マルチモーダル情報に基づく聞き手のバックチャネルの種類推定の基礎検討

大西 俊輝[1] 木下 峻一[1] 東 直輝[2] 石井 亮[3] 深山 篤[3] 中村 高雄[3] 宮田 章裕[2,a)]

概要：対話において，「うんうん」，「はい」，「すごい」などの聞き手の反応を表すバックチャネルは，重要な要素の一つである．これまで，発話や発話前後におけるマルチモーダル情報からバックチャネルの発生を予測する研究が多く行われている．しかし，発話を含んだバックチャネルの種類を推定する取り組みは行われていない．バックチャネルには発話に関する多くの表現や言い回しが含まれるため，マルチモーダル情報からバックチャネルを分類できるようになると，ユーザとの対話内容や振舞いに応じて適切なバックチャネルを行う対話システムの実現が期待できる．そこで本稿では，発話を含んだバックチャネルを言語的・機能的側面から体系的に分類する方法に基づいて，対話におけるマルチモーダル情報からバックチャネルの種類を推定することができるのか明らかにするための初期検討を行う．

An Initial Study of Predicting Listener's Backchannel Type Based on Multimodal Information

1. はじめに

対話において，「うんうん」，「はい」，「すごい」などの聞き手の反応を表すバックチャネルは，重要な要素の一つである．これまで，対話システムの構築を行うために，様々な手法でバックチャネルを予測する研究が多く行われている [1], [2], [3], [4]．これらの研究事例では，発話や発話前後におけるマルチモーダル情報からバックチャネルの発生を予測するタスクを行っている．日本語の対話における発話を含んだバックチャネルは豊富であり，それを表現する言葉も数多く存在している [5]．これより，日本語が対象となるバックチャネルの分析では，機能的な側面に着目してバックチャネルを分類する研究が多く行われている [6]．しかし，バックチャネルを予測する研究事例の中で，バックチャネルの機能的な側面から分類する取り組みは行われていない．

上記を踏まえ，我々はマルチモーダル情報から工学的にバックチャネルを分類できるようにする取り組みを行う．Morikawa らは，発話におけるバックチャネルを言語的・機能的側面から体系的に分類する方法を提案し，対話行為（DA）と呼ばれる話者の発話の意図・種類との関係を明らかにしている [7]．しかし，バックチャネルの種類と DA の関係を明らかにする取り組みにとどまっており，対話におけるマルチモーダル情報からバックチャネルの種類を予測する取り組みは行われていない．そこで本稿では，対話におけるマルチモーダル情報からバックチャネルの種類を推定することができるのか明らかにするための初期検討を行う．

[1] 日本大学大学院総合基礎科学研究科
[2] 日本大学文理学部
[3] 日本電信電話株式会社 NTT 人間情報研究所
[a)] miyata.akihiro@acm.org

2. コーパス

2.1 2 者対話データについて

本研究では，対話コーパスを構築するために，対面における 2 者対話データ [8] を利用する．2 者対話の参加者は，合計で 26 名（異なるペアを 13 組）であり，初対面の日本人男女である．発話を含んだバックチャネルのデータをより多く収集するため，一方の参加者（話し手）が他方の参加者（聞き手）にアニメ「トムとジェリー」の内容を説明するタスクを行っている．発話の単位は Inter-pausal units (IPU)[9] を用いており，沈黙時間が 200ms 未満の連続した音声区間を 1 つとしている．この対話データでは，合計 7,805 件（話し手：4,940 件，聞き手：2,865 件）の IPU が

記録されている．

2.2 バックチャネルの種類について

本研究では，バックチャネルを分類する際に，Morikawa らが提案している 9 種類のラベルを利用する [7]．これらのラベルは，対話システムにおける適切な応答の生成と分類時の判定の揺らぎの低減が考慮されている．

- Positive：「うんうん」，「そうそう」，「それいい」，「なるほど」，「たしかに」など話し手への肯定的な応答．
- Neutral：「うん」，「はい」，「おお」など話し手への感情を含まない応答．
- Non-positive：「うーん」，「ふーん」，「はーん」，「あー」，「へー」，「んー」など話し手への否定的または悩んでいるような応答．
- Emotional word：「すごい」，「ふふ」，「あぁ」，「へぇ」，その他短い感嘆詞など感情の動きを表しているような応答．
- Confirmation：「えっ」，「はっ」，「あっ」，「なんで」など確認を促す，質問するような応答．
- Repetition of Speaker's utterance word：話し手の発言を繰り返す応答．
- Speak a word before speaker's speaking：話し手の話題を先取りしている応答．
- Summarize speaker's speaking：話し手の話の要約、および言い換えをしているような応答．
- Others：聞き手の感想や独り言など，他に該当するラベルが無い応答．

3. 機械学習モデルの検討

本章では，対話におけるマルチモーダル情報を用いて，バックチャネルの種類を推定するための機械学習モデルの検討を行う．

3.1 特徴量抽出について

本研究では，2.1 節の対話データからマルチモーダル情報を特徴量として抽出する．このとき，発話を含んだバックチャネルを分析の対象とするため，聞き手の発話とその前後を含んだ範囲を特徴量を抽出する範囲とする．多くの既存研究では，推定性能を向上させるために ResNet-50[10], VGGish[11], BERT[12] などを利用し，高次元のベクトルに変換した視覚的，韻律的，言語的情報を特徴量として扱っている [2]．一方で，各モダリティの解釈性を担保するために，Action Units[13]，MFCC[14], LIWC[15] など解釈可能な視覚的，韻律的，言語的情報を特徴量として扱っている研究もある [16], [17]．そこで本研究では，視覚的，韻律的，言語的特徴量を高次元のベクトルに変換して抽出する方法と解釈可能なように抽出する方法の 2 つの方法から抽出することを検討している．

3.2 機械学習モデルの構築について

3.1 節で抽出した特徴量を説明変数，2.2 節で付与したラベルを目的変数とし，各発話におけるバックチャネルの種類を推定する機械学習モデルの構築を検討している．既存研究より，視覚的，韻律的，言語的情報に関する特徴量から，バックチャネルの発生の予測が可能であることが示唆されている [2]．さらに，情報提供や自己開示などの DA では，Positive，Neutral，Emotional word に関する反応が得られ，Non-positive などに関する反応は少ない傾向が確認されている [7]．これらより，視覚的，韻律的，言語的情報に関する特徴量からバックチャネルの種類を推定することができるのではないかと考えられる．

4. おわりに

本稿では，発話を含んだバックチャネルを言語的・機能的側面から体系的に分類する方法に基づいて，対話におけるマルチモーダル情報からバックチャネルの種類を推定することができるのか明らかにするための初期検討を示した．今後は，特徴量の抽出方法と機械学習モデルの構築方法を議論する予定である．さらに，特徴量の抽出と機械学習モデルの構築を行い，マルチモーダル情報からバックチャネルの種類を推定することができるのか明らかにしていく予定である．

参考文献

[1] Hara, K., Inoue, K., Takanashi, K. and Kawahara, T.: Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers, *Proc. 19th Annual Conference of the International Speech Communication Association (INTERSPEECH '18)*, pp. 991–995 (2018).

[2] Ishii, R., Ren, X., Muszynski, M. and Morency, L.-P.: Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling?, *Proc. 21st ACM International Conference on Intelligent Virtual Agents (IVA '21)*, pp. 131–138 (2021).

[3] Morency, L.-P., Kok, I. D. and Gratch, J.: Predicting Listener Backchannels: A Probabilistic Multimodal Approach, *International Workshop on Intelligent Virtual Agents (IVA '08)*, pp. 176–190 (2008).

[4] Huang, L., Morency, L.-P. and Gratch, J.: Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior, *Proc. 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '10)*, pp. 1265–1272 (2010).

[5] Maynard, S. K.: On back-channel behavior in Japanese and English casual conversation, *Linguistics*, Vol. 24, No. 6, pp. 1079–1108 (1986).

[6] Mukai, C.: The Use of Back-channels by Advanced Learners of Japanese: Its Qualitative and Quantitative Aspects, *Japanese language education around the globe*, Vol. 9, pp. 197–219 (1999).

[7] Morikawa, A., Ishii, R., Noto, H., Fukayama, A. and Nakamura, T.: Determining Most Suitable Listener

Backchannel Type for Speaker's Utterance, *Proc. 22nd ACM International Conference on Intelligent Virtual Agents (IVA '22)*, pp. 1–3 (2022).

[8] Ishii, R., Higashinaka, R. and Tomita., J.: Predicting Nods by using Dialogue Acts in Dialogue, *Proc. 11th International Conference on Language Resources and Evaluation (LREC '18)*, pp. 2940–2944 (2018).

[9] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y.: An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs, *Language and Speech*, Vol. 41, pp. 295–321 (1998).

[10] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proc. IEEE conference on computer vision and pattern recognition (CVPR '16)*, pp. 770–778 (2016).

[11] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J. and Wilson, K.: CNN architectures for large-scale audio classification, *2017 ieee international conference on acoustics, speech and signal processing (ICASSP '17)*, pp. 131–135 (2017).

[12] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '19)*, pp. 4171–4186 (2019).

[13] Baltrusaitis, T., Zadeh, A., Lim, Y. C. and Morency, L.-P.: OpenFace 2.0: Facial Behavior Analysis Toolkit, *13th IEEE international conference on automatic face and gesture recognition (FG '18)*, pp. 59–66 (2018).

[14] Eyben, F., Weninger, F., Gross, F. and Schuller, B.: Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor, *Proc. 21st ACM international conference on Multimedia (MM '13)*, pp. 835–838 (2013).

[15] Kahn, J. H., Tobin, R. M., Massey, A. E. and Anderson, J. A.: Measuring Emotional Expression with the Linguistic Inquiry and Word Count, *The American journal of psychology*, Vol. 120, No. 2, pp. 263–286 (2007).

[16] Onishi, T., Yamauchi, A., Ogushi, A., Ishii, R., Fukayama, A., Nakamura, T. and Miyata, A.: Modeling Japanese Praising Behavior by Analyzing Audio and Visual Behaviors, *Frontiers in Computer Science*, Vol. 4 (2022).

[17] Park, S., Shim, H., Chatterjee, M., Sagae, K. and Morency, L.: Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach, *Proc. 16th International Conference on Multimodal Interaction (ICMI'14)*, pp. 50–57 (2014).

