ASSIST: Towards Label Noise-Robust Dialogue State Tracking

Anonymous ACL submission

Abstract

The MultiWOZ 2.0 dataset has greatly boosted the research on dialogue state tracking (DST). However, substantial noise has been discovered in its state annotations. Such noise brings about huge challenges for training DST models robustly. Although several refined versions, including MultiWOZ 2.1-2.4, have been published recently, there are still lots of noisy labels, especially in the training set. Besides, it is costly to rectify all the problematic annotations. In this paper, instead of improving the annotation quality further, we propose a general framework, named ASSIST (AAbel noSe-robuSt ddialogue State Tracking), to train DST models robustly from noisy labels. ASSIST first generates pseudo labels for each sample in the training set by using an auxiliary model trained on a small clean dataset, then puts the generated pseudo labels and vanilla noisy labels together to train the primary model. We show the validity of ASSIST theoretically. Experimental results also demonstrate that ASSIST improves the joint goal accuracy of DST by up to 28.16% on the initial version MultiWOZ 2.0 and 8.41% on the latest version MultiWOZ 2.4, respectively.

1 Introduction

Task-oriented dialogue systems play an important role in helping users accomplish a variety of tasks through verbal interactions (Young et al., 2013; Gao et al., 2019). Dialogue state tracking (DST) is an essential component of the dialogue manager in pipeline-based task-oriented dialogue systems. It aims to keep track of users’ intentions at each turn of the conversation (Mrkšić et al., 2017). The state information indicates the progress of the conversation and is leveraged to determine the next system action and generate the next system response (Chen et al., 2017). As shown in Figure 1, the dialogue state is typically represented as a set of (slot, value) pairs (Williams et al., 2014; Henderson et al., 2014).

Therefore, the problem of DST is defined as extracting the values for all slots from the dialogue context at each turn of the conversation.

Over the past few years, DST has made significant progress, attributed to a number of publicly available dialogue datasets, such as DSTC2 (Henderson et al., 2014), FRAMES (El Asri et al., 2017), MultiWOZ 2.0 (Budzianowski et al., 2018), CrossWOZ (Zhu et al., 2020), and SGD (Rastogi et al., 2020). Among these datasets, MultiWOZ 2.0 is the most popular one. So far, lots of DST models have been built on top of it (Lee et al., 2019; Wu et al., 2019; Ouyang et al., 2020; Kim et al., 2020; Hu et al., 2020; Ye et al., 2021b; Lin et al., 2021).

However, it has been found out that there is substantial noise in the state annotations of MultiWOZ 2.0 (Eric et al., 2020). These noisy labels may impede the training of robust DST models and lead to noticeable performance decrease (Zhang et al., 2016). To remedy this issue, massive efforts have been devoted to rectifying the annotations, and four refined versions, including MultiWOZ 2.1 (Eric et al., 2020), MultiWOZ 2.2 (Zang et al., 2020), MultiWOZ 2.3 (Han et al., 2020b), and MultiWOZ 2.4 (Ye et al., 2021a), have been released. Even so, there are still plenty of noisy and inconsistent labels.
For example, in the latest version MultiWOZ 2.4, both the validation set and test set are noise-free, as they have been manually re-annotated. While the training set is still noisy, as it remains intact. In reality, it is costly and labor-intensive to refine existing large-scale noisy datasets or collect new ones with fully precise annotations (Wei et al., 2020), let alone dialogue datasets with multiple domains and multiple turns. In view of this, we argue that it is a necessity to devise particular learning algorithms to train DST models robustly from noisy labels.

Although loads of noisy label learning algorithms (Natarajan et al., 2013; Han et al., 2020a) have been proposed in the machine learning community, most of them target only multi-class classification (Song et al., 2020). However, as illustrated in Figure 1, the dialogue state may contain multiple labels, which makes it unstraightforward to apply existing noisy label learning algorithms to the DST task. In this paper we propose a general framework, named ASSIST (ASIsst to train robust DST tracking), to train DST models robustly from noisy labels. ASSIST first trains an auxiliary model on a small clean dataset to generate pseudo labels for each sample in the noisy training set. Then, it leverages both the generated pseudo labels and vanilla noisy labels to train the primary model. Since the auxiliary model is trained on the clean dataset, it can be expected that the pseudo labels will help us train the primary model more robustly. Note that ASSIST is based on the assumption that we have access to a small clean dataset. This assumption is reasonable, as it is feasible to manually collect a small noise-free dataset or re-annotate a portion of a large noisy dataset.

In summary, our main contributions include:

- We propose a general framework ASSIST to train robust DST models from noisy labels. To the best of our knowledge, we are the first to tackle the DST problem by taking into consideration the label noise.
- We theoretically analyze why the pseudo labels are beneficial and show that a proper combination of the pseudo labels and vanilla noisy labels can approximate the unknown true labels more accurately.
- We conduct extensive experiments on MultiWOZ 2.0 & 2.4. The results demonstrate that ASSIST can improve the DST performance on both datasets by a large margin.

## 2 Problem Definition

In this section, we first provide the conventional definition of DST and then extend the definition to the noisy label learning scenario.

### 2.1 Conventional Dialogue State Tracking

Let $\mathcal{X} = \{(R_1, V_1), \ldots, (R_T, V_T)\}$ denote a dialogue of $T$ turns, where $R_t$ and $V_t$ represent the system response and user utterance at turn $t$, respectively. The dialogue state at turn $t$ is defined as $B_t = \{(s, v_t) | s \in S\}$, where $S$ denotes the set of predefined slots and $v_t$ is the corresponding value of slot $s$. Following previous work (Lee et al., 2019; Hu et al., 2020; Ye et al., 2021b), a slot in this paper refers to the concatenation of the domain name and slot name so as to include the domain information. For example, we use "hotel-name" to represent the slot "name" in the hotel domain.

In general, the issue of DST is defined as learning a dialogue state tracker $\mathcal{F} : \mathcal{X}_t \rightarrow B_t$ that takes the dialogue context $\mathcal{X}_t$ as input and predicts the dialogue state $B_t$ at each turn $t$ as accurately as possible. Here, $\mathcal{X}_t$ represents the dialogue history up to turn $t$, i.e., $\mathcal{X}_t = \{(R_1, U_1), \ldots, (R_t, U_t)\}$.

### 2.2 Dialogue State Tracking with Noisy Labels

Conventionally, all the state labels are assumed to be correct. However, this assumption may not hold. In practice, dialogue state annotations are error-prone (Han et al., 2020b). There are a couple of reasons. First, the states are usually annotated by crowdworkers to improve the labelling efficiency. Due to limited knowledge, crowdworkers cannot annotate all the states with 100% accuracy, which naturally incurs noisy labels (Han et al., 2020a). Second, the dialogue may span multiple domains, which also increases the labelling difficulty. Apparently, the noisy labels are harmful and likely to lead to sub-optimal performance. Therefore, it is crucial to take them into consideration so as to train DST models more robustly.

Let $\tilde{B}_t = \{(s, \tilde{v}_t) | s \in S\}$ denote the noisy state annotations, where $\tilde{v}_t$ is the noisy label of slot $s$ at turn $t$. We use $\hat{B}_t = \{(s, v_t) | s \in S\}$ to denote the noise-free state annotations. Here, $v_t$ represents the true label of slot $s$ at turn $t$, which is unknown. In fact, existing DST approaches are only able to learn a sub-optimal dialogue state tracker $\hat{\mathcal{F}} : \mathcal{X}_t \rightarrow \hat{B}_t$ rather than the optimal state tracker $\mathcal{F} : \mathcal{X}_t \rightarrow B_t$, as none of them have considered the influence of
3 Proposed Approach

We introduce a general framework ASSIST, aiming to train DST models robustly from noisy labels. We assume that a small clean dataset is accessible. Based on this dataset, ASSIST first trains an auxiliary model $\mathcal{A}$. Then, it leverages $\mathcal{A}$ to generate pseudo labels for each sample in the noisy training set. The pseudo state annotations are represented as $\tilde{\mathcal{B}}_t = \{(s, \tilde{v}_t)|s \in \mathcal{S}\}$, where $\tilde{v}_t$ denotes the pseudo label of slot $s$ at turn $t$. Afterwards, both the generated pseudo labels and vanilla noisy labels are exploited to train the primary model $\mathcal{F}^\ast$. That is, we intend to learn $\mathcal{F}^\ast: \mathcal{X}_t \rightarrow C(\tilde{\mathcal{B}}_t, \mathcal{B}_t)$, where $C(\tilde{\mathcal{B}}_t, \mathcal{B}_t)$ is a combination of $\tilde{\mathcal{B}}_t$ and $\mathcal{B}_t$.

Essentially, any existing DST models can be employed as the auxiliary model. However, these models may lead to overfitting due to the small size of the clean dataset. To tackle this issue, we propose a new simple model as the auxiliary model$^1$.

3.1 Auxiliary Model Architecture

Figure 2 shows the architecture, which consists of a dialogue context semantic encoder, a slot attention module, and a slot-value matching module.

Dialogue Context Semantic Encoder

Similar to (Lee et al., 2019; Kim et al., 2020; Ye et al., 2021b), we utilize the pre-trained language model BERT (Devlin et al., 2019) to encode the dialogue context $\mathcal{X}_t$ into contextual semantic representations. Let $Z_t = R_t \oplus U_t$ be the concatenation of the system response and user utterance at turn $t$, where $\oplus$ denotes the operator of sequence concatenation. Then, the dialogue context $\mathcal{X}_t$ can be represented as $\mathcal{X}_t = Z_1 \oplus Z_2 \oplus \cdots \oplus Z_t$.

We also concatenate each slot-value pair and denote the representation of the dialogue state at turn $t$ as $B_t = \bigoplus_{(s, v) \in \mathcal{B}_t, v \neq \text{None}} s \oplus v$, in which only non-none slots are included. $B_t$ can serve as a compact representation of the dialogue history. In view of this, we treat the previous turn dialogue state $B_{t-1}$ as part of the input as well, which can be beneficial when $\mathcal{X}_t$ exceeds the maximum input length of BERT. The complete input sequence to the encoder module is then denoted as:

$$I_t = [CLS] \oplus \mathcal{X}_{t-1} \oplus B_{t-1} \oplus [SEP] \oplus Z_t \oplus [SEP],$$

$^1$We adopt existing DST models as the primary model.

Figure 2: Overall architecture of the auxiliary model. The parameters of the BERT used to encode slots and values are fixed during the training phase.

where [CLS] and [SEP] are the two special tokens introduced by BERT.

Let $H_t \in \mathbb{R}^{|I_t| \times d}$ be the semantic matrix representation of $I_t$. Here, $|I_t|$ and $d$ denote the sequence length of $I_t$ and the BERT output dimension, respectively. Then, we have:

$$H_t = \text{BERT}_{\text{finetune}}(I_t),$$

where $\text{BERT}_{\text{finetune}}$ means that the BERT model will be fine-tuned during the training phase.

For each slot $s$ and its candidate value $v' \in \mathcal{V}_s$, we employ another BERT to encode them into semantic vectors $h^s \in \mathbb{R}^d$ and $h^{v'} \in \mathbb{R}^d$. Here, $\mathcal{V}_s$ denotes the candidate value set of slot $s$. Unlike the dialogue context, we leverage the pre-trained BERT without fine-tuning to embed $s$ and $v'$. Besides, we adopt the output vector corresponding to the special token [CLS] as an aggregated representation of slot $s$ and value $v'$, i.e.,

$$h^s = \text{BERT}_{\text{fixed}}^{[CLS]}([CLS] \oplus s \oplus [SEP]),$$

$$h^{v'} = \text{BERT}_{\text{fixed}}^{[CLS]}([CLS] \oplus v' \oplus [SEP]).$$

Slot Attention

The slot attention module is exploited to retrieve slot-relevant information for all the slots from the same dialogue context. The slot attention is a multi-head attention (Vaswani et al., 2017). Specifically, the slot representation $h^s$ is regarded as the query vector, and the dialogue context representation $H_I$ is taken as both the key matrix and value matrix. The slot attention matches $h^s$ to the semantic vector of each word in the dialogue context and calculates the attention score, based on which the slot-specific information can be extracted. Let $a^s \in \mathbb{R}^d$ denote
a \(d\)-dimensional vector representation of the related information of slot \(s\) at turn \(t\), we obtain:
\[
a_t^s = \text{MultiHead}(h^s, H_t, H_t).
\]

\(a_t^s\) is expected to be close to the semantic vector representation of the true value of slot \(s\).

Considering that the output of BERT is normalized by layer normalization (Ba et al., 2016), we also feed \(a_t^s\) to a layer normalization layer, which is preceded by a linear transformation layer. The final slot-specific vector \(g_t^s \in \mathbb{R}^d\) is calculated as:
\[
g_t^s = \text{LayerNorm}(\text{Linear}(a_t^s)).
\]

### 3.2 Auxiliary Model Training

The slot-value matching module is utilized to predict the value of each slot \(s\). It first calculates the distance between the slot-specific representation \(g_t^s\) and the semantic representation of each candidate value \(v' \in \mathcal{V}_s\), i.e., \(h^{v'}\). Then, the candidate value with the smallest distance is selected as the prediction. The \(\ell_2\) norm is adopted to compute the distance. Denoting \(\hat{v}_t\) as the predicted value of slot \(s\) at turn \(t\), we have:
\[
\hat{v}_t = \text{argmin}_{v' \in \mathcal{V}_s} \|g_t^s - h^{v'}\|_2.
\]

### 3.3 Pseudo Label Generation

Our approach depends on the auxiliary model \(A\) to generate pseudo labels \(\hat{B}_t = \{(s, \hat{v}_t) | s \in S\}\) for each sample in the noisy training set. In this work, we treat each dialogue context \(X_t\) rather than the entire dialogue as a training sample. Without loss of generality, the pseudo label generation process is denoted as follows:
\[
\hat{B}_t = A(X_t, S),
\]
where \(X_t\) belongs to the noisy training set.

### 3.4 Primary Model Training

To reduce the influence of noisy labels, we combine the generated pseudo labels and vanilla noisy labels to train the primary model.

Let \(\tilde{v}_t\) and \(\bar{v}_t\) be the one-hot representation of the pseudo label \(\tilde{v}_t\) and vanilla noisy label \(\bar{v}_t\), respectively. Then, we can define the combined label as:
\[
v_t^c = \alpha \tilde{v}_t + (1 - \alpha) \bar{v}_t,
\]
where \(\alpha(0 \leq \alpha \leq 1)\) is a parameter to balance the pseudo labels and vanilla labels. We calculate the probability of \(v_t^c\) as below:
\[
p(v_t^c | X_t, s) = p(\tilde{v}_t | X_t, s)^{\alpha} p(\bar{v}_t | X_t, s)^{(1-\alpha)}.
\]

Here, \(p(\tilde{v}_t | X_t, s)\) and \(p(\bar{v}_t | X_t, s)\) correspond to the probability of \(\tilde{v}_t\) and \(\bar{v}_t\), respectively.

Let \(C(\tilde{B}_t, \bar{B}_t) = \{(s, v_t^c) | s \in S\}\) represent the combined state annotations. The training objective of the primary model is then defined as:
\[
L_{\text{pri}} = \sum_{(s, v_t^c) \in C(\tilde{B}_t, \bar{B}_t)} - \log p(v_t^c | X_t, s)
\]
\[
= \alpha \sum_{(s, \tilde{v}_t) \in \tilde{B}_t} - \log p(\tilde{v}_t | X_t, s)
\]
\[
+ (1 - \alpha) \sum_{(s, \bar{v}_t) \in \bar{B}_t} - \log p(\bar{v}_t | X_t, s)
\]
\[
= \alpha L_{\text{pseudo}} + (1 - \alpha) L_{\text{vanilla}},
\]
where \(L_{\text{pseudo}}\) and \(L_{\text{vanilla}}\) correspond to the training objective of using only the pseudo labels and using only the vanilla noisy labels, respectively. By minimizing \(L_{\text{pri}}\), the primary model is trained to learn from the vanilla noisy labels and at the same time imitate the predictions of the auxiliary model.

### 3.5 Theoretical Analysis

Since the pseudo labels are generated by the auxiliary model that has been trained on a small clean dataset, it can be expected that the combined labels are able to serve as a better approximation to the unknown true labels. Let \(v_t\) denote the one-hot representation of the unknown true value \(v_t\) of slot \(s\) at turn \(t\). We adopt the mean squared loss to define the approximation error of any corrupted labels \(\tilde{v}_t\) associated with the noisy training set \(D_n\) as:
\[
Y_{\nu} = \frac{1}{|D_n||S|} \sum_{X_t \in D_n} \sum_{s \in S} E_{D_n}[\|\tilde{v}_t - v_t\|_2^2],
\]
where the expectation ranges over different choices of the clean dataset \(D_c\), and \(| \cdot |\) returns the cardinality of a set.
Next, we show that the approximation error of the combined labels can be smaller than that of either the vanilla noisy labels or the generated pseudo labels. The details are presented in Theorem 1.

**Theorem 1.** The optimal approximation error with respect to the combined labels $\tilde{v}_t$ is smaller than that of the vanilla labels $v_t$ and pseudo labels $\tilde{v}_t$, i.e.,
\[
\min_{\alpha} Y_{\tilde{v}_t} < \min\{Y_v, Y_{\tilde{v}_t}\}.
\]
By setting $\alpha = \frac{Y_v}{Y_v + Y_{\tilde{v}_t}}$, $Y_{\tilde{v}_t}$ reaches its minimum:
\[
\min_{\alpha} Y_{\tilde{v}_t} = \frac{Y_v Y_{\tilde{v}_t}}{Y_v + Y_{\tilde{v}_t}}.
\]

**Proof.** The proof is presented in Appendix A. □

Theorem 1 indicates that if $\alpha$ is set properly, the combined labels can approximate the unknown true labels more accurately. Hence, we can potentially train the primary model more robustly. Note that we cannot calculate the optimal value of $\alpha$ directly.

## 4 Experimental Setup

### 4.1 Datasets

We adopt MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.4 (Ye et al., 2021a) as the datasets in our experiments. MultiWOZ 2.0 is one of the largest publicly available multi-domain task-oriented dialogue datasets, including about 10,000 dialogues spanning seven domains. MultiWOZ 2.4 is the latest refined version of MultiWOZ 2.0, in which the annotations of the validation set and test set have been manually rectified and thus are noise-free. However, the training set of MultiWOZ 2.4 is still noisy, as it remains intact and is the same as that of MultiWOZ 2.1 (Eric et al., 2020).

Since the hospital domain and police domain never occur in the test set, we use only the remaining five domains {attraction, hotel, restaurant, taxi, train} in our experiments. These domains have 30 slots in total. Considering that the validation set and test set of MultiWOZ 2.0 are noisy, we replace them with the counterparts of MultiWOZ 2.4. We preprocess the datasets following (Ye et al., 2021b). We use the validation set as the small clean dataset.

### 4.2 Evaluation Metrics

We exploit joint goal accuracy and slot accuracy as the evaluation metrics. The joint goal accuracy is defined as the proportion of dialogue turns in which the values of all slots are correctly predicted. It is the most important metric in the DST task. The slot accuracy is defined as the average of all individual slot accuracies. The accuracy of an individual slot is calculated as the ratio of dialogue turns in which its value is correctly predicted.

We also propose a new evaluation metric, termed as joint turn accuracy. We define joint turn accuracy as the proportion of dialogue turns in which the values of all active slots are correctly predicted. A slot becomes active if its value is mentioned in current turn and is not inherited from previous turns. The advantage of joint turn accuracy is that it can tell us in how many turns the turn-level information is fully captured by the model.

## 4.3 Primary DST Models

To verify the effectiveness of the proposed framework, we apply the generated pseudo labels to three different primary models.

**SOM-DST:** SOM-DST (Kim et al., 2020) is an open vocabulary-based method. It treats the dialogue state as an explicit fixed-sized memory and selectively overwrites this memory at each turn.

**STAR:** STAR (Ye et al., 2021b) is a predefined ontology-based method. It leverages a stacked slot self-attention to model the slot dependencies.

**AUX-DST:** We also test using the proposed auxiliary model as the primary model. For the sake of description, we refer to this model as AUX-DST.

### 4.4 Implementation Details

For the auxiliary model, the pre-trained BERT-base-uncased model$^2$ is utilized as the dialogue context encoder. Another pre-trained BERT-base-uncased model with fixed weights is employed to encode the slots and their candidate values. The maximum input length of the BERT model is set to 512. The number of heads in the slot attention module is set to 4. The output dimension of the linear transformation layer is set to 768, which is the same as the dimension of the BERT outputs. Recall that the previous turn dialogue state is treated as part of the input. The ground-truth one is used during training, and the predicted one is used during testing.

We train the auxiliary model on the clean validation set and the primary model on the noisy training set. When training the auxiliary model, the noisy...
Table 1 presents the performance scores of the three different primary DST models on the test sets of MultiWOZ 2.0 & 2.4 when they are trained using our proposed framework ASSIST. For comparison, we also include the results when only the vanilla labels or only the pseudo labels are used to train the primary models.

As can be seen, ASSIST consistently improves the performance of the three primary models on both datasets. More concretely, compared to the results obtained using only the vanilla labels, ASSIST improves the joint goal accuracy of SOM-DST, STAR, and AUX-DST on MultiWOZ 2.0 by 25.69%, 25.82%, and 28.16% absolute gains, respectively. On MultiWOZ 2.4, ASSIST also leads to 8.41%, 5.79%, and 7.77% absolute joint goal accuracy gains. From Table 1, we further observe that the performance improvements on MultiWOZ 2.4 are lower than on MultiWOZ 2.0. This is because the training set of MultiWOZ 2.4 is the same as that of MultiWOZ 2.1 (Eric et al., 2020), in which lots of annotation errors have been fixed. We also observe that all the primary models demonstrate relatively good performance when only the pseudo labels are used. From these results, it can be concluded that the pseudo labels are beneficial and they can help us train DST models more robustly.

Another observation from Table 1 is that SOM-DST tends to show comparable or even higher joint turn accuracy compared to STAR and AUX-DST, although its performance is worse in terms of joint goal accuracy and slot accuracy. This is because SOM-DST focuses on turn-active slots and copies the values for other slots from previous turns, while both STAR and AUX-DST predict the values of all slots from scratch at each turn. These results show that the joint turn accuracy can help us understand in more depth how different models behave.

5.2 Applying STAR as the Auxiliary Model

Although any existing DST models can be adopted as the auxiliary model, we chose to propose a new simple one to reduce overfitting. In order to verify the superiority of the proposed model, we also apply STAR as the auxiliary model and compare their performance in Figure 3. We chose STAR due to its good performance, as shown in Table 1. From Figure 3, we observe that all three primary models demonstrate higher performance on both datasets when using the proposed auxiliary model than using STAR as the auxiliary model. The results indicate that the proposed auxiliary model is able to generate pseudo labels with higher quality.
5.3 Effects of Parameter $\alpha$

The parameter $\alpha$ adjusts the weights of the pseudo labels and vanilla labels. Here, we study the effects of $\alpha$ by varying its value in the range of 0 to 1 with a step size of 0.1. Figure 4 shows the results of AUX-DST. As can be seen, $\alpha$ plays an important role in balancing the pseudo labels and vanilla labels. The best performance is achieved when $\alpha$ is set to 0.6 on MultiWOZ 2.0 and 0.4 on MultiWOZ 2.4. Since the training set of MultiWOZ 2.0 has more noisy labels than that of MultiWOZ 2.4, more emphasis should be put on its pseudo labels to obtain the best performance. It is also noted that the performance difference between MultiWOZ 2.0 and MultiWOZ 2.4 dwindles away as $\alpha$ increases. This is because the vanilla labels will contribute less to the training of the primary model when $\alpha$ is set to be larger.

5.4 Effects of the Size of the Clean Dataset

Considering that our proposed framework ASSIST relies on a small clean dataset to train the auxiliary model that is further leveraged to generate pseudo labels for the training set, it is valuable to explore the effects of the size of the clean dataset on the performance of the primary model. For this purpose, we vary the number of dialogues in the clean dataset from 500 to 1000\(^3\) to generate different pseudo labels. We then combine these different pseudo labels with the vanilla labels to train the primary model AUX-DST. The results on MultiWOZ 2.4 are reported in Figure 5. For comparison, we also include the results when only the pseudo labels or only the vanilla labels are used to train the primary model. As can be seen, the size of the clean dataset has a great impact on the performance of the primary model. Apparently, fewer clean data will lead to worse performance. Nevertheless, as long as the pseudo labels are combined with the vanilla labels, the primary model can consistently demonstrate the strongest performance.

5.5 Analyses on Pseudo Labels’ Quality

The previous experiments have proven the effectiveness of the generated pseudo labels in training robust DST models. In this part, we provide further analyses on the quality of the pseudo labels to gain more insights into why they can be beneficial.

5.5.1 Quantitative Analysis

We first investigate whether the pseudo labels are consistent with the true labels. To achieve this goal, we can compute the joint goal accuracy and joint turn accuracy of the auxiliary model. Figure 6 presents the results. As shown in Figure 6, the auxiliary model achieves higher performance when more clean dialogues are utilized to train it. If the entire validation set is used, it achieves around 50% joint goal accuracy and around 75% joint turn accuracy. Given that the vanilla noisy labels are regarded as the true labels, we can conjecture that the true performance is actually higher. This experiment shows that the pseudo labels are consistent with the unknown true labels to some extent and are a good complement to the vanilla noisy labels.

5.5.2 Qualitative Analysis

To intuitively understand the quality of the pseudo labels, we show four dialogue snippets with their vanilla labels and the generated pseudo labels in Table 2. As can be seen, the vanilla labels of the first two dialogue snippets are incomplete, while all the missing information is presented in the pseudo labels.

\(^3\)There are 1000 dialogues in total in the validation set.
<table>
<thead>
<tr>
<th>Dialogue Context</th>
<th>Vanilla Labels</th>
<th>Pseudo Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>[sys]: Sure, da vinci pizzeria is a cheap Italian restaurant in the area. [usr]: Would you mind making a reservation for Thursday at 17:15?</td>
<td>(restaurant-name, da vinci pizzeria)</td>
<td>(restaurant-book day, thursday) (restaurant-book time, 17:15) (restaurant-name, da vinci pizzeria)</td>
</tr>
<tr>
<td>[sys]: Do you have a preferred section of town? [usr]: Not really, but I want free wifi and it should be 4 star.</td>
<td>(hotel-internet, free) (hotel-stars, 4)</td>
<td>(hotel-area, dontcare) (hotel-internet, free) (hotel-stars, 4)</td>
</tr>
<tr>
<td>[usr]: I need to find out if there is a train going to stansted airport that leaves after 12:30.</td>
<td>(train-arriveby,13:03) (train-destination, stansted airport) (train-leaveat, 12:30)</td>
<td>(train-destination, stansted airport) (train-leaveat, 12:30)</td>
</tr>
<tr>
<td>[usr]: I am staying in the west part of Cambridge and would like to know about some places to go.</td>
<td>(attraction-area, west)</td>
<td>(attraction-area, west) (hotel-area, west)</td>
</tr>
</tbody>
</table>

Table 2: Four dialogue snippets with their vanilla labels and the generated pseudo labels. These dialogue snippets are chosen from the training set of MultiWOZ 2.4. To save space, we only present turn-active slots and their values.

labels. For the third dialogue snippet, the vanilla labels contain an unmentioned slot-value pair "(train-arriveby, 13:03)". This error has also been fixed in the pseudo labels. For the last dialogue snippet, the vanilla labels are correct. However, the pseudo labels introduce an overconfident prediction of the value of slot "hotel-area". This case study has verified again that the pseudo labels can be utilized to fix certain errors in the vanilla labels. However, the pseudo labels may bring about some new errors. Hence, we should combine the two types of labels.

6 Related Work

In this section, we briefly review related work on DST and noisy label learning.

6.1 Dialogue State Tracking

Recently, DST has got an enormous amount of attention, thanks to the availability of multiple large-scale multi-domain dialogue datasets such as MultiWOZ 2.0 (Budzianowski et al., 2018), MultiWOZ 2.1 (Eric et al., 2020), RiSAWOZ (Quan et al., 2020), and SGD (Rastogi et al., 2020). The most popular datasets are MultiWOZ 2.0 and MultiWOZ 2.1, and lots of DST models have been built on top of them (Lee et al., 2019; Wu et al., 2019; Ouyang et al., 2020; Hosseini-Asl et al., 2020; Kim et al., 2020; Hu et al., 2020; Feng et al., 2020; Ye et al., 2021b; Lin et al., 2021; Liang et al., 2021).

These recent DST models can be grouped into two categories: predefined ontology-based models and open vocabulary-based models. The predefined ontology-based models treat DST as a multi-label classification problem and tend to demonstrate better performance (Chen et al., 2020; Zhang et al., 2020; Shan et al., 2020; Ye et al., 2021b). The open vocabulary-based models leverage either span prediction (Heck et al., 2020; Gao et al., 2020) or sequence generation (Wu et al., 2019; Feng et al., 2020; Hosseini-Asl et al., 2020) to extract slot values from the dialogue context directly.

Although these DST models have made a huge success, they can only achieve sub-optimal performance, due to the lack of handling noisy labels. To the best of our knowledge, we are the first to take into account the noisy labels when tackling DST.

6.2 Noisy Label Learning

Addressing noisy labels in supervised learning is a long-term studied problem (Frénay and Verleysen, 2013; Song et al., 2020; Han et al., 2020a). This issue becomes more prominent in the era of deep learning, as training deep models generally requires a lot of well-labelled data, but it is expensive and time-consuming to collect large-scale datasets with completely clean annotations. This dilemma has sparked a surge of noisy label learning methods (Hendrycks et al., 2018; Zhang and Sabuncu, 2018; Song et al., 2019; Wei et al., 2020a). Even so, these methods mainly focus on multi-class classification (Song et al., 2020), which makes it not straightforward to apply them to the DST task.

7 Conclusion

In this work, we have presented a general framework ASSIST, aiming to train DST models robustly from noisy labels. ASSIST leverages an auxiliary model that is trained on a small clean dataset to generate pseudo labels for the large noisy training set. The pseudo labels are combined with the vanilla labels to train the primary model. Both theoretical analysis and empirical study have verified the validity of our proposed framework. In the future, we intend to explore more advanced techniques to combine the pseudo labels and vanilla noisy labels.
References


A Proof of Theorem 1

Proof. Our proof is based on the bias-variance decomposition theorem\(^4\). For any sample \(X_t\) in the noisy training set \(D_n\), the approximation error with respect to the pseudo label \(\hat{y}_t\) of slot \(s\) is defined as \(E_{D_n}[\|\hat{y}_t - v_t\|_2^2]\), which, according to the bias-variance decomposition theorem, can be decomposed into a bias term and a variance term, i.e.,

\[
E_{D_n}[\|\hat{y}_t - v_t\|_2^2] = (\text{Bias}_{D_n}[\hat{y}_t])^2 + \text{Var}_{D_n}[\hat{y}_t],
\]

where

\[
\text{Bias}_{D_n}[\hat{y}_t] = \|E_{D_n}[\hat{y}_t] - v_t\|_2,
\]

\[
\text{Var}_{D_n}[\hat{y}_t] = E_{D_n}[\|E_{D_n}[\hat{y}_t] - \hat{y}_t\|_2^2].
\]

In our approach, the auxiliary model is a BERT-based model, which has more than 110M parameters. Such a complex model is expected to be able to capture all the samples in the small clean dataset \(D_c\). Therefore, we can reasonably assume that the bias term is close to zero. Then, we have:

\[
\text{Bias}_{D_n}[\hat{y}_t] \approx 0 \Rightarrow E_{D_n}[\hat{y}_t] \approx v_t.
\]

Considering that the pseudo labels are generated by the auxiliary model that is trained on an extra small clean dataset and this clean dataset is independent of the noisy training set, we can regard the pseudo labels and vanilla labels as independent of each other. Consequently, we obtain:

\[
E_{D_n}[(\hat{y}_t - v_t)^T(\hat{y}_t - v_t)] = [E_{D_n}[\hat{y}_t - v_t]]^T E_{D_n}[\hat{y}_t - v_t] = [E_{D_n}[\hat{y}_t - v_t]]^T 0 = 0.
\]

Based on the formula above, we can now calculate the approximation error with respect to the combined label \(v_s\) of slot \(s\) as below:

\[
E_{D_n}[\|v_s - v_t\|_2^2] = E_{D_n}[\|\alpha \hat{y}_t + (1 - \alpha) \hat{y}_t - v_t\|_2^2] = E_{D_n}[\|\alpha \hat{y}_t - v_t\|_2^2] + (1 - \alpha)^2 E_{D_n}[\|\hat{y}_t - v_t\|_2^2],
\]

where the last equality holds because of \(E_{D_n}[(\hat{y}_t - v_t)^T(\hat{y}_t - v_t)] = 0\). Then, we have:

\[
Y_{v_s} = \frac{1}{|D_n||S|} \sum_{X_t \in D_n} \sum_{s \in S} E_{D_n}[\|v_s - v_t\|_2^2] = \frac{\alpha^2}{|D_n||S|} \sum_{X_t \in D_n} \sum_{s \in S} E_{D_n}[\|\hat{y}_t - v_t\|_2^2] + \frac{(1 - \alpha)^2}{|D_n||S|} \sum_{X_t \in D_n} \sum_{s \in S} E_{D_n}[\|\hat{y}_t - v_t\|_2^2] = \alpha^2 Y_\hat{y} + (1 - \alpha)^2 Y_{v_s}.
\]

\(Y_{v_s}\) reaches its minimum when \(\alpha = \frac{Y_\hat{y}}{Y_\hat{y} + Y_{v_s}}\), and

\[
\min_{\alpha} Y_{v_s} = \frac{Y_\hat{y} Y_{v_s}}{Y_\hat{y} + Y_{v_s}},
\]

which concludes the proof. \(\square\)

B Training Details

Note that the proposed auxiliary model is also applied as one primary model in our experiments. In both cases, AdamW (Kingma and Ba, 2014) is adopted as the optimizer, and a linear schedule...
with warmup is created to adjust the learning rate dynamically. The peak learning rate is set to 2.5e-5. The warmup proportion is fixed at 0.1. The dropout (Srivastava et al., 2014) probability and word dropout (Bowman et al., 2016) probability are also fixed at 0.1. When taken as the auxiliary model, the model is trained for at most 30 epochs with a batch size of 8. When taken as the primary model, the batch size and training epochs are set to 8 and 12, respectively. The best model is chosen according to the performance on the validation set.

For SOM-DST and STAR, the default hyperparameters are adopted when they are applied as the primary model.

C Additional Experimental Results

C.1 Error Analysis

We further investigate the error rate with respect to each slot. We adopt AUX-DST as the primary model and use AUX-DST(w/o p) to denote the case when only the vanilla labels are employed to train the model. The results on the test set of MultiWOZ 2.4 are illustrated in Figure 7, from which we can observe that the slot "hotel-type" has the highest error rate. Even though the error rate is reduced with the aid of the pseudo labels, it is still the highest one among all the slots. This is because the labels of this slot are confusing. It is also observed that the "name"-related slots have relatively high error rates. However, when the pseudo labels are used, their error rates reduce remarkably. Besides, we observe that the error rates of some slots are higher when the pseudo labels are leveraged. This is probably due to the fact that we have used the same parameter $\alpha$ to combine the pseudo labels and vanilla labels of all slots. In practice, the noise rate with respect to each slot in the vanilla labels may not be exactly the same. This observation inspires us that more advanced techniques should be developed to combine the pseudo labels and vanilla labels, which we leave as our future work.

C.2 Effects of the Distribution of the Clean Dataset

Except for the size of the clean dataset, the distribution of the clean dataset may also affect the performance of the primary model, especially when the clean dataset has a significantly different distribution from the training set. Thus, it is important to study the effects of the distribution of the clean dataset. However, we are short of clean datasets with different distributions. It is also challenging to model the distribution explicitly since the dialogue state may contain multiple labels. To address this issue, we propose to remove all the dialogues that are related to a specific domain and use only the remaining ones as the clean dataset. As thus, we can create multiple clean datasets with different distributions. The results of AUX-DST on MultiWOZ 2.4 are shown in Figure 8. As can be observed, although different clean datasets indeed lead to different performance, compared to the situation where no clean data is used (i.e., only the vanilla labels are used to train the model), all these clean datasets still bring huge performance improvements.