Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7244–7252
Marseille, 20-25 June 2022

Construction of Responsive Utterance Corpus for Attentive Listening Response Production

Koichiro Ito[†1], Masaki Murata[†2], Tomohiro Ohno[†3], Shigeki Matsubara[†1,4]
[†1]Graduate School of Informatics, Nagoya University
[†2]National Institute of Technology, Toyota College
[†3]Graduate School of Advanced Science and Technology, Tokyo Denki University
[†4]Information & Communications, Nagoya University
[†1,4]Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
[†2]2-1 Eisei-cho, Toyota, 471-8525, Japan
[†3]5 Senjuasahi-cho, Adachi-ku, Tokyo,120-8551, Japan
ito.koichiro.v1@s.mail.nagoya-u.ac.jp, murata@toyota-ct.ac.jp
ohno@mail.dendai.ac.jp, matubara@nagoya-u.jp

Abstract

In Japan, the number of single-person households, particularly among the elderly, is increasing. Consequently, opportunities for people to narrate are being reduced. To address this issue, conversational agents, e.g., communication robots and smart speakers, are expected to play the role of the listener. To realize these agents, this paper describes the collection of conversational responses by listeners that demonstrate attentive listening attitudes toward narrative speakers, and a method to annotate existing narrative speech with responsive utterances is proposed. To summarize, 148,962 responsive utterances by 11 listeners were collected in a narrative corpus comprising 13,234 utterance units. The collected responsive utterances were analyzed in terms of response frequency, diversity, coverage, and naturalness. These results demonstrated that diverse and natural responsive utterances were collected by the proposed method in an efficient and comprehensive manner. To demonstrate the practical use of the collected responsive utterances, an experiment was conducted, in which response generation timings were detected in narratives.

Keywords: spoken language corpus, response timing, listening response, narrative speech

1. Introduction

The act of narration is a fundamental human requirement, and narrating can only be established when there is a listener. However, in Japan, the number of single-person households, particularly among the elderly, is increasing (Statistics Bureau of Japan, 2015; National Institute of Population and Social Security Research, 2018), and there are many situations that listeners cannot be present. Therefore, it is important to increase opportunities for people to narrate.

To address this issue, conversational agents, e.g., communication robots and smart speakers, are expected to listen to narratives. For these agents to be recognized as listeners of narratives, they have to be functionally able to

1. attentively listen to narratives, and
2. convey that they have attentively listened to the narratives.

Function 1 is realized using speech recognition and understanding technologies. An explicit means of realizing function 2 is providing responses to narratives. In particular, producing gestures and utterances to narrative is effective. Here, an utterance to realize function 2, i.e., an utterance made in response to a narrative to convey attentive listening, is referred to as an *attentive listening response*. Representative attentive listening responses are backchannels. In terms of backchannels, data collections and analyses have been performed (Ward and Tsukahara, 2000; Kamiya et al., 2010; Yamaguchi et al., 2016), and backchannel generation methods have been proposed (Noguchi and Den, 1998; Cathcart et al., 2003; Fujie et al., 2004; Kitaoka et al., 2005; Poppe et al., 2010; Morency et al., 2010; Yamaguchi et al., 2016; Ruede et al., 2017). In addition to the backchannel, there are diverse types of attentive listening responses (*Nihongo Kijutsu Bunpo Kenkyukai*, 2009). However, to our knowledge, methods to generate attentive listening responses that include responses other than backchannels have not been extensively investigated to date. To develop such generation methods, it is necessary to collect a wide range of actual responses and observe and analyze the collected response data.

Thus, in this study, we describe the collection of attentive listening responses to realize conversational agents that act as listeners of narratives. In the data collection process, we annotated prerecorded narratives with appropriate attentive listening response expressions and production timings. We obtained the attentive listening responses in an offline manner, and this collection method has the following advantages:

- Multiple attentive listening responses can be collected for a single narrative because the responses do not have any effect on the narratives.
- Only utterances to demonstrate attentive listen-

ing attitudes can be collected because the workers who annotate the narratives with attentive listening responses can focus on producing them.

The remainder of this paper is organized as follows: Section 2 describes the requirements of attentive listening response generation and discusses works related to attentive listening responses. Section 3 explains the attentive listening response collection process, describes the collection results, and classifies the collected responses. Section 4 evaluates the collected responses in terms of response frequency, diversity, coverage, and naturalness. Section 5 describes a response generation timing detection experiment conducted to demonstrate the practical use of the collected responses. Finally, this paper is concluded in Section 6, including suggestions for future work.

2. Attentive Listening Responses

In this study, we attempted to realize a function to convey the attentive listening of narratives to speakers by generating appropriate attentive listening responses.

2.1. Response Generation Requirements

To enable conversational agents to generate responses, it is necessary to select appropriate response expressions and determine effective response generation timings. In terms of enhancing the effect of the response generation, the requirements are summarized as follows:

- The response expressions must be natural and diverse.
- The response frequency is high, and the response generation timings are natural.

To our knowledge, there is no existing response generation method that satisfies these requirements: thus, it is important to collect real data of attentive listening responses and accumulate useful results through observations and analyses.

2.2. Related Work

Several previous dialogue system studies attempted to develop systems that function as a listener (Kobayashi et al., 2010; Meguro et al., 2011; Shitaoka et al., 2017; Lala et al., 2017). In these systems, although the dialogue proceeds in a user-driven manner, it is assumed that the systems aggressively engage with user's story line by asking questions and requesting information.
However, in this study, we assume that the target systems do not take initiative in the dialogue and act only as a listener. Thus, the system's utterances are limited to responses that exhibit attentive listening attitudes, and the users do not respond directly to the system's utterances. Thus, our target communication form differs from that of these previous studies.

speakers	30
speaking time	8:43:55
utterance units	13,234
morphemes	66,897

annotators	11
speaking time	22:16:34
utterance units	148,962
morphemes	232,651

Table 1: Size of narrative speech data (left) and response speech data (right).

3. Data Collection

In order to realize automatic generation of attentive listening responses to narratives, we collect natural and diverse attentive listening response data.

3.1. Collection Policy

It is possible to record real time interactions between a speaker and a listener, and extract the attentive listening responses from the listener's utterances. However, with this method, the versatility of the collected data is limited because the listener's reactions possibly have an effect on the speaker's behavior. Moreover, the diversity of the collected data is limited because only responses given by a single listener can be collected to one narrative.
Therefore, in this study, we had workers annotate the narrative data with response data. Here, the workers produced attentive listening responses that were synchronized with the sound playback of narratives, and the narrative data were annotated with the produced response character data and time data. This collection process allows the workers to focus on producing attentive listening responses effectively because there are no two-way interactions. Furthermore, it is possible to improve the coverage of the collected data because it is relatively easy to record additional responses by additional workers to the same narrative data.

3.2. Data Collection Method and Results

In this study, we used Japanese Elder's Language Index Corpus (JELiCo) (Aramaki, 2016) as narrative data. This corpus includes speech data acquired from 30 elders (average of 20 minutes per person). In this corpus, the elders speak their narratives as a monologue by answering 10 prepared questions.
11 workers with advanced communication skills produced attentive listening responses to the narrative speech data. Here, each worker produced responses to the same narrative data separately. The workers produced responses in real time to the narrative speech. Note that the narrative speech was played only once. The response speech was recorded using a close-talking microphone.
The narrative and response data were manually annotated with the following five tags:

- **(F)**: Filler
- **(G)**: Interjection to express emotions
- **(D)**: Disfluency

Narrative utterance			Responsive utterance		
	Japanese	English translation		Japanese	English translation
01:07:20 – 01:11:09	イタリア旅行をしたことが一番楽しかったです	I enjoyed traveling to Italy the most	01:08:99 – 01:09:25	はい	yes
			01:10:02 – 01:11:55	あーそうですかー	is it so?
			01:11:55 – 01:12:85	素敵ですねー	it is wonderful
01:14:71 – 01:18:86	もう二度と行けないかなと思いながら行ってきましたけど	I went there thinking that I could not go again	01:13:78 – 01:14:30	うん	hmm
			01:14:30 – 01:15:04	イタリア旅行	traveling to Italy
			01:16:54 – 01:18:04	いえいえそんなー	that's not true
			01:18:52 – 01:19:77	ああそうですか	oh, is it?

Figure 1: Example of narrative and responsive utterances.

Name	Description (What does the response demonstrate?)
Backchannel	Successful hearing
Admiration	Admiration, surprise, or attention to the content of the speaker's utterance
Evaluation	Attitude toward the situation described by speaker's utterance
Approval	Approval of the content of the speaker's utterance
Disapproval	Disapproval of the content of the speaker's utterance
Echoic response	Comprehension of the content of the speaker's utterance and a sense of security
Paraphrasing	Attempting to understand and share the content of the speaker's utterance
Satisfaction	The listener's attitude that the content of the speaker's utterance is satisfactory for him/her
Surprise	Strong surprise at the content of the speaker's utterance
Surprise with doubt	Surprise or doubt toward the content of the speaker's utterance
Opinion	The listener's personal experiences, opinions, or feelings
Complement	Eagerly listening to the speaker's utterance
Greeting	Acknowledgement of the speaker's presence and willingness to favorably interact with the speaker
Provoke memory	The listener's memory is provoked by the content of the speaker's utterance
Thinking process	The listener is contemplating the content of the speaker's utterance
Other	Other than the above

Table 2: Types of attentive listening responses.

- **(U)**: Pitch rise at the end
- **(?)**: Uncertainty in perception

Moreover, narrative data were manually annotated with sentence boundaries. Table 1 shows the size of the narrative and collected response speech data. We defined the utterance units as units into which utterances were divided by human perceptible pauses. Moreover, morphological analysis was performed on the narratives and collected responses, and we provide the start and end times of each morpheme. Here, we used MeCab (Kudo et al., 2004) for morphological analysis and the phoneme segmentation kit [1] in Julius (Lee et al., 2001) to identify the start and end times. IPADIC neologd [2] and UniDic (Ver. 2.1.2) [3] were used as narrative and response morphological dictionary, respectively. Figure 1 shows an example of the collected data.

[1] http://julius.osdn.jp/index.php?q=ouyoukit.html

[2] https://github.com/neologd/mecab-ipadic-neologd

[3] https://ja.osdn.net/projects/unidic/releases/58338

3.3. Response Type

In this study, we classified the collected responses into 16 attentive listening response types in reference to the literature (*Nihongo Kijutsu Bunpo Kenkyukai*, 2009). Table 2 details the attentive listening response types considered in this study. All of the collected attentive listening responses were manually annotated with the corresponding types. In the collected data, backchannel, representative attentive listening response type, accounted for 67.96% of the total responses, and the proportion was the largest. The response types except the backchannel type accounted for 32.04% of the total, and the proportions were larger for admiration, echoic response, and evaluation in that order. Figure 2 shows a breakdown of the attentive listening response types (except backchannel type). Moreover, Figure 3 shows examples of narratives and responses for admiration, echoic response, and evaluation. Here, the string in the upper row for each response type show Japanese phrases, and the strings in parentheses show English translations of the Japanese phrases in the upper rows.

4. Evaluation of Collected Response Data

We analyzed and evaluated the collected response data in terms of frequency, diversity, coverage, and natural-

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	average
Appearance rate	1.85	2.82	2.49	2.26	3.22	2.72	3.30	1.82	1.91	1.60	3.15	2.47
Entropy	5.19	4.02	3.62	5.30	3.53	4.87	2.78	4.08	5.25	4.44	5.61	4.43

Table 3: Appearance rate (seconds) and entropy for each worker.

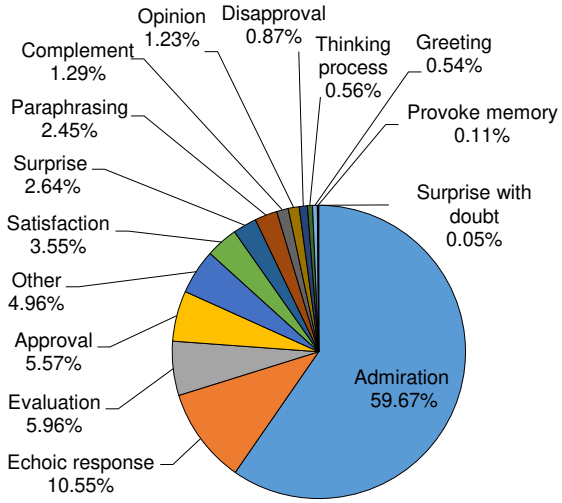

Figure 2: Breakdown of attentive listening response types (except backchannel responses).

Response type	Narrative utterance	Responsive utterance
Admiration	ちょっとあの一腰を痛めたもんで (I've got a little back injury)	あー (uh)
Echoic response	わたくしのこんにちまでの仕事はえーライターです (my current job is a writer)	ライター (writer)
Evaluation	書道も好きで総理大臣賞も頂いたりして (I also like calligraphy and I won the Prime Minister's Award)	凄いですね (that's great)

Figure 3: Example of narrative and responsive utterances for each response type.

ness.

4.1. Response Frequency

We calculated the appearance frequency of the attentive listening responses in the collected data. The top row in Table 3 shows the appearance rate (the interval of attentive listening response generation) for each worker. Here, w_1 – w_{11} represents the 11 workers. In the collected data, the average appearance rate per worker was 2.47 seconds, which is considered a high appearance rate. As reference data, we examined the appearance rate of utterances corresponding to attentive listening responses in Nagoya University Conversation Corpus (Fujimura et al., 2012), which is a representative spoken language resource that contains recorded Japanese chat conversations. The appearance rate in this corpus is 12.4 seconds. This result indicates that the response frequency in the collected data is high even though there were differences among the individual workers.

4.2. Response Diversity

To evaluate the diversity of the attentive listening responses, we measured the diversity index for the response string types, i.e., the types of strings comprising a response. Here, we adopted entropy per response, which is expressed as follows:

$$H = -\sum_{i=1}^{S} p_i \log p_i \quad (1)$$

where S is the number of the response string types in the collected data, and p_i is the proportion of the number of occurrences of response i in the total number of occurrences of all responses. The bottom row in Table 3 shows the entropy for each worker. Here, w_1 – w_{11} represents the eleven workers. We reported that the average entropy per worker was 4.43. We measured entropy in Nagoya University Conversation Corpus in the same manner. The entropy in that corpus was 4.86. Despite the fact that the Nagoya University Conversation Corpus contains free conversations, the entropy in our collected data was slightly less than that of Nagoya University Conversation Corpus. This demonstrates that the diversity of the responses in our data was high.

4.3. Response Coverage

This section evaluates whether the collected response data covers timings that are appropriate in terms of producing effective attentive listening responses. As discussed in Section 3, we collected the response data by having 11 workers produce attentive listening responses in real time to the elder speech separately. Therefore, even if there were multiple responses to the same part of a narrative, their start times did not perfectly match. Here, we considered the *bunsetsu*[4] boundaries in narratives as candidates for timings when responses can be produced, and we mapped the actual produced responses to the bunsetsu boundaries using the narratives and response start times. In particular, we mapped a response to the end boundary of the nearest bunsetsu to the response start time.

[4]Bunsetsu is a linguistic unit in Japanese that roughly corresponds to a basic phrase in English. A bunsetsu comprises a single independent word and zero or more ancillary words.

The following procedure was used to divide the narrative into bunsetsus:

1. We removed the morphemes with (F), (G), (D) and (?) tags explained in Section 3.2.
2. We divided the morpheme sequence obtained in step 1 into bunsetsu sequences using CaboCha (Kudo and Matsumoto, 2002).
3. We inserted the morphemes removed in step 1 into the bunsetsu sequence in step 2. When the position of the inserted morpheme was a bunsetsu boundary defined in step 2, the inserted morpheme was considered to be a new bunsetsu. However, when the consecutive morphemes with the same tag type were inserted at the same bunsetsu boundary, they were grouped together and considered to be a new bunsetsu.

In the following, a bunsetsu boundary is referred to as a *production timing candidate*.
In this examination, the response data by 11 workers were used. As a result of the mapping, any workers produced attentive listening responses at 25,523 out of 29,969 production timing candidates in narrative data, i.e., 85.16% of all production timing candidates. In the following, the set of the timings when any workers produced attentive listening responses is denoted as $T(\mathrm{w_{all}})$. The following procedure was used to analyze the coverage of the timings appropriate to produce attentive listening responses:

1. One of the 11 workers was selected, we obtained the timings when the worker produced attentive listening responses, and we calculated the proportion of these timings in $T(\mathrm{w_{all}})$.
2. We selected one of the remaining workers, and we obtained the timings when the worker produced attentive listening responses.
3. We calculated the union of the timings when the previously selected workers produced attentive listening responses, and we calculated the proportion of this union in $T(\mathrm{w_{all}})$.
4. Steps 2 and 3 were repeated until all 11 workers were selected and processed.

However, there exist $11!(= 39,916,800)$ possible orders to select among the 11 workers. Thus, the above procedure was performed for the 11! orders of worker selection, and we calculated the average of the proportions in Step 1 and 3. The average of the proportions were then used to evaluate the coverage of the timings appropriate to produce attentive listening responses.
Figure 4 shows the results of the coverage analysis. We confirmed that when seven of the 11 workers were used, $> 90\%$ of the timings in all $T(\mathrm{w_{all}})$ were covered. Furthermore, we confirmed that when 10 workers were used, 98.31% of the timings were covered. In other words, only 1.69% of the timings were added by including the eleventh worker. Therefore, it is thought that the number of the response production timings which are newly found is very small even if additional responses by having new workers produce responses to the narrative data used in this study. These results indicate that the collected data covered most of the timings that are appropriate to produce effective attentive listening responses.

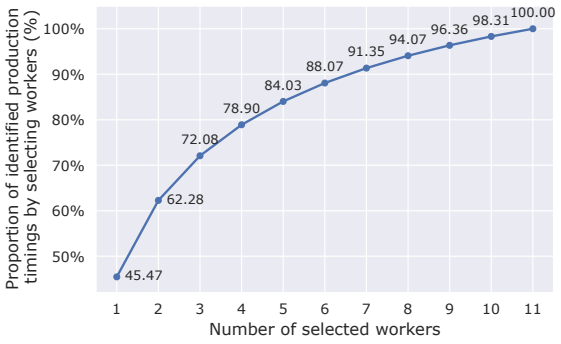

Figure 4: Proportion of identified production timings among all production timings by selecting workers.

4.4. Response Naturalness

A subjective experiment was conducted to evaluate the naturalness of the expressions and production timings of the collected attentive listening responses. Here, stereo sounds of 53 narratives by five elders and 2,191 responses to these narratives were considered. For each narrative, one worker producing responses to the given narrative was randomly selected, and the stereo sound comprising that narrative and the corresponding responses by the selected worker was used. Here, five subjects, all students in their twenties, evaluated the naturalness of each response. As a result, the number of the responses judged unnatural was 47.60 on average per subject, and responses judged natural accounted for 97.75% of the total. Therefore, the naturalness of the collected response data was confirmed.

5. Use of Collected Response Data

Attentive listening responses are expected to encourage speakers to narrate; however, it is necessary for increasing the speaker's narrative motivation to generate responses at appropriate timings. Therefore, to realize automatic generation of attentive listening responses, it is necessary to detect appropriate timings to generate responses. Thus, in this section, we describe an experiment conducted to detect appropriate response generation timings using attentive listening responses collected in this study.

5.1. Experimental Settings

There are individual differences in the production of attentive listening responses, and their production tim-

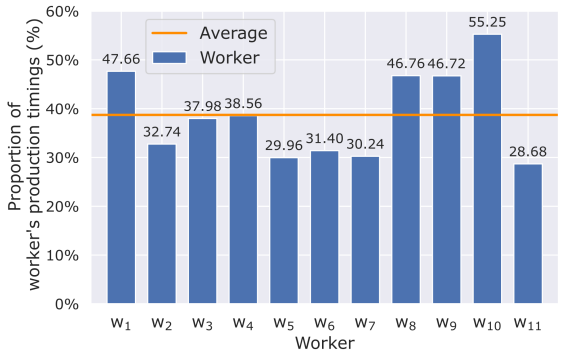

Figure 5: Proportion of worker's production timings among all production timing candidates.

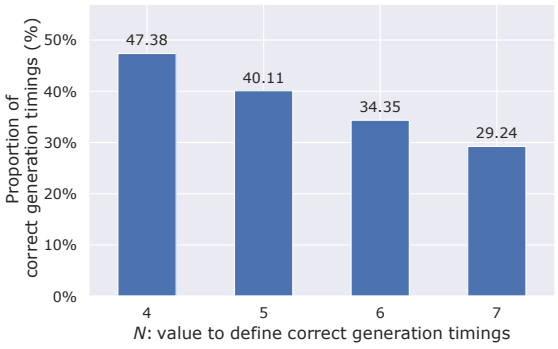

Figure 6: Proportion of correct generation timings among all production timing candidates.

ings vary from listener to listener. When one listener produces a response, another listener does not always produce a response in the same timing. In this study, the attentive listening responses of 11 workers to one narrative were collected. By using the collected responses of the 11 workers, it is possible to obtain the standardized production timings. It is thought that the data of this standard response production timings are useful for the development of an attentive listening response system that maintain generality. This section describes an experiment conducted to detect the standard response production timings, using the response data collected in this study.

In this experiment, we used the mapping results of the production timing candidates (Section 4.3). Note that the attentive listening responses were separately collected by having 11 listeners produce responses to the same narrative data. Therefore, the number of the listeners producing responses mapped to the production timing candidates was 12, ranging from zero to 11. In this experiment, the detection target timings were defined as the production timing candidates which the number of listeners producing responses mapped to was more than N. To realize a system that actively generates attentive listening responses, correct generation timings should be defined by setting N to a small value. Moreover, to realize a system that passively generates attentive listening responses, correct generation timings should be defined by setting N to a large value. In addition to maintaining the generality of the response generation system, it is possible to develop a system with various response generation strategies by changing the value of N as mentioned above. This becomes possible because the response data were collected by having multiple workers produce responses individually to the same narrative speech data.

5.2. Experimental Data

In this experiment, N was set based on the proportion of the attentive listening response production timings of humans. First, we investigated this proportion in the data collected from the 11 workers. Figure 5 shows the proportion of each worker's response production timings among the all production timing candidates. Here, $\mathrm{w}_1 - \mathrm{w}_{11}$ represents the 11 workers. Among the 11 workers, the smallest proportion was 28.68%, and the largest proportion was 55.25%. These two proportions correspond to the proportion of the workers who produced responses in the most active and passive manners, respectively. The average proportion among the 11 workers was 38.72%. Based on these findings, an experiment was conducted to detect timings to generate attentive listening responses for four settings of $N \in [4, 5, 6, 7]$. Figure 6 shows the proportion of correct response generation timings among all production timing candidates for each value of N, where $N = 4$, $N = 5, 6$, and $N = 7$ correspond to active, normal, and passive response generation strategies, respectively.

To construct and evaluate a method to detect effective timings to generate attentive listening responses to be explained in Section 5.3, the production timing candidates were divided into training, development, and test data. The training, development, and test data contained 17,479, 6,308, and 6,182 production timing candidates, respectively.

5.3. Method to Detect Response Generation Timings

A method to detect response generation timings in production timing candidates is outlined in the following: To detect generation timings, we solve binary classification problem to determine whether a production timing candidate is the generation timing. This problem is solved using the character strings immediately before a production timing candidate. The purpose of this section is not to propose a new method to detect generation timings of attentive listening responses, but to demonstrate an example of using the collected response data. Therefore, a standard and simple method in the field of natural language processing is adopted as a method for detecting the generation timing. In this experiment, a model was constructed to solve the binary classification problem by fine-tuning a pre-trained BERT (Devlin et al., 2019).

Here, the character strings immediately before a production timing candidate in narratives were first split into tokens with the BERT tokenizer. [CLS] token and [SEP] token were added at the head and last of this token sequence, respectively. Then, the token sequence were encoded using the pre-trained BERT. The encoded representation corresponding to [CLS] token were input into a classification layer comprising linear transformation and softmax function, and the probability of a production timing candidate belonging to each class was obtained. Finally, if the probability of a production timing candidate belonging to the generation timing class was greater than another class, the candidate was detected as a generation timing.

5.4. Train and Test

Here, we describe the model used to detect generation timings. In this experiment, the character strings from five bunsetsus immediately before a production timing candidate were input to the BERT model. As the pre-trained BERT model, we used cl-tohoku/bert-base-japanese-v2 [5] in huggingface/transformers [6]. Moreover, the training loss was cross entropy loss, and Adam optimizer (Kingma and Ba, 2015) was used. The class weight based on the inverse class frequency in the training data was applied to the training loss. The batch size was set to 128, and the learning rate was set to 1e-5. The detection model was trained for 10 epochs. The model with the lowest development loss in 10 epochs was used for the test data to evaluate detection performance. For comparison, a detection method that randomly detects generation timings according to the correct generation timing proportion in the training data was prepared.

In this experiment, the generation timings were detected among production timing candidates. Here, the precision, recall, and f-measure (the harmonic mean of precision and recall) were used to evaluate detection performance on the test data. Precision P and recall R are calculated as follows:

$$P = \frac{\#\,\text{detected correct generation timings}}{\#\,\text{detected generation timings}} \quad (2)$$

$$R = \frac{\#\,\text{detected correct generation timings}}{\#\,\text{correct generation timings}} \quad (3)$$

5.5. Experimental Results

Table 4 shows the evaluation results for the detection methods. The f-measure values for the BERT-based detection method were 0.705 for $N = 4$, 0.695 for $N = 5$, 0.676 for $N = 6$, and 0.667 for $N = 7$. As can be seen, the BERT-based detection method outperformed the random detection method in terms of precision, recall, and f-measure for all N values. Moreover, we found that the f-measure value decreased as N became large. These results indicate that it was more difficult to detect response generation timings for the more passive response system. It is thought that the simple detection method using the BERT model did not sufficiently capture the characteristics of the generation timings corresponding to large N, i.e., the timings at which a lot of workers produce responses.

		Precision	Recall	F-measure
$N = 4$	BERT	0.684	0.727	0.705
	random	0.455	0.495	0.474
$N = 5$	BERT	0.679	0.712	0.695
	random	0.384	0.421	0.401
$N = 6$	BERT	0.584	0.802	0.676
	random	0.326	0.358	0.342
$N = 7$	BERT	0.595	0.759	0.667
	random	0.280	0.310	0.294

Table 4: Evaluation metrics for each setting and detection method.

In this experiment, we focused on response generation timing detection using only character strings in narratives. However, it is considered that acoustic features, e.g., pitch and power, are also effective factors in this detection task. Moreover, it is necessary for appropriately generating attentive listening responses to consider the history of when previous responses were generated and what previous responses were generated. For example, the elapsed time from the last response generation timing should be considered. Although acoustic features and the response history were not considered in this experiment, it is possible to work on response timing detection considering them by using the collected data in this study. Thus, we would like to incorporate these factors into our work in future.

6. Conclusion

In this study, we describe a method to collect attentive listening responses to narratives. In this collection method, prerecorded narrative speeches were annotated with expressions and the production timings of attentive listening responses. We confirmed that our collection method can efficiently and comprehensively collect natural and diverse responses. Finally, we described a response generation timing detection experiment using the collected responses. In this experiment, the generation timings were detected using character strings in narratives. In future, we would like to detect generation timings using acoustic features extracted from narratives and the listener's response history.

It is necessary for appropriately generating attentive listening responses to not only detect the generation timings, but also determine what response to generate. The response data collected in this study contain both response production timings and the response expres-

[5] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[6] https://github.com/huggingface/transformers

sions by a listener. Moreover, all collected responses were labeled with a response type. In future, we would like to determine which response to generate using the response expressions and response types in the collected responses.

7. Acknowledgements

The narrative corpus was provided by the Social Computing Laboratory of Nara Institute of Science and Technology, Japan. This research was supported in part by a Grant-in-Aid for Challenging Exploratory Research of the JSPS (No. 18K19811) and the Research Expenses of the Nagoya University Interdisciplinary Frontier Fellowship.

8. Bibliographical References

Cathcart, N., Carletta, J., and Klein, E. (2003). A Shallow Model for Backchannel Continuers in Spoken Dialogue. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics (EACL-2003)*, pages 51–58, Budapest, Hungary.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-2019)*, pages 4171–4186, Minneapolis, USA.

Fujie, S., Fukushima, K., and Kobayashi, T. (2004). A Conversation Robot with Back-Channel Feedback Function Based on Linguistic and Nonlinguistic Information. In *Proceedings of the 2nd International Conference on Autonomous Robots and Agents (ICARA-2004)*, pages 379–384, Palmerston North, New Zealand.

Fujimura, I., Chiba, S., and Ohso, M. (2012). Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a Lexical Profiling Approach to Comparing Spoken and Written Corpora. In *Proceedings of the 7th GSCP International Conference. Speech and Corpora*, pages 393–398, Belo Horizonte, Brazil.

Kamiya, Y., Ohno, T., and Matsubara, S. (2010). Coherent Back-Channel Feedback Tagging of In-Car Spoken Dialogue Corpus. In *Proceedings of the 11th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL-2010)*, pages 205–208, Tokyo, Japan.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR-2015)*, San Diego, USA.

Kitaoka, N., Takeuchi, M., Nishimura, R., and Nakagawa, S. (2005). Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems. *Journal of Japanese Society for Artificial Intelligence*, 20(3):220–228.

Kobayashi, Y., Yamamoto, D., Koga, T., Yokoyama, S., and Doi, M. (2010). Design Targeting Voice Interface Robot Capable of Active Listening. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI-2010)*, pages 161–162, Osaka, Japan.

Kudo, T. and Matsumoto, Y. (2002). Japanese Dependency Analysis using Cascaded Chunking. In *Proceedings of the 6th Conference on Natural Language Learning (CONLL-2002)*, pages 63–69, Stroudsburg, USA.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 230–237, Barcelona, Spain.

Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., and Kawahara, T. (2017). Attentive Listening System with Backchanneling, Response Generation and Flexible Turn-Taking. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL-2017)*, pages 127–136, Saarbrücken, Germany.

Lee, A., Kawahara, T., and Shikano, K. (2001). Julius — an Open Source Real-Time Large Vocabulary Recognition Engine. In *Proceedings of 7th European Conference on Speech Communication and Technology (EUROSPEECH-2001)*, pages 1691–1694, Aalborg, Denmark.

Meguro, T., Minami, Y., Higashinaka, R., and Dohsaka, K. (2011). Evaluation of Listening-Oriented Dialogue Control Rules Based on the Analysis of HMMs. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH-2011)*, pages 809–812, Florence, Italy.

Morency, L.-P., de Kok, I., and Gratch, J. (2010). A Probabilistic Multimodal Approach for Predicting Listener Backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84.

National Institute of Population and Social Security Research. (2018). Household Projections for Japan 2015-2040 Outline of Results and Methods. `http://www.ipss.go.jp/pp-ajsetai/e/hhprj2018/hhprj2018_DL.pdf`.

Nihongo Kijutsu Bunpo Kenkyukai, editor, (2009). *Gendai Nihongo Bunpo 7*, pages 165–182. *Kuroshio Shuppan*. (In Japanese).

Noguchi, H. and Den, Y. (1998). Prosody-based Detection of the Context of Backchannel Responses. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, pages 487–490, Sydney, Australia.

Poppe, R., Truong, K. P., Reidsma, D., and Heylen, D. (2010). Backchannel Strategies for Artificial Listen-

ers. In *International Conference on Intelligent Virtual Agents (IVA-2010)*, pages 146–158, Philadelphia, USA.

Ruede, R., Müller, M., Stüker, S., and Waibel, A. (2017). Enhancing Backchannel Prediction Using Word Embeddings. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH-2017)*, pages 879–883, Stockholm, Sweden.

Shitaoka, K., Tokuhisa, R., Yoshimura, T., Hoshino, H., and Watanabe, N. (2017). Active Listening System for a Conversation Robot. *Journal of Natural Language Processing*, 24(1):3–47. (In Japanese).

Statistics Bureau of Japan. (2015). Chapter VIII: Household Status. `http://www.stat.go.jp/english/data/kokusei/2015/poj/pdf/2015ch08.pdf`.

Ward, N. and Tsukahara, W. (2000). Prosodic Features which Cue Back-Channel Responses in English and Japanese. *Journal of pragmatics*, 32(8):1177–1207.

Yamaguchi, T., Inoue, K., Yoshino, K., Takanashi, K., G. Ward, N., and Kawahara, T. (2016). Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents. In *Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS-2016)*, pages 1–12, Saariselkä, Finland.

9. Language Resource References

Aramaki, Eiji. (2016). *Japanese Elder's Language Index Corpus v2*. `https://doi.org/10.6084/m9.figshare.2082706.v1`.