depends on the song, but if I was in good shape, I could get a score of 98 or something like that.)
B	す、すごい、、生まれ変わったら歌上手い人間になりたい、、 (That's great. If I were born again, I would want to be a good singer.)

3.1.2. Japanese version of MSC

Multi-Session Chat (MSC) (Xu et al., 2021) is composed of English text chat data of up to five sessions with two crowdworkers acting as virtual speakers defined by persona statements. We collected a three-session and a five-session version of Japanese Multi-Session Chat using the original MSC procedure.

In the first session, the personas were chosen from the Japanese version of PersonaChat (Sugiyama et al., 2023)[3]. The interval between sessions was randomly chosen from either 1-7 hours or 1-7 days, as in the original MSC. This interval is virtual, not a real time lapse. Since it is difficult for workers to read multiple and long past dialogues, the original MSC provides workers with response summaries of past dialogues written by other workers for each session. Our Japanese MSC also provides such response summaries. We instructed each worker to chat with another worker naturally as a continuation of the past sessions and provided them with personas, past dialogue history, and summaries of the history. The same persona was not always played by the same worker. Each session consisted of 12 utterances, six by each of the two

[2]https://slack.com

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

Table 2: Example of MSC. Polite expressions "*desu/masu (*です/ます*)*" are underlined.

Spk	Utterance
A	さっそく、百均で DIY グッズを買ってきたよ。 (I immediately went to the hundred-yen store and bought some DIY goods.)
B	お、いいですね!何か作ってみました? (Oh, that's nice! Have you tried making something?)
A	ブックスタンドを作ろうと思ったんだけど、なかなかうまくいかない。傾いちゃうんだ。 (I was going to make a book stand, but I can't seem to get it right. It would tilt.)
B	あはは、傾いちゃうのも味があっていいじゃない。私も最初は失敗ばっかりでしたよ。 (Haha, in DIY, you have to enjoy failure. I made a lot of mistakes at first, too.)

Table 3: Data statistics of comparison data: numbers of sentences, dialogs, and different pairs. One dialog refers to one session in MSC and one day in LAC.

Week	MSC			LAC		
	sent	dialog	pair	sent	dialog	pair
1	8,526	336	100	10,617	420	60
2	3,440	139	73	9,584	419	60
3	620	25	17	8,860	420	60
Total	12,586	500	-	29,061	1,259	-

participating workers. Each utterance was limited to 100 characters or fewer in Japanese. The five-session version of the Japanese MSC was used for the analysis in this study. The number of pairs was 100, and the total number of utterances was 6,000 (12,586 sentences, 231,345 Japanese characters). A dialog example is displayed in Table 2.

Since MSC is not composed of chat data from real speakers, the true value of subjective impressions cannot be determined. Therefore, subjective impression data, such as intimacy score, were not collected in the MSC.

3.1.3. Extracting data for comparison

The LAC collected 56 days of dialogue data, while the MSC is expected to be up to 28 days long (= 7 days + 7 days + 7 days + 7 days). In order to compare the two datasets with the same length of dialogue periods, we extracted data for comparison from each dataset. We first extracted data from the MSC only for pairs that completed five sessions. The maximum time lapse in the extracted data was 20 days (= 7 days + 6 hours + 6 days + 6 days). We thus extracted data from the LAC through week 3 (21 days). One session of MSC consisted of 12 utterances. For ease of comparison, we divided the LAC by each day (= 10 utterances or more expected). The resulting statistics are shown in Table 3.

3.2. Evaluation

3.2.1. Manipulation Check

To clarify whether LAC represents dialogue data that bring speakers closer together, we compared week 1 intimacy scores to those of week 3.

3.2.2. Honorifics

Honorifics are linguistic expressions that convey relationship consideration to the dialogue partner. They tend to occur more frequently in dialogues of parties for whom the relationship is shallow and decrease in dialogues of parties for whom the relationship is deep. Since MSC includes acting data, it may not adequately express relationship consideration compared to LAC, which is actual data, and the rate of use of honorifics may differ. Therefore, we attempted to compare the proportions of honorifics.

Extraction of honorifics is rule-based. Here, each utterance is separated into sentences by a morphological analysis tool[4]. If a sentence contains the Japanese honorific expression desu/masu, it is considered an honorific sentence. The proportion of honorific sentences was calculated for each dialogue (one session for MSC and one day for LAC), and the median value for each week was also calculated. The Mann-Whitney U Test was used in a comparison to confirm whether a difference existed between the medians of MSC and LAC.

3.2.3. Dialogue acts

Dialogue acts such as self-disclosure and questioning depend on the relationship with the interaction partner. For example, self-disclosure becomes more extensive and deeper with a closer partner (Altman and Taylor, 1973). Questions are more reduced in friend-to-friend interactions than in first encounters (Yamazaki et al., 2020). To compare whether there are differences in the proportions of dialog acts that occur in LAC and MSC, we calculated the percentage of each dialogue act in each dialogue sample and used the Mann-Whitney U Test to analyze whether the median differs between MSC and LAC. We also conducted a comparison to clarify whether the proportions of questions differed between week 1 and week 3 within each corpus.

In order to identify differences in the main dialogue acts, the analysis focused on questioning, self-disclosure, confirmation, empathy, and providing information, which had a frequency of 5% or more in both corpora. For the estimation of dialogue acts, the utterances were segmented into sentence units, and the Japanese dialogue acts were estimated using conventional methods (Meguro et al., 2013; Higashinaka et al., 2014).

[4]https://megagonlabs.github.io/ginza/

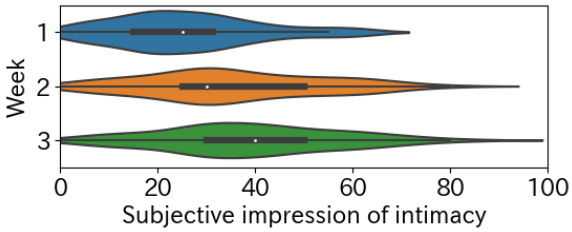

Figure 1: Intimacy score of LAC

Table 4: Comparison results. Bonferroni correction was used on p-values.

	Week	Median		p-value
		MSC	LAC	
honorific	1	0.69	**0.86**	$p < 0.01$
	2	0.59	**0.86**	$p < 0.01$
	3	0.78	**0.85**	$p < 0.01$
self disclosure	1	0.43	**0.46**	$p < 0.01$
	2	0.41	**0.49**	$p < 0.01$
	3	0.40	**0.47**	$p < 0.05$
empathy	1	0.15	0.14	n.s.
	2	0.16	0.15	n.s.
	3	0.17	0.17	n.s.
question	1	**0.14**	0.10	$p < 0.01$
	2	**0.13**	0.08	$p < 0.01$
	3	**0.12**	0.07	$p < 0.01$
inform	1	**0.09**	0.06	$p < 0.01$
	2	**0.12**	0.07	$p < 0.01$
	3	0.12	0.08	n.s.
confirm	1	0.08	0.06	n.s.
	2	0.07	0.08	n.s.
	3	0.08	0.07	n.s.

4. Results

4.1. Manipulation check

LAC intimacy scores from week 1 to week 3 are shown in Figure 1. The results of a Wilcoxon signed-rank test show that the score at week 3 was significantly higher than at week 1 (week1=25, week3=40, W=217, $p<.01$). Accordingly, LAC was confirmed to be a text chat that enhances the speaker's intimacy with the interlocutor.

4.2. Honorifics

The percentage of honorific sentences in each week for LAC and MSC is shown in Table 4. All weeks showed that LAC speakers used significantly more honorifics than those in MSC. Neither LAC nor MSC showed a significant difference between weekly 1 and weekly 3.

4.3. Dialogue acts

The comparison results for the percentage of each dialogue act across corpora are shown in Table 4. The test results reveal that utterances in LAC were significantly more self-disclosing, asking fewer questions and providing more information compared to MSC.

A comparison of the percentages of dialogue acts between week 1 and week 3 within each corpus reveals that change in the percentage of questions was a significant phenomenon in LAC (week1=0.10, week3=0.07, U=107102.5, $p<.01$). On the other hand, no significant difference was identified for MSC (week1=0.14, week3=0.12, U=4708.05, p=.31).

5. Disscussion

The results of the manipulation check confirm that LAC speakers subjectively felt that they had gradually become closer to their dialog partners. This finding confirms that the present collection method is capable of obtaining long-term chat data that increase the intimacy of the speakers.

A comparison of LAC and MSC honorifics shows that MSC speakers used fewer honorific sentences than LAC speakers in every week. This finding suggests that MSC speakers treat their interlocutors as if they were familiar. A comparison of dialogue acts reveals that MSC speakers asked more questions and provided more information but made less self-disclosure than LAC speakers. In addition, LAC showed a decrease in the rate of questioning, a phenomenon observed when comparing first-time acquaintances and friends (Yamazaki et al., 2020), whereas this was not significantly observed in MSC. These results indicate that MSC speakers did not naturally increase intimacy as LAC speakers did.

Therefore, MSC showed a different tendency from LAC in terms of the intimacy process. Furthermore, MSC showed contradictory characteristics: One was as if it were a highly intimate dialogue with few honorifics, and the other was as if it were a less intimate dialogue with few self-disclosures but many questions. This may be due to the fact that MSC is an acted dataset; it is possible that the reason why MSC speakers were less respectful is that they believed their dialogue partners were also acting and would not be offended if they did not show due consideration to their counterparts. Moreover, the finding that MSC speakers asked more questions and provided more information than LAC speakers, but made self-disclosure less frequently, suggests that MSC speakers preferred efficient information exchange rather than relationship-oriented behavior with their interlocutors.

5.1. Limitations

This study found that when examining intimacy processes among Japanese speakers, there were differences in the proportion of honorifics and dia-

logue acts between LAC, which is collected from actual chat, and MSC, which is collected from participants acting out roles. LAC speakers received monetary rewards, but they still developed more intimacy with their partner through long-term text chats. These results may differ depending on the speaker's cultural region and language. Japanese is a language that explicitly distinguishes between honorific and non-honorific forms, which may have made it easier to distinguish the difference between LAC and MSC.

6. Conclusion

This study analyzed the difference between actual long-term chats and acting-based chats, in terms of the speaker intimacy process. The results of the speech level and dialogue act analyses suggest that MSC, compared to actual long-term chats, showed two contradictory and unnatural tendencies: first, participants behaved as if they had a close relationship by using non-polite speech levels; at the same time, their utterances implied a more shallow relationship by asking more questions and making less self-disclosure. Therefore, these findings suggest that LAC can provide more suitable datasets than MSC for generating responses that take into account the human intimacy process.

Ethics Statement

The Japanese MSC and LAC data in this study were collected with due consideration to ethical and privacy issues. Workers were informed in advance about the work to be done. They participated voluntarily in the data collection and received fair compensation. Participants had the option to stop during the collection process and were compensated according to the amount of work they had completed up to that point. MSC and LAC were composed solely of text chat data and did not collect workers' personal information, images, or voices. Japanese MSC workers chatted under fictitious personas, masking their own real identities. LAC workers chatted from their own perspective, without fictitious personas, but were instructed never to share personal information. The experimenter manually verified that the LAC did not contain any personally identifiable information. LAC workers were also instructed not to harass their conversation partners and to contact the work manager if they had any problems. The work manager monitored chats in all rooms to ensure that privacy and harassment issues did not arise. In recent years, preventing users' addiction to dialogue agents (Xie et al., 2023) and aggressive attitudes toward dialogue agents (Chin et al., 2020) have become issues. To solve these problems, chatbots need to recognize and control their relationship with users. This research can contribute to the development of such technology.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP19H05690 and JP19H05693.

8. Bibliographical References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Yuya Chiba, Yoshihiro Yamazaki, and Akinori Ito. 2021. Speaker intimacy in chat-talks: analysis and recognition based on verbal and non-verbal information. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA. Association for Computing Machinery.

William Coon, Charles Rich, and Candace L Sidner. 2013. Activity planning for long-term relationships. In *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013, Proceedings*, volume 8108, page 425. Springer.

Emmelyn A J Croes and Marjolijn L Antheunis. 2021. Can we be friends with mitsuku? a longitudinal study on the process of relationship formation between humans and a social chatbot. *J. Soc. Pers. Relat.*, 38(1):279–300.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939.

Yukiko Kageyama, Yuya Chiba, Takashi Nose, and Akinori Ito. 2018. Improving user impression in spoken dialog system with gradual speech form control. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 235–240.

Toyomi Meguro, Yasuhiro Minami, Ryuichiro Higashinaka, and Kohji Dohsaka. 2013. Learning to control listening-oriented dialogue using partially observable markov decision processes. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(4):1–20.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Subaru Ouchi, Kazuki Mizumaru, Daisuke Sakamoto, and Tetsuo Ono. 2019. Should speech dialogue system use honorific expression? In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 232–233.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691. IEEE.

Tianling Xie, Iryna Pentina, and Tyler Hancock. 2023. Friend, mentor, lover: does chatbot engagement lead to psychological dependence? *Journal of Service Management*, 34(4):806–828.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 443–448.

Tong Zhang, Yong Liu, Boyang Li, Zhiwei Zeng, Pengwei Wang, Yuan You, Chunyan Miao, and Lizhen Cui. 2022. History-aware hierarchical transformer for multi-session open-domain dialogue system. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3395–3407.