06

Chat collection for developing long-term chat model

Real dialogue data tells how people get close

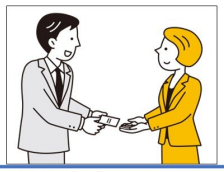
Abstract

Demand is growing for long-term chat data with which chatbots can be developed that gradually become acquainted with a user in a manner that resembles human interaction. In this study, we collected long-term text chat data over eight weeks by recording text chats between speakers who met for the first time. These pairs gradually became more acquainted with each other through repeated interactions. Conventional long-term chat data are acting-based data (i.e. simulated data) crafted by workers in virtual speaker settings. Our comparative analysis showed that the actual data collected in this study are more natural in terms of speech level and dialogue acts than conventional acting-based chat data. Our future challenge is developing a long-term chat dialogue model using these data to maintain consistency with a dialogue's history and to reflect the establishment of ongoing speaker relationships.

Conversation and Intimacy

■ Human conversations vary depending on intimacy

- Even in casual chats, topics/speaking manners depend on whether talking with someone new or a close acquaintance



(First meeting)
・More polite language
・Superficial topics

(Acquaintance/friend)
・Less formal language
・Deeper topics (inner thoughts)

To create dialogue systems that engage in prolonged conversations, such systems must learn from humans how to become gradually acquainted through conversations.

Purpose

■ Our research generates chat data that capture natural intimacy processes of humans, an essential step for creating dialogue systems that can engage naturally with humans over long-term periods.

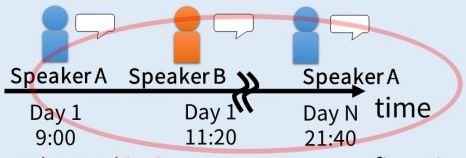
■ Conventional long-term chat: Acting-based

Persona setting	Student, Born in the US, ...
Time lapse setting	Two days after their previous conversation

Acting setup → Actor

Intimacy process may not be perfectly captured in acting.

■ Our long-term chat: Real



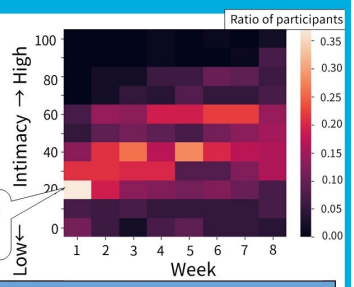
We record actual intimacy processes as first-time speakers became acquainted through text chat.

Long-term collection and analysis of chats between first-time pairs

■ Data collection

- Japanese participants
- 60 pairs in their 20s
- Same-sex first-meeting pairs
- 56-day conversations (max)
- About 10 utterances per day

Intimacy between participants actually increased



Example of topic transitions

Day	Topic	Day	Topic
1	Recent weather	23	School course about criminal and civil law
2	Recent crowd levels	24	Approaches for studying criminal/ civil law
3	Profiles and School Exams	25	Study methods
4	Report writing for school exams	26	Future desire for children
5	Speaker A's exam schedule	27	Numbers of brothers and sisters
6	Speaker B's part-time job, Speaker A's study	28	Gender of child

Shallow topics on Day 1 — Deeper topics on Day 26

■ Evaluation of linguistic features

We compared linguistic features over a three-week period between actual and acting-based data* (Japanese version).

→ Proportion of honorifics and speech acts appeared significantly different.

	Week	Act	Real
Honorifics	1	0.69	0.86
	2	0.59	0.86
	3	0.78	0.85
Self-disclosures	1	0.43	0.46
	2	0.41	0.49
	3	0.40	0.47
Questions	1	0.14	0.10
	2	0.13	0.08
	3	0.12	0.07
Information providing	1	0.09	0.06
	2	0.12	0.07
	3	0.12	0.08

※ These numbers represent the median frequency of occurrences per dialogue for the relevant sentences.

We confirmed that actual, long-term chat data more clearly capture trends with which humans speak as they become acquainted with partners than acting-based data.

Acknowledgements: Part of our work is supported by JSPS KAKENHI Grant Number **JP19H05690, JP19H05693**

* Xu, Jing., et al. "Beyond goldfish memory: Long-term open-domain conversation." *arXiv preprint arXiv:2107.07567* (2021).

References

[1] T. Arimoto, H. Sugiyama, H. Narimatsu, M. Mizukami, "Comparison of the Intimacy Process between Real and Acting-based Long-term Text Chats," in *Proc. The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024.

Contact

Tsunehiro Arimoto
Interaction Research Group, Innovative Communication Laboratory