気の利く対話AIのための「空気を読む」技術

～マルチモーダル情報を用いた対話の場・関係の理解とインクリメンタル応答生成～

Techniques for "reading the room" in attentive conversational AI
– Understanding dialogue context through multimodal information and incremental response generation –

NTT コミュニケーション科学基礎研究所 協創情報研究部
実世界インタラクション研究グループ 主任研究員
千葉　祐弥
Yuya Chiba

●プロフィール

2015年東北大学大学院工学研究科博士後期課程修了。博士（工学）。2016年から2020年まで、東北大学大学院工学研究科助教。2020年NTT入社。研究分野は音声対話システム、マルチモーダル対話システム、対話分析など。ISCA、日本音響学会、電子情報通信学会、情報処理学会、言語処理学会各会員。

大規模言語モデルの進展により、対話システムの応答の自然さは飛躍的に向上しました。ChatGPTをはじめとする多くのAIアプリケーションでは、すでに会話による操作が一般的になっています。しかしながら、対話システムが単なるインターフェースの枠を超え、日常生活の中で人々と自然に会話し、人間と共生するパートナーとなるためにはまだ多くの課題が残っています。例えば、今その場で起こっている対話の性質や、対話に関与する人々の人間関係が正確に把握できなければ、システムは突然脈略のない発話をしたり、目上の人に失礼な発言をしたりするなど、対話の状況や相手の気持ちを無視したふるまいをしてしまいます。また、人間同士の素早い対話の間合いに巧みに追随して、対話の文脈に合致した適切な応答を行うことも必要です。私たちの研究グループでは、言語情報だけでなく声の調子や表情、仕草などの非言語的情報を用いて会話の特徴や話者の関係を理解する対話状況認識技術（図1）と、人間のような間合いで会話を行うための基盤技術である逐次応答生成技術（図2）に取り組んできました。本講演では、これらの技術について紹介します。

対話状況の理解（図1①）

人間同士の対話では、対話に参加する人たちの関係や会話の形態、場所に応じて特徴が異なることがわかっています[1]。2023年度のオープンハウスでは、日常会話の分析から得られた対話の特徴を説明する7つの要素を紹介しました。これらの要素に基づいて対話を分析することで、会議における会話では説明が多い、同僚との会話では丁寧な話し方が多い、友人との会話では自分の気持ちを率直に伝える発話が多いといったような、様々な対話の性質が説明できます。この分析結果に基づくと、対話システムも状況を理解することで、より場面にあった適切な会話ができるようになると期待できます。そこで私たちは、対話状況を理解する上で重要な手がかりとなる会話中の音声や動画像などのマルチモーダル情報を用いた対話状況認識に取り組みました。この手法では、音声・画像・言語情報をそれぞれ事前学習済みモデルで表現し、Transformerで統合することで対話の状況を認識します[2]。

精緻な対話参加者の関係の理解（図1②）

話者のふるまいは相手に抱く親しみの度合いによっても変化します。例えば相手に対する親しみが強いほど、より素直な価値観や意見を伝えるようになり、姿勢や声の調子などの非言語的なふるまいも相手に同調することが知られています[3]。システムがこのような人間の感情の機微をとらえてふるまいを切り替えるには、話者間・システム間の関係の精緻な認識が必要になります。私たちは、東北大学と共同で、人間同士の対話における話し手の聞き手に対する親しみを推定する技術の開発に取り組みました。この研究では、学術的に知られている対人関係に応じた話者のふるまいが現在の情報処理技術によって抽出できることを確認し、それらの情報を統合して親しみの度合いを推定します[4]。

人間の間合いに追随する応答生成（図2）

人間同士の会話のようなテンポでシステムが応答するために、どのタイミングで発話を開始するべきかを判断するターン交替予測技術が考案されています。これまでの研究では、応答タイミングの適切性が対話とは独立に評価されており、ユーザの満足度やタスク達成率のような対話体験や印象に与える影響が明らかになっていませんでした。私たちの研究グループでは大規模言語モデルを活用した汎用的な逐次応答生成技術を開発し、対話実験による評価を行いました。この結果、単に高度なターン交替手法を用いるだけではユーザ評価の向上にはつながらないことが示され、現在の音声対話システムの課題が浮き彫りになりました[5]。この過程で名古屋大学、名古屋工業大学と開発した対話システム構築ツールキット[6]は、多くの人が利用できる形で公開しており、様々な基礎研究で用いられています。

今後の展望

状況・関係認識技術を用いることで、どのような環境や相手であっても自然でスムーズな会話が成立する対話システムの実現をめざします。その要素技術として、近年の大規模基盤モデルを活用したマルチモーダル情報の効果的な統合も必要になると考えられます。逐次応答生成においては、断片的な発話からユーザの意図を正確に推定する技術を検討します。加えて、ターン交替の自然性をさらに向上させるためには、人間の対話におけるやりとりを直接システムの学習に用いることも必要になるでしょう。いずれの方法においても、対話文脈の正確な理解とそれに応じた応答生成が課題になります。今後も、このような技術を取り入れ、より人間らしく、様々な場面で適切なふるまいのできる対話システムの研究に取り組んでいきます。

図１：対話状況・話者関係の認識

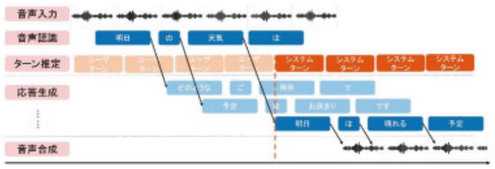

図２：逐次応答生成

●参考文献

[1] Y. Chiba, R. Higashinaka, "Analyzing variations of everyday Japanese conversations based on semantic labels of functional expressions," ACM Transactions on Asian and Low-Resource Language Information Processing (ACM TALLIP), Vol. 22, No. 2, pp. 1-26, 2023.
[2] Y. Chiba, R. Higashinaka, "Dialogue situation recognition in everyday conversation from audio, visual, and linguistic information," IEEE Access, pp. 70819-70832, 2023.
[3] 辻幸夫, 菅井三実, 佐治伸郎編, "第4章 ことばと対話の多層性," inことばのやりとり（シリーズ〈ことばの認知科学〉), 朝倉書店, 2024, pp.76-94.
[4] Y. Chiba, A. Ito, "Speaker intimacy estimation in chat-talks based on verbal and non-verbal information," IEEE Access, pp. 184592-184606, 2024.
[5] Y. Chiba, R. Higashinaka, "Investigating the impact of incremental processing and voice activity projection on spoken dialogue systems," in Proc. International Conference on Computational Linguistics (COLING), pp. 3687-3696, 2025.
[6] 東中竜一郎, 光田航, 千葉祐弥, 李晃伸, Pythonと大規模言語モデルで作るリアルタイムマルチモーダル対話システム, 科学情報出版, 2024.