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

Ryu Hirai, Ao Guo, Ryuichiro Higashinaka (Graduate School of Informatics, Nagoya University)

Nagoya University Dialogue System Research Group

Overview

Background

- Task-oriented dialogue systems are widely used in our daily lives

Problem

- Due to some users having limited knowledge about the system, not all users can fully accomplish their tasks

Goal

- Construct a system that can **estimate the user's task success ability** so that the system can adapt to that ability

Proposed Method

- Estimate user's task success ability by **item response theory** (IRT)
  - Item response theory is a measurement theory that quantifies examinees' abilities on tests and commonly used in educational fields

Experiment

- We collected dialogues between the MultiWOZ-based systems and users and predicted the probability of a correct answer to each slot
- The proposed method **significantly outperformed baselines**

1 Proposed Method

1 Dialogue Collection

- Present dialogue goals to users and engage them in dialogue

Dialogue goal

	Domain	Slot	Value
Inform	Restaurant	Area	East
Inform	Restaurant	Food	Pizza
…	…	…	…

2 Judging Correctness of Each Slots

- We regard each dialogue as a single test
- We consider whether each slot is filled in correctly as a problem
- We compare the dialogue goal and the belief state to judge the correctness of each slot

Belief state at the end of dialogue

Domain	Slot	Value
Restaurant	Area	West
	Food	Pizza
	Price	

Compare

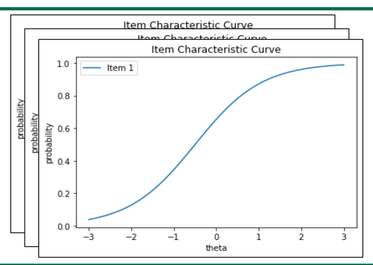
Dialogue goal

	Domain	Slot	Value	Label
Inform	Restaurant	Area	East	incorrect
Inform	Restaurant	Food	Pizza	correct
…	…	…	…	

Labels denoting if each slot was correctly filled or not

3 Estimating Item Characteristics

- We use IRT to estimate item characteristics by means of marginal maximum likelihood estimation



4 Dialogue with New User

- New user engages in a dialogue for a given dialogue goal
- Judge whether each slot is correctly filled

Dialogue goal with correct/incorrect labels for slots

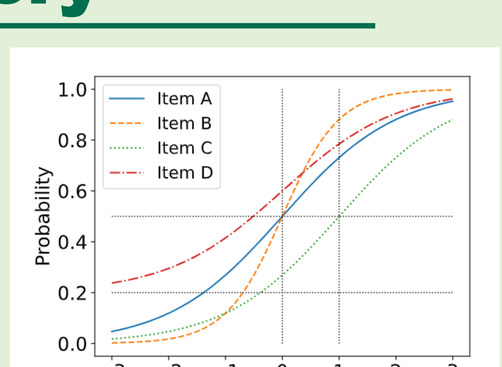
	Domain	Slot	Value	
Inform	Hotel	Area	North	correct
Inform	Hotel	Price	Moderate	incorrect
…	…	…	…	

5 Estimating Users' Task Success Abilities

- We estimate the task success ability by using item characteristics by expected a posteriori estimation

Item Response Theory

- We can estimate both users' abilities and item characteristics (discrimination, difficulty and guessing)
  - The relationship between the examinee's abilities θ and the probabilities of correct answers to questions $prob$ is calculated for each question



Examples of item characteristics curves

2 Experiment

Experimental Procedure

- We collected dialogues from 477 users via crowdsourcing
  - We built two systems using the **MultiWOZ 2.1** dataset [Eric+ 2019]
    - **Pipeline** [Zhang+ 2020] and **SimpleTOD** [Hosseini-Asl+ 2020]
  - Each user engaged in three consecutive dialogues with Pipeline or SimpleTOD with randomly generated dialogue goals
- We predicted the probabilities of correct answers
  - 5-fold cross validation ($train$: $test$ = 4: 1)
    - Estimate item characteristics from the train folds
    - Estimate user's task success abilities from the first dialogue of the test fold
    - Predict the probabilities of correct answers to each slot in the second and third dialogue of the test fold

Baselines

- **Baseline (Slot)** uses the average probability of a correct answer for a target slot as the probability of a correct answer for the slot
- **Baseline (User)** uses the average 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

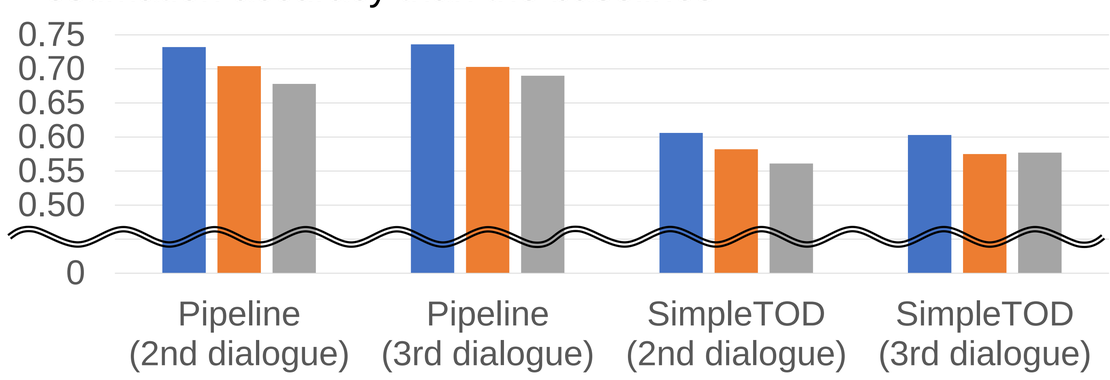
Evaluation Metric

- Accuracy of estimating the probabilities of correct answers
  - $prob$: probability of a correct answer to each slot
  - $ans \in \{0, 1\}$: actual correctness of the slot

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

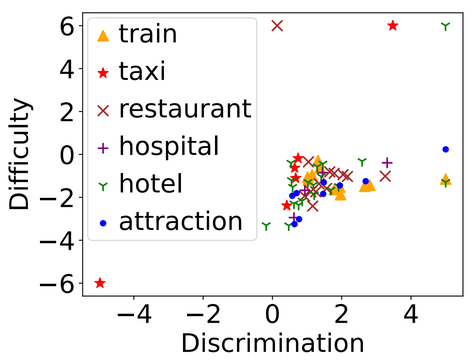
3 Result

- The proposed method achieved a significantly higher estimation accuracy than the baselines

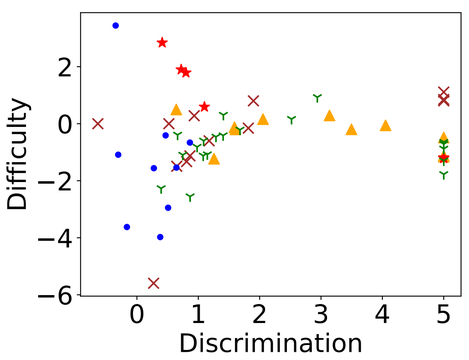


Accuracy of estimating probabilities of correct answers

- Different slots have different item characteristics
- We can create appropriate tests by selecting slots with high discrimination



(a) Pipeline



(b) SimpleTOD