Applying Item Response Theory to Task-oriented Dialogue Systems for Accurately Determining User's Task Success Ability

Anonymous ACL submission

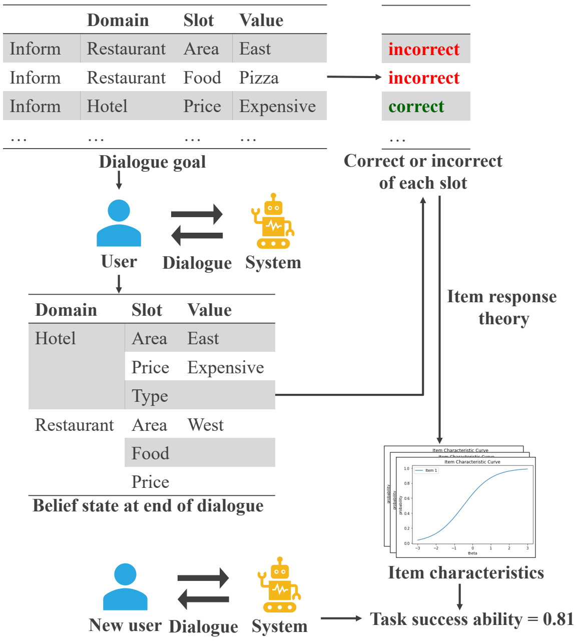
Abstract

While task-oriented dialogue systems have improved, not all users can fully accomplish their tasks. Users with limited knowledge about the system may experience dialogue breakdowns or fail to achieve their tasks because they do not know how to interact with the system. For addressing this issue, it would be desirable to construct a system that can estimate the user's task success ability and adapt to that ability. In this study, we propose a method that estimates this ability by applying item response theory (IRT), commonly used in education for estimating examinee abilities, to task-oriented dialogue systems. Through experiments predicting the probability of a correct answer to each slot by using the estimated task success ability, we found that the proposed method significantly outperformed baselines.

1 Introduction

Although task-oriented dialogue systems have improved, not all users can accomplish their tasks (Takanobu et al., 2020). Even in dialogue systems built using large language models such as OpenAI's ChatGPT[1], the system's performance is not always satisfactory (Hudeček and Dušek, 2023). In particular, users with limited knowledge about the system may experience dialogue breakdowns or fail to achieve their tasks because they do not know how to communicate with the system. One solution would be for the system to estimate the user's task success ability and then engage in dialogue in accordance with that ability, for example, by changing the expressions in utterances or interaction strategies.

We therefore propose a method (shown in Figure 1) that estimates the user's task success ability by utilizing item response theory (IRT) (Lord, 1980), which is commonly used in the field of education. Specifically, we first collect dialogues between the system and users, where each user is presented with a unique dialogue goal and must engage in dialogue based on that goal. Next, considering correctly filling in each designated slot as a problem, we estimate the item characteristics of the slots by using IRT. Finally, we let the user engage in a dialogue on the basis of a given dialogue goal, and the user's task success ability is estimated by using the item characteristics of the filled or unfilled slots.



	Domain	Slot	Value
Inform	Restaurant	Area	East
Inform	Restaurant	Food	Pizza
Inform	Hotel	Price	Expensive
…	…	…	…

incorrect
incorrect
correct
…

Domain	Slot	Value
Hotel	Area	East
	Price	Expensive
	Type	
Restaurant	Area	West
	Food	
	Price	

Dialogue goal
Correct or incorrect of each slot
User
Dialogue
System
Item response theory
Belief state at end of dialogue
Item Characteristic Curve
Item 1
Item characteristics
New user
Dialogue
System
Task success ability = 0.81

Figure 1: Overview of proposed method.

Our experimental results showed that the proposed method significantly outperformed the baselines in accurately predicting the probabilities of correct answers to slots. In addition, our analysis of the item characteristics of slots in the MultiWOZ dataset (Eric et al., 2020) revealed further insights about how the dialogue goals should be determined for predicting task success abilities. The contributions of this paper are as follows.

- This is the first work to apply IRT for understanding users' task success abilities in task-

[1] https://openai.com/blog/chatgpt/

oriented dialogue systems.

- We reveal item characteristics such as slot difficulty and discrimination in the MultiWOZ dataset.

2 Item Response Theory

We first explain item response theory (IRT), which is a measurement theory that quantifies examinees' abilities on tests (Lord, 1980). In traditional tests, the total score of the correctly answered questions represents the examinee's score. However, in such tests it is necessary to predetermine the score of each problem, and since there is rarely a rational basis for setting scores, the scores may not accurately represent the examinees' ability.

In tests that utilize IRT, the relationship between the examinee's abilities θ and the probabilities of correct answers to questions $prob$ is calculated for each question using a large amount of user response data. The relationship is described by item characteristics such as discrimination a, difficulty b, and guessing c, as shown in the following equation.

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

Then, the ability at which the examinee's response patterns are most likely to occur is estimated. We explain the item characteristics in more detail in Appendix A.

3 Related Work

3.1 Modeling User Characteristics

In the field of human-computer interaction, Ghazarian and Noorhosseini (2010) constructed an automatic skill classifier using mouse movements in desktop applications. Lo et al. (2012) identified students' cognitive styles and developed an adaptive web-based learning system.

In the area of voice user interfaces (VUIs) and spoken dialogue systems, Ward and Nakagawa (2002) proposed a system that adjusts the system's speaking rate on the basis of that of the user's. Myers et al. (2019) clustered user behaviors in interactions with VUIs. Komatani et al. (2003) proposed a method that estimates user attributes such as skill level to the system, knowledge level to the target domain, and degree of hastiness to adapt the system's behavior for a bus information system. However, these studies did not exploit the characteristics of problems, which should be considered when estimating the task success ability.

3.2 Application of IRT

Sedoc and Ungar (2020) introduced IRT to the evaluation of chatbots and conducted tests to determine which of two chatbots provided appropriate responses during dialogues. This research considered the pairs of chatbots as examinees and input utterances as the problems in IRT. This allowed for the evaluation of both the input utterances and the chatbots. Lalor et al. (2016) applied IRT to the textual entailment recognition task and compared system performance with human performance. This research considered the systems or humans as examinees and textual entailment recognition tasks as the problems in IRT. However, these studies did not aim to estimate users' ability to interact with systems.

4 Proposed Method

In our method, we first collect dialogues between the system and users. Next, we calculate the correctness of each slot by comparing the dialogue goal and the belief state at the end of the dialogue. We use IRT to estimate item characteristics (difficulty, discrimination, and guessing for each slot) by means of marginal maximum likelihood estimation. Then, we let the user whose task success ability we want to estimate engage in a dialogue for a given dialogue goal, judge whether each slot is correctly filled, and estimate the task success ability by expected a posteriori estimation based on Bayesian statistics (Fox, 2010). In this study, we regard each dialogue as a single test and consider whether each slot is filled in correctly as the problem of IRT.

In task-oriented dialogue systems, the dialogue goal includes the content of the slots that the user should convey to the system (inform goals) and the slots that the user should ask about (request goals).

For an inform goal slot, it is considered correct if the user can appropriately convey their slot values to the system. Let v and $b[d][s]$ denote the value of the goal and the belief state at the end of the dialogue for a domain d and slot s. The correctness $ans \in \{0,1\}$ is defined as follows.

$$ans = \begin{cases} 1 & (v = b[d][s]) \\ 0 & (\text{otherwise}) \end{cases} \qquad (2)$$

For a request goal slot, it is considered correct if the user can appropriately obtain the information from the system. Let s and $S[d]$ denote the slot of the goal and the set of slots of the domain d for which the system has conveyed information in the dialogue. The correctness $ans \in \{0,1\}$ is defined as follows.

$$ans = \begin{cases} 1 & (s \in S[d]) \\ 0 & (\text{otherwise}) \end{cases} \tag{3}$$

5 Experiment

We collected dialogue data and estimated users' task success abilities using IRT. We then evaluated the accuracy of estimating the probabilities of correct answers to slots utilizing the users' estimated task success abilities. We also analyzed the estimated item characteristics.

5.1 Dialogue Systems

We built the systems using the MultiWOZ 2.1 dataset (Eric et al., 2020), an English dialogue dataset between a tourist and a clerk at a tourist information center in seven domains: restaurant, hotel, attraction, taxi, train, hospital, and police.

Since item characteristics may differ depending on the system configuration, we used two dialogue systems: a pipeline (Zhang et al., 2020), which consists of four modules, and SimpleTOD (Hosseini-Asl et al., 2020), an end-to-end system. The pipeline system consists of a natural language understanding module based on BERT (Devlin et al., 2019), a rule-based dialogue state tracking module, a rule-based policy module (Schatzmann et al., 2007), and a template-based natural language generation module. To construct the pipeline system, we utilized the ConvLab-2 toolkit (Zhu et al., 2020; Liu et al., 2021), which enables the development of task-oriented dialogue systems. SimpleTOD is a GPT2-driven language model fine-tuned for MultiWOZ dialogues. We trained the model using the public source code on GitHub[2]. Appendix B provides the details of the training settings.

5.2 Experimental Procedure

First, we collected dialogues through Amazon Mechanical Turk, a crowdsourcing platform. We presented different randomly generated dialogue goals, including two through four domains containing ten through 20 slots, to 377 workers and engaged them in dialogue with the systems. Each worker was presented with a randomly generated dialogue goal and engaged in three consecutive dialogues with the same dialogue system, either pipeline or SimpleTOD, but with different dialogue goals. The experiment was approved by the ethical review committee of our organization.

	Pipeline	SimpleTOD
No. of users	179	198
No. of dialogues	537	594
No. of utterances	24,340	20,532
No. of tokens	311,043	233,760
Task success rate	47.5%	28.3%
Slot correct rate	77.6%	62.0%

Table 1: Statistics of collected dialogues.

The statistics of the collected dialogues are shown in Table 1. We used NLTK[3], a Python library, for counting the number of tokens. As we can see, the dialogues of the pipeline system have a moderate success rate (47.5%), whereas those of SimpleTOD are lower (28.3%), as expected from (Zhu et al., 2020).

We utilized 5-fold cross-validation to evaluate the results. We selected one fold as test data and the remaining four as training data. We made sure there was no overlap of users between the folds. First, we estimated item characteristics using IRT for each slot in the training data. For this purpose, we utilized the GIRTH library[4], a Python library for IRT. Then, using the estimated item characteristics from the training data and the estimated user's task success abilities from the first dialogue of the test data, we predicted the probabilities of correct answers for each slot in the second and third dialogues of the test data. This process was repeated for each fold.

5.3 Baselines

We prepared two baselines with different approaches for estimating probabilities of correct answers to the slots.

Baseline (Slot) A method that uses the average probability of a correct answer for a target slot as the probability of a correct answer for the slot. That is, the probability of a correct answer to slot s over all users in the training data is used for the probability for slot s for users in the test data.

Baseline (User) A method that uses the average

[2] https://github.com/salesforce/simpletod/

[3] https://www.nltk.org/

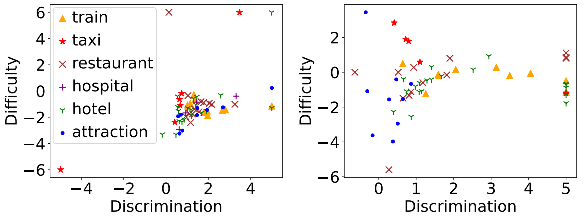
[4] https://github.com/eribean/girth

	2nd dialogue	3rd dialogue
Proposed	**0.732**	**0.736**
Baseline (Slot)	0.704	0.703
Baseline (User)	0.678	0.690

Table 2: Accuracy of estimating the probabilities of correct answers (pipeline).

	2nd dialogue	3rd dialogue
Proposed	**0.606**	**0.603**
Baseline (Slot)	0.582	0.575
Baseline (User)	0.561	0.577

Table 3: Accuracy of estimating the probabilities of correct answers (SimpleTOD).



(a) Pipeline

(b) SimpleTOD

Figure 2: Distribution of discrimination and difficulty estimated for all slots. (a): Results of pipeline system. (b): Results of SimpleTOD system.

probability of a correct answer from the target user in the test data's first dialogue as the probability of a correct answer for the slot. That is, the probability of a correct answer to slot s is the averaged probability of a correct answer to all slots of that user in previous dialogues.

5.4 Evaluation Metrics

We set the accuracy of estimating the probabilities of correct answers as the evaluation metric. Specifically, if the estimated probability of a correct answer is denoted as $prob$, and the actual correctness of the user is denoted as $ans \in \{0, 1\}$, then the accuracy of estimating the probabilities of the correct answers is the average for all slots where each slot's accuracy is calculated by:

$$acc = |1 - ans - prob| \qquad (4)$$

This is equivalent to the average estimation accuracy when performing infinite trials that involve predicting the correctness of each slot as correct with probability $prob$.

5.5 Results

Tables 2 and 3 show the results for the pipeline system and the SimpleTOD system, respectively. Wilcoxon signed-rank tests with Bonferroni correction revealed that the proposed method achieved a significantly higher estimation accuracy than the other methods ($p < .01$).

Comparing the results for the second and third dialogues, we found almost no difference in estimation accuracy for all methods, indicating that the nature of the dialogue does not significantly vary with the number of dialogues. Note that, since imbalanced data with more correct answers than incorrect ones lead to higher accuracy (Table 1), we cannot compare the absolute score of the accuracy between the pipeline and the SimpleTOD system. Appendix C provides examples of dialogues between users and the pipeline system and the users' estimated task success abilities.

5.6 Analysis of Item Characteristics of Slots

Figure 2 shows the distribution of discrimination and difficulty of the slots. In both systems, almost all slots exhibited discrimination values greater than 0 and had the power to estimate the user's task success ability. While the pipeline system showed minimal differences in discrimination and difficulty among slots, the SimpleTOD system revealed substantial variations in discrimination and difficulty across slots, making it possible to appropriately select slots with high discrimination for appropriate testing.

6 Conclusion and Future Work

We proposed a method for estimating users' task success abilities with task-oriented dialogue systems utilizing item response theory. Experiments on predicting the probability of a correct answer for each slot showed that the proposed method significantly outperformed the baselines.

Various challenges need to be addressed in future work, such as the dependence of slots; to this end, we want to explore methods that consider multiple slots as a single problem. We also want to estimate the task success ability using deep learning-based IRT methods that may achieve higher accuracy (Yeung, 2019; Tsutsumi et al., 2021). Additionally, we aim to investigate methods for estimating task success abilities more quickly, that is, using less than a single dialogue. Finally, we want to construct dialogue systems that can adapt their behavior on the basis of the users' estimated task success abilities.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *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)*, pages 4171–4186.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428.

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

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

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191.

Vojtěch Hudeček and Ondřej Dušek. 2023. Are LLMs All You Need for Task-Oriented Dialogue? *arXiv preprint arXiv:2304.06556*.

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. 2003. User Modeling in Spoken Dialogue Systems for Flexible Guidance Generation. In *Proceedings of 8th European Conference on Speech Communication and Technology*, pages 745–748.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an Evaluation Scale using Item Response Theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness Testing of Language Understanding in Task-Oriented Dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2467–2480.

Jia-Jiunn Lo, Ya-Chen Chan, and Shiou-Wen Yeh. 2012. Designing an adaptive web-based learning system based on students' cognitive styles identified online. *Computers & Education*, 58(1):209–222.

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

Chelsea M. Myers, David Grethlein, Anushay Furqan, Santiago Ontañón, and Jichen Zhu. 2019. Modeling Behavior Patterns with an Unfamiliar Voice User Interface. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, page 196–200.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152.

João Sedoc and Lyle Ungar. 2020. Item Response Theory for Efficient Human Evaluation of Chatbots. In *Proceedings of the 1st Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310.

Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. 2021. Deep-IRT with Independent Student and Item Networks. *International Educational Data Mining Society*, pages 510–517.

Nigel Ward and Satoshi Nakagawa. 2002. Automatic User-Adaptive Speaking Rate Selection for Information Delivery. In *Proceedings of 7th International Conference on Spoken Language Processing*, pages 549–552.

Chun-Kit Yeung. 2019. Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. *arXiv preprint arXiv:1904.11738*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 142–149.

Speaker	Utterance
	...
User	**Is there a train leaving after 18:15 from Broxbourne to Cambridge on Sunday?**
System	We can book you for the train leaving at 18:15. Would you like to take the train on Sunday? Would you like me to book a train to Cambridge for you?
User	I need tickets for two.
	...

(a) Dialogue with estimated task success ability of 0.195

Speaker	Utterance
	...
User	I want train service
System	When would you like to leave by? Is there a time you would like to get there by? On what day will you be traveling?
User	**11.30**
System	Sorry I do not understand, can you paraphrase?
	...

(b) Dialogue with estimated task success ability of -1.116

Figure 3: Example dialogues from pipeline system with estimated task success abilities.

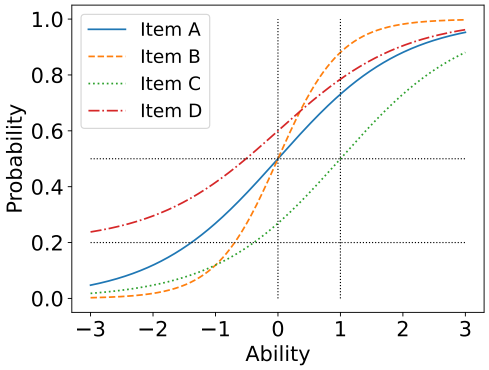


Figure 4: Example of item characteristic curves for four different questions (item A, item B, item C, item D) with distinct characteristics.

A Item Characteristics

Figure 4 provides examples of item characteristic curves that represent the characteristics of each particular question, where the horizontal axis of each curve represents the examinee's ability value θ and the vertical axis represents the probability $prob$ of a correct answer to the item. Generally, the item characteristic curve shows that the probability of a correct answer is small when the ability is small, increases around the medium ability value, and reaches a high probability for large ability values. It forms an upward-sloping curve.

Discrimination represents the extent to which a question distinguishes between examinees of different abilities. In Figure 4, items A and B differ only in their discrimination parameters. Difficulty indicates an item's difficulty level. In Figure 4, items A and C differ only in their difficulty parameters. Guessing represents the probability of a chance guess resulting in a correct response for an examinee with no ability. In multiple-choice questions, the reciprocal number of choices can be used to estimate the guessing parameter. In Figure 4, items A and D differ only in their guessing parameters.

B Training Settings for the SimpleTOD system

As the best hyperparameters for SimpleTOD were unknown, we trained it by using the public source code on GitHub with different hyperparameter values (e.g., the batch size from two to eight, learning rate from $1e-5$ to $1e-4$, and maximum sequence length from 256 tokens to 1,024 tokens); then, we selected the most optimized model. We further modified the lexicalization rules to ensure the legibility of the generated system responses.

C Examples

Figure 3 presents examples of dialogues between users and the pipeline system. The user's estimated task success ability for the dialogue in (a) is 0.195 while that for the dialogue in (b) is -1.116. In the dialogue shown in (a), the system responds appropriately to the user's utterance, indicating that the user understands what to say to the system. Specifically, when the user conveys their preferred departure time for the train to the system, they provide the information in a complete sentence rather than just a single word, thus enabling the system to understand the user's intent. In contrast, in the dialogue shown in (b), the user provides only a single word to convey the desired time for the train, and the system fails to understand the user's intent.