Taiga Mori

National Institute of Advanced Industrial Science and Technology, AI Research Center, Tokyo, Japan

Chiba University, Chiba, Japan

mori.taiga.0412@gmail.com

1 Research Interests

My research interest lies at the intersection of **cognitive science** and **dialog system research**; more specifically, I am interested in the cognitive process of listener response generation and aim to implement my model in a dialogue system to validate its effectiveness and build more human-like dialog systems.

1.1 Listener response studies

In everyday conversation, it is generally the principle that one person speaks at a time (Sacks et al., 1974). However, in reality, listeners do not just listen passively; they respond with short utterances such as "yeah," nods, or laughter. These listener responses are referred to as back-channels (Yngve, 1970), continuers (Schegloff, 1982), response tokens (Gardner, 2001), or reactive tokens (Clancy et al., 1996), and contribute to smooth turn-taking and the deepening of relationships.

It has been found that the frequency of listener responses is especially high in Japanese (Maynard, 1986; Clancy et al., 1996), indicating their important role especially in Japanese conversations. Additionally, Japanese listeners use a variety of responses. Den and Yoshida (2011) extended Gardner's (2001) response tokens to Japanese and categorized Japanese response tokens into six types: responsive interjections, expressive interjections, lexical reactive expressions, repetitions, assessments, and completions.

Dialogue systems that primarily focus on listening to the user's speech and providing appropriate responses are called attentive listening dialogue systems, and they have been the subject of active research (Bevacqua et al., 2012; Lara, 2017). There is also a significant amount of research on detecting the timing for producing listener responses (Ward and Tsukahara, 2000; Morency et al., 2010; Kawahara et al., 2015). However, current attentive listening dialogue systems still face challenges regarding the diversity and consistency of responses. We believe that the cognitive approach is an effective way to address these challenges.

1.2 Cognitive listener response generation model

According to Clark's (1996) grounding model, human communication consists of four hierarchical levels, which he calls *action ladders*. According to this model, at the lowest level of communication, Level 1, the speaker executes a behavior such as vocalization or movement, and the listener pays attention to it. At Level 2, the listener recognizes the signal, such as words or gestures, produced by the speaker. At Level 3, the listener understands what the speaker means. At Level 4, the listener considers the joint action proposed by the speaker. Allwood et al. (1992) also proposed four feedback functions similar to these: *contact*, *perception*, *understanding*, and *attitudinal reaction*.

Based on these theory, we hypothesize that the cognitive process of generating listener responses in everyday conversation also consists of four levels, with different types of responses used depending on the level.
Attention level: Responses at this lowest level indicate that the listener is listening to and paying attention to the speaker's speech, and are typically observed immediately after disfluencies such as fillers or pauses. This is almost synonymous with traditional back-channels. Responses at the attention level include responsive interjections (e.g., "yeah" or "uh-huh" in English).
Word level: This level of responses indicate the listener's understanding or recognition of a certain word produced by the speaker and are observed after devices that induce listener responses, such as rising intonation, lengthening, pauses, or eye contact. This includes responses to try-markers (Sacks and Schegloff, 1979). Responses at the word level include not only responsive interjections but also expressive interjections and repetitions (e.g., "Oh, Mr. Yamada").
Propositional information level: Responses at this level indicate the listener's understanding, empathy, or emotions to a propositional information and are used at a position where the propositional information is complete or predictable. While this partially overlaps with the continuer (Schegloff, 1982), it differs in that it can also be seen within the TCU (Turn Constructional Unit). Responses at this level include responsive interjections,

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 40–42
September 16–17, 2024.

expressive interjections, repetitions, lexical reactive expressions (e.g., "right" or "I see"), and assessments (e.g., "scary" or "interesting").

Activity level: Responses at this highest level also indicate the listener's understanding, empathy, emotions, etc. but are oriented towards activities rather than single propositional information. Since responses at this level are used at the endpoint of the activity, they overlap with sequence-closing devices. Responses at the activity level include responsive interjections, expressive interjections, repetitions, lexical reactive expressions, and assessments.

However, as with Clark's action ladder, these levels are hierarchical, with higher-level reactions encompassing lower-level ones. For example, a response at the conclusion of a storytelling not only serves as a response to the entire story but also retrospectively indicates that the listener has been attentive to the speaker's talk and has correctly understood the individual propositional information and words that make up the story.

Traditional studies on predicting listener responses have primarily focused only on attention level responses (Morency et al., 2010; Kawahara et al., 2015). The lowest attention-level responses can be generated using these traditional prediction methods based on the speaker's speech and body movements as features. However, generating higher-level responses will require matching with the system's knowledge base and some form of reasoning.

1.3 Listener response generation using knowledge graph and LLMs

Currently, we are working on implementing the aforementioned cognitive model as a system. In particular, we are focusing on developing an architecture that generates responses based on the listener's knowledge. Our proposed architecture consists of system knowledge in the form of a knowledge graph and three modules using LLMs.

Information extraction module: This module extracts information from the user's utterance and converts it into structured data using an LLM. The LLM extracts information from the user's utterances and converts it into triples consisting of subject, predicate, and object.

Knowledge comparison module: In this module, the user's knowledge is compared with the system's knowledge, and the system's knowledge state is determined. There are five types of system knowledge states: *complete match* when the system has the same triple of knowledge as the user, *partial match* when the system does not have the same knowledge but has related knowledge that aligns with it, *no knowledge* when the system lacks any related knowledge, *partial conflict* when the system has related knowledge that contradicts the user's knowledge, and *complete conflict* when the system holds contradictory knowledge. Whether related knowledge aligns with or contradicts the user's knowledge is determined by the LLM. For example, even if the system doesn't know the exact temperature, knowing that it is snowing would be considered having related knowledge about the temperature.

Response generation module: This module generates a response using the system's knowledge based on the determined knowledge state. If the knowledge state is a complete match/complete conflict, the module generates an *agreement*/*disagreement* response. If the knowledge state is a partial match/partial conflict, it converts the related knowledge into a natural language sentence using the LLM and generates a *noticing*/*surprise* response. If the knowledge state is a no knowledge, it generates an *acceptance* response.

2 Future of Spoken Dialog Research

Interaction is a topic that spans multiple fields and there is a wealth of knowledge available on it. However, collaboration between these fields has been yet sufficient. One reason is the technical challenge of implementing the higher-order cognitive processing models constructed by linguistics, sociology, psychology and cognitive science into actual systems. However, with the advent of LLMs and other technological innovations, this issue is gradually being resolved. For example, it has become easier for cognitive science researchers like myself to create simple dialogue systems to prove their hypotheses. In the future, further integration is desirable to allow for the effective utilization of each other's insights.

3 Suggestions for discussion

- Multimodality: How can speech be integrated with other modalities such as paralinguistic information, gestures, facial expressions, and eye gaze?
- Explainability: To what extent should the dialogue system be able to explain its own actions? How best to use LLMs?
- Collaboration with other fields: How can we contribute to other fields? What do we expect from other fields?

Acknowledgement

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 127-136.

Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Maat, Gary Mckeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Etienne De Sevin, Michel Valstar, Martin Wollmer, Marc Shroder, and Bjorn Schuller. 2012. Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing*, 3(2):165-183.

Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In: Tannen, D. (Ed.), *Analyzing Discourse: Text and Talk* (Georgetown University Round Table on Language and Linguistics, 1981). Georgetown University Press, Washington, DC.

Harvey Sacks, Emanuel A. Schegloff. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696-735.

Harvey Sacks, Emanuel A. Schegloff. 1979. Two preferences in the organization of reference to persons in conversation and their interaction. In *Everyday Language: Studies in Ethnomethodology*. New York.

Herbert H. Clark. 1996. *Using language*. Cambridge university press.

Jens Allwood, Joakim Nivre, Elisabeth Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1-26.

Louis-Philippe Morency, Iwan de Kok, Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70-84.

Nigel G. Ward, Olac Fuentes, and Alejandro Vega. 2010. Dialog prediction for a general model of turn-taking. In *INTERSPEECH*, Citeseer, pages 2662-2665.

Patricia M. Clancy, Sandra A. Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26:355-387.

Rod Gardner. 2001. *When Listeners Talk*. John Benjamins, Amsterdam.

Senko K. Maynard. 1986. On back-channel behavior in Japanese and English casual conversation. *Linguistics*, 24:1079-1108.

Tatsuya Kawahara, Miki Uesato, Koichiro Yoshino, Katsuya Takanashi. 2015. Toward adaptive generation of backchannels for attentive listening agents. In *International Workshop Series on Spoken Dialogue Systems Technology*, pages.1-10.

Victor H. Yngve. 1970. On getting a word in edgewise. In: Campbell, M.A. (Ed.), *Papers from the Sixth Regional Meeting of Chicago Linguistic Society*, Chicago Linguistic Society, Chicago, pages 567-577.

Yasuharu Den, Nao Yoshida, Katsuya Takanashi, Hanae Koiso. 2011. Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, IEEE, pages 168-173.

Biographical Sketch

Taiga Mori is a PhD student at Chiba University and research assistant at the Artificial Intelligence Research Center (AIRC), National Institute of National Institute of Advanced Industrial Science and Technology (AIST). He is generally interested in multimodal interaction and currently working on modeling multimodal listener response generation such as verbal response tokens and head nodding. He uses both quantitative methods such as statistical modeling and qualitative methods such as conversation analysis to build models, and then implement them in dialogue systems to verify the effectiveness and validity of the models.