Enhancing the generalization for Intent Classification and Out-of-Domain Detection in SLU

Yilin Shen, Yen-Chang Hsu, Avik Ray, Hongxia Jin
Samsung Research America
{yilin.shen,yenchang.hsu,avik.r,hongxia.jin}@samsung.com

Abstract

Intent classification is a major task in spoken language understanding (SLU). Since most models are built with pre-collected in-domain (IND) training utterances, their ability to detect unsupported out-of-domain (OOD) utterances has a critical effect in practical use. Recent works have shown that using extra data and labels can improve the OOD detection performance, yet it could be costly to collect such data. This paper proposes to train a model with only IND data while supporting both IND intent classification and OOD detection. Our method designs a novel domain-regularized module (DRM) to reduce the overconfident phenomenon of a vanilla classifier, achieving a better generalization in both cases. Besides, DRM can be used as a drop-in replacement for the last layer in any neural network-based intent classifier, providing a low-cost strategy for a significant improvement. The evaluation on four datasets shows that our method built on BERT and RoBERTa models achieves state-of-the-art performance against existing approaches and the strong baselines we created for the comparisons.

1 Introduction

Spoken language understanding (SLU) systems play a crucial role in ubiquitous artificially intelligent voice-enabled personal assistants (PA). SLU needs to process a wide variety of user utterances and carry out user's intents, a.k.a. *intent classification*. Many deep neural network-based SLU models have recently been proposed and have demonstrated significant progress (Guo et al., 2014; Liu and Lane, 2016; Zhang and Wang, 2016; Wang et al., 2018; Goo et al., 2018; Chen et al., 2019) in classification accuracy. These models usually apply the closed-world assumption, in which the SLU model is trained with predefined domains, and the model expects to see the same data distribution
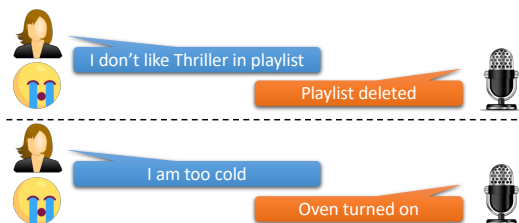

Figure 1: Failure Examples of Unsupported Skills in AI Voice Assistants. The user's utterances are out of the designed domains of the assistant.

during both training and testing. However, such an assumption is not held in the practical use case of PA systems, where the system is used under a dynamic and open environment with personal expressions, new vocabulary, and unknown intents that are out of the design scope.

To address the challenges in open-world settings, previous works adopt varied strategies. Shen et al. (2018a, 2019c) use a cold-start algorithm to generate additional training data to cover a larger variety of utterances. This strategy relies on the software developers to pre-build all possible skills. Shen et al. (2019b,a) introduce a SkillBot that allows users to build up their own skills. Recently, Ray et al. (2018, 2019); Shen et al. (2018b, 2019d) enables an SLU model to incorporate user personalization over time. However, the above approaches do not explicitly address unsupported user utterances/intents, leading to catastrophic failures illustrated in Figure 1. Thus, it is critically desirable for an SLU system to classify the supported intents (**in-domain (IND)**) and reject unsupported ones (**out-of-domain (OOD)**) correctly.

A straightforward solution is to collect OOD data and train a supervised binary classifier on both IND data and OOD data (Hendrycks et al., 2018). However, collecting a representative set of OOD data could be impractical due to the infinite

compositionality of language. Arbitrarily selecting a subset could incur the selection bias, causing the learned model might not generalize to unseen OOD data. Ryu et al. (2017, 2018) avoid learning with OOD data by using generative models (*e.g.,* autoencoder and GAN) to capture the IND data distribution, then judge IND/OOD based on the reconstruction error or likelihood. Recently, Tan et al. (2019) utilizes a large training data to enable the meta-learning for OOD detection. Zheng et al. (2020) generates pseudo OOD data to learn the OOD detector. The above-discussed approaches require additional data or training procedures beyond the intent classification task, introducing significant data collection effort or inference overhead.

This paper proposes a strategy based on neural networks to use only IND utterances and their labels to learn both the intent classifier and OOD detector. Our strategy modifies the structure of the classifier, introducing an extra branch as a regularization target. We call the structure a *Domain-Regularized Module* (DRM). This structure is probabilistically motivated and empirically leads to a better generalization in both intent classification and OOD detection. Our analysis focuses more on the latter task, finding that DRM not only outputs a class probability that is a better indicator for judging IND/OOD, but also leads to a feature representation with a less distribution overlap between IND and OOD data. More importantly, DRM is a simple drop-in replacement of the last linear layer, making it easy to plug into any off-the-shelf pretrained models (*e.g.* BERT (Devlin et al., 2019)) to fine-tune for a target task. The evaluation on four datasets shows that DRM can consistently improve upon previous state-of-the-art methods.

2 Problem Definition & Background

2.1 Problem Definition

In the application of intent classification, a user utterance will be either an in-domain (IND) utterance (supported by the system) or an out-of-domain (OOD) utterance (not supported by the system). The classifier is expected to correctly (1) predict the intent of supported IND utterances; and (2) detect to reject the unsupported OOD utterances.

The task is formally defined below. We are given a closed world IND training set $D_{IND} = \{\mathbf{x}, \mathbf{y}\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$. Each sample $(\boldsymbol{x}_i, y_i)$, an utterance $\boldsymbol{x}_i$ and its intent class label $y_i \in \{1 \dots C\}$ for C predefined in-domain classes, is drawn from a fixed but unknown IND distribution $P_{IND}(\mathbf{x}, \mathbf{y})$. We aim to train an intent classifier model *only* on IND training data D_{IND} such that the model can perform: (1) **Intent Classification:** classify the intent class label y of an utterance $\boldsymbol{x}$ if $\boldsymbol{x}$ is drawn from the same distribution P_{IND} as the training set D_{IND}; (2) **OOD Detection:** detect an utterance $\boldsymbol{x}$ to be an abnormal/unsupported sample if $\boldsymbol{x}$ is drawn from a different distribution P_{OOD}.

2.2 Related Work

Intent Classification is one of the major SLU components (Haffner et al., 2003; Wang et al., 2005; Tur and De Mori, 2011). Various models have been proposed to encode the user utterance for intent classification, including RNN (Ravuri and Stoicke, 2015; Zhang and Wang, 2016; Liu and Lane, 2016; Kim et al., 2017; Wang et al., 2018; Goo et al., 2018), Recursive autoencoders (Kato et al., 2017), or enriched word embeddings (Kim et al., 2016). Recently, the BERT model (Devlin et al., 2019) was explored by (Chen et al., 2019) for SLU. Our work also leverages the representation learned in BERT.

OOD Detection has been studied for many years (Hellman, 1970). Tur et al. (2014) explores its combination with intent classification by learning an SVM classifier on the IND data and randomly sampled OOD data. Ryu et al. (2017) detects OOD by using reconstruction criteria with an autoencoder. Ryu et al. (2018) learns an intent classifier with GAN and uses the discriminator as the classifier for OOD detection. Zheng et al. (2020) leverages extra unlabeled data to generate pseudo-OOD samples using GAN via auxiliary classifier regularization. Tan et al. (2019) further incorporates the few-shot setting, learning the encoding of sentences with a prototypical network that is regularized with the OOD data outside a learning episode. Other researchers developed methods in computer vision based on the rescaling of the predicted class probabilities (ODIN) (Liang et al., 2017) or building the Gaussian model with the features extracted from the hidden layers of neural networks (Mahalanobis) (Lee et al., 2018). Recently, (Hsu et al., 2020) proposed Generalized-ODIN with decomposed confidence scores. However, both approaches also heavily depend on the image input perturbation to achieve good performance. Unfortunately, such perturbation cannot be applied to discrete utterance data in SLU.

3 Our Method

Our method is inspired by the decomposed confidence of Generalized-ODIN (Hsu et al., 2020), but we leverage the fact that the training data are all from IND to introduce an extra regularization. This regularization leads to a better generalization (lower classification error) on the intent classification. The improvement is in contrast to the original Generalized-ODIN, which has its classification error slightly increased. Since the improved generalization is likely due to a more generalizable feature representation, we leverage this observation, providing a modified Mahalanobis (Lee et al., 2018), which we called L-Mahalanobis, for a transformer-based model to detect OOD data. In the following sections, we first describe the DRM and then elaborate on using the outputs of a DRM-equipped model to detect OOD data.

3.1 Domain-Regularized Module (DRM)

The motivation begins with introducing the domain variable d ($d = 1$ means IND, while $d = 0$ means OOD) following the intuition in (Hsu et al., 2020), then rewrite the posterior of class y given x with domain d as follows:

$$
\begin{aligned}
\widehat{p}(\mathrm{y}|\mathrm{d} = 1, \boldsymbol{x}) \quad &= \frac{\widehat{p}(\mathrm{y}, \mathrm{d} = 1|\boldsymbol{x})}{\widehat{p}(\mathrm{d} = 1|\boldsymbol{x})} \\
&= \frac{\widehat{p}(\mathrm{y}|\boldsymbol{x})}{\widehat{p}(\mathrm{d} = 1|\boldsymbol{x})} - \frac{\widehat{p}(\mathrm{y}, \mathrm{d} = 0|\boldsymbol{x})}{\widehat{p}(\mathrm{d} = 1|\boldsymbol{x})} \\
&\approx \frac{\widehat{p}(\mathrm{y}|\boldsymbol{x})}{\widehat{p}(\mathrm{d} = 1|\boldsymbol{x})}
\end{aligned}
\tag{1}
$$

where the last step holds since $\widehat{p}(\mathrm{y}, \mathrm{d} = 0|\boldsymbol{x})$ is close to 0 with the intrinsic conflict between IND classes y and random variable d = 0 for OOD.

3.1.1 DRM Design

Motivated by the above Equation 1, we design the DRM to mitigate overconfidence by decomposing the final logits f into two branches. Figure 2 illustrates the architecture.

Domain Logits f_d models $\widehat{p}(\mathrm{d} = 1|\boldsymbol{x})$ before normalization. It projects from hidden state h to a scalar w.r.t. d:

$$f_d = \mathbf{W}_d \mathrm{h} + \mathrm{b}_d \tag{2}$$

where $\mathbf{W}_d \in \mathcal{R}^{|h| \times 1}$. Since $\widehat{p}(\mathrm{d} = 1|\boldsymbol{x})$ is a probability between 0 and 1, Section 3.1.2 will describe the training details of domain loss via the sigmoid function.
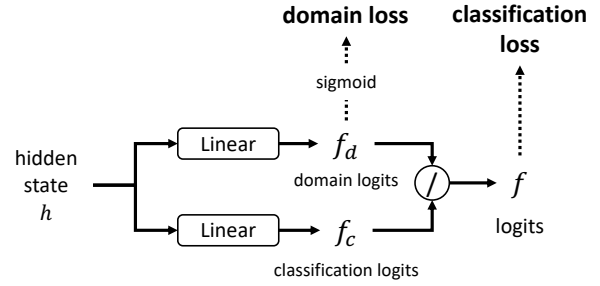

Figure 2: The DRM involves a domain component and a classification components for the IND classes.

Classification Logits f_c models the probability posterior $\widehat{p}(\mathrm{y}|\boldsymbol{x})$ before normalization. It follows the conventional linear projection from hidden state h to the number of classes:

$$f_c = \mathbf{W}_c \mathrm{h} + \mathrm{b}_c \tag{3}$$

where $\mathbf{W}_d \in \mathcal{R}^{|h| \times C}$ with C classes.

At the end, we obtain the final logits f to represent $\widehat{p}(\mathrm{y}|\mathrm{d} = 1, \boldsymbol{x})$ by putting f_d and f_c together following the dividend-divisor structure of Equation 1:

$$f = f_c / f_d \tag{4}$$

where each element of f_c is divided by the same scalar f_d.

3.1.2 DRM Training

We propose two training loss functions to train a model with DRM. The first training loss aims to minimize a cross-entropy between the predicted intent class and ground truth IND class labels.

$$\mathcal{L}_{classification} \triangleq - \sum_{i=1}^{C} \mathrm{y}^i \log \widehat{p}(f)^i \tag{5}$$

where $\widehat{p}(f)$ is the softmax of logits f:

$$\widehat{p}(f) = softmax(f)$$

The second training loss aims to ensure that the domain component f_d is close to 1 since all utterances in the training set are IND.

$$\mathcal{L}_{domain} \triangleq (1 - sigmoid(f_d))^2 \tag{6}$$

We first restrict f_d between 0 and 1 by using *sigmoid* activation function. Then, this loss function encourages $sigmoid(f_d)$ close to 1 for training on IND utterances. In order to avoid f_d to be very large values and affect the training convergence,

we further apply clamp function on f_d before it feeds to Equation 4:

$$f_d = \begin{cases} f_d & \text{if } -\delta < f_d < \delta \\ \delta & \text{if } f_d <= -\delta \text{ or } f_d >= \delta \end{cases}$$

Thus, we sum them up to optimize the model:

$$\mathcal{L} = \mathcal{L}_{\text{classification}} + \mathcal{L}_{\text{domain}} \tag{7}$$

Remarks: It is important to note that the design of $\mathcal{L}_{\text{domain}}$ is to introduce extra regularization to mitigate the overconfidence in standard posterior probability $\hat{p}(f)$. sigmoid(f_d) is not used to directly predict if an utterance is IND or OOD.

3.2 IND Intent Classification Method

Following Equation 1 and our DRM design, it is straightforward to use the confidence score of *softmax*(f) to predict the IND intent class.

3.3 OOD Detection Methods

There are two types of strategies to utilize the outputs of a classifier to perform OOD detection. One is based on the confidence which is computed from logits, the other is based on the features. In the below, we describe how to compute different OOD scores with our DRM.

3.3.1 Confidence-based Methods

Recent works (Liang et al., 2017) has shown that the softmax outputs provide a good scoring for detecting OOD data. In our DRM model, we use the *decomposed softmax outputs* for the score. The logits f_c w.r.t. the true posterior distribution in open-world can be combined with varied approaches:

DRM Confidence Score:

$$Conf_{DRM} = softmax(f_c) \tag{8}$$

DRM ODIN Confidence Score:

$$ODIN_{DRM} = softmax(f_c/T) \tag{9}$$

with large $T = 1000$ (Liang et al., 2017).

DRM Entropy Confidence Score:

$$ENT_{DRM} = Entropy[softmax(f_c)] \tag{10}$$

The OOD utterances have low $Conf_{DRM}$, $ODIN_{DRM}$ scores and high ENT_{DRM} score.

3.3.2 Feature-based Method

While our DRM confidence already outperforms many existing methods (later shown in experiments), we further design the feature-based Mahalanobis distance score, inspired by the recent work (Lee et al., 2018) for detecting OOD images.

We first recap the approach in (Lee et al., 2018) which consists of two parts: Mahalanobis distance calculation and input preprocessing. Mahalanobis distance score models the class conditional Gaussian distributions w.r.t. Gaussian discriminant analysis based on both low- and upper-level features of the deep classifier models. The score on layer ℓ is computed as follows:

$$S^\ell_{Maha}(\boldsymbol{x}) = \max_i -(f^\ell(\boldsymbol{x}) - \mu_c^\ell)^T \Sigma_\ell^{-1}(f^\ell(\boldsymbol{x}) - \mu_c^\ell)$$

where $f^\ell(\boldsymbol{x})$ represents the output features at the ℓth-layer of neural networks; μ_i and Σ are the class mean representation and the covariance matrix. Thus, the overall score is their summation:

$$S_{Maha}(\boldsymbol{x}) = \sum_\ell S_{Maha}(f^\ell(\boldsymbol{x}))$$

In addition, the input preprocessing adds a small controlled noise to the test samples to enhance the performance.

Although Mahalanobis distance score can be applied only to the last feature layer without input preprocessing $S^{last}_{Maha}(\boldsymbol{x})$, the analysis (Table 2 in (Lee et al., 2018)) shows that either input preprocessing or multi-layer scoring mechanism is required to achieve decent OOD detection performance. Unfortunately, neither of the above two mechanisms is applicable in the intent classifier for SLU. First, unlike image data, noise injection into discrete natural language utterances has been shown not to perform well. Second, in most cutting-edge intent classifier models, low- and upper-level network layers are quite different. The direct application of multi-layer Mahalanobis distance leads to much worse OOD detection performance.

Since BERT-based models showed significant performance improvement for intent classification in SLU (Chen et al., 2019), we focus on designing the multi-layer Mahalanobis score for BERT-based classifier models. In existing BERT-based text classification models, such as BERT, RoBERTa, Distil-BERT, ALBERT, etc., there are different designs between the last transformer layer and the classification layer. Figure 3 shows our generic design of

Table 1: SLU Benchmark and In-House Dataset Statistics

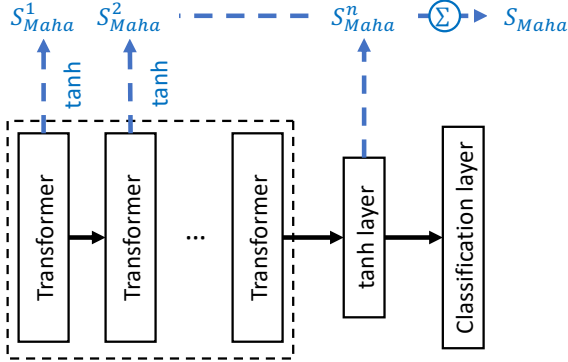| Dataset	Domain	#Intents	#Train	#Dev	#Test
CLINC (Larson et al., 2019)	various domains in voice assistants other out-of-scope domains	150 -	15,000 100	3,000 100	4,500 1,000
ATIS (Hemphill et al., 1990)	airline travel information domain	18	4,478	500	893
Snips (Coucke et al., 2018)	music, book, and weather domains	7	13,084	700	700
Movie (in-house)	movie QA domain	38	39,558	4,897	4,926

Figure 3: Multi-layer Mahalanobis Score Design for BERT-based Classifier Model

Mahalanobis score computation (blue) for various BERT-based models.

Our design is based on our extensive experiments and understanding of the common insights in different BERT-based models. Specifically, we use the features from different layers between the last transformer layer and the classification layer. We empirically found that the nonlinear tanh layer plays an important role. Thus, to map the features of each transformer layer and last layer into the same semantic space, we pass the features of each layer through tanh function and sum them up to compute our Mahalanobis score:

$$
\begin{aligned}
S_{L-Maha}(\boldsymbol{x}) &= S_{Maha}(f^n(\boldsymbol{x})) \\
&+ \sum_{1 \le \ell < n} S_{Maha}(tanh(f^\ell(\boldsymbol{x})))
\end{aligned}
\tag{11}
$$

where f^ℓ and f^n are the features of each layer ℓ and last layer n in a BERT-based intent classifier model. We refer to our proposed approach as **L-Mahalanobis**.

4 Experimental Evaluation

4.1 Datasets

We evaluate our proposed approach on three benchmark SLU datasets and one in-house SLU dataset. Table 1 provides an overview of all datasets.

Among all these datasets, the recently released CLINC dataset serves as a benchmark for OOD detection in SLU. For the other three datasets, we treat them mutually OOD due to non-overlapping domains.

We crowdsourced the in-house Movie dataset containing common questions that users may ask regarding movies. This dataset mainly consists of queries a user may ask in the movie domain. The dataset consists of 38 different intents (e.g. rating information, genre information, award information, show trailer) and 20 slots or entities (e.g., director, award, release year). This dataset was collected using crowdsourcing as follows. At first, some example template queries were generated by linguistic experts for each intent, along with intent and slot descriptions. Next, a generation crowdsourcing job was launched where a crowd worker was assigned a random intent, a combination of entities, and few slots generally associated with the intent. To better understand the intent and slots, the worker was asked to review the intent and slot descriptions, and example template utterances. The first task of the worker was to provide 3 different queries corresponding to the given intent, which also contains the provided entities. The second task of the worker was to provide additional entities corresponding to the same slot type. A subsequent validation crowdsourcing job was launched where these crowdsourced queries were rated by validation workers in terms of their accuracy with the provided intent and entities. Each query was rated by 5 different validation workers, and the final validated dataset contains a subset of crowdsourced queries with high accuracy score and high interrater agreement.

4.2 Implementation and Training Details

We implemented our method using PyTorch on top of the Hugging Face transformer library (Wolf et al., 2019). We follow the hyperparameters in the original models. For the only hyperparameter δ, we experimented only on CLINC dataset

from 2.2 to 4 with uniform interval 0.2 (we try 10 values of δ) based on $sigmoid(2.2) \approx 0.9$ and $sigmoid(4) \approx 0.982$. We used $\delta = 3$ which gives the best performance in our experiment for all datasets. We train each model with 3 epochs using 4 NVIDIA Tesla V100 GPUs (16GB) for each training. We conducted experiments on two transformer-based models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019).

Remarks: All experiments only use IND data for both training and validation. We use the same hyperparameters in all datasets and validate the generalizability of our method.

4.3 Baselines

4.3.1 IND Intent Classification Baselines

We consider the strongest baseline **BERT-Linear** (the last layer is linear) fine-tuned on the pre-trained BERT-based models (Chen et al., 2019).

4.3.2 OOD Detection Baselines

We consider the existing OOD detection methods:

ConGAN (Ryu et al., 2018): a GAN-based model based on given sentence representations to generate OOD features with additional feature matching loss. OOD utterances are expected to have low discriminator confidence scores.

Autoencoder (AE) (Ryu et al., 2017): first uses an LSTM based classifier model to train sentence representations; then train an autoencoder on the above sentence embeddings. OOD utterances are expected to have high reconstruction error.

ODIN (Liang et al., 2017): we only use the temperature scaling on logits. OOD utterances are expected to have a low scaled confidence score.

Generalized-ODIN (G-ODIN) (Hsu et al., 2020): we fine-tune on pre-trained BERT models with replaced last layer and only use the decomposed confidence. We evaluate all three variations proposed in the paper h^I, h^E and h^C and report the best one. OOD utterances are expected to have low scaled confidence score.

Mahalanobis (Lee et al., 2018): we only use the feature of BERT's last layer to compute Mahalanobis distance score. OOD utterances are expected to have a low scaled confidence score.

For ConGAN and AE, we evaluate the model in the original paper as well as customized BERT-based backbone models as strong baselines. Specifically, we customize **En-ConGAN** and **En-AE** as follows: **En-ConGAN** uses BERT sentence representation as input; **En-AE** applies a BERT classi-

fier model to train the sentence representation and then use them to further train an autoencoder. *Thus, En-ConGAN and En-AE are not existing baselines.*

Note that ERAEPOG (Zheng et al., 2020) and O-Proto (Tan et al., 2019) are not comparable since they require additional unlabeled data and labels. We only put the ERAEPOG results on CLINC dataset (from the original paper) for reference.

4.4 Evaluation Metrics

4.4.1 IND Intent Classification Metrics

We evaluate IND performance using the **classification accuracy** metric as in literature (Liu and Lane, 2016; Wang et al., 2018; Chen et al., 2019).

4.4.2 OOD Detection Metrics

we follow the evaluation metrics in literature (Ryu et al., 2018) and (Liang et al., 2017; Lee et al., 2018). Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative. We use the following OOD evaluation metrics:

EER (lower is better): (Equal Error Rate) measures the error rate when false positive rate (FPR) is equal to the false negative rate (FNR). Here, FPR=FP/(FP+TN) and FNR=FN/(TP+FN).

FPR95 (lower is better): (False Positive Rate (FPR) at 95% True Positive Rate (TPR)) can be interpreted as the probability that an OOD utterance is misclassified as IND when the true positive rate (TPR) is as high as 95%. Here, TPR=TP/(TP+FN).

Detection Error (lower is better): measures the misclassification probability when TPR is 95%. Detection error is defined as follows:

$$\min_{\delta}\{P_{IND}(s \leq \delta)p(\boldsymbol{x} \in P_{IND})$$
$$+P_{OOD}(s > \delta)p(\boldsymbol{x} \in P_{OOD})\}$$

where s is a confidence score. We follow the same assumption that both IND and OOD examples have an equal probability of appearing in the testing set.

AUROC (higher is better): (Area under the Receiver Operating Characteristic Curve) The ROC curve is a graph plotting TPR against the FPR=FP/(FP+TN) by varying a threshold.

AUPR (higher is better): (Area under the Precision-Recall Curve (AUPR)) The PR curve is a graph plotting the precision against recall by varying a threshold. Here, precision=TP/(TP+FP) and recall=TP/(TP+FN). AUPR-IN and AUPR-OUT is AUPR where IND and OOD distribution samples are specified as positive, respectively.

Table 2: Comprehensive OOD Detection Results on CLINC Dataset (CLINC Train/OOD)

Model	Last Layer	OOD Method	EER(↓)	FPR95(↓)	OOD Evaluation Detection Error(↓)	AUROC(↑)	AUPR In(↑)	AUPR Out(↑)
ConGAN	-	-	78.90[§]	94.40[§]	52.04[§]	52.22[§]	82.79[§]	23.54[§]
AE	-	-	18.13[§]	58.50[§]	23.94[§]	87.78[§]	96.98[§]	54.12[§]
ERAEPOG	-	-	12.04[§]	23.70[§]	11.67[§]	95.83[§]	99.05[†]	83.98[§]
BERT		En-ConGAN	75.20[§]	98.72[§]	49.95[§]	22.36[§]	69.86[§]	11.27[§]
		En-AE	8.70[§]	13.03[§]	8.47[§]	96.12[§]	98.89[§]	88.38[§]
		ODIN	9.01[§]	16.52[§]	8.66[§]	96.24[§]	98.73[§]	87.34[§]
		G-ODIN	8.91[§]	12.99[§]	8.40[§]	95.81[§]	98.75[†]	88.81[§]
	Linear	Confidence	11.31[§]	21.98[§]	11.00[§]	94.96[§]	98.52[§]	84.59[§]
		Entropy	10.33[§]	17.99[§]	10.10[§]	95.65[§]	98.73[§]	87.20[§]
		Mahalanobis	8.31[§]	12.68[§]	8.02[§]	96.90[§]	99.14[†]	88.19[§]
		L-Mahalanobis*	7.21	10.18	6.92	97.52	99.41	89.37
	DRM*	Confidence*	8.50	12.85	7.85	96.34	98.95	87.51
		Entropy*	8.31	12.53	8.14	96.67	99.01	89.68
		Mahalanobis*	7.01	10.88	6.88	97.43	99.37	90.36
		L-Mahalanobis*	**6.70**	**10.12**	**6.62**	**97.77**	**99.46**	**91.55**
RoBERTa		En-ConGAN	80.26[§]	99.34[§]	49.95[§]	15.20[§]	66.64[§]	10.58[§]
		En-AE	8.56[§]	12.38[§]	8.29[§]	96.82[§]	99.08[†]	90.06[§]
		ODIN	9.11[§]	15.12[§]	8.68[§]	96.11[§]	98.84[§]	88.72[§]
		G-ODIN	8.85[§]	12.26[§]	8.53[§]	96.74[§]	99.12[†]	89.95[§]
	Linear	Confidence	10.81[§]	22.35[§]	10.38[§]	95.23[§]	98.58[§]	86.46[§]
		Entropy	9.31[§]	14.81[§]	8.93[§]	95.89[§]	98.73[§]	88.70[§]
		Mahalanobis	8.40[§]	11.82[§]	8.13[§]	96.92[§]	99.06[§]	90.37[§]
		L-Mahalanobis*	6.90	9.53	6.71	97.94	99.50	92.47
	DRM*	Confidence*	8.35	11.76	8.02	97.10	99.25	90.46
		Entropy*	8.29	11.51	7.86	97.17	99.27	90.69
		Mahalanobis*	6.31	7.80	6.13	98.07	99.53	92.86
		L-Mahalanobis*	**6.11**	**7.63**	**5.98**	**98.16**	**99.56**	**92.96**

Our best method (DRM+L-Mahalanobis) is significantly better than each baseline model (without *) with *p-value* < 0.01 (marked by §) and *p-value* < 0.05 (marked by †) using t-test. All methods with * are our proposed methods.

Note that EER, detection error, AUROC, and AUPR are threshold-independent metrics.

4.4.3 Statistical Significance

We also evaluate the statistical significance between all baselines and our best result (DRM + L-Mahalanobis) on all the above metrics. We train each model 10 times with different PyTorch random seeds. We report the average results and t-test statistical significance results.

4.5 Results

4.5.1 IND Classification Results

Table 3 reports the IND intent classification results on each dataset finetuned using BERT and RoBERTa pre-trained models. It is interesting to observe that all DRM combined models consistently achieve better classification accuracy with up to 0.8% improvement (reproduced "No joint" row in Table 3 in (Chen et al., 2019) on Snips dataset). This is because the domain loss forces $sigmoid(f_d)$ close to 1 and therefore also slightly mitigates its impact to IND classification. Thus, the true posterior distribution of IND data is also modeled more precisely. For both BERT and RoBERTa

Table 3: IND Intent Classification Results

Model	Last Layer	Datasets CLINC	ATIS	Snips	Movie
BERT	Linear	96.19[†]	97.76[†]	97.97[†]	97.26[†]
	DRM*	**96.66**	**98.21**	**98.23**	**97.87**
RoBERTa	Linear	96.82[†]	97.64[†]	98.07[†]	98.07[†]
	DRM*	**97.15**	**98.31**	**98.87**	**98.63**

Our DRM methods (marked by *) are significantly better than baseline model on all datasets with *p-value* < 0.05 (marked by †) using t-test.

backbones, DRM models are significantly better than conventional BERT-linear classification models with *p-value* < 0.05.

4.5.2 OOD Detection Results

Results on CLINC Dataset: Table 2 reports the OOD detection results on CLINC dataset. This result covers all existing work and our enhanced baselines. We focus on analyzing the contribution by each of our proposed techniques, DRM and L-Mahalanobis. The first three rows report the performance of existing approaches based on the original designs in their papers (ERAEPOG in grey uses additional unlabeled data). Unfortunately, we observe that their performance is even worse than the simple confidence-based approach via BERT

Table 4: OOD Detection Results on Snips/ATIS/Movie Datasets (RoBERTa Model Finetuning)

OOD Method	EER($\downarrow$)	FPR95($\downarrow$)	OOD Evaluation Detection Error($\downarrow$)	AUROC($\uparrow$)	AUPR In($\uparrow$)	AUPR Out($\uparrow$)
			IND dataset: Snips; OOD Datasets: CLINC_OOD/ATIS/Movie			
En-ConGAN	54.50§/63.05§/54.22§	99.16§/99.87§/99.10§	42.61§/49.10§/37.32§	39.03§/30.88§/45.64§	37.15§/34.47§/30.03§	51.23§/45.70§/52.59§
Confidence	9.91§/17.83§/22.22§	14.94§/47.43§/51.85§	9.18§/11.17§/19.34§	96.09§/92.03§/87.44§	94.78§/92.65§/97.67§	97.21§/92.29§/55.16§
Entropy	10.21§/18.05§/23.15§	14.54§/45.04§/52.68§	9.25§/10.77§/19.58§	96.32§/92.44§/87.12§	94.90§/92.94§/97.60§	97.53§/92.99§/52.27§
ODIN	10.01§/16.93§/23.15§	14.22§/39.04§/58.33§	9.43§/9.64§/23.01§	96.46§/93.81§/83.58§	94.59§/93.99§/96.63§	97.75§/94.53§/47.36§
G-ODIN	9.65§/15.16§/22.02§	13.31§/37.86§/55.67§	8.32§/8.55§/21.82§	97.21§/94.73§/85.60§	95.70§/95.04§/97.73§	98.02§/95.44§/50.38§
En-AE	4.40§/4.37§/3.59§	4.18§/3.59§/3.08§	4.25§/4.00§/3.64†	98.56§/98.12§/88.96§	97.41§/98.92†/94.39§	98.12§/95.34§/86.84§
Maha	3.90§/1.81/11.11§	2.66§/2.23§/5.58§	3.47§/1.36†/10.21§	98.79†/99.74§/95.61§	97.73§/99.75†/99.22§	99.21§/99.77†/76.61§
DRM+L-Maha*	**3.00/1.79/2.78**	**1.95/0.00/2.78**	**2.63/1.16/3.16**	**98.90/99.79/98.53**	**98.15/99.79/99.76**	**99.24/99.80/87.02**
			IND dataset: ATIS; OOD Datasets: CLINC_OOD/Snips/Movie			
En-ConGAN	21.60§/19.74§/23.28§	81.52§/86.33§/93.77§	15.51§/15.54§/16.03§	82.34§/81.79§/79.32§	84.52§/89.35§/58.36§	72.74§/60.20§/89.14§
Confidence	10.21§/8.52§/10.19§	20.50§/12.92§/17.59§	9.28§/8.36§/9.33§	96.99§/97.84§/96.62§	97.19§/98.57§/99.56†	97.04§/96.99§/84.27§
Entropy	9.91§/8.84§/10.12§	21.67§/13.75§/17.59§	9.11§/8.16§/9.38§	97.06§/97.93§/96.68§	97.25§/98.62§/99.57†	97.11§/97.14§/85.02§
ODIN	9.11§/8.36§/10.08§	21.32§/14.39§/18.52§	7.50§/6.15§/9.37§	97.16§/98.00§/96.73§	97.39§/98.68§/99.58†	97.16§/97.18§/84.88§
G-ODIN	8.75§/8.01§/9.97§	20.87§/13.44§/17.76§	7.31§/6.02§/8.98§	97.27§/98.11§/96.85§	97.46§/98.76§/99.59†	97.28§/97.32§/85.90§
En-AE	4.00§/**2.09§**/3.69§	2.20†/**0.00**/0.35†	3.45†/**1.33**/1.97†	99.41§/**99.83**/99.63§	99.43†/**99.89**/98.72§	99.43†/**99.74**/97.93§
Maha	4.00§/3.85§/6.48§	12.13§/8.06§/11.64§	3.76§/2.94§/5.04§	99.18§/99.47§/98.72†	98.78§/99.45§/99.71§	99.46§/99.49§/95.45§
DRM+L-Maha*	**2.70/2.09/1.85**	**1.30/0.32/0.00**	**2.55/2.01/1.23**	**99.48/99.70/99.78**	**99.51/99.82/99.97**	**99.47/99.50/98.22**
			IND dataset: Movie; OOD Datasets: CLINC_OOD/ATIS/Snips			
En-ConGAN	45.90§/15.12§/41.09§	44.05§/14.35§/39.59§	22.95§/7.56§/20.55§	43.85§/57.44§/45.78§	85.21§/88.23§/90.40§	14.68§/17.56§/10.09§
Confidence	19.22§/16.70§/18.81§	36.81§/47.94§/47.52§	18.51§/15.15§/18.53§	91.65§/91.99§/90.53§	98.11§/98.50§/98.68§	76.78§/67.58§/59.63§
Entropy	19.12§/17.26§/19.13§	34.64§/44.24§/44.80§	18.25§/16.12§/18.87§	91.79§/92.14§/90.72§	98.11§/98.50§/98.69§	78.66§/70.87§/63.96§
ODIN	19.42§/18.95§/19.94§	34.43§/39.91§/39.38§	18.24§/18.38§/19.33§	91.34§/91.40§/90.03§	97.96§/98.29§/98.53§	78.56§/71.62§/65.18§
G-ODIN	18.61§/18.23§/19.25§	34.19§/36.42§/37.03§	18.15§/17.27§/18.91§	91.86§/91.97§/90.63§	98.21§/98.34§/98.70§	78.98§/72.07§/66.79§
En-AE	13.70§/7.28§/16.05§	43.42§/16.05§/32.29§	11.00§/4.46§/11.87§	94.57§/93.56§/92.23§	98.91§/99.58†/99.01†	77.12§/76.13§/68.75§
Maha	3.90§/3.41†/6.11§	6.02§/2.35§/15.40§	3.72§/3.02§/6.02§	99.37§/99.43§/98.63§	99.81†/99.89§/99.82†	97.82†/97.27§/91.44§
DRM+L-Maha*	**3.70/3.36/4.66**	**2.56/1.01/4.34**	**3.61/2.85/4.58**	**99.48/99.53/99.06**	**99.89/99.92/99.88**	**97.90/97.38/93.85**

In each OOD method for an IND dataset, "/" separates the results for different OOD datasets.
Our method (*) is significantly better than baseline models with *p-value* < 0.01 (marked by §) and *p-value* < 0.05 (marked by †) using t-test in most cases.

finetuning baseline (row 5). Thus, we mainly focus on comparing our method with strong baselines with BERT and RoBERTa models.

For a given OOD detection method, we find that their combinations with DRM consistently perform better than those with standard models. The improvement is at least 1-2% for all metrics against our enhanced baselines. Among all OOD detection approaches, our proposed L-Mahalanobis OOD detection approach achieves the best performance for both linear and DRM combined BERT and RoBERTa models. It is not surprising to observe that our DRM method combined with a better pretrained RoBERTa model achieves larger OOD detection performance improvement. Note that our customized En-AE performs much better than most other methods since we incorporated the enhanced reconstruction capability with pre-trained BERT models. However, En-AE cannot utilize all BERT layers as our proposed L-Mahalanobis method, resulting in worse performance.

In addition, DRM+L-Mahalanobis models are significantly better than existing methods and enhanced baselines with *p-value* < 0.01 on most metrics for both BERT and RoBERTa backbones.

Ablation Study on CLINC Dataset: We analyze how our two novel components, DRM model and L-Mahalanobis, impact the performance.

The rows with "DRM" in "Last Layer" column of Table 2 show the performance of DRM model. As one can see, for all OOD methods, DRM consistently performs better than the conventional "Linear" last layer. Specifically, the DRM and Confidence combo also outperforms its closest baseline G-ODIN. This validates the effectiveness of our disentangled logits design in DRM based on the mathematical analysis of overconfidence. It also shows that our new domain loss can indeed enhance the model awareness that all training data is IND.

The rows with "L-Mahalanobis" in "OOD Method" column of Table 2 outperform other OOD methods with the same model and last layer. Compared with its closest baseline Mahalanobis, the better performance of L-Mahalanobis validates the usefulness of all layers' features in various models.

Results on ATIS/Snips/Movie Datasets: Since our strong baselines on pre-trained RoBERTa model showed better results on CLINC, we next evaluate other results finetuned on RoBERTa

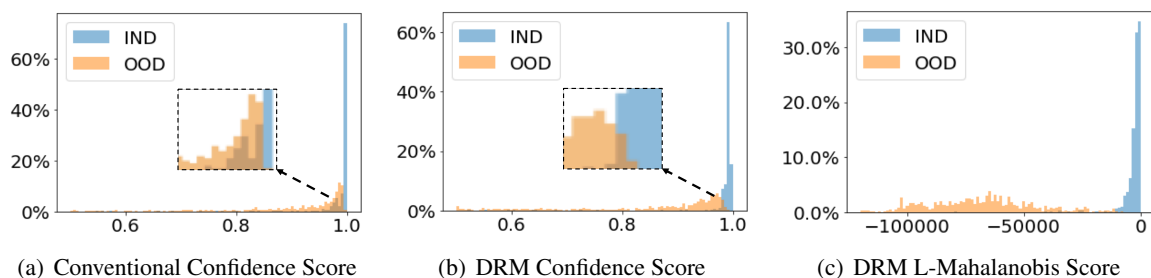(a) Conventional Confidence Score   (b) DRM Confidence Score   (c) DRM L-Mahalanobis Score

Figure 4: Histogram of Detection Scores using Various Methods (Snips IND, ATIS OOD) (We choose this IND/OOD combination to provide the best visualization for analysis)

model. When taking each dataset as IND, we use the other two mutually exclusive datasets and CLINC_OOD as OOD datasets for evaluating OOD detection performance. As one can see in Table 4, our method outperforms other approaches on both Snips and movie IND datasets. For ATIS IND dataset, En-AE for Snips OOD dataset achieves almost perfect performance. This is because ATIS and Snips are almost completely non-overlapping and ATIS is well designed with carefully selected varieties and entities in the airline travel domain. When taking Snip as IND and ATIS as OOD, it is interesting to see that our method achieves better performance than En-AE. This is because that Snips contains a large number of entities such that the reconstruction error will be lower and become less separable than that in ATIS OOD utterances.

For both Snips and Movie IND datasets, DRM+L-Mahalanobis are significantly better than baseline methods with *p-value* < 0.01 in most cases for all OOD datasets. For ATIS IND dataset, DRM+L-Mahalanobis shows similar behavior except En-AE since it is easier to train an autoencoder model for ATIS IND dataset due to its carefully collected clean training utterances.

4.6 Qualitative Analysis

We provide a quantitative analysis by visualizing our two methods, DRM and L-Mahalanobis.

4.6.1 Detection Score Distribution

Figure 4 plots the histograms of detection scores for OOD and IND data. Compared with Figure 4(a), DRM significantly reduces the overlap between OOD and IND in Figure 4(b). L-Mahalanobis utilizes features from all layers to further reduce the overlap in Figure 4(c). Moreover, the score distributions from left to right in Figure 4, imply that a larger entropy of all score reflects a better uncertainty modeling.

4.6.2 Feature Distribution Visualization

Figure 5 visualizes the utterance representations learned with or without DRM. The red IND data are tightly clustered within classes (totally 150 CLINC IND classes), while the blue OOD data spread arbitrarily. As one can see, the blue dots in Figure 5(b) have less overlap with red dots, indicating the DRM helps to learn the utterance representation to better disentangle IND and OOD data.
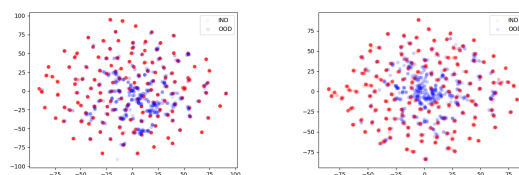
(a) Conventional RoBERTa   (b) DRM RoBERTa

Figure 5: t-SNE Visualization of Utterance Representations on CLINC Dataset (Red: IND, Blue: OOD)

5 Conclusion

This paper proposes using only IND utterances to conduct intent classification and OOD detection for SLU in an open-world setting. The proposed DRM has a structure of two branches to avoid overconfidence and achieves a better generalization. The evaluation shows that our method can achieve state-of-the-art performance on various SLU benchmark and in-house datasets for both IND intent classification and OOD detection. In addition, thanks to the generic of our DRM design and with the recent extensive use of BERT on different data modalities, our work can contribute to improving both in-domain classification robustness and out-of-domain detection robustness for various classification models such as image classification, sound classification, vision-language classifications.

Impact Statement

Our proposed method in this paper has been deployed in the domain classification SLU model for Samsung Bixby voice assistant. In addition to SLU, our work could have a broader impact on other applications, which can be benefited from having a more robust classification system. For example, our method can help the robot to detect objects more accurately or stop safely by correctly identifying unknown objects, classify environmental sounds or detect anomaly sounds, and so on. Moreover, by better detecting the OOD samples that are different from the training data distribution, our method can facilitate to handle distributional shifts between training data and practical usage data.

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. *CoRR*, abs/1902.10909.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *NAACL-HLT*, pages 753–757.

Daniel Guo, Gokhan Tur, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *SLT*, pages 554–559.

Patrick Haffner, Gokhan Tur, and Jerry H Wright. 2003. Optimizing svms for complex call classification. In *ICASSP*, volume 1.

Martin E Hellman. 1970. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6(3):179–185.

Charles T Hemphill, John J Godfrey, George R Doddington, et al. 1990. The atis spoken language systems pilot corpus. In *Proceedings of the DARPA speech and natural language workshop*, pages 96–101.

Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. 2018. Deep anomaly detection with outlier exposure. In *ICLR*.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2020. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, pages 10948–10957.

Tsuneo Kato, Atsushi Nagai, Naoki Noda, Ryosuke Sumitomo, Jianming Wu, and Seiichi Yamamoto. 2017. Utterance intent classification of a spoken dialogue system with efficiently untied recursive autoencoders. In *SIGDIAL*, pages 60–64.

Joo-Kyung Kim, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016. Intent detection using semantically enriched word embeddings. In *SLT*, pages 414–419.

Young-Bum Kim, Sungjin Lee, and Karl Stratos. 2017. ONENET: joint domain, intent, slot prediction for spoken language understanding. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop*, pages 547–553.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *EMNLP-IJCNLP*, pages 1311–1316.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177.

Shiyu Liang, Yixuan Li, and R Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *INTERSPEECH*, pages 685–689.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Suman Ravuri and Andreas Stoicke. 2015. A comparative study of neural network models for lexical intent classification. In *ASRU*, pages 368–374.

Avik Ray, Yilin Shen, and Hongxia Jin. 2018. Learning out-of-vocabulary words in intelligent personal agents. In *IJCAI*, pages 4309–4315.

Avik Ray, Yilin Shen, and Hongxia Jin. 2019. Fast domain adaptation of semantic parsers via paraphrase attention. In *DeepLo@EMNLP-IJCNLP*, pages 94–103.

Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. 2017. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recogn. Lett.*, 88(C):26–32.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *EMNLP*, pages 714–718.

Yilin Shen, Sandeep Nama, and Hongxia Jin. 2019a. Teach once and use everywhere – building ai assistant eco-skills via user instruction and demonstration (poster). In *MobiSys*, pages 606–607.

Yilin Shen, Avik Ray, Hongxia Jin, and Sandeep Nama. 2019b. SkillBot: Towards automatic skill development via user demonstration. In *NAACL-HLT, System Demonstrations*, pages 105–109.

Yilin Shen, Avik Ray, Abhishek Patel, and Hongxia Jin. 2018a. CRUISE: cold-start new skill development via iterative utterance generation. In *ACL, System Demonstrations*, pages 105–110.

Yilin Shen, Yu Wang, Abhishek Patel, and Hongxia Jin. 2019c. Sliqa-i: Towards cold-start development of end-to-end spoken language interface for question answering. In *ICASSP*, pages 7195–7199.

Yilin Shen, Xiangyu Zeng, and Hongxia Jin. 2019d. A progressive model to enable continual learning for semantic slot filling. In *EMNLP-IJCNLP*, pages 1279–1284.

Yilin Shen, Xiangyu Zeng, Yu Wang, and Hongxia Jin. 2018b. User information augmented semantic frame parsing using progressive neural networks. In *INTERSPEECH*, pages 3464–3468.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *EMNLP-IJCNLP*, pages 3564–3570.

G. Tur and R. De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley.

Gokhan Tur, Anoop Deoras, and Dilek Hakkani-Tür. 2014. Detecting out-of-domain utterances addressed to a virtual personal assistant. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5):16–31.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. In *NAACL-HLT*, pages 309–314.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, pages 2993–2999.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:1198–1209.