Active Discovering New Slots for Task-oriented Conversation

Yuxia Wu*, Tianhao Dai*, Zhedong Zheng, Lizi Liao†

***Abstract*—Existing task-oriented conversational systems heavily rely on domain ontologies with pre-defined slots and candidate values. In practical settings, these prerequisites are hard to meet, due to the emerging new user requirements and ever-changing scenarios. To mitigate these issues for better interaction performance, there are efforts working towards detecting out-of-vocabulary values or discovering new slots under unsupervised or semi-supervised learning paradigms. However, overemphasizing on the conversation data patterns alone induces these methods to yield noisy and arbitrary slot results. To facilitate the pragmatic utility, real-world systems tend to provide a stringent amount of human labeling quota, which offers an authoritative way to obtain accurate and meaningful slot assignments. Nonetheless, it also brings forward the high requirement of utilizing such quota efficiently. Hence, we formulate a general new slot discovery task in an information extraction fashion and incorporate it into an active learning framework to realize human-in-the-loop learning. Specifically, we leverage existing language tools to extract value candidates where the corresponding labels are further leveraged as weak supervision signals. Based on these, we propose a bi-criteria selection scheme which incorporates two major strategies, namely, uncertainty-based and diversity-based sampling to efficiently identify terms of interest. We conduct extensive experiments on several public datasets and compare with a bunch of competitive baselines to demonstrate the effectiveness of our method.**

***Index Terms*—New slot discovery, Task-oriented conversation, Active learning, Language processing**

I. Introduction

With the development of smart assistants (*e.g.*, Alexa, Siri), conversational systems play an increasing role in helping users with tasks, such as searching for restaurants, hotels, or general information. Slot filling has been the main technique for understanding user queries in deployed systems, which heavily relies on pre-defined ontologies [1, 2, 3] However, many new places, concepts or even application scenarios are springing up constantly [4]. Existing ontologies inevitably fall short of hands, which hurts the system performance and reliability. As one of the foundation blocks in ontology learning, new slot discovery is particularly crucial in those deployed systems. It not only discovers potential new concepts for later stage ontology construction or update, but also helps to avoid incorrect answers or abnormal actions.

Generally speaking, new slot discovery requires handling two situations properly as illustrated in Figure 1: to recognize out-of-vocabulary values for pre-defined slots, and to group certain related values into new slots (as in dotted circles). Existing works tend to **separate** these two situations into two independent tasks for ease of modeling: (1) In the first new value discovery task, several pioneering works leverage character embeddings to handle the unseen words during training [5] while others harness the copy mechanism for selection [6]. There are also methods making use of background knowledge [7, 8]. The core of such methods lies in finding the patterns or relations of existing values among predefined slots. (2) For the second new slot scenario, it is more complicated and requires grouping the values into different slot types even without knowing the exact number of new slots. To simplify the problem, Wu et al. [9] proposed a novel slot detection task without differentiating the exact new slot names. For a more realistic setting, other researchers adapt transfer learning to leverage the knowledge in the source domain to discover new slots in the target domain [10]. They assume that the slot descriptions or even some example values are available. However, such availability is still less likely in practice. Hence, another line of research efforts seek help from existing tools such as semantic parser or other information extraction tools to gain knowledge [11]. Nonetheless, such methods suffer from the noisy nature of dialogue data and require intensive human decisions in various processing stages and settings.

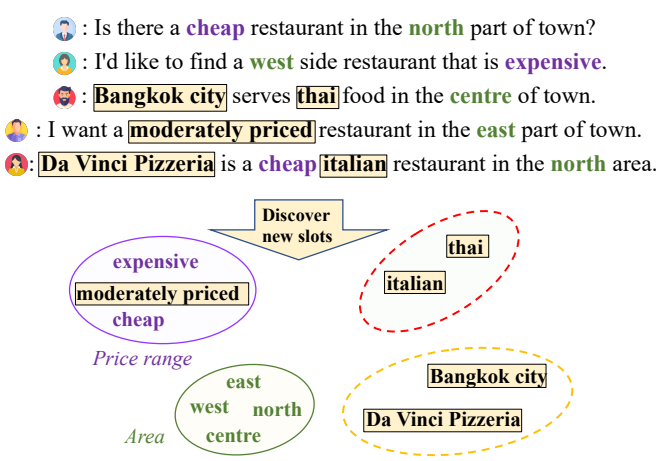


Fig. 1: Illustration of the new slot discovery task. It not only finds new values for predefined slots (e.g., *Price range*, *Area* in solid circles), but also discovers new slots with values (in dotted circles). The phrases w/o rectangular represent the known values and new values, respectively.

Yuxia Wu and Lizi Liao are with the Singapore Management University (e-mail: yieshah2017@gmail.com, lzliao@smu.edu.sg).

Tianhao Dai is with Wuhan University (e-mail: tianhao.dai@outlook.com).

Zhedong Zheng is with the Faculty of Science and Technology, and Institute of Collaborative Innovation, University of Macau (e-mail: zhedongzheng@um.edu.mo).

* Co-first authors with equal contribution.

† Corresponding author.

The current popular sequence labeling way emphasizes the relationship patterns in word or token sequences and labels, which is less sufficient for out-of-scope slots. As the new value and new slot discovery are inherently intertwined, we propose to adopt an Information Extraction (IE) fashion to tackle them concurrently as a general new slot discovery task. Candidate values are extracted firstly, which are then leveraged to find group structures. Nonetheless, if we obtain group structures purely based on data patterns, the resulting slots will tend to be noisy and arbitrary. Fortunately, a stringent amount of human labeling quota is usually available to facilitate the pragmatic utility, which offers an authoritative way to obtain accurate and meaningful slot assignments. To utilize such quota efficiently, a viable way is to adopt the active learning (AL) scheme [12] to progressively select and annotate data to expand our slot set. In general, existing active learning methods can be categorized into two major groups based on the sample selection strategy: uncertainty-based, diversity-based [13]. The former tries to find hard examples using heuristics like highest entropy or margin and so on [14, 15], while the latter aims to select a diverse set to alleviate the redundancy issue [16, 17]. Although there are works combining these two kinds of strategies and working well on the sequence tagging task [18, 12], their success is not directly applicable to our setting, because one sequence might contain multiple different slots and the goal of finding new slots is less emphasized in these sequence labeling models when label sets are known.

In this work, we formulate the general new slot discovery task in an information extraction fashion and design a Bi-criteria active learning scheme to efficiently leverage limited human labeling quota for discovering high-quality slot labels. The IE task can naturally fit the proposed active learning procedure. It allows our method to focus on only one of the slots in the input sentence during the sample selection. Specifically, we make use of the existing well-trained language tools to extract value candidates and corresponding weak labels. Being applied as weak supervision signals, these weak labels are integrated into a BERT-based slot classification model via multi-task learning to guide the training process. With the properly trained model, we further design a Bi-criteria sample selection scheme to efficiently select samples of interest and solicit human labels. In particular, it incorporates both uncertainty-based sampling and diversity-based sampling strategies via maximal marginal relevance calculation, which strives to reduce redundancy while maintaining uncertainty levels in selecting samples.

To sum up, our contributions are three-fold:

- We formulate a general new slot discovery task that wraps up the new value and new slot scenarios. Formatted in an IE fashion, it benefits from existing language tools as weak supervision signals.
- We propose an efficient Bi-criteria active learning scheme to identify new slots. In particular, it incorporates both uncertainty and diversity-based strategies via maximal marginal relevance calculation.
- Extensive experiments verify the effectiveness of the proposed method and show that it can largely reduce human labeling efforts while maintaining competitive performance.

II. Related Work

A. Out-of-Vocabulary Detection

New slot discovery aims to discover potential new slots for conversation ontology construction or update. It is closely related to the Out-of-Vocabulary (OOV) detection task that aims to find new slot values for existing slots. Under this task setting, the slot structures are predefined. For example, given the slots such as *Price range* and *Area*, it aims to find new values such as *moderately priced* to enrich the value set. Liang et al. [5] combined the word-level and character-level representations to deal with the out-of-vocabulary words. They treated the characters as atomic units which can learn the representations of new words. Zhao and Feng [6] leveraged the copy mechanism based on pointer network. The model is learned to decide whether to copy candidate words from the input utterance or generate a word from the vocabulary. Chen et al. [19] trained BERT [20] for slot value span prediction which is also capable of detecting out-of-vocabulary values. He et al. [7] proposed a background knowledge enhanced model to deal with OOV tokens. The knowledge graph provides explicit lexical relations among slots and values to help recognize the unseen values. More recently, Coope et al. [8] regarded the slot filling task as span extraction problem. They integrate the large-scale pre-trained conversational model to few-shot slot filling which can also handle the OOV values.

B. New Slot Discovery

Finding new slots requires proper estimation of the number and structural composition of new slots. As this is hard, there are efforts assuming that the slot descriptions of new slots or even some example values for these slots are available. These slot description or example values are directly interacted with user utterances to extract the values for each new slot individually [21, 22]. However, the over-reliance on slot descriptions hinders the generality and applicability of such methods. There are works trying to ignore such information. For example, Wu et al. [9] proposed a novel slot detection task to identify whether a slot is new or old without further grouping them into different classes. The Out-of-Distribution detection algorithms (such as MSP [23] and GDA [24]) are leveraged to fulfill the task. However, they only worked on simulated datasets and the task scenario is oversimplified.

Hence, researchers proposed a two-stage pipeline which first extracts slot candidates and values using information extraction tools, and then utilizes various ranking or clustering methods to pick out salient slots and corresponding values. For example, Chen et al. [11] combined semantic frame parsing with word embeddings for slot induction. In the same line, they further constructed lexical knowledge graphs and performed a random walk to get slots. Although the language tools provide useful clues for the later stage slot discovery, such methods suffer from the noisy nature of dialogue data and the selection, ranking process requires intensive human involvement. To mitigate this issues, Hudeček et al. [25] extended the ranking into an iterative process and built a slot tagger based on sequence labeling model for achieving higher recall. Nonetheless, the

model still relies on obtained slots in the former iterative process which requires intensive human decisions.

C. Active Learning

Active learning (AL) [26] reduces the demand for abundant labeled data by intelligently selecting unlabeled examples for iterative expert annotation, demonstrating its value in various natural language processing tasks [18]. There are two major sample selection strategies for active learning, namely, uncertainty-based and diversity-based sampling [17]. Uncertainty-based sampling selects new samples that maximally reduce the uncertainty the algorithm has on the target classifier [27]. However, a previous work points out that focusing only on the uncertainty leads to a sampling bias [14]. It creates a pathological scenario where selected samples are highly similar to each other. This may cause problems, especially in the case of noisy and redundant real-world datasets. Another approach is diversity-based sampling, wherein the model selects a diverse set such that it represents the input space without adding considerable redundancy [28]. Certain recent studies for classification tasks adapt the algorithm BADGE [17]. It first computes embedding for each unlabeled sample based on induced gradients, and then geometrically picks the instances from the space to ensure their diversity.

More recently, several existing approaches support a hybrid of uncertainty-based sampling and diversity-based sampling [13]. For instance, Hazra et al. [12] proposed to leverage sample similarities to reduce redundancy on top of various uncertainty-based strategies as a two-stage process. Better performances achieved signal a potential direction to further reduce human labeling efforts. At the same time, Shelmanov et al. [29] investigated various pre-trained models and applied Bayesian active learning to sequence tagging tasks. Experiments also showed better performance as compared to those single strategy based ones. In our work, we take advantage of pre-trained models such as BERT, and design a Bi-criteria active learning scheme to possess the benefits of both uncertainty-based and diversity-based sampling strategy.

The main differences between the proposed method and the related work are: 1) Our method only needs a few annotated data rather than extra prior knowledge such as slot descriptions or example values. 2) Compared with the new slot detection method, our model further organizes the new slots into different categories. 3) Compared with the weak supervised or unsupervised methods, our method mitigates human efforts such as selecting and ranking the candidate slots. Besides, we formulate slot discovery as an information extraction task to better capture the relationship among different values.

III. Problem Formulation

A. Background

Current task-oriented dialogue systems heavily rely on slot filling where an ontology O is usually provided with slots S and some candidate values. Existing approaches typically model it as a sequence labeling problem using RNN [30, 31] or pre-trained language models such as BERT [32, 33]. Given an utterance $X = \{x_1; x_2; \quad ; x_N\}$ with N tokens, the target of slot filling is to predict a label sequence $L = \{l_1; l_2; \quad ; l_N\}$ using BIO format. Each l_n belongs to three types: B-slot_type, I-slot_type, and O, where B- and I- represent the beginning and inside of one candidate value, respectively, and O means the token does not belong to any slot.

B. New Slot Discovery in an IE Fashion

Though popular [9, 25], the sequence labeling framework does not naturally fits the new slot discovery task well. First, the label set is not known beforehand in realistic settings. Second, sequence labeling models rely heavily on the linguistic patterns in utterance and the dependencies among the labels in label sequence. In fact, the candidate values are diverse in nature, they may reside in rather different dialogue contexts and show various linguistic patterns. Hence, it might be hard for sequence labeling models to take the dependencies between labels in the sequence into account [34]. Last but not the least, one utterance usually contains semantics about multiple slots. Thus the sample selection in active learning methods has to consider the scores of all tokens in a sentence, which leads to a mixed measure of the mutual interaction between different slots.

From another perspective, the general new slot discovery task covers the new value and new slot scenarios, which naturally fits the information extraction framework where we first extract value candidates, then dispatch or group them into different slots. Under this framework, there are many off-the-shelf language tools available to assist the candidate values extraction and provide weak supervision signals to further assist the grouping stage [25].

1) Candidate Value Extraction and Filtering: To reduce the labeling effort, we first extract candidate values which can be a single word or a span of words conveying important semantics. Inspired from [25], we adopt a frame semantic parser SEMAFOR [16, 35] and named entities recognition (NER) to extract candidate values[1]. The tools also provide labels for the candidate values which can be regarded as weak signals for further model design. Trained on a general corpus, the semantic tools produce some irrelevant values for conversational search. We filter these using rules, excluding stop words, low-frequency words, and less useful terms like 'then', 'please' and so on.

2) Our New Slot Discovery Formulation: Our method tackles the limited labeled training data challenge in new slot discovery, reflecting real-world scenarios for developing conversational agents in novel domains or new task settings. We work with a set of limited labeled data D_l and a large amount of unlabeled data D_u containing new slot types. We design an active learning scheme to efficiently make use of limited human labeling resources for accurate new slot discovery.

Formally, given a candidate value $X_i^{i+k} = \{x_i; \quad ; x_{i+k}\}$ with $k+1$ tokens extracted from the utterance X, our goal is to identify the slot type y of X_i^{i+k}. Although we only have limited labeled data D_l which contains a set of (X_i^{i+k}, X, y) tuples at the beginning, we will iteratively select and annotate a sample set S from D_u to enrich the data D_l in our active learning scheme. Besides, we also have the weak label y_{weak}

[1]If the same token span is labeled multiple times by different annotation sources, the span is more likely to be considered as a candidate term.

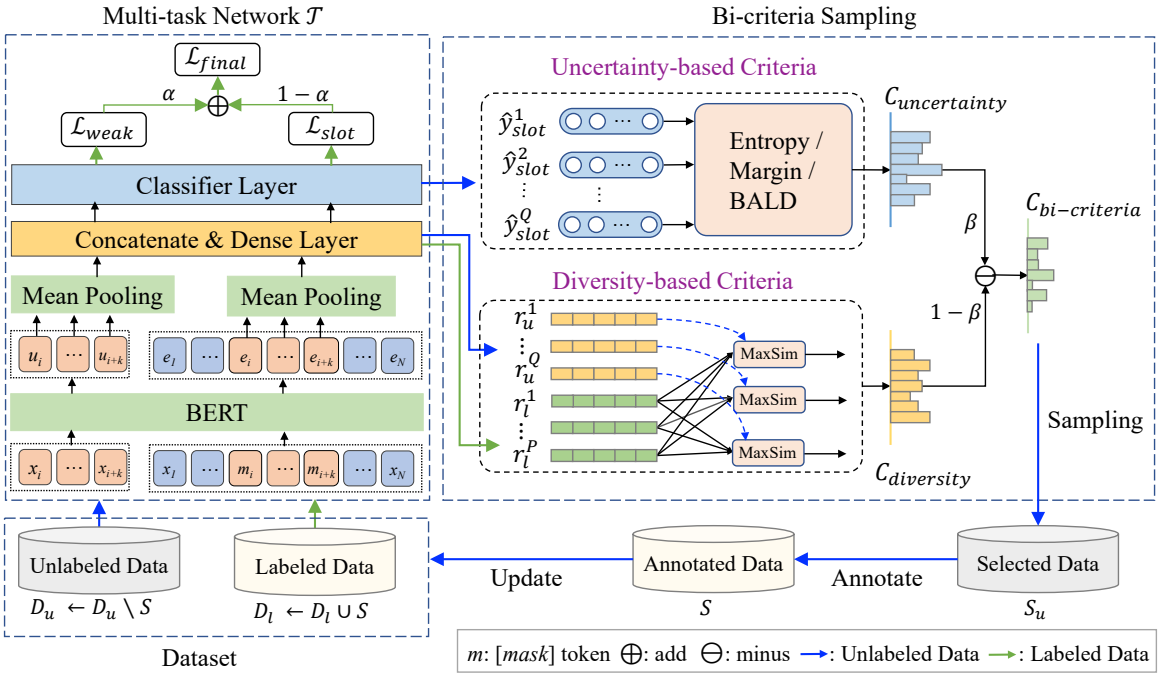


Fig. 2: The framework of the proposed Bi-criteria active learning scheme. For each iteration, the labeled data is utilized to train the multi-task network T. The unlabeled data is applied to select samples via Bi-criteria sampling strategy containing both uncertainty and diversity criteria, where BALD is an abbreviation for Bayesian Active Learning by Disagreement. Then the selected samples are annotated and applied to update the dataset for the next loop.

for the candidate value X_i^{i+k} which provides additional useful semantics for our model training and sample selection.

Note that the general new slot discovery task not only covers existing ontology update which identifies new candidate values and label them correctly to existing ontology O_{old}, but also includes ontology expansion where new slots are added to O_{old} to get O_{new}.

IV. Bi-criteria Active Learning Scheme

The proposed Bi-criteria active learning method is illustrated in Figure 2. The dataset contains labeled data and unlabeled data which is updated iteratively via active learning scheme. There are two stages in the iteration loop: multi-task network T training via labeled data and bi-criteria sampling from unlabeled data. The network T contains a BERT-based feature extractor and classifier layer. For feature extraction, we concatenate the representations of the candidate values and their context (the original sentence with the slot-value span replaced with [mask] per token). We train the multi-task network under the supervision of the weak signals from the NLP tools and the ground truth slot types of the candidate values. For the second stage, we first obtain the distributions of classification probabilities and representation features via the trained model T. Then a Bi-criteria strategy is specially designed to incorporate both uncertainty and diversity to select samples. The uncertainty is measured by the characteristics of the probability $\hat{\mathbf{y}}_{slot}$ via different strategies. The diversity is computed based on the representations of each sample. Two criteria are integrated by a balanced weight. Finally, the selected samples S_u are annotated and are applied to update the dataset for the next loop. We will introduce more details about our framework in the following parts.

The active learning loop is illustrated in Algorithm 1 for better understanding. The classification model T is first trained on 5% of the whole training dataset denoted as D_l. And then different active learning strategies can be applied to select unlabeled samples. After that, we annotate the selected samples S and add them into the labeled dataset D_l. The iteration will stop when $jD_uj = 0$ which means there is no more unlabeled data (or stop when model performance no longer increases).

Algorithm 1: Active Learning Scheme

Data: D: training dataset
Input: D_l 5% of dataset D ; // labeled data
D_u D D_l ; // unlabeled data
Output: Well-trained model T for new slots discovery
Initialization: D_l
T TRAIN(D_l) ; // train with init labeled data
/* Now starts active learning */
while *not converged and* $jD_uj>0$ **do**
  /* selection */
  S_u $Select_{Bi\ Criteria}(D_u; T)$
  S $Annotate(S_u)$
  D_l D_l [S ; // update labeled data
  D_u D_u n S; // update unlabeled data
  T Re-TRAIN(D_l)
return T

A. Multi-task Network

We first explain the base classification model. As mentioned before, we have some limited labeled data with ground truth values extracted from the input utterance and the corresponding slot types. We also obtain the candidate values and their weak labels by language tools. To effectively utilize the weak labels, we introduce a multi-task network to integrate them. Generally speaking, the model contains a feature extractor and a classifier layer with two branches, one for ground truth slot label prediction and the other for weak label prediction. Both branches share the same parameters of the feature extractor and are trained simultaneously. We also try an alternative way of using weak labels where two tasks are conducted chronologically. We show the comparison results and detailed analysis in Section V-F3. In the following parts, we introduce the feature extractor, classifier layer, and the loss function of the multi-task network.

1) Feature Extractor: Feature extraction for candidate value is the foundation of the subsequent processing for new slot discovery. Both the exact value and its context are essential for a task-oriented conversation system to understand the intents of users. Therefore, we integrate the two kinds of representations for each candidate value to facilitate further slot discovery. Specifically, we apply the pre-trained BERT model as the backbone for feature extraction. For the inherent representation, we only consider the token sequence in the candidate value X_i^{i+k}. For the context representation, we learn the pure contextual semantics in the input utterance with the candidate value masked to avoid its influence. The detailed process is introduced as follows.

Given a candidate value $X_i^{i+k} = \{x_i, \ldots, x_{i+k}\}$ with $k+1$ tokens in the utterance X, the inherent representation is the mean pooling of all the tokens in X_i^{i+k}:

$$u_i, \ldots, u_{i+k} = \text{BERT}(x_i, \ldots, x_{i+k}); \quad (1)$$

$$r_{\text{inherent}} = \text{mean\_pooling}(u_i, \ldots, u_{i+k}); \quad (2)$$

where u_i represents the embedding of the token x_i obtained from the BERT model.

For the context representation of the candidate value, we assume that if two values have the same context, they should have similar representations for slot discovery. Therefore we replace the tokens belonging to a specific value span in the original utterance X with a special token [mask]. In this way, the utterance is reconstructed as $X' = \{x_1, \ldots, h[\text{mask}]_i, \ldots, [\text{mask}]_{i+k}i, \ldots, x_n\}$[2]. We also adopt the BERT model to obtain the representation of each token in X'. With the self-attention mechanism in BERT, the [mask] tokens aggregate the contextual semantics of the corresponding values. Hence, we adopt mean pooling on the output of these [mask] tokens to obtain the context representation:

$$e_1, \ldots, e_i, \ldots, e_{i+k}, \ldots, e_n = \text{BERT}(X'); \quad (3)$$

$$r_{\text{context}} = \text{mean\_pooling}(e_i, \ldots, e_{i+k}); \quad (4)$$

[2]Special tokens such as [CLS] in beginning and [SEP] at end are omitted for easy illustration.

where $h e_i, \ldots, e_{i+k} i$ denotes the embeddings of the [mask] tokens in the last hidden layer of BERT.

We concatenate the inherent and context representation and apply one linear layer followed by tanh activation as the final representation of the candidate value as follows:

$$r = \tanh(W_1[r_{\text{inherent}}; r_{\text{context}}]^T + b_1); \quad (5)$$

where W_1 and b_1 represent the learnable weights and bias.

2) Classifier Layer: Weak labels are derived from existing language tools, and they serve as a form of indirect or partial supervision. These tools extract value candidates from conversational data, providing insights into potential terms of interest. While supervised labels offer authoritative guidance, weak labels contribute by capturing nuanced patterns and variations in user expressions that may not be explicitly annotated in the supervised data. By incorporating both sources of information, our method gains a more comprehensive understanding of the conversational domain. The weak labels act as a supplementary source, helping the model generalize better to out-of-vocabulary terms and diverse conversational patterns. This synergy between weak and supervised labels enables our approach to navigate the challenges of real-world conversational search systems more effectively, striking a balance between authoritative guidance and adaptability to diverse user expressions. Thus we design two classifiers in the multi-task network. Specifically, we introduce two independent fully-connected layers to map the representation into the probabilities over ground truth slot labels and weak labels given by language tools, i.e.,

$$\hat{y}'_{\text{slot}} = \text{Softmax}(W_2 r^T + b_2); \quad (6)$$

$$\hat{y}_{\text{weak}} = \text{Softmax}(W_3 r^T + b_3); \quad (7)$$

where W_2, W_3, b_2 and b_3 represent the learnable weight matrices and biases. $\hat{y}'_{\text{slot}}$ and $\hat{y}_{\text{weak}}$ represent the predicted probability over all slot labels and weak labels respectively.

3) Loss Function: It is worth noticing that not all the slot labels may have been discovered during training so we apply a label mask to prevent the leakage of unknown labels:

$$\hat{y}_{\text{slot}} = \hat{y}'_{\text{slot}} \odot m; \quad (8)$$

where m is a vector with the same dimension as $\hat{y}'_{\text{slot}}$. Each element $m^{(i)}$ equals 1 if slot label i is known and 0 if unknown. The symbol $\odot$ denotes element-wise multiplication.

Finally, for each sample, given two predicted probability distributions $\hat{y}_{\text{slot}}$ and $\hat{y}_{\text{weak}}$, the final loss is constructed as:

$$L_{\text{final}} = (1-\alpha) L(\hat{y}_{\text{slot}}, y_{\text{slot}}) + \alpha L(\hat{y}_{\text{weak}}, y_{\text{weak}}); \quad (9)$$

where L represents the cross-entropy loss; y_{slot} and y_{weak} represent one-hot vectors of the slot label and the weak label of the sample respectively; the hyperparameter α adjusts how much weak supervision loss contributes to the final loss.

B. Uncertainty-based Criteria

In this section, we introduce three commonly-used uncertainty-based active learning strategies. We test the performance of each and integrate them into the proposed Bi-criteria active learning scheme to find the best setting.

Entropy Sampling: Given the predicted probability distribution $\hat{y}_{slot}$, the entropy score will be:

$$C_{entropy} = \sum_i \hat{y}_{slot}^{(i)} \log(\hat{y}_{slot}^{(i)}); \quad (10)$$

where i denotes each dimension of these vectors. We select samples where $C_{entropy} \quad _e$, where $_e$ is a hyperparameter.

Margin Sampling: Margin score is defined as the difference between the highest probability $\bar{y}_{slot}$ and the second highest probability $\bar{y}_{slot}$ obtained from the predicted $\hat{y}_{slot}$, i.e.,:

$$C_{margin} = \bar{y}_{slot} \quad \bar{y}_{slot}: \quad (11)$$

This strategy tries to find hard samples where $C_{margin} \quad _m$, where $_m$ is a hyperparameter.

Bayesian Active Learning by Disagreement (BALD): As discussed in [37], models with activating dropout produce a different output during multiple inferences. BALD [38] computes model uncertainty by exploiting the variance among different dropout results. Suppose $\bar{y}^t_{slot}$ is the best scoring output for X in the t th forward pass and T is the number of forward passes with a fixed dropout rate, and then we have:

$$C_{BALD} = 1 \quad \frac{\text{count}(\text{mode}(\bar{y}^1_{slot}; \bar{y}^2_{slot}; \quad ; \bar{y}^T_{slot}))}{T}; \quad (12)$$

where the $mode( )$ operation finds the output which is repeated most times, and the $count( )$ operation counts the number of times this output was repeated. This strategy selects unlabeled samples with $C_{BALD} \quad _b$, where $_b$ is a hyperparameter.

C. Infusing Diversity

As shown in Figure 3(a), simply relying on uncertainty-based criteria would invite the redundancy problem where samples of similar semantics and context are selected. To address this, we infuse diversity into the sampling strategy. Inspired by Maximal Marginal Relevance (MMR) in Information Retrieval [39], we develop a Bi-criteria sampling method which selects those unlabeled samples with high uncertainty and also diverse in meaning at the same time (Figure 3(c)). If we adopt the margin score as the uncertainty score, then the Bi-criteria score for each unlabeled sample q should be:

$$C_{bi \ criteria} = C_{margin}(q) \quad (1 \quad) \max_{p2P} Sim(r_l^p; r_u^q); \quad (13)$$

where P is the set of all labeled samples and p is the index of the labeled sample, r is the representation of the sample obtained from Equation (5); Sim stands for the cosine similarity between two representation vectors; is the hyperparameter that controls the contribution of uncertainty and diversity. Specially, when is set to 0, we get the purely diversity-based score as:

$$C^0_{diversity} = \max_{p2P} Sim(r_l^p; r_u^q): \quad (14)$$

Intuitively, the Diversity Sampling selects unlabeled samples by their distances from the nearest labeled sample in the feature space (Figure 3(b)). The larger that distance is, the more different in meaning the sample is from labeled sample sets. On the other hand, when is set to 1, Bi-criteria will be reduced to Margin Sampling, where diversity is no longer taken into account.

V. Experiments

A. Datasets

We adopt the datasets from [25], excluding CamRest and Cambridge SLU due to limited slot numbers. The statistics are detailed in Table I, with the known slots derived from the initial 5% randomly labeled data.

B. Implementation Details

Our model employs the "bert-base-cased" BERT version [30], optimizing with Adam [40] and a base learning rate of 5e-5. Linear decay is applied following [20]. The number of max initial training epochs is 30 and the batch size is 128. For each follow-up active learning iteration, we fine-tune the model on the updated labeled training set for two epochs following

(a) Uncertainty (b) Diversity (c) Bi-criteria

Fig. 3: The selected samples via different criteria. (Adapted from [36, Figure 3.9, Figure 4.1 and Figure 5.1].)

TABLE I: The statistic information of three datasets

Dataset	Domain	#Samples	#Slots		
			Known	New	Total
ATIS	Flight	4,978	54	25	79
WOZ-attr	Attraction	7,524	4	4	8
WOZ-hotel	Hotel	14,435	4	5	9

[12]. Each dataset is divided into training / testing / validation sets (0.8/0.1/0.1). A random 5% of the whole training set is chosen as a warm-up dataset. At each active learning iteration, 2% of new training samples are selected for annotation. For selection strategies based on the Monte Carlo dropout, we make five stochastic predictions.

C. Evaluation Metric

We evaluate the performance via the widely used classification metric, i.e., F1-score [41]. Suppose the ground-truth slot values are $M_1, M_2, ..., M_n$, where n denotes the number of slots. The predicted values are $E_1, E_2, ..., E_n$. For each slot type i, we first calculate the precision and recall score as:

$$P_i = \frac{|M_i \cap E_i|}{|E_i|}, \tag{15}$$

$$R_i = \frac{|M_i \cap E_i|}{|M_i|}. \tag{16}$$

Then the final F1 score is computed as:

$$F1 = \frac{2PR}{P + R} \tag{17}$$

$$P = \sum_{i=1}^{n} \frac{|E_i|}{\sum_{j=1}^{n} |E_j|} P_i, \tag{18}$$

$$R = \sum_{i=1}^{n} \frac{|M_i|}{\sum_{j=1}^{n} |M_j|} R_i. \tag{19}$$

Since the F1 score is calculated based on slot value spans, it is also called Span-F1 in this paper.

D. Competitive Methods

We compare our method with two groups of baselines: active learning methods and semi-supervised methods with 21% randomly labeled data. We utilize the same backbone for different methods for fair comparison.

Active learning methods (1) Random: The active learning method with random sampling strategy. (2) Uncertainty-based sampling: We compare our methods with several uncertainty-based strategies mentioned before including Entropy, Margin and BALD sampling. (3) Diversity-based sampling: As mentioned before, we set β as 0 to achieve the pure diversity sampling method. (4) Hybrid sampling: Active^2 Learning [12] is a two-stage hybrid sampling method. It first utilizes an uncertainty-based criterion to select a coarse sample set. Then an external corpus is adopted to assist the clustering step in order to ensure the diversity. To adapt [12] for a fair comparison, we choose the Margin Sampling as the uncertainty-based criterion. Then we naturally apply the weak labels obtained from the language tools to replace the extra clustering step in the second stage.

Semi-supervised methods: We compare our method with a semi-supervised new slot discovery model named SID [42] which is designed by incremental clustering. As we formulate the new slot discovery in an IE fashion, we actually transform the problem into an instance (one value candidate and its context) class discovery task which is rather close to the intent discovery setting. Hence, we also compare with the semi-supervised intent discovery methods CDAC+ [43] and DeepAligned [44]. We adapt them to our new slots discovery task since they are designed as a classification scheme.

LLM-based methods: In the era of large language model (LLM), models such as ChatGPT [45][3] have shown excellent performance in open-domain information extraction. Considering that we reformulate the new slot discovery in an IE fashion, we compare our method with several LLM models including ChatGPT and LLaMA [46] with few-shot learning. We employ the ChatGPT-3.5 and LLaMa-7B versions with few-shot settings for the task of new slot discovery with instruction template followed by [47]. To establish a fair comparison, we randomly select the same number of slots as known slots and proceed to randomly select 5 samples as prompt samples.

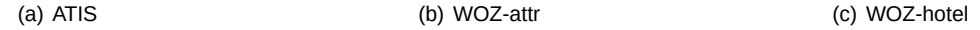
E. Quantitative Results

1) Active v.s. Semi-supervised: We report the results compared with semi-supervised methods in Table II. We can observe that our method outperforms all the semi-supervised methods on all three datasets. The proposed method surpasses the second-best method on ATIS, WOZ-attr, WOZ-hotel by 21.5%, 7.84%, and 20.69% respectively. The result demonstrates the effectiveness of using active learning and the strength of human labeling efforts. It is also shown that the SIC method has better performance than CDAC+ and DeepAligned. Specifically, SIC outperforms DeepAligned by 3.16%, 3.69%, 4.39% respectively on ATIS, WOZ-attr, and WOZ-hotel. It is worth noticing that there is a huge performance drop for the two methods on WOZ-hotel dataset. We suspect it is attributed to the fact that the distribution of the WOZ-hotel dataset is difficult to fit, especially for the CDAC+ method which overemphasizes the pairwise similarity as prior knowledge.

We note that the performance of the LLM models lags behind that of other methods. This disparity can be attributed to the LLM model's reliance on pre-training on general corpora, making it less tailored for the specific challenges posed by the novel slot discovery task within the target-oriented conversation domain. This highlights the importance of domain-specific fine-tuning to enhance the model's efficacy. ChatGPT consistently outperforms LLaMA across various datasets. This underscores the effectiveness of ChatGPT, particularly evident in the ATIS dataset characterized by a complex schema ontology.

2) Bi-Criteria v.s. Other Active Strategies: The results of experiments on three public datasets with different active learning strategies are presented in Figure 4. Due to the intrinsic

[3] https://chat.openai.com



(a) ATIS (b) WOZ-attr (c) WOZ-hotel

Fig. 4: The results of different active learning strategies on the three public datasets. All methods start from the same initial training checkpoint over 5% randomly sampled instances. These plots have been magni ed to highlight the regions of interest.

TABLE II: Comparison with other competitive semi-supervised methods. Here we provide the Span-F1 score.

Method	ATIS	WOZ-attr	WOZ-hotel
CDAC+ [43]	60.07	58.00	16.51
DeepAligned [44]	63.30	66.72	43.86
SIC [42]	66.46	70.41	48.25
LLaMA [46]	18.19*	44.17	30.86
ChatGPT [45]	71.23*	56.12	43.09
Ours (Bi-Criteria)	87.96	78.25	68.94

* Note that due to the complexity of ATIS dataset, we supply the LLM with the set of slot names in the instruction template and request it to select the appropriate slot name.

discrepancy among datasets, we set the in Equation(9) for each dataset differently (0.05 on ATIS and WOZ-attr, 0.1 on WOZ-hotel). As seen, the F1 scores signi cantly vary among different active learning strategies, and Bi-criteria generally performs the best on all three datasets in terms of accuracy and stability. The mean of differences between the best scores of bi-criteria and the best scores among other sampling strategies over all sampling steps is 0.61% on ATIS and 0.95% on WOZ-attr. On WOZ-hotel, though surpassed by BALD and Hybrid strategy at the 17 and 21 percent stages, Bi-criteria exhibits performance with less uctuation thus better stability.

As expected, Random Sampling strategy is generally overwhelmed by most active learning strategies most of the time, since neither redundancy nor diversity is concerned during the data selection. However, this tendency is less conspicuous on WOZ-attr, where Entropy Sampling and BALD perform worst.

Note that Margin Sampling and Diversity Sampling are the special cases of the Bi-criteria strategy when in Equation(13) is set to = 1 and = 0 respectively. It is easily observed from Figure 4 that Bi-criteria strategy outperforms both of the strategies in Span-F1 and stability. As the mixture of Margin Sampling and Diversity Sampling, Bi-criteria takes advantage of both uncertainty and diversity. It indicates that these two strategies are both essential components in terms of active selection and impact the results in a cooperating way to some extent. Further analysis could be found in Subsection V-F2.

F. Ablation Studies and Further Analysis

1) Effect of hyperparameter: We x the in Equation(13) and adjust in Equation(9) to see its effect on the performance of Bi-criteria active learning strategy. The hyperparameter indicates the proportion of weak supervision loss in the nal loss. According to our observation, the tends to have a relatively small effect on the performance compared with other parameters and therefore only four value settings are tested and shown here in Figure 5.

As is seen from the line charts in Figure 5, tuningto 0.05 leads to the performance with both better Span-F1 and stability compared with other settings on ATIS and WOZ-attr while at 0.1 results in the best stability and relatively high Span-F1 on WOZ-hotel. Moreover, method with at 0 does not perform best on all three datasets, which validates the usefulness of weak supervision. The mean of differences between the Span-F1 of the selected (red line in the graphs) and the Span-F1 of at 0 over all active learning steps is 0.36%, 1.61%, 0.36% on ATIS, WOZ-attr, and WOZ-hotel respectively. This result proves that weak supervision indeed boosts the performance of model on our task, thus necessitating the adoption of our multi-task network structure.

However, the performance does not necessarily improve as the proportion of weak supervision grows higher. This tendency is easily observed from the results on ATIS, where the performance declines as grows bigger from 0.05. Therefore, nding an appropriate weight for weak supervision is critical to our multi-task network.

2) Effect of hyperparameter: We also study the effect of the in Equation(13). Note that = 0 and = 1 are equivalent to purely Diversity Sampling and Margin Sampling respectively, performances of which have been shown in Figure 4. In general, the Bi-criteria method incorporating both uncertainty-based and diversity-based strategies tends to yield better results compared to using either strategy alone. Moreover, the weights of these two aspects also exert certain in uence on the performance. As Figure 6 shows, the Bi-criteria method achieves satisfying results whenis set to 0.9 on ATIS and WOZ-hotel and 0.7 on WOZ-attr. Experiments with below or equal to 0.5 generally achieve poor results compared to settings with higher . This indicates that uncertainty actually

(a) ATIS (b) WOZ-attr (c) WOZ-hotel

Fig. 5: Ablation study of on three public datasets. All methods start from the same initial training checkpoint over 5% randomly sampled instances. These plots have been magni ed to highlight the regions of interest.

(a) ATIS (b) WOZ-attr (c) WOZ-hotel

Fig. 6: Ablation study of on the three public datasets. All methods start from the same initial training checkpoint over 5% randomly sampled instances. These plots have been magni ed to highlight the regions of interest.

contributes more to the overall performance. However, the diversity signal is still indispensable since it helps to achieve results that Margin Sampling itself cannot.

3) Comparison with different kinds of weak supervision: We also explore different ways of using weak supervision. The key goal of weak supervision is to make use of existing weak labels to facilitate our task. In our proposed method, weak supervision is implemented in a multi-task fashion. Labels given by language tools are adopted to conduct an individual classi cation task, whose loss contributes to the nal loss. The alternative way is to pre-train the BERT model with these labels for classi cation rst, and then ne-tune the BERT parameters for the new training phase for our new slot discovery task with new classi er head.

TABLE III: Comparison with different kinds of weak supervision on three datasets. Here we provide the Span-F1 score.

Method	ATIS		WOZ-attr		WOZ-hotel	
	Start	End	Start	End	Start	End
no weak.	73.21	87.71	58.14	75.38	59.61	68.26
pretrain	74.61	87.85	59.21	74.15	61.86	67.95
multi-task	73.34	87.96	59.84	78.25	58.60	68.94

Table III shows the results under different kinds of weak supervision. These results represent the Span-F1 at the start point (5% labeled data) and the endpoint (21% labeled data) of the active learning process on three datasets. It can be seen that methods with weak supervision (pretrain and multi-task) achieve Span-F1 higher than the method without it both at the beginning and the end of the active learning process in all three datasets, which again demonstrates the effectiveness of weak supervision. It is worth noticing that when 5% training data are labeled, the pretraining method achieves Span-F1 higher than the second best method by 1.27% and 2.25% on ATIS and WOZ-hotel respectively. However, when 21% of training data are labeled, the multi-task method prevails. We can therefore infer that weak supervision as pre-training may enhance the starting point but tend to converge at a lower level than weak supervision as multi-task does in our setting.

4) Case study and error analysis: Upon meticulous examination of problematic samples, a signi cant portion of errors in slot lling arise from slot name misclassi cations, falling into three main categories. In the ATIS dataset, challenges emerge with directional slots like “fromloc.city_name” and “toloc.city_name”, highlighting the importance of contextual information and suggesting potential improvements through varying weights for values and context. Similar issues occur with numerical values, where the same value may belong to different slots. For instance, the model may misclassify numerical values, such as in “7 people and 5 nights” where “5” pertains to a new slot labeled "stay." Errors also arise

with specific location information, involving incorrect value boundaries or correct values assigned to the wrong slot. For example, in the utterance "I am looking for information regarding Magdalene College", our model annotates only 'college" instead of the correct location name "Magdalene College". Exploring the integration of external knowledge graph information about locations may address such errors.

5) Time efficiency analysis: As we adopt the BERT architecture in our work, the time complexity is $O(n^2)$ where n is the sequence length. To evaluate time efficiency, we conducted experiments on training and testing time across different methods using a machine equipped with an NVIDIA RTX 3090 GPU and 16 CPU cores. The results are presented in Table IV for three datasets. The result demonstrates that our method's training time is comparable to SIC, and faster than CDAC+ and DeepAligned. This efficiency is attributed to the absence of pair-wise similarity computation and clustering refinement during training. Notably, our method's testing time is considerably faster than other approaches, as it doesn't require clustering during testing. Despite increasing dataset size, our method maintains efficiency, showcasing its effectiveness.

TABLE IV: Comparison of the training and testing time (hours) of different methods on three datasets.

Method	ATIS		WOZ-attr		WOZ-hotel	
	Train	Test	Train	Test	Train	Test
CDAC+ [43]	8.01	0.15	10.01	0.09	9.29	0.26
DeepAligned [44]	8.06	0.14	10.05	0.09	16.10	0.25
SIC [42]	**2.34**	0.14	**1.32**	0.10	3.57	0.30
Ours (Bi-Criteria)	2.53	**0.05**	1.44	**0.02**	**2.72**	**0.03**

VI. Conclusion and Future Work

In this work, we formulated a general new slot discovery task for task-oriented conversational systems. We designed a bi-criteria active learning scheme for integrating both uncertainty-based and diversity-based active learning strategies. Specifically, to alleviate the limited labeled data problem, we leverage the existing language tools to extract the candidate values and pseudo labels as weak signals. Extensive experiments show that it effectively reduces human labeling effort while ensuring relatively competitive performance.

Future work involves exploring signals from abundant responses in dialogue datasets for guiding the sample selection process in active learning. Besides, during the training of AL, we fine-tune the model at each epoch with newly added samples. With the increase of the trained data, it could be computationally expensive. A possible way is to solely fine-tune the model on the newly labeled examples to avoid re-training from scratch. However, this will encounter the catastrophic forgetting problem which hurts the performance of the previously seen examples due to the shifting distribution of newly added samples. In future work, we will explore a more flexible training strategy to handle this issue.

Acknowledgments

This research is supported by the Ministry of Education, Singapore, under its AcRF Tier 2 Funding (Proposal ID: T2EP20123-0052). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

[1] Y. Wu, L. Liao, G. Zhang, W. Lei, G. Zhao, X. Qian, and T.-S. Chua, "State graph reasoning for multimodal conversational recommendation," *TMM*, 2022.

[2] L. Liao, T. Zhu, L. H. Long, and T. S. Chua, "Multi-domain dialogue state tracking with recursive inference," in *WWW*, 2021, pp. 2568–2577.

[3] J. Liu, M. Yu, Y. Chen, and J. Xu, "Cross-domain slot filling as machine reading comprehension: A new perspective," *TASLP*, vol. 30, pp. 673–685, 2022.

[4] B. Li, H. Fei, F. Li, Y. Wu, J. Zhang, S. Wu, J. Li, Y. Liu, L. Liao, T.-S. Chua *et al.*, "Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis," *ACL Findings*, 2023.

[5] D. Liang, W. Xu, and Y. Zhao, "Combining word-level and character-level representations for relation classification of informal text," in *Rep4NLP@ACL*, 2017, pp. 43–47.

[6] L. Zhao and Z. Feng, "Improving slot filling in spoken language understanding with joint pointer and attention," in *ACL*, 2018, pp. 426–431.

[7] K. He, Y. Yan, and W. Xu, "Learning to tag OOV tokens by integrating contextual representation and background knowledge," in *ACL*, 2020, pp. 619–624.

[8] S. Coope, T. Farghly, D. Gerz, I. Vulic, and M. Henderson, "Span-convert: Few-shot span extraction for dialog with pretrained conversational representations," in *ACL*, 2020, pp. 107–121.

[9] Y. Wu, Z. Zeng, K. He, H. Xu, Y. Yan, H. Jiang, and W. Xu, "Novel slot detection: A benchmark for discovering unknown slot types in the task-oriented dialogue system," in *ACL*, 2021, pp. 3484–3494.

[10] L. Wang, X. Li, J. Liu, K. He, Y. Yan, and W. Xu, "Bridge to target domain by prototypical contrastive learning and label confusion: Re-explore zero-shot learning for slot filling," in *EMNLP*, 2021, pp. 9474–9480.

[11] Y.-N. Chen, W. Y. Wang, and A. I. Rudnicky, "Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing," in *ASRU*, 2013, pp. 120–125.

[12] R. Hazra, P. Dutta, S. Gupta, M. A. Qaathir, and A. Dukkipati, "Active2 learning: Actively reducing redundancies in active learning methods for sequence tagging and machine translation learning," in *NAACL*, 2021, pp. 1982–1995.

[13] Y. Kim, "Deep active learning for sequence labeling based on diversity and uncertainty in gradient," in *ACL*, 2020, pp. 1–8.

[14] S. Dasgupta, "Two faces of active learning," *Theor. Comput. Sci.*, pp. 1767–1781, 2011.

[15] Z. Zheng and Y. Yang, "Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation," *IJCV*, pp. 1106–1120, 2021.

[16] D. Das, N. Schneider, D. Chen, and N. A. Smith, "Probabilistic frame-semantic parsing," in *NAACL*, 2010, pp. 948–956.

[17] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in *ICLR*, 2020.

[18] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," in *ICLR*, 2018.

[19] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," *arXiv preprint arXiv:1902.10909*, 2019.

[20] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[21] D. Shah, R. Gupta, A. Fayazi, and D. Hakkani-Tur, "Robust zero-shot cross-domain slot filling with example values," in *ACL*, 2019, pp. 5484–5490.

[22] Y. Hou, W. Che, Y. Lai, Z. Zhou, Y. Liu, H. Liu, and T. Liu, "Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network," in *ACL*, 2020, pp. 1381–1393.

[23] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR (Poster)*, 2016.

[24] H. Xu, K. He, Y. Yan, S. Liu, Z. Liu, and W. Xu, "A deep generative distance-based classifier for out-of-domain detection with mahalanobis space," in *COLING*, 2020, pp. 1452–1460.

[25] V. Hudeček, O. Dušek, and Z. Yu, "Discovering dialogue slots with weak supervision," in *IJCNLP*, 2021, pp. 2430–2442.

[26] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, no. 2, pp. 133–168, 1997.

[27] A. Siddhant and Z. C. Lipton, "Deep bayesian active learning for natural language processing: Results of a large-scale empirical study," in *EMNLP*, 2018, pp. 2904–2909.

[28] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *ICLR (Poster)*, 2018.

[29] A. Shelmanov, D. Puzyrev, L. Kupriyanova, N. Khromov, D. Dylov, A. Panchenko, D. Belyakov, D. Larionov, E. Artemova, and O. Kozlova, "Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates," in *EACL*, 2021, pp. 1698–1712.

[30] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *NAACL*, 2018, pp. 753–757.

[31] S. Zhu, Z. Zhao, R. Ma, and K. Yu, "Prior knowledge driven label embedding for slot filling in natural language understanding," *TASLP*, vol. 28, pp. 1440–1451, 2020.

[32] L. Liao, L. H. Long, Y. Ma, W. Lei, and T.-S. Chua, "Dialogue state tracking with incremental reasoning," *TACL*, vol. 9, pp. 557–569, 2021.

[33] J. Liang and L. Liao, "Clusterprompt: Cluster semantic enhanced prompt learning for new intent discovery," in *EMNLP Findings*, 2023.

[34] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," in *ACL*, 2016, pp. 1064–1074.

[35] D. Das, D. Chen, A. F. T. Martins, N. Schneider, and N. A. Smith, "Frame-semantic parsing," *Comput. linguistics*, vol. 40, no. 1, pp. 9–56, 2014.

[36] R. M. Monarch, *Human-in-the-loop machine learning: active learning and annotation for human-centered AI*, 2021.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *JMLR*, pp. 1929–1958, 2014.

[38] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *stat*, vol. 1050, p. 24, 2011.

[39] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *SIGIR*, 1998, pp. 335–336.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[41] E. T. K. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *CoNLL*, 2003, pp. 142–147.

[42] Y. Wu, L. Liao, X. Qian, and T.-S. Chua, "Semi-supervised new slot discovery with incremental clustering," in *EMNLP Findings*, 2022, pp. 6207–6218.

[43] T.-E. Lin, H. Xu, and H. Zhang, "Discovering new intents via constrained deep adaptive clustering with cluster refinement," in *AAAI*, 2020, pp. 8360–8367.

[44] H. Zhang, H. Xu, T.-E. Lin, and R. Lyu, "Discovering new intents with deep aligned clustering," in *AAAI*, 2021, pp. 14 365–14 373.

[45] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[46] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[47] Y. Feng, Z. Lu, B. Liu, L. Zhan, and X.-M. Wu, "Towards llm-driven dialogue state tracking," in *EMNLP*, 2023, pp. 739–755.

Yuxia Wu is a research scientist at Singapore Management University. She received the Ph.D. degree from Xi'an Jiaotong University in 2023. Her research interests include natural language processing, social multimedia mining, graph mining and recommender systems. She has served as the reviewer and program committee member for multiple conferences and journals, including TPAMI, TKDE, ACL, EMNLP, ACM MM etc.

Tianhao Dai is currently pursuing his undergraduate degree in the School of Cyber Science and Engineering at Wuhan University. From August 2022 to January 2023, he served as a visiting research student at Singapore Management University under the supervision of Assistant Professor Lizi Liao. His research interests include the field of natural language processing and computational linguistics.

Zhedong Zheng is an assistant professor with the University of Macau. He received the Ph.D. degree from the University of Technology Sydney, Australia, in 2021 and the B.S. degree from Fudan University, China, in 2016. He was a postdoctoral research fellow at the School of Computing, National University of Singapore. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He has served as the reviewer and program committee member for multiple conferences and journals, including TPAMI, IJCV, NeurIPS, CVPR and ICCV.

Lizi Liao is an assistant professor at the Singapore Management University. She obtained her Ph.D. from National University of Singapore in 2019. Dr. Liao's research interests center on task-oriented dialogues, proactive conversational agents, and multimodal conversational search and recommendation as the application target. She publishes regularly in top conferences and journals such as ACL, SIGIR, WWW, ACM MM, TKDE etc. One of her work was nominated in the ACM MM Best Paper Final List in 2018. She serves as senior PC member or area chair of these prestigious conferences and organizing committee members of SIGIR 2024, WWW 2024, WSDM 2023 and ACM MM 2019 etc.