タスク指向型対話システムへの項目反応理論の適用によるユーザのタスク達成能力の推定

平井龍　　郭傲　　東中竜一郎
名古屋大学情報学研究科
{hirai.ryu.k6@s.mail, guo.ao.i6@f.mail, higashinaka@i}.nagoya-u.ac.jp

概要

タスク指向型対話システムの性能は改善しているものの，すべてのユーザが自身のタスクを完全に達成できるわけではない．システムについての知識の少ないユーザは，システムに対しての話し方が分からず，対話破綻を引き起こしたり，タスクが達成できなかったりする．この問題を解決するためには，システムはユーザのタスク達成能力を推定し，ユーザの能力に合わせて対話することが望ましい．本研究では，教育分野で受験者の能力推定によく用いられる項目反応理論をタスク指向型対話システムに適用し，ユーザのタスク達成能力を推定する手法を提案する．推定したタスク達成能力を用いてスロットの正答確率を予測する実験を行った結果，提案手法はベースラインよりも高い精度でスロットの正答確率を予測できることが分かった．

1　はじめに

タスク指向型対話システムの性能は改善しているものの，すべてのユーザが自身のタスクを完全に達成できるわけではない [1]．例えば，OpenAI 社の GPT-4 [2] などの大規模言語モデルを用いたタスク指向型対話システムにおいてもその性能は高くはない [3, 4]．特に，システムについての知識の少ないユーザは，システムに対しての話し方が分からず，対話破綻 [5] を引き起こしたり，タスクが達成できなかったりする．システムがユーザのタスク達成能力を推定し，ユーザの能力に合わせて対話することは，この問題の解決策の一つである．

本研究では，教育分野でよく用いられる項目反応理論 (Item Response Theory, IRT) を用いて，ユーザのタスク達成能力を推定する手法を提案する．具体的には，まず，タスク指向型対話システムとユーザの対話を収集する．対話収集時には，ユーザに対話ゴールを提示し，ユーザには当該対話ゴールに基づいて対話をさせる．次に，所定のスロットを適切に埋められるかどうかを問題だと見做し，IRT を用いてスロットの項目特性を推定する．最後に，能力を推定したいユーザに対話ゴールに基づいて対話をさせ，IRT を適用することで，ユーザのタスク達成能力を推定する．

MultiWOZ データセット [6] を用いて構築したタスク指向型対話システムにおいて，推定したタスク達成能力からスロットの正答確率を予測する実験を行った結果，提案手法はベースラインよりも有意に高精度でスロットの正答確率を予測可能であることが分かった．また，タスク達成能力を推定するための適切な対話ゴールの作成のため，推定した項目特性を用いて MultiWOZ データセットにおけるスロットの困難度や識別力などの特徴の分析を行った．

2　項目反応理論

本研究で用いる項目反応理論 (以降，IRT) を簡単に説明する．IRT とはテストで受験者の能力を数値化する測定の理論である [7]．従来のテストでは受験者が正答した問題の配点を足し合わせた点数を受験者の得点とする．しかし，あらかじめ決められた配点によって正しく受験者の能力を表すことができるとは限らない．

IRT を用いたテストでは，大量のユーザの正誤データを使用して，問題ごとに受験者の能力 θ と正答率 $prob$ の関係を計算する．受験者の能力と正答率の関係は，次式の通り識別力 a，困難度 b，当て推量 c の項目特性で表される．

$$prob = c + \frac{1-c}{1+e^{-a(\theta-b)}} \tag{1}$$

識別力は問題が受験者の能力の高低を識別する度合いを表す．困難度は項目の難しさの度合いを表す．当て推量は受験者が偶然に正解できる確率を表し，多肢選択問題では選択肢数の逆数が当て推量の目安

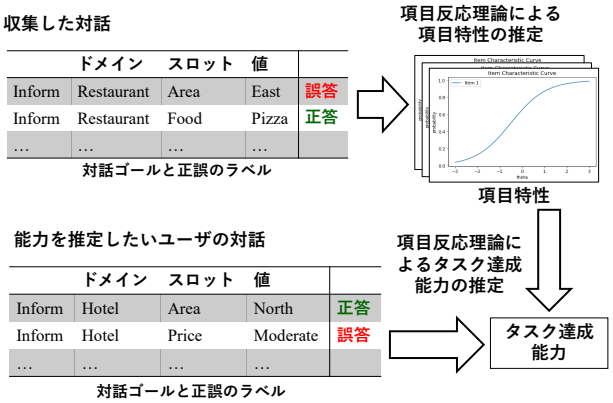


図 1 提案手法の全体像

となる．項目特性を用いて，受験者の回答の正誤パターンが最も起きやすい能力の値が受験者の能力として推定される．

3 関連研究

ユーザの能力を推定する研究として，human-computer interaction の分野では，Ghazarian らがユーザのスキルに適応的なインタフェースの開発のため，デスクトップアプリケーションのマウスの動きなどを用いた自動スキル分類器を構築した [8]．また，Lo らは学習者の認知スタイルを推定し，認知スタイルに適応的なウェブベースの学習システムを開発した [9]．

Voice user interface (VUI) や音声対話システムにおいては，Ward らがユーザの発話の速度に合わせてシステムの発話の速度を変化させるシステムを提案した [10]．Myers らはユーザの VUI との対話における行動をクラスタ化した [11]．駒谷らはユーザの「システムに対する習熟度」，「ドメインに関する知識レベル」，「性急度」を推定してユーザに応じてシステムの振る舞いを変化させるシステムを提案した [12]．しかし，これらの研究ではユーザの能力の推定時に対話内容を考慮していない．

対話システムにおける IRT の利用については，Sedoc らがチャットボットの評価に IRT を導入した [13]．この研究では，チャットボットのペアを IRT における受験者，入力文を IRT における問題と見做し，入力文とモデルの両方の評価を可能とした．しかし，この研究はモデルの性能や入力文の質を推定しており，ユーザの能力は推定していない．

4 提案手法

タスク指向型対話システムに IRT を適用し，ユーザのタスク達成能力を推定する手法を提案する．図 1 に提案手法の全体像を示す．まず，対話ゴールに基づいたシステムとユーザの対話を収集する．次に，対話ゴールと対話終了時の信念状態を比較してスロットの正誤判定を行う．そして，周辺最尤法 [14, 15] を用いてスロットごとに項目特性 (困難度，識別力，当て推量) を推定する．周辺最尤法は受験者の能力の分布を標準正規分布と仮定して項目特性のみを推定する手法であり，受験者数が増えても安定した結果が得られることが知られている．

タスク指向型対話システムの対話ゴールには，ユーザがシステムに伝えるべきスロットの内容 (inform ゴール)，および，尋ねるべきスロット (request ゴール) が含まれる．本研究では，1 つの対話を 1 つのテストとし，所定のスロットを適切に埋められるかということを問題だと見做し，問題ごとに正誤判定を行うことで IRT を適用する．

inform ゴールのスロットでは，ユーザがスロットの値を適切にシステムに伝えることができれば正答とする．v をゴールにおけるスロットの値，$b[d][s]$ を対話終了時におけるドメイン d，スロット s の信念状態とすると，正誤 $ans \in \{0, 1\}$ は $v = b[d][s]$ のとき $ans = 1$ で，$v \neq b[d][s]$ のとき $ans = 0$ となる．

request ゴールのスロットでは，ユーザがシステムから適切に情報を聞き出せたら正答とする．s を対話ゴールに含まれるスロット，$S[d]$ を対話中でドメイン d についてシステムがユーザに情報を伝えたスロットの集合とすると，正誤 $ans \in \{0, 1\}$ は $s \in S[d]$ のとき $ans = 1$ で，$s \notin S[d]$ のとき $ans = 0$ となる．

タスク達成能力を推定したいユーザには，所定の対話ゴールに基づいて対話をさせる．そして，各スロットの正誤を判定し，Expected A Posteriori (EAP) 推定 [16] によってタスク達成能力を推定する．式 (1) を用いることで，タスク達成能力からスロットの正答確率を求めることができる．

5 実験

実際に対話を収集し，IRT を用いてユーザの能力を推定した．そして，推定したユーザの能力を用いてスロットの正答確率の予測精度を評価した．スロットを正しく埋める能力がタスク達成能力に対応するとした仮定の下で，提案手法が各スロットの正

答確率を正しく推定できれば，提案手法はユーザのタスク達成能力を正確に推定可能といえる．

また，タスク達成能力を推定するための適切な対話ゴールの作成のため，推定したスロットの困難度や識別力といった特徴や，ユーザの能力とタスク達成回数の関係の分析を行った．なお，本実験における対話データ収集については，事前に倫理的観点における承認を所属組織から得ている．

5.1 対話システム

実験に使用する対話システムは MultiWOZ 2.1 データセット [6, 17] を用いて構築した．MultiWOZ は旅行案内タスクにおける，人間同士の英語の対話データセットである．ユーザがシステムに観光地を尋ねたり，ホテルや飲食店などの予約を依頼したりする対話が含まれている．ドメインは restaurant, hotel, attraction, taxi, train, hospital, police の 7 つで，対話数は 10,438 である．

対話システムには，4 つのモジュールからで構成されるパイプラインシステムを使用した．パイプラインシステムは BERT [18] をベースとした発話認識モジュール (NLU)，ルールベースの対話状態追跡モジュール (DST)，ルールベースの行動決定モジュール (Policy) [19]，それに，テンプレートを用いた発話生成モジュール (NLG) が連結されている．システムの構築にあたっては，タスク指向型対話システムの実装のためのツールキットである ConvLab-2 [20, 21] を使用した．パイプラインシステムは ConvLab-2 によって実装することが可能なシステムの中で最もタスク達成率が高い [20]．

5.2 実験手順

クラウドソーシングのプラットフォームである Amazon Mechanical Turk[1] 上で対話の収集を実施した．369 人のワーカに対話ゴールを提示し，システムと対話させた．1 人のワーカは 3 回連続で対話を行った．369 人のうち，190 人には対話ごとにランダムで生成された対話ゴール (ランダムゴール) を提示し，179 人にはあらかじめ決められた 3 つの対話ゴール (固定ゴール) をランダムな順で提示した．1 つの対話ゴールに含まれるドメイン数は 3 以上 4 以下で，スロット数は 15 以上 20 以下である．

収集した対話の統計量は表 1 の通りである．トークン数の計測にあたっては，Python のライブラリである NLTK [22] を使用した．固定ゴールは平均タスク達成率がランダムゴールよりも高く，ゴールの難易度が若干低かった．

表 1 収集した対話の統計量．発話数，トークン数はシステムとユーザの合計の値である．

	ランダムゴール	固定ゴール
ユーザ数	179	190
対話数	537	570
発話数	24,340	23,474
トークン数	311,043	289,792
平均タスク達成率	47.5%	66.0%

5 分割交差検証で結果を評価した．具体的には，収集した対話データをユーザの重複がないように 5 分割し，そのうち 1 つをテストデータ，残りの 4 つを学習データとした．まず，IRT を用いて学習データの各スロットの項目パラメータを推定した．推定には Python の IRT のライブラリである GIRTH [2] を用いた．続いて，学習データから推定した項目パラメータとテストデータの 1 対話目から推定したユーザのタスク達成能力を用いて，テストデータの 2 対話と 3 対話目の各スロットの正答確率を予測した．この操作は各分割に対して実施した．

5.3 ベースライン

スロットの正答確率の求め方が異なる 2 種類のベースラインを用意した．

ベースライン (スロット) 対象スロットの正答率の平均をスロットの正答確率とする手法．すなわち，学習データにおける全ユーザのスロット s の正解確率がテストデータにおけるユーザ u のスロット s の正答確率となる．

ベースライン (ユーザ) テストデータの 1 対話目における対象ユーザの正答率の平均をスロットの正答確率とする手法．すなわち，1 対話目におけるユーザ u の全スロットの正答確率の平均値が，ユーザ u のスロット s の正答確率となる．

5.4 評価指標

評価指標には，正答確率の推定精度を用いた．これは，推定された正答確率でスロットの正誤を予測する試行を無限回行った時の平均推定精度に相当する．具体的には，推定したスロットの推定確率を $prob$，実際のユーザによる正誤を $ans \in \{0, 1\}$ とす

1） https://www.mturk.com/

2） https://github.com/eribean/girth

表 2　正答確率の予測精度 (ランダムゴール)．太字は各カラムで最も値が大きいスコアを示す．

	2 対話目	3 対話目
提案手法	**0.732**	**0.736**
ベースライン (スロット)	0.704	0.703
ベースライン (ユーザ)	0.678	0.690

表 3　正答確率の予測精度 (固定ゴール)．太字は各カラムで最も値が大きいスコアを示す．

	2 対話目	3 対話目
提案手法	**0.762**	**0.759**
ベースライン (スロット)	0.701	0.699
ベースライン (ユーザ)	0.663	0.658

ると，各スロットの正答確率の推定精度は次式の通りである．

$$acc = \begin{cases} prob & (ans = 1) \\ 1 - prob & (ans = 0) \end{cases} \quad (2)$$

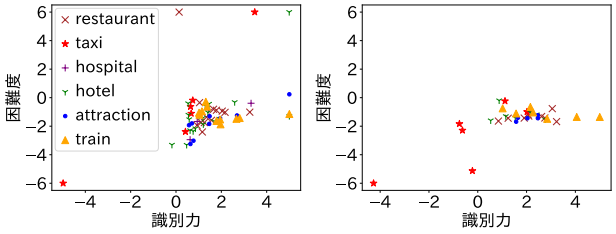
5.5　結果

ランダムゴールの結果を表 2 に，固定ゴールの結果を表 3 に示す．Bonferroni 補正を用いた Wilcoxon の符号順位和検定を行ったところ，ランダムゴールと固定ゴールのいずれにおいても，提案手法は他の手法よりも推定精度が有意に大きいことが分かった ($p < .01$)．

また，2 対話目と 3 対話目の結果を比べると，いずれの手法においても推定精度の差はほぼなく，対話回数による対話の性質の違いはほとんどないことが分かった．さらに，ランダムゴールと固定ゴールの結果を比べると，固定ゴールの方が推定精度が高かった．これは，固定ゴールでは学習データ内に同一のスロットについての正誤の情報を多く含み，より正確にスロットの項目特性を推定できているためであると考えられる．タスク指向型対話システムに IRT を適用して項目特性を得るときには，1 つのスロットあたりの正誤の情報を多く収集し項目特性を推定する方が良いといえる．

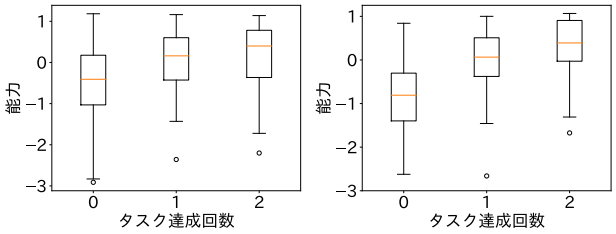
5.6　項目特性の分析

図 2 にスロットごとの識別度と困難度の分布を示す．ランダムゴールと固定ゴールのいずれにおいても，ほぼ全てのスロットは識別力が 0 より大きく，ユーザのタスク達成能力を推定に有用である．このため，識別力の高いスロットを適切に選択することでユーザの能力を推定するための適切な対話ゴールを作成することが可能になると考えられる．



(a) ランダムゴール　(b) 固定ゴール

図 2　全てのスロットについての推定された識別力と困難度



(a) ランダムゴール　(b) 固定ゴール

図 3　1 対話目から推定されたユーザのタスク達成能力と 2，3 対話目でタスク達成した回数の関係

5.7　ユーザの能力の分析

図 3 に 1 対話目から推定されたユーザのタスク達成能力と 2 対話目と 3 対話目におけるタスク達成回数の関係を表す．ランダムゴールと固定ゴールのいずれにおいてもタスク達成する回数が多いユーザほどタスク達成能力が高い傾向にあり，推定した能力がユーザのタスク達成能力を適切に表しているといえる．

6　まとめと今後の課題

本研究ではユーザの能力に適応的なタスク指向型対話システムの構築に向けて，タスク指向型対話システムに IRT を適用しユーザのタスク達成能力を推定する手法を提案した．予測したタスク対話能力を用いてスロットの正答確率を予測する実験を行った結果，提案手法のユーザの能力を推定する手法はベースラインよりも有意に高精度でスロットの正答確率を予測することができた．

今後は，英語のデータセットだけでなく日本語のデータセット [23, 24] にも提案手法を適用し，適切にユーザのタスク達成能力を推定できるかを検証したい．さらに，推定したユーザのタスク達成能力に基づいてユーザに適応的な振る舞いをする対話システムの構築を行いたい．

謝辞

本研究は，科研費「モジュール連動に基づく対話システム基盤技術の構築」(課題番号 19H05692) の支援を受けた．

参考文献

[1] Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. In **Proc. of SIGDIAL**, pp. 297–310, 2020.

[2] OpenAI. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023.

[3] Vojtěch Hudeček and Ondrej Dusek. Are Large Language Models All You Need for Task-Oriented Dialogue? In **Proc. of SIGDIAL**, pp. 216–228, 2023.

[4] Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. Clarifying the Dialogue-Level Performance of GPT-3.5 and GPT-4 in Task-Oriented and Non-Task-Oriented Dialogue Systems. In **Proc. of AI-HRI**, 2023.

[5] Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. Integrated taxonomy of errors in chat-oriented dialogue systems. In **Proc. of SIGDIAL**, pp. 89–98, 2021.

[6] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In **Proc. of EMNLP**, pp. 5016–5026, 2018.

[7] Frederic M Lord. **Applications of Item Response Theory to Practical Testing Problems**. Routledge, 1980.

[8] Arin Ghazarian and S Majid Noorhosseini. Automatic detection of users' skill levels using high-frequency user interface events. **User Modeling and User-Adapted Interaction**, Vol. 20, pp. 109–146, 2010.

[9] Jia-Jiunn Lo, Ya-Chen Chan, and Shiou-Wen Yeh. Designing an adaptive web-based learning system based on students' cognitive styles identified online. **Computers & Education**, Vol. 58, No. 1, pp. 209–222, 2012.

[10] Nigel Ward and Satoshi Nakagawa. Automatic User-Adaptive Speaking Rate Selection for Information Delivery. In **Proc. of ICSLP**, pp. 549–552, 2002.

[11] Chelsea M. Myers, David Grethlein, Anushay Furqan, Santiago Ontañón, and Jichen Zhu. Modeling Behavior Patterns with an Unfamiliar Voice User Interface. In **Proc. of UMAP**, p. 196–200, 2019.

[12] Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. User Modeling in Spoken Dialogue Systems for Flexible Guidance Generation. In **Proc. of EUROSPEECH**, pp. 745–748, 2003.

[13] João Sedoc and Lyle Ungar. Item Response Theory for Efficient Human Evaluation of Chatbots. In **Proc. of Eval4NLP**, pp. 21–33, 2020.

[14] R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. **Psychometrika**, Vol. 46, No. 4, pp. 443–459, 1981.

[15] Michael R. Harwell, Frank B. Baker, and Michael Zwarts. Item Parameter Estimation Via Marginal Maximum Likelihood and an EM Algorithm: A Didactic. **Journal of Educational Statistics**, Vol. 13, No. 3, pp. 243–271, 1988.

[16] Jean-Paul Fox. **Bayesian item response modeling: Theory and applications**. Springer, 2010.

[17] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In **Proc. of LREC**, 2020.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, pp. 4171–4186, 2019.

[19] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In **Proc. of NAACL**, pp. 149–152, 2007.

[20] Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. In **Proc. of ACL**, pp. 142–149, 2020.

[21] Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. Robustness Testing of Language Understanding in Task-Oriented Dialog. In **Proc. of ACL-IJCNLP**, pp. 2467–2480, 2021.

[22] Steven Bird, Edward Loper, and Ewan Klein. **Natural Language Processing with Python**. O'Reilly Media, Inc., 2009.

[23] 大橋厚元, 平井龍, 飯塚慎也, 東中竜一郎. JMultiWOZ: 日本語タスク指向型対話データセットの構築. 言語処理学会 第 29 回年次大会 発表論文集, pp. 3093–3098, 2023.

[24] 大橋厚元, 平井龍, 飯塚慎也, 東中竜一郎. JMultiWOZ に対する対話状態アノテーションの付与と対話システムの実装評価. 言語処理学会 第 30 回年次大会 発表論文集, 2024.