Automatically Generating Head Nods with Linguistic Information

Ryo Ishii, Ryuichiro Higashinaka(✉), Kyosuke Nishida(✉), Taichi Katayama(✉), Nozomi Kobayashi(✉), and Junji Tomita(✉)

NTT Media Intelligence Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, Japan
{ishii.ryo,higashinaka.ryuichiro,nishida.kyosuke,katayama.taichi, kobayashi.nozomi,tomita.junji}@lab.ntt.co.jp

Abstract. In addition to verbal behavior, nonverbal behavior is an important aspect for an embodied dialogue system to be able to conduct a smooth conversation with the user. Researchers have focused on automatically generating nonverbal behavior from speech and language information of dialogue systems. We propose a model to generate head nods accompanying an utterance from natural language. To the best of our knowledge, previous studies generated nods from the final words at the end of an utterance, i.e. bag of words. In this study, we focused on various text analyzed using linguistic information such as dialog act, part of speech, a large-scale Japanese thesaurus, and word position in a sentence. First, we compiled a Japanese corpus of 24 dialogues including utterance and nod information. Next, using the corpus, we created a model that generates nod during a phrase by using dialog act, part of speech, a large-scale Japanese thesaurus, word position in a sentence in addition to bag of words. The results indicate that our model outperformed a model using only bag of words and chance level. The results indicate that dialog act, part of speech, the large-scale Japanese thesaurus, and word position are useful to generate nods. Moreover, the model using all types of linguistic information had the highest performance. This result indicates that several types of linguistic information have the potential to be strong predictors with which to generate nods automatically.

Keywords: Nod · Generation · Japanese dialogue
Linguistic information

1 Introduction

Nonverbal behavior in human communication has important functions of transmitting emotions and intentions in addition to verbal behavior [2]. This means that an embodied dialogue system should be able to express nonverbal behavior according to the utterance to communicate smoothly with the user [10,28,35]. Against such a background, researchers have focused on constructing automatic

G. Meiselwitz (Ed.): SCSM 2018, LNCS 10914, pp. 383–391, 2018.
https://doi.org/10.1007/978-3-319-91485-5_29

generation models of nonverbal behavior from speech and linguistic information. Among nonverbal behaviors, nodding of the head is very important for emphasizing speech, giving and receiving speech authority, giving feedback, expressing conversational engagement, and intention of starting to speak [12,14,31,33,34]. It has been shown that nodding improves the naturalness of avatars and dialog systems and promotes conversation.

Nodding accompanying an utterance has the effect of strengthening the persuasive power of speech and making it easier for the conversational partner to understand the content of the utterance [27]. Researchers have tried to generate nods during speaking from speech and natural language. In particular, they used several acoustic features, such as sound pressure and prosody, for generating nods [1,4,6,10,24,25,37]. However, it has been difficult to accurately generate nods at an appropriate time according to an utterance from only speech.

A few studies have tackled the problem of generating nods from natural language. These studies focused on the final word in the phrase of an utterance and analyzed the co-occurrences with nods. They found that morphemes related to the interjections, feedback, questionnaire, and conjunctions appearing in turn-keeping [8,9] tend to co-occur with nods. On the basis of this information, a simple automatic nod-generation model was proposed [10,32]. It was found that the behavior of humanoid robots and avatars that generated nods with the model gave a better impression of naturalness. It is thought that if a model that can generate nodding more accurately is constructed, it will lead to smoother communication between the dialog system and user. Therefore, a more accurate nod-generation model should be constructed by clarifying the relevance of more detailed language information and nodding. It is also known that the relevance of a speech feature to nodding and vice versa depends on the language; for instance this is weaker in Japanese [8]. A detailed examination of a nod-generation model using language information is thus considered important.

In this research, we constructed a highly accurate head-nod-generation model using natural Japanese language by focusing on the various text analyzed linguistic information such as dialog act, part of speech, a large-scale Japanese thesaurus, and word position in a sentence, which has not been investigated. A dialogue act is information indicating the intention of a speaker in a whole utterance, and it is believed that the occurrences of nods change in accordance with the intention. We hypothesized that words in phrases other than the final phrase and lexicons of utterances had strong relationships with head nodding.

We collected a corpus consisting of 24 Japanese dialogues including utterances and head-nod information. Next, we used the corpus to create our model that generates a nod during a phrase by using bag of words, dialog act, part of speech, a large-scale Japanese thesaurus, and word position in a sentence in addition to the bag of words. The results indicate that our model using dialog act, part of speech, the large-scale Japanese thesaurus, and word position outperformed a model using only bag of words and chance level. The results indicate that dialog act, part of speech, the large-scale Japanese thesaurus, and word position are useful to generate nods. Moreover, the model using all types

of language information had the highest performance. This result indicates that several types of linguistic information have the potential to be strong predictors with which to generate nods automatically.

Fig. 1. Photograph of two participants having dialogue

2 Corpus

To collect a Japanese conversation corpus including verbal and nonverbal behaviors for generating nods in dialogue, we recorded 24 face-to-face two-person conversations (12 groups of two different people). The participants were Japanese males and females in their 20s to 50s who had never met before. They sat facing each other (Fig. 1). To gather more data on nodding accompanying utterances, we adopted the explanation of an animation participants have not seen as the conversational content. Before the dialogue, they watched a famous popular cartoon called "'Tom & Jerry" in which the characters do not speak. In each dialogue, one participant explained the content of the cartoon to the conversational

partner within ten minutes. At any time during this period, the partner could freely ask questions about the content.

We recorded the participants' voices with a pin microphone attached to the chest and videoed the entire discussion. We also took bust (chest, shoulders, and head) shots of each participant (recorded at 30 Hz). In each dialogue, the data on the utterances and nodding behaviors of the person explaining the cartoon were collected in the first half of the ten-minute period (120 min in total) as follows.

- Utterances: We built an utterance unit using the inter-pausal unit (IPU) [26]. The utterance interval was manually extracted from the speech wave. A portion of an utterance followed by 200 ms of silence was used as the unit of one utterance. We collected 2965 IPUs. Moreover, we used J-tag [5] which is a general morphological analysis tool for Japanese to divide IPU into phrases. We collected 11877 phrases in total.
- Head nod: A head nod is a gesture in which the head is tilted in alternating up and down arcs along the sagittal plane. A skilled annotator annotated the nods by using bust/head and overhead views in each frame of the videos. We regarded nodding continuously in time as one nod event.
- Gaze: The participants wore a glass-type eye tracker (Tobii Glass2). The gaze target of the participants and the pupil diameter were measured at 30 Hz.
- Hand gesture and body posture: The participants' body movements, such as hand gestures, upper body, and leg movements, were measured with a motion capture device (Xsens MVN) at 240 Hz.

All verbal and nonverbal behavior data were integrated at 30 Hz for display using the ELAN viewer [36]. This viewer enabled us to annotate the multimodal data frame-by-frame and observe the data intuitively. In this research, we only handled utterance and head-nod data in the corpus we constructed. Nods occurred in 1601 out of the 2965 IPUs.

3 Head-Nod-Generation Model

The goal of our research was to demonstrate that bag-of-words, dialog acts, parts of speech, a large-scale Japanese thesaurus, and word position in a sentence is useful for generating nods. We evaluated our proposed model for generating nods from several types of linguistic information and the previously constructed estimation model using only the final word at the end of an utterance [8,9]. We constructed another estimation model using all types of linguistic information to evaluate the effectiveness of this fusion (All model). The feature values of linguistic information for each phrase were as follows.

- Length of phrase (LP): Number of characters in a phrase.
- Word position (WP): Word position in a sentence.

- Bag of words (BW): The word injunctions related to feedback (e.g., "en", "ee", "aa", "hi", etc.) and particles related to questioning and turn-keeping (e.g., "de", "kara", "kedo", "'kana", "janai", etc.) co-occurring with the nod is used for estimation of noding the previous studies [8,9]. To deal with more generic word information in addition to them, we examined the number of occurrences of all words, not some morphemes. We used J-tag [5], a general morphological analysis tool for Japanese.
- Dialogue act (DA): A dialogue act was extracted using an estimation technique for Japanese [7,29]. The technique can estimate a dialogue act using the word N-grams, semantic categories (obtained from a Japanese thesaurus Goi-Taikei), and character N-grams. The dialog acts and number of IPUs are listed in Table 1.
- Part of speech (PS): Number of occurrences of parts of speech of words in a phrase. We used J-tag [5] to extract part-of-speech information.
- Large-scale Japanese thesaurus (LT): Large-scale Japanese thesaurus is a large lexical database of Japanese. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

We constructed the nod-generation models by using J.48 [], which implements a decision tree in Weka [3] and evaluated the accuracy of the models and the effectiveness of each type of linguistic information. The class was a binary value as to whether a nod occurred.

We used 24-fold cross validation using a leave-one-person-out technique with the data for the 24 participants. We evaluated how well a participant's nods could be estimated with an estimator generated only from data of other people. As shown in Table 2, the performance of the model using only LP, WP, DA, PS, or LT was higher than the chance level. However, the performance of the model using BW with an F-score of 0.423 was lower than chance level. This suggests that BW, which was used in a previous study [8,9], was not useful for generating nods in our experiment. The model using LT had the highest performance among the models using only LP, WP, BM, DA, PS, or LT, with an F-score of 0.588. This suggests that LT is most useful in generating nods. In addition, the performance of mode using all information was higher than that using LT. This suggests that using several types of language information is useful to generate nods.

4 Discussion

The experimental results indicate that BW is not useful to generate nods. The reason is that the amount of data is not large, and it is thought that learning cannot be done well because the frequency of each word included in the learning data is too small. Because it is costly and difficult to collect a massive amount of multimodal data, BW is not effective. On the other hand, LT is most effective since such information is super classified rather than the word; therefore, the possibility that it could be learned well even with a relatively small amount of

Table 1. Dialogue act labels

Label	Dialogue Act	Label	Dialogue Act
DA0	Greeting	DA15	Question (habit)
DA1	Provision	DA16	Question (desire)
DA2	Self-disclosure (fact)	DA17	Question (plan)
DA3	Self-disclosure (experience)	DA18	Question (evaluation)
DA4	Self-disclosure (habit)	DA19	Question (other)
DA5	Self-disclosure (positive preference)	DA20	Question (Yourself)
DA6	Self-disclosure (negative preference)	DA21	Sympathy
DA7	Self-disclosure (neutral preference)	DA22	Non-sympathy
DA8	Self-disclosure (desire)	DA23	Confirmation
DA9	Self-disclosure (plan)	DA24	Proposal
DA10	Self-disclosure (other)	DA25	Repeat
DA11	Acknowledgment	DA26	Paraphrase
DA12	Question (information)	DA27	Approval
DA13	Question (fact)	DA28	Thanks
DA14	Question (experience)	DA29	Apology
		DA30	Filler
		DA31	Admiration
		DA32	Other

data can be considered. All linguistic information is useful to generate nods. This suggests that using several type of language information has the potential to generate nonverbal behaviors.

In this research, we used language information extracted from a unit of phrase and tried to determine whether nodding occurs in the phrase. We did not consider

Table 2. Evaluation result of generation models.

Used feature values	Precision	Recall	F-score
Chance level	0.500	0.500	0.500
LP	0.561	0.556	0.558
WP	0.526	0.528	0.527
BW	0.353	0.527	0.423
DA	0.513	0.533	0.523
PS	0.521	0.528	0.524
LT	0.614	0.538	0.578
ALL	0.578	0.599	0.590

the time-sequential information as a feature. We plan to focus on time-sequential linguistic information to generate nods. Furthermore, we would like to work on constructing a model that can generate the detailed parameters of nods such as number and depth.

5 Conclusion

We constructed a highly accurate head-nod-generation model using natural Japanese language. In this research, we focused on various text analyzed linguistic information such as dialog acts, parts of speech, a large-scale Japanese, and word positions in sentences. In an experiment, we found that our estimation model these types of information outperformed that using bag-of-words information alone. We also found that a model using all types of linguistic information is most useful to generate nods. These results indicate that several types of linguistic information have the potential to be strong predictors to generate nods automatically.

In the future, we will focus on time-sequential linguistic information to generate nods. We would like to work on constructing a model that can generate the detailed parameters of nods such as number and depth. Furthermore, we plan to construct a model for generating the occurrence timing of nods within an utterance and a model for generating nonverbal behaviors such as gaze, which is important for turn management [11–13,19–22], expression of conversational engagement [15–18,30], and body posture.

References

1. Beskow, J., Granstrom, B., House, D.: Visual correlates to prominence in several expressive modes. In: Proceedings of INTERSPEECH (2006)
2. BirdWhistell, R.L.: Kinesics and Context. University of Pennsylvania Press, Philadelphia (1970)
3. Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA-experiences with a java open-source project. J. Mach. Learn. Res. **11**, 2533–2541 (2010)
4. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: analysis and synthesis. In: IEEE Transactions on Audio, Speech, and Language Processing, pp. 1075–1086 (2007)
5. Fuchi, T., Takagi, S.: Japanese morphological analyzer using word cooccurrence -Jtag. In: Proceedings of International Conference on Computational Linguistics, pp. 409–413 (1998)
6. Graf, H.P., Cosatto, E., Strom, V., Huang, F.J.: Visual prosody: facial movements accompanying speech. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp. 381–386 (2002)
7. Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing. In: Proceedings of International Conference on Computational Linguistics, pp. 928–939 (2014)

8. Ishi, C.T., Haas, J., Wilbers, F.P., Ishiguro, H., Hagita, N.: Analysis of head motions and speech, and head motion control in an android. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 548–553 (2007)
9. Ishi, C.T., Ishiguro, H., Hagita, N.: Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. In: Proceedings of International Conference of Speech and Language, pp. 2006–2009 (2006)
10. Ishi, C.T., Ishiguro, H., Hagita, N.: Head motion during dialogue speech and nod timing control in humanoid robots. In: Proceedings of ACM/IEEE International Conference on Human-Robot Interaction, pp. 293–300 (2010)
11. Ishii, R., Kumano, S., Otsuka, K.: Multimodal fusion using respiration and gaze behavior for predicting next speaker in multi-party meetings. In: Proceedings of the International Conference on Multimodal Interaction (ICMI 2015), pp. 99–106 (2015)
12. Ishii, R., Kumano, S., Otsuka, K.: Predicting next speaker using head movement in multi-party meetings. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015), pp. 2319–2323 (2015)
13. Ishii, R., Kumano, S., Otsuka, K.: Analyzing gaze behavior during turn-taking for estimating empathy skill level. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017), pp. 365–373 (2017)
14. Ishii, R., Kumano, S., Otsuka, K.: Prediction of next-utterance timing using head movement in multi-party meetings. In: Proceedings of the 5th International Conference on Human Agent Interaction (HAI 2017), pp. 181–187 (2017)
15. Ishii, R., Miyajima, T., Fujita, K., Nakano, Y.: Avatar's Gaze Control to Facilitate Conversational Turn-Taking in Virtual-Space Multi-user Voice Chat System. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, p. 458. Springer, Heidelberg (2006). https://doi.org/10.1007/11821830_47
16. Ishii, R., Nakano, Y.I.: Estimating user's conversational engagement based on gaze behaviors. In: Prendinger, H., Lester, J., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 200–207. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85483-8_20
17. Ishii, R., Nakano, Y.I.: An empirical study of eye-gaze behaviors: towards the estimation of conversational engagement in human-agent communication. In: Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction (EGIHMI 2010), pp. 33–40 (2010)
18. Ishii, R., Nakano, Y.I., Nishida, T.: Gaze awareness in conversational agents: estimating a user's conversational engagement from eye gaze. ACM Trans. Interact. Intell. Syst. **3**(2), 1–25 (2013). Article No. 11
19. Ishii, R., Otsuka, K., Kumano, S., Yamamoto, J.: Predicting of who will be the next speaker and when using gaze behavior in multiparty meetings. ACM Trans. Interact. Intell. Syst. **6**(1), 4 (2016)
20. Ishii, R., Otsuka, K., Kumano, S., Yamamoto, J.: Using respiration to predict who will speak next and when in multiparty meetings. ACM Trans. Interact. Intell. Syst. **6**(2), 20 (2016)
21. Ishii, R., Otsuka, K., Kumano, S., Matsuda, M., Yamato, J.: Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In: Proceedings of the International Conference on Multimodal Interaction (ICMI 2013), pp. 79–86 (2013)

22. Ishii, R., Otsuka, K., Kumano, S., J., Yamato, S.: Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 694–698 (2014)
23. Ishii, R., Shinohara, Y., Nakano, Y., Nishida, T.: Combining multiple types of eye-gaze information to predict user's conversational engagement. In: Proceedings of the 2011 Workshop on Eye Gaze in Intelligent Human Machine Interaction (EGIHMI 2011) (2011)
24. Iwano, Y., Kageyama, S., Morikawa, E., Nakazato, S., Shirai, K.: Analysis of head movements and its role in spoken dialogue. In: Proceedings of International Conference on Spoken Language, pp. 2167–2170 (1996)
25. Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol. Sci. **15**(2), 133–137 (2004)
26. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y.: An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. Lang. Speech **41**, 295–321 (1998)
27. Lohse, M., Rothuis, R., Gallego-Pérez, J., Karreman, D.E., Evers, V.: Robot gestures make difficult tasks easier: the impact of gestures on perceived workload and task performance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2014), pp. 1459–1466 (2014)
28. McBreen, H.M., Jack, M.A.: Evaluating humanoid synthetic agents in e-retail applications. IEEE Trans. Syst. Man Cybern. A Syst. Hum. **31**, 5 (2001)
29. Meguro, T., Higashinaka, R., Minami, Y., Dohsaka, K.: Controlling listening-oriented dialogue using partially observable Markov decision processes. In: Proceedings of International Conference on Computational Linguistics, pp. 761–769 (2010)
30. Nakano, Y.I., Ishii, R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI 2010), pp. 139–148 (2010)
31. Ooko, R., Ishii, R., Nakano, Y.I.: Estimating a user's conversational engagement based on head pose information. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS (LNAI), vol. 6895, pp. 262–268. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23974-8_29
32. Sakai, K., Ishi, C.T., Minato, T., Ishiguro, H.: Online speechdriven head motion generating system and evaluation on a tele-operated robot. In: Proceedings of IEEE International Symposium on Robot and Human Interactive Communication, pp. 529–534 (2015)
33. Maynard, S.: Interactional functions of a nonverbal sign: head movement in Japanese dyadic casual conversation. J. Pragmat. **11**, 589–606 (1987)
34. Maynard, S.: Japanese Conversation: Self-contextualization Through Structure and Interactional Management. Ablex Publishing Corporation, Norwood (1989)
35. Watanabe, T., Danbara, R., Okubo, M.: Effects of a speech-driven embodied interactive actor interactor on talker's speech characteristics. In: Proceedings of IEEE International Workshop on Robot-Human Interactive Communication, pp. 211–216 (2003)
36. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: a professional framework for multimodality research. In: Proceedings of International Conference on Language Resources and Evaluation (2006)
37. Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E.: Linking facial animation, head motion and speech acoustics. J. Phonetics **30**(3), 555–568 (2002)