Tabidachi: 旅行代理店タスク対話コーパス

稲葉通将 [1]　千葉祐弥 [2]　斉志揚 [1]　東中竜一郎 [3]
駒谷和範 [4]　宮尾祐介 [5]　長井隆行 [4]
[1] 電気通信大学　[2] 日本電信電話株式会社
[3] 名古屋大学　[4] 大阪大学　[5] 東京大学
m-inaba@uec.ac.jp　yuya.chiba@ntt.com　qizhiyang@uec.ac.jp
higashinaka@i.nagoya-u.ac.jp　komatani@sanken.osaka-u.ac.jp
yusuke@is.s.u-tokyo.ac.jp　nagai@sys.es.osaka-u.ac.jp

概要

人は他者とコミュニケーションを行う際，使用語彙や話す速さ，表情やボディランゲージなどを相手に応じて使い分けている．しかし，現在の対話システムがユーザに応じて話し方や対話戦略を変更することはほとんどない．対話システムがより効率的にタスクを達成したり，ユーザの満足度をより高めるためには，ユーザに応じて対話戦略などを変更できることが望ましい．そこで我々は話し方の変化に大きく影響を与える要素として話者の年齢に着目し，幅広い年齢層の話者によるマルチモーダル対話コーパス Tabidachi を構築した．本コーパスは国立情報学研究所情報学研究データリポジトリ（IDR）[1)] にて公開している．

1　はじめに

タスク指向型対話システムは対話システム研究において常に活発な研究対象である [1, 2, 3, 4]．近年ではニューラルネットワークを用いた応答生成 [5, 6, 7] や対話状態追跡 [8, 9, 10] に関する研究が活発に行われている．これらの研究はいずれもユーザの入力に対し，適切な応答を返すことにフォーカスしている．

人は他者とコミュニケーションを行う際，使用語彙や話す速さ，表情やボディランゲージなどを相手に応じて使い分けている [11]．例えば，子供相手であればより簡単な語彙を使用し，感情を込めて話すかもしれないし，高齢者相手であればゆっくりと話すかもしれない．しかし，ユーザに応じて話し方や対話戦略を変更する対話システムは少数であり，システムがより効率的にタスクを達成したり，ユーザの満足度をより高めるためには，ユーザに応じて対話戦略などを変更できることが望ましい．

図 1　旅行代理店タスク対話コーパス Tabidachi

そこで本研究では，話し方の変化に大きく影響を与える要素として話者の年齢に着目し，児童から高齢者まで幅広い年齢層の話者によるマルチモーダル対話コーパス Tabidachi を構築した (図 1)．

収集する対話として，幅広い年代が興味を持つ旅行に着目し，旅行代理店におけるオペレータとカスタマーによる 2 者間の観光相談を元にした旅行代理店タスクを設定した．オペレータは対話中に観光情報検索システムを使用し，観光地に関する情報を取得することができる．その際，システムにどのタイミングでどのようなクエリを投げ，検索結果が返ってきたかという情報も収集した．また，人手による音声書き起こしを実施し，対話行為タグおよび検索

1)　https://www.nii.ac.jp/dsc/idr/rdata/Tabidachi/

システムのログと発話を対応付けるアノテーションを実施した．本稿では Tabidachi の構築方法およびアノテーションの詳細について述べる．

2 関連研究

本研究と同じく，2 者による対話を収録したマルチモーダル対話コーパスはこれまでに複数構築されている．Cardiff Conversation Database (CCDb) [12] は聞き手や話し手などの役割やシナリオの無い自由な対話を収録した動画が収録されている．また，そのうちの一部の動画に対して対話行為・感情・頭の動きに関するアノテーションが付与されている．会話は 1 回 5 分，合計 300 分の対話が含まれており，参加者の年齢は 25 歳から 57 歳である．The Emotional Dyadic Motion CAPture (IEMOCAP) データセット [13] は，音声とジェスチャーの分析のために収集されたコーパスであり，10 人の話者が顔，頭，手にマーカーをつけて対話を行った様子を収録している．合計収録時間は約 12 時間である．NoXi コーパス [14] は主に英語，フランス語，ドイツ語での対話を収録している．また，頭の動き，笑顔，視線，エンゲージメントなどのアノテーションが付与されている．合計収録時間は約 25 時間であり，参加者の年齢は 21 から 50 歳である．CANDOR コーパス [15] は合計 850 時間のマルチモーダル対話コーパスであり，本研究で収集したコーパスよりも長時間の対話を収録している．一方で対話の書き起こしは自動で行われたものが付与されている．参加者の年齢は 19 歳から 66 歳である．Hazumi [16] は日本語のマルチモーダル対話コーパスである．WOZ (Wizard of Oz) 方式で操作されるシステムと参加者による合計約 65 時間の対話を収録しており，20 代から 70 代までの 214 名が参加した．上記のコーパスと我々が構築した Tabidachi の最大の違いは参加者の年齢層の広さ (7〜72 歳) であり，特に 10 歳未満の未成年が参加している対話コーパスは希少である．また，人手で音声書き起こし・アノテーションが行われたコーパスの中では最大規模のコーパスである．

3 旅行代理店タスク

収集した対話は，オペレータ（店員）役の話者とカスタマー（客）役の話者の 2 名による旅行代理店における観光相談を模した対話である．対話は Zoom[2] を用いたビデオ通話により行い，1 回の対話時間は 20 分である．カスタマー役の話者は対話前に旅行の状況設定 (3.1) を作成し，その状況設定に基づいてオペレータと相談しながら行きたい観光地を決定する．オペレータ役の話者はカスタマー役の話者から要望を聞き出すとともに，観光情報検索システム (3.2) を活用して観光地の推薦を行う．

3.1 旅行の状況設定

対話前にカスタマー役の話者は相談する旅行の状況設定を行う．状況は話者自身の人間関係や実際に行きたい旅行先などを考慮して各自で設定する．

本研究では，状況設定の方式が異なる 2 種類の対話シナリオを用いた．対話シナリオ 1 は具体的な状況設定を行うシナリオである．旅行先 (都道府県名もしくは地方)，時期 (春，夏，秋，冬)，人数，メンバー構成 (友人，両親など) の 4 項目をあらかじめ決定した上で対話を行い，観光情報検索システムに含まれる観光地から旅行中の行き先として 3 箇所決定する．対話シナリオ 2 は自由記述によるおおまかな状況設定を行うシナリオである．どういう旅行がしたいかについて 20 文字程度で記述したものを状況設定とし，対話中に観光情報検索システムに含まれる観光地から旅行中の行き先を 1 件以上決定する．シナリオ 2 の状況設定の例としては「温泉地でゆっくりしたい」「昼は神社仏閣を巡って，夜は名物料理を食べたい」などである．

3.2 観光情報検索システム

オペレータ役の話者が操作する観光情報検索システムは，株式会社 JTB パブリッシングが提供する「るるぶ DATA」[3] を用いて独自に構築した．るるぶ DATA には約 45,000 件の国内の観光地に関する情報が収録されている．

システムのスクリーンショットを図 2 に示す．システムの画面の左側で検索条件が指定でき，地域・エリアの設定，自由記述のキーワード検索，ジャンルに基づく検索（例：「見る - 建物・史跡 - 歴史的建造物」，「食べる - 外国料理 - フランス料理」），予算や喫煙の可否などによる絞り込みが可能である．画面の右側には検索結果が表示され，観光地の説明文，地図，画像，住所，アクセス方法などの情報が閲覧できる．なお，観光情報検索システムから得た情報を活用した対話とするため，オペレータ役の話者に対し，自身の記憶に基づく観光地情報の提供は

2） https://zoom.us/

3） https://solution.jtbpublishing.co.jp/service/domestic/

図 2　観光情報検索システム

できる限り行わず，システムの検索結果を用いた情報提供を行うように指示した．

4　データ収集

4.1　対話の収録

対話の収録は 2020 年 11 月 10 日から 2021 年 2 月 25 日に実施した．なお，本収録は「電気通信大学ヒトを対象とする実験に関する倫理委員会」で承認を受けた上で実施した（管理番号：19061(2) 号）．

カスタマー役の話者は未成年 20 名 (7〜17 歳)，一般 25 名 (21〜59 歳)，高齢者 10 名 (65〜72 歳) の合計 55 名の男女である．カスタマー役の話者は 1 名につき 6 回の対話を行った．すなわち，収集した対話数は $55 \times 6 = 330$ 対話である．前半の 3 回の対話はそれぞれの Web カメラの動画を表示したギャラリービューによる対話であり，後半の 3 回は観光情報検索システムの画面を共有しながら実施した．旅行の状況設定の対話シナリオはシナリオ 1 →シナリオ 1 →シナリオ 2 の順で 2 回繰り返した．

オペレータ役の話者は 5 名であり，そのうち 3 名は旅行代理店における窓口業務経験者，残りの 2 名は接客業経験者である．旅行代理店における窓口業務経験者 3 名は全体の 78.2%の対話を担当した．

5　アノテーション

収集した対話は人手による音声書き起こしを実施した．書き起こした対話に対しては，以下の 4 種類のアノテーションを実施した．

1. ISO 24617-2 で定義された対話行為タグ (汎用対話行為タグ)
2. 独自に定義した旅行代理店タスク専用の対話行為タグ (専用対話行為タグ)
3. 話者が言及した観光地とシステム上の観光地 ID との対応付け（言及アノテーション）
4. オペレータが観光情報検索システムで発行したクエリと発話の対応付け（クエリアノテーション）

表 1　汎用対話行為タグの一部

タグ名	説明
Inform	情報提供
Stalling	フィラー
SetQuestion	5W1H 質問
ChoiceQuestion	選択肢から選ばせる質問
Answer	質問に対する回答
Agreement	相手の意見や確認に対する同意

すべてのアノテーションは ISO 24617-2 で定義されたセグメント (Functional segment) と呼ばれる発話よりも小さな単位に対して行った．セグメントはコミュニケーション機能を持つ最小限の範囲と定義される．そこで，アノテーションに先立ち，書き起こした発話を人手でセグメントへ分割した．

5.1　汎用対話行為タグ

汎用対話行為タグとして ISO 24617-2 annotation scheme [17] で定義された対話行為タグのサブセットを用いた．ISO 24617-2 annotation scheme では，タグは 9 つの dimension に分類されており，異なる dimension のタグであれば，1 つのセグメントに対して重複して付与することができる．本研究では他の多くの対話行為タグで行われているように 1 つのセグメントに対して 1 つのタグのみを付与できるようにするため，4 つの dimension の 12 種類のタグを除外した．最終的に，書き起こしの際に聞き取りが困難であったセグメントに主に付与する Other タグを加えた，合計 44 タグを用いた．なお，本研究で除外したタグを追加で付与することで，ISO 24617-2 に準拠したアノテーションにすることも可能である．表 1 にタグの一部を示す．

5.2　専用対話行為タグ

本コーパス用に独自に対話行為タグを設計し，付与した．設計時には，セグメントを内容ごと，および検索システムの検索条件や検索結果に含まれる情報に関連付けてグルーピングして仮のタグセットを

表 2　専用対話行為タグの一部

タグ名	説明
NameInform	観光地の名称を伝える情報提供
SearchResultInform	検索結果に関する情報提供
RequestQuestion	要望を尋ねる質問
PeopleQuestion	旅行人数を尋ねる質問
SpotRequirement	観光地に関する要望
SpotImpression	観光地に対する印象

作成し，そのタグセットで一貫したアノテーションができるかを評価した．そして評価に基づいてタグセットの改良を繰り返すことで合計 37 種類のタグを設計した．表 2 にタグの一部を示す．

5.3　言及アノテーション

オペレータは観光情報検索システムに表示された複数の観光地の中から，カスタマーに適した観光地を選び，推薦する．その際，どの観光地が選ばれたかについては明確には示されない．そこでオペレータのセグメントに対し，システムに含まれるどの観光地に言及したかという情報を付与した．観光地には一意な ID が付与されており，本アノテーションでは観光地に言及したセグメントに観光地 ID を付与した．1 セグメントにおいて複数の観光地に言及している場合は，複数の観光地 ID を付与した．

加えて，観光地に言及する際，どのようにシステムを活用したかという言及レベル (レベル 0, 1, 2) についても観光地ごとに付与した．この情報により，オペレータが候補として言及したのか，それとも推薦対象として言及したのかという点が明確になる．言及レベル 2 は「営業開始は 10 時からのようです.」のように，観光情報検索システムに表示された観光地の詳細な情報に基づいた内容であることを意味する．言及レベル 1 は「北海道植物園というのがあります」のように表示された観光地の詳細な情報を含まず，検索結果に基づいた内容であることを意味する．言及レベル 0 は主に対話の最後に行われることが多いオペレータによる対話の振り返りなど，発話時にはシステム上に表示されていない観光地に言及した場合に付与される．観光情報検索システムでは，クエリを発出すると 10 件の観光地名と写真が表示され，各観光地の詳細な情報は「Read more」ボタンをクリックすることで閲覧可能となる．システムの操作ログには，どのタイミングで「Read more」ボタンをクリックしたのかという情報も含まれており，アノテータは操作ログと発話の内容から言及レベルを判断した．

表 3　コーパスの統計情報

対話数	330
収録時間 (分)	6,948
発話数	120,140
- オペレータ発話数	66,224
- カスタマー発話数	53,916
セグメント数	245,543
- オペレータセグメント数	152,500
- カスタマーセグメント数	93,043
観光地 ID 付与セグメント数	43,768
観光地 ID 数	46,330
- 言及レベル 0	2,948
- 言及レベル 1	24,124
- 言及レベル 2	19,258
クエリアノテーションセグメント数	5,164

5.4　クエリアノテーション

オペレータは対話中に任意のタイミングで観光情報検索システムを操作することができ，データ収集時にはオペレータの操作ログと表示された内容を収集した．本アノテーションでは，システムへの検索クエリに着目し，オペレータがどのセグメントに基づいてクエリを決定したかという情報を付与した．

5.5　統計情報

データ収集，およびアノテーションの結果と統計情報を表 3 に示す．1 回の対話時間は 20 分であるが，収録では自動的に対話を打ち切っているわけではない．そのため $330 \times 20 = 6,600$ よりも 5%ほど長くなっている．またオペレータ役の話者が対話を主導することから，オペレータのセグメント数はカスタマーに比べて 1.5 倍ほど多いことが確認できる．

6　おわりに

本稿では旅行代理店タスク対話コーパス Tabidahci の構築方法について述べた．本コーパスは Zoom のビデオ会議を使用して収録した合計 115 時間のマルチモーダル対話コーパスである．すべての対話は人手で書き起こしされており，4 種類のアノテーションを実施した．Tabidachi は年齢層ごとの対話行為の頻度の違いや表情の分析にも利用されている [18].

謝辞

本研究は JSPS 科研費 19H05692 の助成を受けたものです.

参考文献

[1] Daniel G Bobrow, Ronald M Kaplan, Martin Kay, Donald A Norman, Henry Thompson, and Terry Winograd. GUS, a frame-driven dialog system. **Artificial intelligence**, Vol. 8, No. 2, pp. 155–173, 1977.

[2] Victor Zue, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff. Integration of speech recognition and natural language processing in the MIT VOYAGER system. In **IEEE International Conference on Acoustics, Speech, and Signal Processing**, pp. 713–716, 1991.

[3] Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. Let's go public! taking a spoken dialog system to the real world. In **Proceedings of the Interspeech**, 2005.

[4] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. MultiWOZ-a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 5016–5026, 2018.

[5] TH Wen, D Vandyke, N Mrkšíc, M Gašíc, LM Rojas-Barahona, PH Su, S Ultes, and S Young. A network-based end-to-end trainable task-oriented dialogue system. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics**, Vol. 1, pp. 438–449, 2017.

[6] Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3696–3709, 2019.

[7] Yichi Zhang, Zhijian Ou, and Zhou Yu. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 9604–9611, 2020.

[8] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. Neural belief tracker: Data-driven dialogue state tracking. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1777–1788, 2017.

[9] Victor Zhong, Caiming Xiong, and Richard Socher. Global-locally self-attentive encoder for dialogue state tracking. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1458–1467, 2018.

[10] Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 7521–7528, 2020.

[11] Howard Giles, Tania Ogay, et al. Communication accommodation theory. **Explaining communication: Contemporary theories and exemplars**, pp. 293–310, 2007.

[12] Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. Cardiff Conversation Database (CCDb): A database of natural dyadic conversations. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops**, pp. 277–282, 2013.

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. **Language resources and evaluation**, Vol. 42, No. 4, pp. 335–359, 2008.

[14] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The NoXi database: multimodal recordings of mediated novice-expert interactions. In **Proceedings of the 19th ACM International Conference on Multimodal Interaction**, pp. 350–359, 2017.

[15] Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. **Science Advances**, Vol. 9, No. 13, p. eadf3197, 2023.

[16] Kazunori Komatani and Shogo Okada. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In **Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII)**, pp. 1–8, 2021.

[17] Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. Dialogue act annotation with the ISO 24617-2 standard. In **Multimodal interaction with W3C standards**, pp. 109–135. Springer, 2017.

[18] Michimasa Inaba, Yuya Chiba, Zhiyang Qi, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers. Vol. 23, No. 9, 2024.

A 付録

表 4 対話例 (未成年カスタマー)

話者	発話	汎用 / 専用対話行為タグ
オペレータ	んとねー、	Stalling / None
	市場の周りとかにー、お寿司とか食べられるお店がたくさんあるんだけどー。	Inform / IntroductionInform
カスタマー	はい。	AutoPositive / None
オペレータ	どんなお寿司が食べたいとかってある？	PropositionalQuestion / RequestQuestion
カスタマー	えっと、	Stalling / None
	いくら丼みたいにいくらがたっぷーりあるお寿司屋さん。	Answer / SpotRequirement
オペレータ	あ、	Stalling / None
	いくらがたくさん乗ってるお寿司屋さん？	CheckQuestion / RequestConfirm
カスタマー	はい。	Answer / None
オペレータ	うん、	AutoPositive / None
	じゃ、それを探してみるね。	Inform / SearchInform

表 5 対話例 (高齢者カスタマー)

話者	発話	汎用 / 専用対話行為タグ
オペレータ	やっぱり歩くのはつらいですかね。	PropositionalQuestion / None
カスタマー	うーん、	AutoPositive / None
	やっぱり歳と共にちょっと膝もきてるのであまり、若いころは山に登ったりするのも好きだったけど。	Answer / None
オペレータ	あー、	Stalling / None
	色々やはり	CheckQuestion / SearchResultInform
	あの、	Stalling / None
	歌舞伎屋根の家とかやっぱあるんですけども、いかがですかね、	CheckQuestion / SearchResultInform
	なんか	Stalling / None
	ミュージアムみたいなところもあるんですけども、そういった古い	Inform / SearchResultInform
	あの	Stalling / None
	家並みを展示しているみたいなんですけども。	Inform / SearchResultInform

表 6 言及アノテーション例

話者	発話	観光地 ID	言及レベル
オペレータ	えー、		
	カニと道産料理雪華亭というお店となっております。	80000202	1
カスタマー	はい		
オペレータ	こちらが、純和風の店内、庭を眺めながら、カニの懐石料理、料理や道産料理が堪能できるお店となっております。	80000202	2

表 7 クエリアノテーション例

話者	発話	クエリ
オペレータ	お客様、どのようなご旅行をお考えでしょうか？	
カスタマー	えーと、	
	京都に。	方面=近畿, 県=京都
オペレータ	はい。	
カスタマー	あの、	
	紅葉のときに行きたいんですけど。	季節=秋
オペレータ	はい。	