them in terms of **Recall@N**, which reflects how many intents among all potential intents of q are retrieved among the N intents emitted by the model.

The key desideratum for label recommendation models is to cover as many potential questions as possible. It is relatively fair to compare the recall of potential intents recommended by different methods on the annotated data set. For label trajectory τ_N, the recall can be computed as

$$\operatorname{recall}(q, \tau_N) = \frac{\sum_{x \in \tau_N} |\mathcal{M}(x) \bigcap \mathcal{Q}(q)|}{|\mathcal{Q}(q)|} \tag{13}$$

where $\mathcal{Q}(q)$ is the set of potential intents for ambiguous query q, and $\mathcal{M}(x)$ is the set of all intents mapped to intent x in the intent inventory. The upper bound is calculated inversely from the results of annotated corpora:

$$\tau_N^*(q) = \underset{\tau_N}{\operatorname{argmax}} \sum_{x \in \tau_N} |\mathcal{M}(x) \cap \mathcal{Q}(q)| \tag{14}$$

$\tau_N^*(q)$ denotes the set of N best labels covering the potential intents. Thus, the upper bound recall of q would be recall$(\tau_N^*(q))$.

Evaluation metrics for online experiments. In our subsequent online experiments, our key metrics are the rate of transferal to human agents (THA) and the click through rate (CTR). In our experiments, every time a question is classified as an ambiguous question, six labels are provided to the user, who may select one of them or just ignore the selection. Given t as the number of times we output labels, and c as the number of times the user selected one of them, we define $\mathrm{CTR} = \frac{c}{t}$. Note that the user may opt to select *none of the above options*. In this case, the pipeline equals intent retrieval without clarification. The CTR reflects how useful the recommended labels are to users.

Evaluation metrics for complementary experiments. We compare the repetition rate at the word piece level of labels generated by two methods as an experiment to evaluate the diversity. The diversity is quantified as:

$$\operatorname{div}(\tau_N) = \frac{|\mathcal{W}(\tau_N)|}{\sum\limits_{w \in \mathcal{W}(\tau_N)} C(w)}, \tag{15}$$

where $\mathcal{W}(\tau_N)$ is the set of word pieces tokenized from the labels, and $C(w)$ denotes the number of times word piece w appears among the labels. We also count the overlap rate:

$$\operatorname{overlap}(\tau_N, q) = \frac{\sum\limits_{x \in \tau_N} \sum\limits_{t \in \mathcal{T}(x) \bigcap \mathcal{T}(q)} C(t)}{\sum\limits_{x \in \tau_N} \sum\limits_{t \in \mathcal{T}(x)} C(t)} \tag{16}$$

Here, $\mathcal{T}(x_t)$, $\mathcal{T}(q)$ denote the tokens sets of x_t and q, respectively. The overlap thus essentially reflects the number of tokens of labels appearing in a query.

5.3 Baselines

Several methods for label clarification serve as baselines for the offline experiments, while our method is denoted as *RL (ours)*.

5.3.1 Label Clarification Methods

Supervised. Given a query and a set of potential intents, there are limited labels related to the potential intents set. Traverse all possible label sequences over the limited labels set and choose the one with the highest rewards as the ground truth. If there are multiple sequences corresponding to the highest reward, pick one randomly.

Greedy. Given a user question, we train a classification model on the annotated corpus of ambiguous questions and the corresponding potential intents by minimizing the loss function

$$\mathcal{L} = \sum_q D_{\mathrm{KL}}[f_\theta(\cdot|q) \parallel P(\cdot|q)] \tag{17}$$

The classification model f_θ is used to estimate the probability distribution $P(\cdot|q)$ of the potential intents. Through this greedy method, our goal is to find a set of intents for which the sum of the probabilities of intents they cover is as high as possible. The greedy rule is given by $\mathrm{Score}(x_t) = \sum_{s \in \mathcal{D}(x_t)} f_\theta(s, q)$, where $\mathcal{D}(x_t)$ is the marginal recall of intents described in Section 4. At each time step t, we select the label with the highest score as x_t. Thus, the label set is generated by the rule.

RL (no state transition). As another baseline, we explore the implication of not taking recommended labels into account. This is a BERT classification model which outputs the intent with the highest probability at time step t and masks it at the next time step.

5.3.2 Ablation Study

Top-K intents. To contrast the truncated interface with the original, full interface, we retrieve the m most similar intents in terms of semantic similarity without interacting with users. The detail of intent retrieval is described below. (Note that for a label-oriented interface, after the user selects one label, the original query is concatenated with the label phrase as a new query and relevant intents are retrieved by the same model.)

Intent retrieval. For each query, a list of potential intents can be retrieved and ranked by BM25. We re-rank the candidates by applying BERT model to estimate the semantic similarity between query and each candidate. The model is a 12-layer BERT, which takes the concatenation of two sentences as input. Considering the display limitation of the dialogue bot environment, the top three results are presented to the user.

5.4 Results

Offline experiments. The experimental results in Table 2 show that our method significantly outperforms others. The greedy method has limited recall due to its reliance on the accuracy of its classification model. We observed that it is difficult to achieve a satisfactory recall by estimating potential intent probabilities through a classification model.

Our policy model also significantly outperforms the model without state transitions, confirming the need for considering the action history. The labels recommended by simple classification models do not yield sufficient diversity, resulting in very low recall. By modeling the problem as a seq2seq one, our model learns to recommend a next label that differs from previous ones, thereby improving the recall of potential intents.

It is worth noting that the supervised method outperforms all other baselines except ours. We believe that it does not explore the training data sufficiently. In most cases, there are multiple label sequences that can get similar rewards, and the supervised method can only consider one of them as the ground truth, remaining unable to explore equally good or second-best paths, which leads to insufficient exploration of labels. Thus, the search of the supervised method is not as exhaustive as our method's. Our results are close to the theoretical upper bound, which is further corroborates the effectiveness of our method.

	labels=3	**labels=6**
Greedy	19.47%	32.01%
Supervised	45.72%	51.53%
RL (no state transition)	17.23%	29.83%
RL (ours)	**52.45%**	**57.22%**
Upper bound	60.46%	67.34%

Table 2: Offline experimental results.

Online experiments. The offline experimental results show that *RL (no state transition)* and the *Greedy* method do not perform well, leaving only *RL (ours)* for the online experiments. Here we mainly compare the performance of two rewards: recall only and reward + entropy. We compare label recommendation methods and perform an ablation study using real online user clicks. For this, we collected data over a period of two weeks in our real deployment. The experimental results, illustrated in Table 3, show that the CTR of *RL (ours)* is significantly higher than for *RL (recall)*. We believe that this gap objectively

	THA	**CTR**
Top-K intents	15.40%	-
RL (recall)	14.51%	62.61%
RL (ours)	**14.20%**	**66.36%**

Table 3: Online experimental results.

reflects the importance of entropy to improve the quality of the label set. Furthermore, *RL (ours)* also outperforms *RL (recall)* with regard to the rate of transferal to human agents (THA). The Top-k intents method directly retrieves the most relevant three questions without interacting with users. The THA gap between Top-K intents and RL based methods reflects the contribution of label clarification. The experiments show that our method has a positive effect with regard to the system's ability to clarify ambiguous questions, reducing the workload of human agents.

5.5 Complementary Evaluation

How to apply	
RL (recall)	apply, register, credit card
RL (ours)	credit card, loan, QR code
How to claim insurance?	
RL (recall)	claim, health insurance, medical insurance
RL (ours)	health insurance, medical insurance, homeowners insurance
What was the payment just now?	
RL (recall)	transaction records, inquire, transfer money
RL (ours)	billing details, transition records, inquire records

Table 4: Excerpts of outputs from different methods. For simplicity, only the first three are displayed.

By inspecting specific cases, we find that the main difference between *RL (recall)* and *RL (ours)* is the complementarity with the user's question. Taking "How to apply" in Table 4 as an example, *RL (recall)* selects "apply", "register", which exhibit semantic overlap with the question itself. Though these may lead to improved recall of potential intents, they do not enable any further clarification. The results of *RL (ours)* include products that one can apply for, helping to establish the user's underlying intent. For a recall-only approach, the labels that yield the highest rewards must be the ones with the highest semantic overlap. Hence, it is inevitable that repetitive information will be chosen, thereby making a part of the label set redundant.

To verify our conjecture, we compare the diversity and complementarity using the indicators introduced in Section 5.2. Although the two indicators are not precise metrics for diversity and semantic overlap, they help to assess the gap of the models trained with the two different reward mechanisms. As we can see from Table 5, the reinforcement learning methods significantly surpass the *Greedy* method on diversity, but the two RL methods are comparable to each other. This illustrates that recall as a reward is a major contribution to diversity. On its own, the overlap indicator is not meaningful, as it can be reduced to 0 by recommending irrelevant labels. But along with the recall, the difference in overlapping rate illustrates the effectiveness on reducing semantic repetition. Therefore, the proposed reward is superior to all other compared methods.

	Diversity	**Overlap**
Greedy	75.27%	6.39%
RL (recall)	79.92%	9.69%
RL (ours)	**80.10%**	**7.69%**

Table 5: Complementarity evaluation: The lower the overlap, the better the complementarity.

6 Conclusion

We present an end-to-end model to resolve ambiguous questions in dialogue by clarifying them using label suggestions. We cast the question clarification problem as a collection partition problem. In order to improve the quality of the interactive labels as well as reduce the semantic overlap of the labels and the user's question, we propose a novel reward based on recall of potential intents and information gain. We establish its effectiveness in a series of experiments, which suggest that this novel notion of clarification may as well be adopted for other kinds of disambiguation problems.