多面的なユーザ意欲を考慮した セールス対話データセットおよび対話システムの構築と評価

邊土名 朝飛 [1] 馬場 惇 [1] 佐藤 志貴 [1] 赤間 怜奈 [2]
[1] サイバーエージェント [2] 東北大学
{hentona_asahi,baba_jun,sato_shiki}@cyberagent.co.jp
akama@tohoku.ac.jp

概要

購買意欲を向上させるセールス対話システムを実現するためには多面的なユーザの意欲を考慮したデータセットが必要だが、既存データセットにはシステムの想定運用環境で収集された信頼性の高いユーザの意欲に関するデータが含まれていない。本研究では、想定運用環境に基づいた対話データ収集環境を開発し、3 種類のユーザ意欲データを含む日本語セールス対話データセットを構築した。ユーザ評価実験では、発話レベルでユーザの意欲を考慮し、さらにデータセット分析で得られたセールス対話戦略の知見を組み込むことが対話システムによる対話成功率の向上につながることが示唆された。
github.com/CyberAgentAILab/salestalk-dataset

1 はじめに

購買意欲を向上させるセールス対話システムを実現するためには、(1) 対話継続意欲、(2) 情報提供意欲、(3) 目標受容意欲の少なくとも 3 種類のユーザ意欲を考慮しつつ対話を進行することが重要だと考えられる。そのようなシステムの開発には、生態学的妥当性 (ecological validity) [1, 2] の高いセールス対話データセットの整備が必要不可欠である。生態学的妥当性とは、現実世界へのユーザへの実験結果の適用可能性を意味し、主に心理学や Human-Computer-Interaction の分野において用いられる概念である。しかし、既存のセールス対話データセット [3, 4] は、人間同士の対話設定であったり、販売対象となる商品が事前に固定されているなど、実際のセールス対話システムの想定運用環境とは乖離した設定でデータが収集されているため生態学的妥当性が高いとはいえない。

本研究では、実用的なセールストーク対話システムの実現を目指し、ユーザの多面的な意欲を考慮した日本語セールス対話データセットを構築する。データセットの生態学的妥当性を高めるため、実際のセールス対話システムの想定運用環境を可能な限り再現した実験設定下で、自然なユーザの対話および意欲データを収集することを試みる。具体的には、ユーザの自然なエンゲージメントを可能な限り正確に計測するため、実験参加者であるユーザ役に対して、任意のタイミングで対話を離脱することを許可し、対話システムのふりをしたセールス経験者（セールス役）がユーザ役の対話相手となる Wizard-of-Oz (WOZ) 法 [5] の設定で対話データを収集する。さらに、ユーザの意欲の評価をユーザ役自身に行ってもらい、かつ対話単位のみならず発話単位での評価も収集する。本データセットにより、ユーザの購買意欲を高めるセールス対話戦略のきめ細やかな分析や、ユーザの反応に応じて柔軟に対話戦略を切り替える効果的なセールス対話システムの開発につながることが期待される。

本稿では、構築したデータセットを概説したうえで、本データセット上でユーザの購買意欲を向上させるセールス対話戦略を分析した結果を報告する。さらに、分析により得られた対話戦略を組み込んだ大規模言語モデルベースのセールス対話システムが、ユーザ評価実験において高い評価を得ることを通して，本データセットの有用性を示す。

2 データセット概要

2.1 データセット構築

図 1 に、セールス対話データセットの構築プロセスの概要を示す。本研究では、我々の過去の収集 [8] で用いた設定に従い、3 種類の架空のワイヤレスイヤホンの情報を掲載した Web ページに訪れた

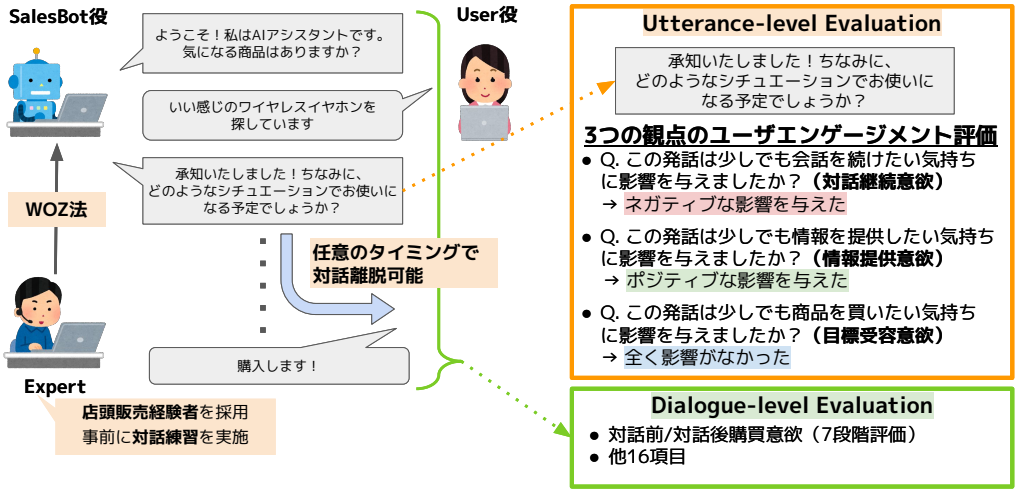


図 1: セールス対話データセット構築プロセスの概要図

表 1: セールス対話データセットの統計量。トークナイザーには MeCab[6]+UniDic[7] を使用した。

	Total	Mean	Max	Min
# dialogues	109	-	-	-
# success dialogues	63	-	-	-
# utterances	3289	30.2	92	11
— User	1144	10.5	41	3
— Sales	2145	19.7	52	7
# tokens	54301	498.2	1406	153

ユーザと、その Web ページ上に設置されたセールス対話システムがテキストチャットを行うシナリオで対話データを収集した。構築したデータセットの統計情報を表 1 に示す。

2.2 ユーザ意欲

本研究では、対話レベルのユーザ意欲データとして、対話実験の前後に 7 段階リッカート尺度評価の購買意欲データを収集した。また、発話レベルのユーザ意欲データとして (1) 対話継続意欲 (Continuing dialogue; CD)、(2) 情報提供意欲 (Providing information; PI)、(3) 目標受容意欲 (Goal-oriented acceptance; GA) の 3 種類の意欲データを収集した。対話継続意欲は、ユーザがシステムとの対話を続けたいと感じる意欲であり、主にオープンドメイン対話の分野で扱われてきた [9, 10, 11, 12]。情報提供意欲は、ユーザが自分のニーズや要望をシステムに伝える意欲を指す。この意欲を高めることで、システムはユーザの嗜好に合わせたセールス対話が可能となり、説得力とユーザ満足度が向上することが期待できる [13, 14, 15]。目標受容意欲は、ユーザがシステムの対話目標、すなわちセールス対話の最終目標である商品購入 [16] を受け入れたいと感じる意欲を指す。発話レベルのユーザ意欲は、それぞれ 3 段階 (Positive、Neutral、Negative) で評価してもらった。

3 データセット分析

上述の手順で構築した日本語セールス対話データセットを用いて、ユーザの購買意欲向上に寄与するセールス対話戦略について分析する。

3.1 対話レベル分析

各対話に占める各ユーザ意欲評価ラベルの割合と、ユーザの購買意欲の向上度合い（対話前後の 7 段階リッカート尺度評価の変化量）との間での相関分析結果を図 2 に示す。相関係数を見ると、Positive および Neutral 評価は購買意欲の向上とほぼ相関がなく、反対に Negative 評価との間には負の相関が示された。この結果から、ユーザの購入意欲を向上させるためには、各種意欲を高めるような発話を試みるよりも、対話を通じて意欲を低下させる発話を避けることが重要であることが示唆された。

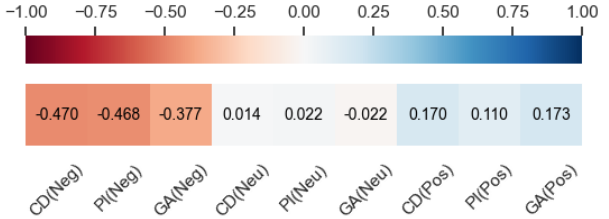


図 2: 各対話中の意欲評価ラベルの割合と購買意欲向上度合い間の相関分析結果

表 2: ユーザ評価実験結果

モデル	学習データ	ユーザ意欲の考慮	対話戦略の考慮	対話成功率	平均ターン数
GPT-3.5	成功対話 (63 対話)	-	-	0.23	9.35
GPT-3.5W	全対話 (109 対話)	✓	-	0.33	9.23
GPT-3.5WD	全対話 (109 対話)	✓	✓	0.44	9.08
GPT-4o (reference)	-	-	-	0.58	5.81

3.2 発話レベル分析

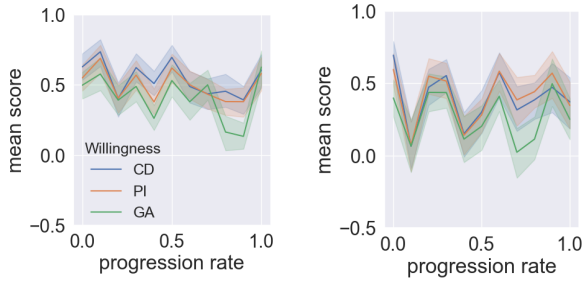
対話ターンごとの意欲の推移 図 3a、3b は、それぞれ成功対話（ユーザの購入意向が向上した対話）と失敗対話（ユーザの購入意向が向上しなかった対話）の平均意欲スコアの推移を示した図である。ここで、平均意欲スコアは、各発話に付与された評価ラベルについて Positive: +1、Neutral: 0、Negative: −1 のスコアを付与し、意欲スコアの移動平均を計算することで求めた。成功対話の平均意欲スコアの推移を見ると、対話開始直後と終了直前に各種意欲スコアが上昇していることがわかる。特に、対話終盤にかけて目標受容意欲スコアが大幅に増加している。一方、失敗対話では、対話開始直後および終了直前に意欲スコアが減少していることがわかる。特に、対話開始直後に意欲スコアが大きく減少している。

図 4a、 4b は、それぞれ成功対話、失敗対話における累計 Negative ラベルの推移を表している。成功対話と失敗対話を比較すると、全ての種類の意欲について成功対話の Negative 評価数が低く抑えられている。特に、ユーザ意欲の中でも情報提供意欲の Negative 評価数は低く抑えられており、対話中盤においてこの傾向が顕著であることがわかる。

効果的なセールス対話戦略 以上の分析から、ユーザの購買意欲を高めるためには、対話序盤、中盤、終盤ごとに以下の対話戦略が有効であると考えられる。(1) 対話序盤：対話継続意欲に Positive な印象を与え、早期の対話離脱を防ぐ対話の実施、(2) 対話中盤：情報提供意欲に Negative な印象を与えないユーザヒアリングの実施、(3) 対話終盤：目標受容意欲に Positive な影響を与える対話（例えば商品推薦）の実施。

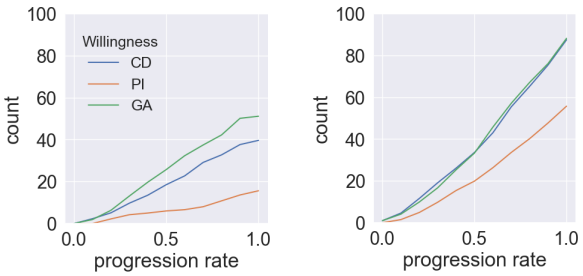
4 ユーザ評価実験

セールス対話データセットの有用性を検証するために、データセットを用いてセールス対話システムを構築し、ユーザ評価実験を行った。



(a) 成功対話　(b) 失敗対話

図 3: 対話中のユーザ意欲スコアの推移



(a) 成功対話　(b) 失敗対話

図 4: 各意欲ラベルの累計 Negative 評価数の推移

4.1 実験設定

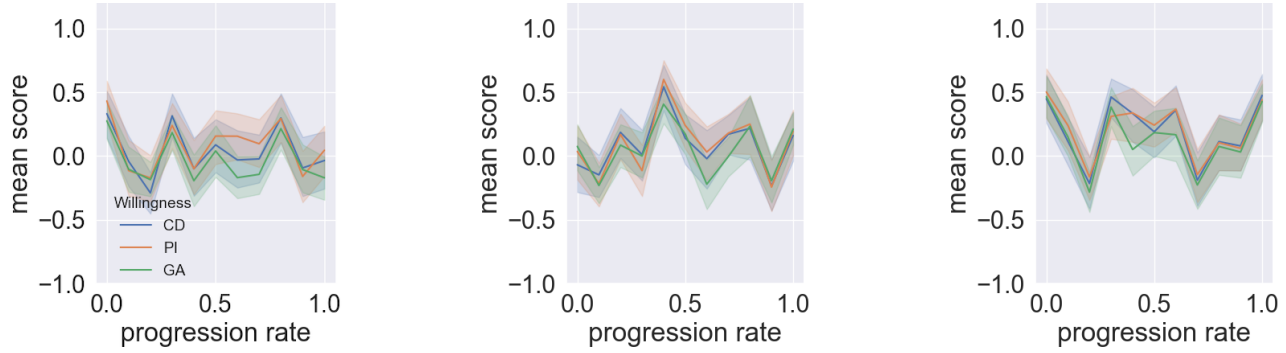
クラウドソーシングを通じて 48 名の作業者を募集し、複数のシステムと対話および評価を行うよう指示した。各システムとの対話方法および評価内容については、データセット構築時の設定と同様である。なお、各システムとの対話順序によってモデルの評価に順序バイアスが生じる可能性があるため、作業者間でシステムと対話する順序をランダムに割り振りバイアスの軽減を図った。

4.2 モデル

以下の 3 種類のシステムを構築した。各モデルは、OpenAI の GPT-3.5 (`gpt-3.5-turbo-0125`) および Fine-tuning API[1] を使用して開発した。

GPT-3.5 (baseline): 発話レベルのユーザ意欲を考慮せず、構築したデータセット中の成功対話 (計

1) https://platform.openai.com.



(a) GPT-3.5　(b) GPT-3.5W　(c) GPT-3.5WD

図 5: ユーザ評価実験における各モデルのユーザ意欲スコアの推移

63 件) のみを学習データとして使用して Fine-tuning したモデルである。

GPT-3.5 with willingness (GPT-3.5W): baseline とは異なり、ユーザの発話レベルの意欲ラベルを考慮して Fine-tuning したモデルである。Fine-tuning のアプローチとして、Attribute-conditioned Supervised Fine-Tuning [17] を採用した。具体的には、各ユーザ発話の末尾に、次に発話されるシステム発話の各意欲ラベルの属性名、属性値のペアを CONTINUING_DIALOGUE:1 のように付与して学習することで、属性値に応じたシステム発話が出力されることを試みる[2)]。生成時の各意欲ラベルの属性値は全て 1(Positive) に設定した。

GPT-3.5 with willingness + dialogue strategy (GPT-3.5WD): 上記の GPT-3.5W をベースに、§3.2 のセールス対話戦略を組み込んだモデルである。具体的には、1-3 ターン目は対話継続意欲、4-6 ターン目は情報提供意欲、7 ターン目以降は目標受容意欲がそれぞれ Positive になるよう属性値を指定した[3)]。

4.3 実験結果・議論

ユーザ評価実験の結果を表 2 に示す。対話成功率を見ると、発話レベルのユーザ意欲を考慮した GPT-3.5W、GPT-3.5WD が baseline モデルよりも高い。また、§3.2 のセールス対話戦略を反映した GPT-3.5WD が最も対話成功率が高くなっている。

図 5 に、モデルごとの平均意欲スコアの推移を示す。最も低い成功率を獲得した baseline モデルは、対話開始時点の意欲スコアは GPT-3.5WD と並んでいるものの、その後スコアは減少し、対話終盤も向上することなく終了している。2 番目に対話成功率が高かった GPT-3.5W は、対話開始時点の意欲スコアは GPT-3.5、GPT-3.5WD と比較して低いものの、対話終盤には各種意欲スコアが向上している。最も対話成功率が高かった GPT-3.5WD は、対話序盤の意欲スコアは減少傾向を示しているものの開始時点のスコアは GPT-3.5W よりも高くなっている。対話中盤においては情報提供意欲のスコアが Positive 傾向で維持されており、対話終盤には各種意欲スコアが大幅に向上していることがわかる。

これらの結果から、発話レベルのユーザの意欲を考慮することで、ユーザの購買意欲向上につながるセールス対話を実現できることが示唆された。また、GPT-3.5WD の各種意欲スコアの推移から、§3.2 のセールス対話戦略がシステムにある程度反映されており、その結果最も高い対話成功率を示したと考えられる。

5 おわりに

本研究では、生態学的妥当性に基づいた対話データ収集環境を開発し、3 種類のユーザ意欲データを含む日本語セールス対話データセットを構築した。構築したデータセットの分析により、購買意欲向上につながると考えられるセールス対話戦略の知見が得られた。また、ユーザ評価実験では、発話レベルでユーザの意欲を考慮し、さらにデータセット分析で得られたセールス対話戦略の知見を組み込むことが対話成功率向上につながることが示された。今後は、ユーザの応答に応じて動的にセールス対話戦略を切り替えることが可能なシステムを開発し、その有効性を実証実験によって検証する予定である。

2) 属性値は Positive=1、Neutral=0、Negative=-1 と設定した。
3) その他の意欲ラベルの属性値は 0(Neutral) とした。

謝辞

本研究の一部は、JSPS 科研費 JP22K17943 の支援を受けたものである。

参考文献

[1] Egon Brunswik. Thing constancy as measured by correlation coefficients. **Psychological Review**, Vol. 47, No. 1, p. 69, 1940.

[2] E. Brunswik. **The Conceptual Framework of Psychology**. International encyclopedia of unified science. University of Chicago Press, 1952.

[3] Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Construction and analysis of a persuasive dialogue corpus. **Situated Dialog in Speech-Based Human-Computer Interaction**, pp. 125–138, 2016.

[4] Abhisek Tiwari, Abhijeet Khandwe, Sriparna Saha, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. Towards personalized persuasive dialogue generation for adversarial task oriented dialogue setting. **Expert Systems with Applications**, Vol. 213, p. 118775, 2023.

[5] J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. **ACM Trans. Inf. Syst.**, Vol. 2, No. 1, p. 26–41, jan 1984.

[6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[7] Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[8] 邊土名朝飛, 馬場惇, 赤間怜奈. セールストークを対象とするエンゲージメントを考慮した目標指向対話データセット. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 3Xin240–3Xin240, 2024.

[9] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander Rudnicky. A Wizard-of-Oz study on a non-task-oriented dialog systems that reacts to user engagement. In Raquel Fernandez, Wolfgang Minker, Giuseppe Carenini, Ryuichiro Higashinaka, Ron Artstein, and Alesia Gainer, editors, **Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 55–63, Los Angeles, September 2016. Association for Computational Linguistics.

[10] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 1702–1723, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Sarik Ghazarian, Ralph Weischedel, Aram Galstyan, and Nanyun Peng. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 7789–7796, Apr. 2020.

[13] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Shiwei Sun, Jin Zhang, Yiwei Zhu, Mian Jiang, and Shuhui Chen. Exploring users' willingness to disclose personal information in online healthcare communities: The role of satisfaction. **Technological Forecasting and Social Change**, Vol. 178, p. 121596, 2022.

[15] Shlomo Berkovsky, Jill Freyne, and Harri Oinas-Kukkonen. Influencing individually: Fusing personalization and persuasion. **ACM Trans. Interact. Intell. Syst.**, Vol. 2, No. 2, jun 2012.

[16] Yeonkwon Jung. **Sales Talk**, pp. 87–114. Springer Nature Singapore, Singapore, 2022.

[17] Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 11275–11288, Singapore, December 2023. Association for Computational Linguistics.