Towards Multi-label Unknown Intent Detection

Yawen Ouyang Zhen Wu* Xinyu Dai Shujian Huang Jiajun Chen
National Key Laboratory for Novel Software Technology, Nanjing University, China
ouyangyw@smail.nju.edu.cn
{wuz,daixinyu,huangsj,chenjj}@nju.edu.cn

Abstract

Multi-class unknown intent detection has made remarkable progress recently. However, it has a strong assumption that each utterance has only one intent, which does not conform to reality because utterances often have multiple intents. In this paper, we propose a more desirable task, *multi-label unknown intent detection*, to detect whether the utterance contains the unknown intent, in which each utterance may contain multiple intents. In this task, the unique utterances simultaneously containing known and unknown intents make existing multi-class methods easy to fail. To address this issue, we propose an intuitive and effective method to recognize whether **A**ll **I**ntents contained in the utterance are **K**nown (**AIK**). Our high-level idea is to predict the utterance's intent number, then check whether the utterance contains the same number of known intents. If the number of known intents is less than the number of intents, it implies that the utterance also contains unknown intents. We benchmark AIK over existing methods, and empirical results suggest that our method obtains state-of-the-art performances. For example, on the MultiWOZ 2.3 dataset, AIK significantly reduces the FPR95 by 12.25% compared to the best baseline.[1]

1 Introduction

Intent classification is a crucial component of task-oriented dialogue systems, which aims to map the utterance to the known intent set. In an open environment, it is nearly impossible that dialogue systems are only exposed to utterances with known intents, i.e., in-distribution (IND) utterances. Therefore, unknown intent detection is proposed to identify the out-of-distribution (OOD) utterance, which contains the unknown intent (Hendrycks and Gimpel, 2017). It can prevent dialogue systems from generating unrelated responses to ensure good user experiences. The detected OOD utterances can also provide future direction for developers.

Utterance	I am looking to stay at the Lovell Lodge hotel and to see the areas local attractions.
Intent	*Inform-Hotel-Name*, *Request-Attraction-Area*

Table 1: An example of utterance with multiple intents from MultiWOZ 2.3 (Han et al., 2020). For dialogue systems designed for the hotel domain, the utterance is mixed OOD as it contains known intent *Inform-Hotel-Name* and unknown intent *Request-Attraction-Area*.

Recent works follow the assumption that each utterance has only one intent and focus on *multi-class unknown intent detection* (Podolskiy et al., 2021; Ouyang et al., 2021; Lin and Xu, 2019; Shu et al., 2017). Based on this assumption, a rather popular strategy to perform OOD detection relies on the maximum classifier output. For example, Shu et al. (2017) propose using the maximum binary classifier output. An utterance will be regarded as containing the known intent and classified as IND if the output is larger than the predefined threshold, otherwise it will be classified as OOD.

Nevertheless, the above assumption is too strong: several intents are usually expressed in an utterance in a real-world scenario. For example, Gangadharaiah and Narayanaswamy (2019) show that 52% of utterances include multi-label intents in the amazon internal dataset. It is obvious that the multi-class unknown intent detection research cannot fully meet the needs of dialogue systems.

In this work, we propose a more practical task, *multi-label unknown intent detection*, which is to detect whether the user utterance contains unknown intents, where each utterance may contain multiple intents. We summarize three types of utterances for unknown intent detection in the multi-label setting: 1) **IND utterances**, only containing known intents; 2) **pure OOD utterances**, only containing unknown intents; and 3) **mixed OOD utterances** simultaneously containing known and unknown in-

* Corresponding author.

[1]Code and data are available at https://github.com/yawenouyang/AIK.

Proceedings of the 29th International Conference on Computational Linguistics, pages 626–635
October 12–17, 2022.

tents (see Table 1 for an example). Note that mixed OOD utterances are unique to multi-label because utterances in multi-class can only have one intent.

The existence of mixed OOD utterances brings a great challenge for multi-label unknown intent detection, which makes the existing strategy easy to fail. As shown in Figure 1, such methods will regard the mixed OOD utterances as IND utterances once detecting the known intents.

To address the above issue, we propose a novel method, by recognizing whether **A**ll **I**ntents of the utterance are **K**nown (**AIK**), to detect both pure and mixed OOD utterances for multi-label unknown intent detection. Overall, we first predict the number of intents contained in the utterance. Then we check whether the utterance contains the same number of known intents by measuring the probability density of the utterance's known intent-wise representations. Specifically, we assume the known intent-wise representation can be fitted well by a conditional Gaussian distribution, then we can estimate its probability density via the Mahalanobis distance. We empirically demonstrate that AIK can significantly improve OOD detection performance, especially for mixed OOD utterances.

To summarize, the key contributions of the paper are as follows:

- We propose a new task: *multi-label unknown intent detection*, which is desirable for practical dialogue systems.
- We propose a novel and effective method AIK to detect OOD utterances in multi-label setting. By discerning whether all intents contained in the utterance are known, AIK can naturally distinguish IND from pure and mixed OOD utterances.
- We show that AIK outperforms existing methods on two multi-label benchmarks, validating the effectiveness of our method.

2 Task Formulation

Multi-label unknown intent detection breaks the assumption that each utterance only contains one intent, allowing each utterance contain multiple intents. It aims to detect OOD utterances that contain unknown intents.

Formally, given a training dataset $\mathcal{D} = \{(\mathbf{u}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$ where $\mathbf{u}^{(i)}$ is an utterance, $\mathbf{y}^{(i)}$ is a set of intent expressed in $\mathbf{u}^{(i)}$, and it belongs to the known intent set $Y_{\text{in}} = \{c_1, c_2, ..., c_k\}$, i.e., $\mathbf{y}^{(i)} \subseteq Y_{\text{in}}$. When testing, given an utterance, we consider it to be OOD if not *all* intents in its intent set $\mathbf{y}$ belong to Y_{in}. Furthermore, if an utterance is OOD and $\mathbf{y} \cap Y_{\text{in}} \neq \varnothing$, i.e., it also contains known intent(s), we call it mixed OOD utterance. If an utterance is OOD and $\mathbf{y} \cap Y_{\text{in}} = \varnothing$, we call it pure OOD utterance. The task goal is to train a score function $S(\mathbf{u})$ based on the training dataset $\mathcal{D}$ to detect OOD utterances (pure and mixed).

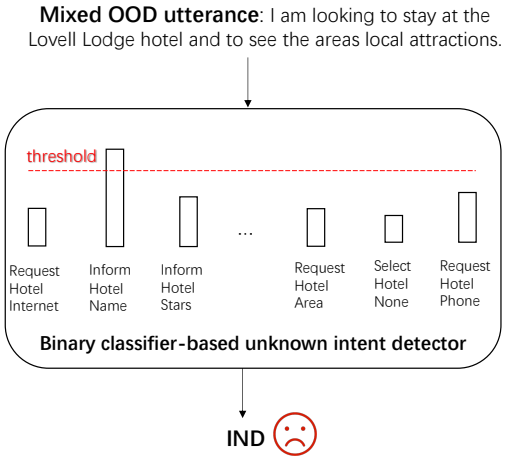

Figure 1: If dialogue systems are equipped with the existing unknown intent detector, such as binary classifier-based detector (Shu et al., 2017), they will misclassify the above utterance as IND as the output of the binary classifier for Inform-Hotel-Name is greater than the threshold.

3 Approach

To perform multi-label unknown intent detection, we propose a novel method AIK. In this section, we first introduce the overall idea of AIK, then present its model architecture and training objective.

3.1 Overall Description

As aforementioned, AIK aims to recognize whether all intents contained in the test utterance are known. Formally, given a test utterance $\mathbf{u}$, we first predict its intent number r. Then we extract its known intent-wise representation $\mathbf{v}_c$ for each known intent $c \in Y_{\text{in}}$, and estimate $\mathbf{v}_c$'s probability density. Suppose that the representations of known intents follow the conditional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_c$ is the center vector and $\boldsymbol{\Sigma}$ is the corvariance matrix [2]. The $\mathbf{v}_c$'s probability

[2]For calculation convenience, we assume all known intents share the same covariance matrix, which is also assumed in Yan et al. (2020) and Lee et al. (2018).

Algorithm 1 OOD detection using AIK

Input: Test utterance $\mathbf{u}$; threshold τ; Known intent set Y_{in} and each known intent c's distribution $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$.

```
Predict u's intent number r
D = {}
for c ∈ Y_in do
    Extract u's known intent-wise representation v_c
    Calculate the Mahalanobis distance d_c between v_c and N(μ_c, Σ)
    Add −d_c into D
end for
S(u) = r-th maximum D
if S(u) < τ then
    return OOD
else
    return IND
end if
```

density can be denoted as $\mathcal{N}(\mathbf{v}_c; \boldsymbol{\mu}_c, \boldsymbol{\Sigma})$, and estimated by its Mahalanobis distance d_c[3] with respect to $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ (Murphy, 2022). After calculating the Mahalanobis distance for each known intent-wise representation, we take the negative of them and aggregate them into $\mathbf{D} = \{-d_{c_1}, -d_{c_2}, ..., -d_{c_k}\}$. Finally, we take the r-th maximum $\mathbf{D}$ as $S(\mathbf{u})$ to measure whether the utterance contains r known intents. An utterance with low $S(\mathbf{u})$ (e.g., lower than the pre-defined threshold) indicates its contained known intent number is less than r. Namely, it also contains unknown intent(s), thus can be classified as OOD. We present the pseudo-code of the above process in Algorithm 1, and provide interpretation below.

Interpretation: If an utterance $\mathbf{u}$ contains the known intent c, $\mathbf{v}_c$ will fit the distribution $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$, $\mathcal{N}(\mathbf{v}_c; \boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ will be large and $-d_c$ will be large, otherwise $-d_c$ will be small (Murphy, 2022). So if $\mathbf{u}$ is IND, i.e., all r intents contained in $\mathbf{u}$ are known intents, there will be r large $-d$ in $\mathbf{D}$, thus the r-th maximum $\mathbf{D}$ should be large. If $\mathbf{u}$ is pure or mixed OOD, i.e., intents contained in $\mathbf{u}$ are not all known intents, there will be less than r large $-d$ in $\mathbf{D}$, thus the r-th maximum $\mathbf{D}$ should be small.

Although AIK is proposed from a multi-label perspective, it has a strong connection with OOD detection methods in multi-class. In multi-class, utterances are assumed to have one intent and methods always take the maximum score, such as maximum softmax probability (Hendrycks and Gimpel, 2017), maximum logit (Shu et al., 2017), to detect OOD utterances, which is equivalent to a special case of AIK that is r equals 1.

[3]Mahalanobis distance d_c can be calculated as: $d_c = (\mathbf{v}_c - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1} (\mathbf{v}_c - \boldsymbol{\mu}_c)$.

3.2 Model Architecture

Figure 2 show the model architecture of AIK.

Utterance encoding. We first employ a pre-trained BERT to encode the utterance $\mathbf{u} = \{w_1, w_2, ..., w_n\}$, where n is the number of tokens. Each token is encoded into a fix-length vector $\mathbf{h}$, and $\mathbf{h}_0$ is the hidden state for [CLS] token. We choose BERT due to its powerful capability of feature extraction. The utterance encoder can also be other models, such as GRU (Chung et al., 2014) or CNN (Kim, 2014).

Intent number prediction. Similar to other sentence-level tasks (Sun et al., 2019), we use $\mathbf{h}_0$ as the overall utterance representation, and predict the intent number of the utterance $\mathbf{u}$:

$$\hat{r} = f_{mlp}(\mathbf{h}_0), \tag{1}$$

where $\hat{r}$ is the predicted intent number, f_{mlp} is a multi-layer perceptron (MLP) network that maps $\mathbf{h}_0$ to a single scalar.

Known intent-wise representation extraction. Inspired by Mullenbach et al. (2018), we utilize a label-wise attention mechanism to get the known intent-wise representations. Specifically, we randomly initialize a trainable query $\mathbf{q}_c$ for each known intent c, and apply the query to calculate attention over hidden states. After that, we aggregate them to get the intent-wise utterance representation $\mathbf{v}_c$ for the intent c:

$$a_t = \frac{\exp(\mathbf{q}_c^T \mathbf{h}_t)}{\sum_{j=1}^{n} \exp(\mathbf{q}_c^T \mathbf{h}_j)}, \tag{2}$$

$$\mathbf{v}_c = \sum_{t=1}^{n} a_t \, \mathbf{h}_t, \tag{3}$$

where exp is the exponential function.

3.3 Training Objective

Intent number loss is mean-squared error (MSE) between the model's predicted intent number and golden intent number:

$$\mathcal{L}_{\text{int}} = \mathbb{E}_{(\mathbf{u},\mathbf{y})\sim\mathcal{D}} (\hat{r}_{\mathbf{u}} - r_{\mathbf{u}})^2, \tag{4}$$

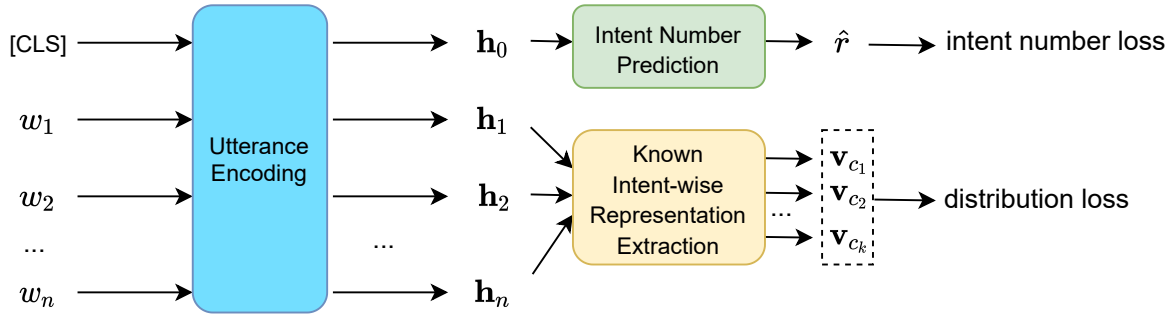

Figure 2: The model architecture of AIK.

where $\hat{r}_{\mathbf{u}}$ is the predicted intent number for utterance $\mathbf{u}$, $r_{\mathbf{u}}$ is the golden intent number, i.e., the size of set $\mathbf{y}$.

Distribution loss drives the known intent-wise representations toward the trainable conditional Gaussian distribution. For known intents contained in the utterance, we maximize the corresponding probability density, i.e., minimize the following loss:

$$\mathcal{L}_{\text{pos}} = -\mathbb{E}_{(\mathbf{u},\mathbf{y})\sim\mathcal{D}}\mathbb{E}_{c\sim\mathbf{y}}\,\mathcal{N}(\mathbf{v}_c;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}). \quad (5)$$

For known intents not contained in the utterance, we make the corresponding probability density less large by setting a margin t:

$$\mathcal{L}_{\text{neg}} = -\mathbb{E}_{(\mathbf{u},\mathbf{y})\sim\mathcal{D}}\mathbb{E}_{c\sim Y_{\text{in}}\setminus\mathbf{y}}\max(0, t-\mathcal{N}(\mathbf{v}_c;\boldsymbol{\mu}_c,\boldsymbol{\Sigma})). \quad (6)$$

Overall Loss: Finally, we train the AIK model by minimizing the following loss:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{pos}} + \lambda_2\mathcal{L}_{\text{neg}} + \lambda_3\mathcal{L}_{\text{int}}, \quad (7)$$

where λ_1, λ_2 and λ_3 are loss weights.

4 Experimental Setup

4.1 Datasets

To evaluate the effectiveness of AIK for multi-label unknown intent detection, we build two benchmark datasets from the existing multi-label intent classification datasets MixSNIPS (Qin et al., 2020) and MultiWOZ 2.3 (Han et al., 2020). The construction details are as follows:

- MixSNIPS (Qin et al., 2020) is collected from the SNIPS personal voice assistant (Coucke et al., 2018). We randomly select two intents as unknown intents for the validation set and another two intents as unknown intents for the test set. We conduct experiments with five different splits.

Statistic	MixSNIPS	MultiWOZ 2.3
Train-IND	6998	20319
Validation-IND	389	2531
Validation-OOD	664	2236
Test-IND	398	2530
Test-OOD	671	2418
Test-Mixed OOD	489	64
Test-Pure OOD	182	2354
Number of known intents	3	52
Average intent number per utterance	1.6	1.5

Table 2: Statistics of multi-label unknown intent detection datasets MixSNIPS and MultiWOZ 2.3. Test-Mixed OOD (or Test-Pure OOD) indicates mixed (or pure) OOD utterances in the test set.

- MultiWOZ 2.3 (Han et al., 2020) hosts more than 10K dialogues across eight different domains. For this dataset, we randomly select intents from two domains as unknown intents for the validation set and intents from another two domains as unknown intents for the test set. We also conduct experiments with five different splits.

Table 2 provides average summary statistics of all five splits on two datasets. Note that the training set does not contain OOD utterances.

4.2 Metrics

Similar to multi-class unknown intent detection, we adopt widely used metrics, including AUROC, FPR95, AUPR In, AUPR Out, to measure the performance of different methods in multi-label unknown intent detection.

- AUROC: the area under the true positive rate-false positive rate curve.
- FPR95: The false positive rate(FPR) when the true positive rate(TPR) is 95%. OOD data are treated as positive samples here.
- AUPR In: the area under the precision-recall curve. IND data are treated as positive samples.

- AUPR Out: the area under the precision-recall curve. OOD data are treated as positive samples.

Note that the larger AUROC, AUPR In, AUPR Out mean better performance, and the lower FPR95 indicates better performance.

4.3 Baselines

The multi-label and multi-class unknown intent detection have the same goal, i.e., identifying OOD utterances, thus some competitive OOD detection methods for multi-class can also be chosen as baselines for multi-label. In this work, we compare our AIK method with the generative-based method **Likelihood**, **Likelihood Ratio (LLR)** (Gangal et al., 2020; Ren et al., 2019) and the classifier-based method **Energy** (Ouyang et al., 2021; Liu et al., 2020), **Logit** (Shu et al., 2017), and **LOF** (Lin and Xu, 2019):

- Likelihood trains a language model with IND utterances, and OOD utterances tend to have a lower likelihood.
- LLR trains an extra language model with perturbed utterances to eliminate the unrelated factor in the likelihood for OOD detection.
- Energy uses the sum of exponential of binary classifier output to detect OOD.
- Logit uses the maximum binary classifier output to detect OOD.
- LOF uses local outlier factor (Breunig et al., 2000) in the utterance representation from the binary classifier to detect OOD.

4.4 Implementation Details

The encoder for all classifiers used in the baselines and ours are pre-trained BERT (Devlin et al., 2018). For a fair comparison, we also equip Energy and Logit with the label-wise attention mechanism. We select parameter values based on AUROC on the validation set. For LOF method, we set nearest neighbor number to 20. For LLR method, we follow Gangal et al. (2020), using UNIGRAM to introduce noise and setting p_{noise} to 0.5.

For our AIK method, we simply set λ_1, λ_2, λ_3 to 1, and τ can be set according to FPR95. In the training process, we randomly initialize known intent centers, and set the covariance matrix as identity matrix for reducing the difficulty of optimizing. So we can optimize the probability density by optimizing the Euclidean distance between the known intent-related representation and the known intent centers. The margin t is set to 300 when we optimize the distance. In the testing process, we follow Lee et al. (2018) to compute the empirical center and covariance for known intents as their conditional Gaussian distributions. We use rounding to ensure the predicted intent numbers are integers.

For all methods, we conduct five experiments with different seeds {0, 1, 2, 3, 4} on each split. As each dataset has five splits, we report the average results of 25 experiments.

5 Results and Analysis

5.1 Main Results

Table 3 shows the main results of different methods on multi-label unknown intent detection. From the results, we can observe that:

- AIK can achieve state-of-the-art results on all datasets and metrics. In particular, compared to the best baselines, AIK significantly reduces FPR95 by 15.29% on MixSNIPS dataset and 12.25% on MultiWOZ2.3. Figure 3 further provides the ROC curves of different methods.
- The method Logit and Energy perform poorly on MixSNIPS. The reason is that most OOD utterances in MixSNIPS are mixed, i.e., they also contain known intents, which makes the maximum binary classifier and energy easy to fail. We will discuss more on this in Section 5.2.
- LOF and AIK perform better on MixSNIPS than MultiWOZ 2.3. Note that both methods are based on utterance representation, a good representation space, such as making the representations of utterances with the same intent compact, is critical for them. Considering that the number of known intents is larger on MultiWOZ 2.3 (see Table 2), it is more difficult for the model to learn a good representation space.
- Likelihood and LLR perform stably and obtain appreciable results on two datasets. One bottleneck of such methods is that generative models are difficult to fit the the more complex distribution well for multi-label utterances.

Methods	MixSNIPS				MultiWOZ 2.3			
	AUROC ↑	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**	**AUROC ↑**	**FPR95↓**	**AUPR In↑**	**AUPR Out↑**
Likelihood	87.29	55.12	91.17	83.73	89.52	76.44	89.18	90.73
LLR	89.40	45.54	92.96	82.73	85.90	54.31	85.01	85.71
Energy	68.85	84.66	55.27	78.56	89.25	44.31	88.80	89.35
Logit	67.83	84.16	54.81	77.41	89.44	43.99	89.06	89.47
LOF	92.79	30.73	88.84	95.14	80.68	72.11	78.50	74.56
AIK	**96.29**	**15.44**	**94.93**	**97.46**	**92.22**	**31.74**	**93.33**	**91.01**

Table 3: AUROC, FPR95, AUPR In, AUPR Out on MixSNIPS, MultiWOZ 2.3 datasets. All results are percentages. Best results are in bold. Our method is significantly better than baselines with *p-value* < 0.01 using t-test.

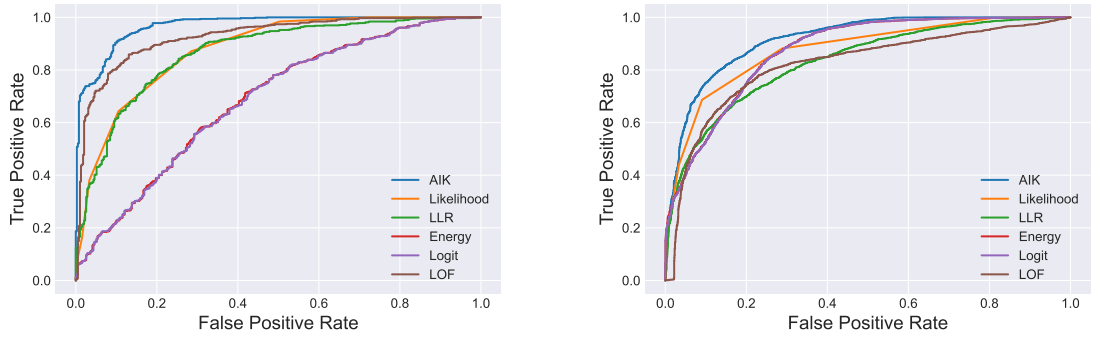

Figure 3: The ROC curves under different methods on MixSNIPS (left) and MultiWOZ 2.3 (right). For the convenience of visualization, we only choose one split for each dataset. The curves indicate AIK always performs better at different thresholds.

Method	**MixSNIPS** Pure/Mixed	**MultiWOZ 2.3** Pure/Mixed
Likelihood	95.52/84.19	89.69/83.98
LLR	93.21/87.96	86.44/67.72
Energy	93.37/59.76	89.55/79.10
Logit	92.47/58.69	89.73/79.25
LOF	93.09/92.66	80.78/75.84
AIK	**96.84/96.06**	**92.34/86.15**

Table 4: AUROC on two type OOD of utterances.

Split	**MixSNIPS** IND/OOD	**MultiWOZ 2.3** IND/OOD
Train	99.46/-	92.79/-
Validation	99.41/85.82	91.83/53.28
Test	99.47/82.70	91.79/53.87

Table 5: Intent number prediction accuracy.

5.2 Performance on Pure and Mixed OOD Utterances

To gain further insights, we measure the OOD performance of all methods in the cases of pure OOD utterances and mixed OOD utterances, respectively.

As shown in Table 4, compared with pure OOD utterances, we observe that all methods obtain lower performance on mixed OOD utterances, which demonstrates mixed OOD utterances are more challenging to detect. Specifically, for Likelihood and LLR, the known intent parts of mixed OOD utterances lie in high-density regions, resulting in a higher likelihood for the whole utterance. For Logit and Energy, the known intent parts will result in a higher binary classifier output. For LOF, the known intent parts will pull the whole utterance representation towards IND utterances, causing it to be misclassified as IND.

For AIK, detecting mixed OOD utterances performs comparably to pure OOD utterances on MixSNIPS. However, the performance gap on MultiWOZ 2.3 still exists. We will reveal that this is caused by the intent number prediction accuracy.

5.3 Analysis for Intent Number Prediction

The performance of AIK is depended on the intent number prediction accuracy. As described in Section 3.1, if an IND utterance has r intent(s), the r-th maximum $\mathbf{D}$ will be large. But once the predicted intent number $\hat{r}$ is greater than r, then $\hat{r}$-th maximum $\mathbf{D}$ might be small, causing the utterance to be misclassified as OOD. Similarly, for an OOD utterance with r intent(s), once the predicted intent number $\hat{r}$ is less than r, then $\hat{r}$-th maximum $\mathbf{D}$ might be large, causing the utterance to be misclassified as IND.

Table 5 shows the intent prediction accuracy of AIK. For IND utterances, we observe that both

Method	MixSNIPS Pure/Mixed	MultiWOZ 2.3 Pure/Mixed
AIK	96.84/96.06	92.34/86.15
AIK with Golden intent number	**96.94/97.26**	**93.66/94.46**

Table 6: Effect of using golden intent number. Values are AUROC.

	$\mathcal{L}_{\text{pos}}$	$\mathcal{L}_{\text{neg}}$	$\mathcal{L}_{\text{int}}$	MixSNIPS Pure/Mixed	MultiWOZ 2.3 Pure/Mixed
1	√			96.22/79.39	92.04/85.25
2	√	√		**97.02**/82.42	**92.86**/85.88
3	√		√	95.14/94.72	91.19/85.83
4	√	√	√	96.84/**96.06**	92.34/**86.15**

Table 7: AUROC results of ablation study of the objective function.

Method	AUROC↑
AIK	92.22
AIK with HM	**94.31**

Table 8: Effect of using HM centers for AIK on MultiWOZ 2.3.

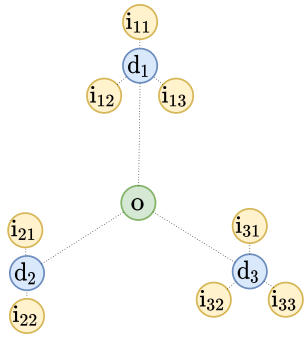

Figure 4: A toy example to show the HM centers. o represents the root node. d_i represents the domain node. i_{ij} represents the intent node from the domain d_i.

MixSNIPS and MultiWOZ 2.3 can reach high accuracy on validation and test set, such as greater than 90%. For OOD utterances, the accuracy is still maintained at a high level on MixSNIPS, while the accuracy is only about 50% on MultiWOZ 2.3. We conjecture that MixSNIPS is constructed manually, and there are some explicit features in utterances, such as "and", to indicate the number of intents. For the more challenging dataset MultiWOZ 2.3, predicting the number of intents is not so easy. Considering that we only take the hidden state of [CLS] to predict, there might be a great potential for improvement. For example, one can explicitly consider some intent number-related features, such as utterance length, number of verbs, etc.

We also test the OOD performance of AIK with the golden intent number. Namely, we use the utterance's golden intent number directly instead of predicting it. Table 6 shows the unknown intent detection performance can be further improved, and more performance improvement can be achieved on the low accuracy dataset MultiWOZ 2.3.

5.4 Ablation Study

We perform an ablation study to investigate the contribution of different losses for AIK. Table 7 shows the corresponding AUROC results.

Effect of $\mathcal{L}_{\text{neg}}$. We can observe that $\mathcal{L}_{\text{neg}}$ brings better results (Line 2 vs. Line 1, Line 3 vs. Line 4). This is because $\mathcal{L}_{\text{pos}}$ is not good at optimizing the inter-class dispersion, i.e., for intents not contained in the utterance, which makes the intent-wise representations low probability density. Ignoring inter-class dispersion might cause the encoder to learn a degenerate solution that all utterances have the same representation, resulting in the indistinguishability of IND and OOD utterances. $\mathcal{L}_{\text{neg}}$ makes up for this by making the corresponding probability density less high.

Effect of $\mathcal{L}_{\text{int}}$. Without intent number loss, we directly take the maximum $\mathbf{D}$ instead of the r-th maximum $\mathbf{D}$ as the score function. We observe that $\mathcal{L}_{\text{int}}$ brings improvement on mixed OOD utterances detection but leads to slight performance degradation on pure OOD utterances (Line 3 vs. Line 1, Line 2 vs. Line 4). For pure OOD utterances, their maximum $\mathbf{D}$ is small as they do not contain any known intent, so maximum $\mathbf{D}$ is enough for detecting them. However, for mixed OOD utterances, their maximum $\mathbf{D}$ is large due to the contained known intents, so using maximum $\mathbf{D}$ would result in these utterances being misclassified as IND.

5.5 Consider the Intent Relation for the Center Initialization

For AIK, at the beginning of the training stage, we initialize each known intent center randomly. However, for the complicated dataset (e.g., MultiWOZ 2.3), randomly initialized centers may ignore the relation between the intents. For example, for intents from the same domain, we might expect their intent centers to be closer.

To achieve this goal, we follow Pang et al. (2020) to preset untrainable hierarchical Max-Mahalanobis (HM) centers for each known intent. HM centers adaptively craft the class centers according to their tree structure. In our scenario, the

Method	AUROC↑
AIK	**96.29**
AIK with logit	95.21

Table 9: Effect of AIK with logit on MixSNIPS dataset.

tree structure is root-domain-intent. Specifically, we first generate the center for the root node (e.g., the origin). Next, we generate centers for each domain node by considering the root. Finally, we generate centers for each intent node by considering its domain node. Figure 4 shows a toy example to illustrate the generated intent centers.

As Table 8 shows, after considering the intent relation, the AIK performance can be further improved. We only consider the domain relation between intents here. It is also very interesting to consider correlation between intents in future work.

5.6 Use Logit to Check Known Intent Number

In our method, we choose to use the probability density, more specifically r-th maximum negative Mahalanobis distance, to check whether the utterance contains the same number of known intents. As an extension, we explore the effectiveness of using the logit. Particularly, we add the intent number prediction to the baseline method Logit, and use the r-th maximum binary classifier output to detect OOD. Ideally, an utterance with small r-th maximum output indicates its contained known intent number is less than its intent number, which can be classified as OOD.

Table 9 shows the performance degradation using logit. This is because logit always suffers from the label-overfitted problem and is not as reliable as probability density (Lee et al., 2018).

6 Related Work

6.1 Unknown Intent Detection

Classifier-based unknown intent detection depends on scores derived from the intent classifier trained with IND utterances and their labels. Hendrycks and Gimpel (2017) propose using the softmax score, which has become a common baseline. Nevertheless, some work (Louizos and Welling, 2017) demonstrates that the softmax score for OOD data can be arbitrarily high. Liu et al. (2020) propose using the energy score because it is theoretically aligned with the density of the input. Ouyang et al. (2021) extends the energy score for unknown intent detection. Some other works attempt to use distance-based scores to detect OOD utterances (Podolskiy et al., 2021; Yan et al., 2020; Lin and Xu, 2019). Although achieving significant results, the mixed OOD utterances cause that the above methods essentially might not generally apply to the more practical multi-label setting. Different from these methods, our method AIK, by recognizing whether all intents are known, is competent for the multi-label setting.

Generative-based unknown intent detection depends on scores derived from the generative model. These methods train the generative model to directly approximate the distribution of IND utterances, then use likelihood or likelihood ratio to detect OOD utterances (Ren et al., 2019; Gangal et al., 2020). These methods are more generalized as they do not rely on utterance labels. However, on more complex multi-label datasets, the generative model might be more challenging to train.

Unknown intent detection with auxiliary OOD utterances depends on scores derived from the model trained with both IND and OOD utterances. Ryu et al. (2018) directly train a discriminator with IND and OOD utterances. Zheng et al. (2019) use OOD utterances to calibrate the softmax score. Ouyang et al. (2021) use OOD utterances to shape the energy gap between IND and OOD utterances. Our method may also use the OOD utterances to further improve the performance by optimizing their probability density less high.

6.2 Multi-label Classification

The multi-label classification task aims to assign multiple non-exclusive labels to each sample. For text, many promising approaches have been proposed to address this problem, such as Binary relevance (Boutell et al., 2004), Classifier chains (Read et al., 2011), seq2seq models (Yang et al., 2018). Multi-label intent classification has also attracted interest recently (Qin et al., 2020; Hou et al., 2021). However, these methods make the closed world assumption, meaning that all classes of the test samples are known. In this paper, we consider an open world environment and detect samples with unknown classes.

7 Conclusion

In this paper, we propose a valuable and practical research task *multi-label unknown intent detection*. It aims to detect OOD utterances that may contain multiple intents. We further propose a novel

AIK method to perform multi-label unknown intent detection, by recognizing whether all intents contained in the utterance are known. Experimental results on two datasets validate the effectiveness of our method. We also analyze the challenges of detecting mixed OOD utterances for multi-label unknown intent detection through experiments.

Acknowledgments

Yawen would like to thank Yuanhang Tang for his constructive suggestions. This work was supported by NSFC Projects (Nos. 61936012 and 61976114).

References

Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. 2004. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 564–569.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. 2020. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7764–7771. AAAI Press.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2020. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv:2010.05594*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*.

Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. 2021. Few-shot learning for multi-label intent detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13036–13044. AAAI Press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7167–7177. Curran Associates, Inc.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496, Florence, Italy. Association for Computational Linguistics.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33.

Christos Louizos and Max Welling. 2017. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2218–2227. JMLR. org.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans,

Louisiana. Association for Computational Linguistics.

Kevin P Murphy. 2022. *Probabilistic machine learning: an introduction*. MIT press.

Yawen Ouyang, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Energy-based unknown intent detection with data manipulation.

Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. 2020. Rethinking softmax cross-entropy loss for adversarial robustness. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13675–13682. AAAI Press.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1807–1816, Online. Association for Computational Linguistics.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14680–14691.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916, Copenhagen, Denmark. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam. 2020. Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060, Online. Association for Computational Linguistics.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2019. Out-of-domain detection for natural language understanding in dialog systems. *arXiv preprint arXiv:1909.03862*.