非言語資源としての『日本語日常会話コーパス』の活用

言語処理学会第31回年次大会 併設ワークショップ JLR2025 出島メッセ長崎

森大河（千葉大学）・伝康晴（千葉大学）

背景

- 対面のインタラクションでは言語情報と, 表情, 視線, ジェスチャーなどの非言語情報が統合されて情報や感情の伝達が行われる[1-3]
- 非言語情報が持つ意味や頻度は言語によって異なる[4]

→言語資源だけでなく, 日本語の非言語資源の充実も重要な課題

- センサーデータや身体動作のアノテーションが付与された日本語コーパスはいくつか存在する（例.『千葉大3人会話コーパス』[5], 『Hazumi』[6], 『AICO』[7]）

→実際の日常会話を対象としたものは少ない

- 豊富な言語情報のアノテーションを含む映像付き大規模日常会話コーパス『日本語日常会話コーパス』[8]に画像認識やVLLMを用いて非言語情報を付与することを計画

Introduction

- In face-to-face interactions, linguistic information is integrated with nonverbal information such as facial expressions, gaze, and gestures to facilitate the transmission of information and emotions [1-3]
- The meaning and frequency of nonverbal cues vary across languages [4]

→ Therefore, enhancing not only linguistic resources but also nonverbal resources for Japanese is an important challenge

- Several Japanese corpora include sensor data and annotations of body movements (e.g., "Chiba Three-Party Conversation Corpus" [5], "Hazumi" [6], "AICO" [7])

→ However, few of them focus on actual daily conversations

- We plan to incorporate nonverbal information into the large-scale daily conversation corpus with video and rich linguistic annotations, the "Corpus of Everyday Japanese Conversation" [8], using image recognition and VLLM

日本語日常会話コーパス

- 日常生活の中で自然に生じた多様な場面の会話を収録した映像付き会話コーパス
- 総時間：200時間（現在『子ども版日本語日常会話コーパス』[9]も作成中）
- 話者数（述べ）：1675名
- 語数（短単位）：約240万語
- アノテーション：転記テキスト, 形態論情報, 談話行為情報, 韻律情報

Corpus of Everyday Japanese Conversation

- A video-equipped conversation corpus that captures dialogues from diverse naturally occurring situations in daily life
- Total duration: 200 hours (Currently, children’s version [9] is also being developed)
- Number of speakers (total): 1,675
- Number of words (short units): Approximately 2.4 million words
- Annotations: Transcribed text, morphological information, dialogue act information, prosodic information

計画

第一段階：動画中の話者の位置座標のアノテーション

第二段階：表情と姿勢のアノテーション

第三段階：ジェスチャーのアノテーション

Plan

Phase 1: Annotation of the speaker's position coordinates in the video

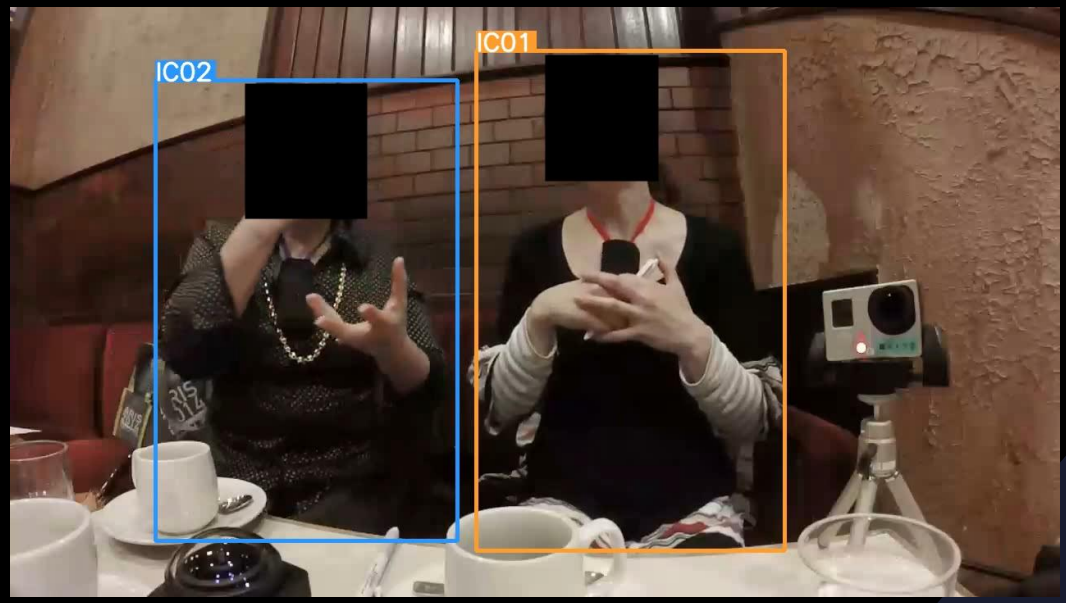
Phase 2: Annotation of facial expressions and posture

Phase 3: Annotation of gestures

進捗状況（第一段階）

- YOLOv8を用いて人物の検出とトラッキング
- トラッキング結果の修正と話者IDの紐付け

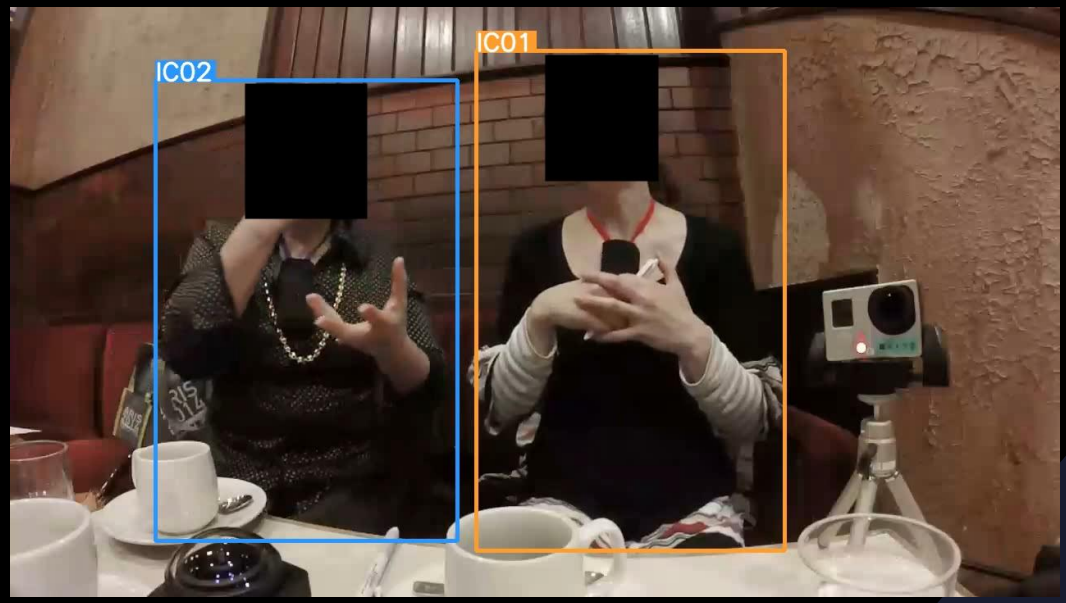
→表情や身体動作などの特徴量を抽出し，言語情報と統合して分析することが可能になる



Progress (Phase 1)

- Detection and tracking of individual speakers using YOLOv8
- Correction of tracking results and linking speaker IDs

→ It becomes possible to extract features such as facial expressions and body movements and integrate them with linguistic information for analysis



参考文献

[1] Goldin-Meadow, S. (1997). "When Gestures and Words Speak Differently." Current Directions in Psychological Science, 6(5), pp. 138-143. SAGE Publications Sage CA: Los Angeles, CA.

[2] Cassell, J., McNeill, D., and McCullough, K.-E. (1999). "Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information." Pragmatics & Cognition, 7(1), pp. 1-34. John Benjamins.

[3] de Gelder, B., Vroomen, J., de Jong, S. J., Masthoff, E. D., Trompenaars, F. J., and Hodiamont, P. (2005). "Multisensory Integration of Emotional Faces and Voices in Schizophrenics." Schizophrenia Research, 72(2-3), pp. 195-203. Elsevier.

[4] Moris, D. (1977). "Man Watching." Jonathan Gape.

[5] Den, Y. and Enomoto, M. (2007). "A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation." In Conversational Informatics: An Engineering Approach, pp. 305-330. Wiley Online Library.

[6] 駒谷和範 (2022). マルチモーダル対話コーパスの設計と公開. 日本音響学会誌, 78(5), pp. 265-270. 一般社団法人 日本音響学会.

[7] Jokinen, K. (2020). The AICO Multimodal Corpus--Data Collection and Preliminary Analyses. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 559-564.

[8] 小磯花絵, 天谷晴香, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 伝康晴, 西川賢哉, 渡邊友香 (2023). 『日本語日常会話コーパス』 設計と構築. 国立国語研究所論集, 24, pp. 153-168.

[9] 小磯花絵, 天谷晴香, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 滕越, 西川賢哉 他 (2023). 『子ども版日本語日常会話コーパス』 の構築. 言語資源ワークショップ発表論文集= Proceedings of Language Resources Workshop, 1, pp. 103-108.

Reference

[1] Goldin-Meadow, S. (1997). "When Gestures and Words Speak Differently." Current Directions in Psychological Science, 6(5), pp. 138-143. SAGE Publications Sage CA: Los Angeles, CA.

[2] Cassell, J., McNeill, D., and McCullough, K.-E. (1999). "Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information." Pragmatics & Cognition, 7(1), pp. 1-34. John Benjamins.

[3] de Gelder, B., Vroomen, J., de Jong, S. J., Masthoff, E. D., Trompenaars, F. J., and Hodiamont, P. (2005). "Multisensory Integration of Emotional Faces and Voices in Schizophrenics." Schizophrenia Research, 72(2-3), pp. 195-203. Elsevier.

[4] Moris, D. (1977). "Man Watching." Jonathan Gape.

[5] Den, Y. and Enomoto, M. (2007). "A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation." In Conversational Informatics: An Engineering Approach, pp. 305-330. Wiley Online Library.

[6] 駒谷和範 (2022). マルチモーダル対話コーパスの設計と公開. 日本音響学会誌, 78(5), pp. 265-270. 一般社団法人 日本音響学会.

[7] Jokinen, K. (2020). The AICO Multimodal Corpus--Data Collection and Preliminary Analyses. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 559-564.

[8] 小磯花絵, 天谷晴香, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 伝康晴, 西川賢哉, 渡邊友香 (2023). 『日本語日常会話コーパス』 設計と構築. 国立国語研究所論集, 24, pp. 153-168.

[9] 小磯花絵, 天谷晴香, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 滕越, 西川賢哉 他 (2023). 『子ども版日本語日常会話コーパス』 の構築. 言語資源ワークショップ発表論文集= Proceedings of Language Resources Workshop, 1, pp. 103-108.