Are Pre-trained Transformers Robust in Intent Classification? A Missing Ingredient in Evaluation of Out-of-Scope Intent Detection

Anonymous ACL submission

Abstract

Pre-trained Transformer-based models were reported to be robust in intent classification. In this work, we first point out the importance of in-domain out-of-scope detection in few-shot intent recognition tasks and then illustrate the vulnerability of pre-trained Transformer-based models against samples that are in-domain but out-of-scope (ID-OOS). We construct two new datasets, and empirically show that pre-trained models do not perform well on both ID-OOS examples and general out-of-scope examples, especially on fine-grained few-shot intent detection tasks. To figure out how the models mistakenly classify ID-OOS intents as in-scope intents, we further conduct analysis on confidence scores and the overlapping keywords, as well as point out several prospective directions for future work.

1 Introduction

Intent detection, which aims to identify intents from user utterances, is a vital task in goal-oriented dialog systems (Gao et al., 2019). However, the performance of intent detection has been hindered by the data scarcity issue, as it is non-trivial to collect sufficient examples for new intents. In practice, the user requests could also be not expected or supported by the tested dialog system, referred to as out-of-scope (OOS) intents. Thus, it is important to improve OOS intents detection performance while keeping the accuracy of detecting in-scope intents in the few-shot learning scenario.

Recently, several approaches (Zheng et al., 2019; Zhang et al., 2020; Wu et al., 2020; Cavalin et al., 2020; Zhan et al., 2021; Xu et al., 2021) have been proposed to improve the performance of identifying in-scope and OOS intents in few-shot scenarios. Previous experiments have shown that a simple confidence-based out-of-distribution detection method (Hendrycks and Gimpel, 2017; Hendrycks et al., 2020) equipped with pre-trained BERT can improve OOS detection accuracy. However, there is a lack of further study of pre-trained Transformers on few-shot fine-grained OOS detection where the OOS intents are more relevant to the in-scope intents. Besides, those studies mainly focus on the CLINC dataset (Larson et al., 2019), in which the OOS examples are designed such that they do *not* belong to any of the known intent classes. Their distribution is dissimilar to each other, and thus they are easy to be distinguished from the known intent classes. Moreover, CLINC is not enough to study more challenging few-shot fine-grained OOS detection as it lacks such semantically similar OOS examples to in-scope intents, and other popular used datasets such as BANKING77 (Casanueva et al., 2020) do not contain OOS examples.

In this paper, we aim to investigate the following research question: "*Are pre-trained Transformers robust in intent classification w.r.t. general and relevant OOS examples?*". We first define two types of OOS intents: out-of-domain OOS (**OOD-OOS**) and in-domain OOS (**ID-OOS**). We then investigate *how* robustly state-of-the-art pre-trained Transformers perform on these two OOS types. The OOD-OOS is identical to the OOS in the CLINC dataset, where the OOS and in-scope intents (e.g., requesting an online TV show service in a banking system) are topically rarely overlapped. We construct an ID-OOS set for a domain, by separating semantically-related intents from the in-scope intents (e.g., requesting a banking service that is not supported by the banking system).

Empirically, we evaluate several pre-trained Transformers (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and ELECTRA (Clark et al., 2020)) in the few-shot learning scenario, as well as pre-trained ToD-BERT (Wu et al., 2020) on task-oriented dialog system. The contributions of this paper are two-fold. First, we constructed and released two new datasets for OOS intent detection based on the

single-domain CLINC dataset and the large fine-grained BANKING77 dataset. Second, we reveal several interesting findings through experimental results and analysis: 1) the pre-trained models are much less robust on ID-OOS than on the in-scope and OOD-OOS examples; 2) both ID-OOS and OOD-OOS detections are not well tackled and require further explorations on the scenario of fine-grained few-shot intent detection; and 3) it is surprising that pre-trained models can predict undesirably confident scores even when masking keywords shared among confusing intents.

2 Evaluation Protocol

Task definition We consider a few-shot intent detection system that handles pre-defined K in-scope intents. The task is, given a user utterance text u, to classify u into one of the K classes or to recognize u as OOS (i.e., OOS detection). To evaluate the system, we adopt in-scope accuracy $A_{\text{in}} = C_{\text{in}}/N_{\text{in}}$, and OOS recall $R_{\text{oos}} = C_{\text{oos}}/N_{\text{oos}}$, following Larson et al. (2019) and Zhang et al. (2020). We additionally report OOS precision, $P_{\text{oos}} = C_{\text{oos}}/N'_{\text{oos}}$. C_{in} and C_{oos} are the number of correctly predicted in-scope and out-of-scope examples, respectively; N_{in} and N_{oos} are the total number of the in-scope and out-of-scope examples evaluated, respectively; if an in-scope example is predicted as OOS, it is counted as wrong. N'_{oos} ($\leq N_{\text{in}} + N_{\text{oos}}$) is the number of examples predicted as OOS.

Inference We use a confidence-based method (Hendrycks et al., 2020) to evaluate the five pre-trained Transformers. We compute a hidden vector $h = \text{Encoder}(u) \in \mathbb{R}^{768}$ for u, where Encoder $\in$ {BERT, RoBERTa, ALBERT, ELECTRA, ToD-BERT}, and compute a probability vector $p(y|u) = \text{softmax}(Wh + b) \in \mathbb{R}^K$, where W and b are the model parameters. We first take the class c with the largest value of $p(y = c|u)$, then output c if $p(y = c|u) > \delta$, where $\delta \in [0.0, 1.0]$ is a threshold value, and otherwise we output OOS. δ is tuned by using the development set, so as to maximize $(A_{\text{in}} + R_{\text{oos}})$ averaged across different runs (Zhang et al., 2020).

Training To train the model, we use training examples of the in-scope intents, without using any OOS examples. This is reasonable as it is nontrivial to collect sufficient OOS data to model the large space and distribution of the unpredictable OOS intents (Zhang et al., 2020; Cavalin et al., 2020).

3 Dataset Construction

We describe the two types of OOS (i.e., OOD-OOS and ID-OOS), using the CLINC dataset (Larson et al., 2019) and the fine-grained BANKING77 dataset (Casanueva et al., 2020). The CLINC dataset covers 15 intent classes for each of the 10 different domains, and it also includes OOS examples. We randomly select two domains, i.e., the "Banking" and "Credit cards", out of the ten domains for models evaluation. The BANKING77 dataset is a large fine-grained single banking domain intent dataset with 77 intents, and it initially does not include OOS examples. We use these two datasets since CLINC dataset focuses on the OOS detection task, and we can evaluate models on the large single fine-grained banking domain on BANKING77 dataset.

OOD-OOS We use the initially provided OOS examples of CLINC dataset as OOD-OOS examples for both datasets. To justify our hypothesis that the CLINC's OOS examples can be considered as out of domains, we take 100 OOS examples from the development set, and check whether the examples are related to each domain. Consequently, only 4 examples are relevant to "Banking", while none of them is related to "Credit cards". There are also no overlaps between the added OOS examples and the original BANKING77 dataset. These findings show that most of the OOS examples are not related to the targeted domains, and we cannot effectively evaluate the model's capability to detect OOS intents within the same domain.

ID-OOS Detecting the OOD-OOS examples is important in practice, but we focus more on how the model behaves on ID-OOS examples. For the ID-OOS detection evaluation, we separate 5 intents from the 15 intents in each of the domains and use them as the ID-OOS samples for the CLINC dataset, following the previous work (Shu et al., 2017). In contrast to the previous work that randomly splits datasets, we intentionally design a confusing setting for each domain. More specifically, we select 5 intents that are semantically similar to some of the 10 remaining intents. As for the BANKING77 dataset, we randomly separate 27 intents from the 77 intents and use them as the ID-OOS samples, following the above process.

Table 1 shows the dataset statistics. More details, including released datasets, are in the GitHub.[1]

[1] github.com/aclsubmission2022/arr

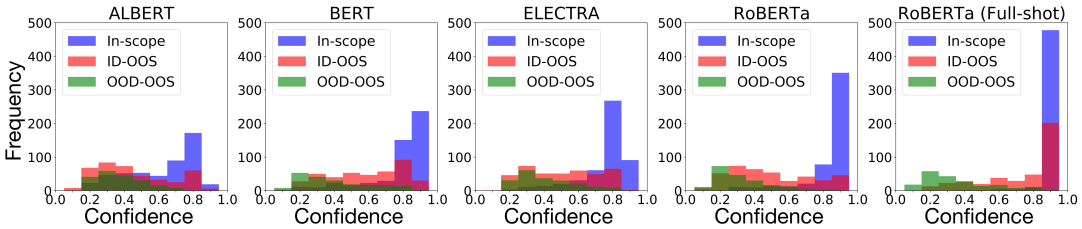

Figure 1: Model confidence on the development set of "Banking" domain in CLINC-Single-Domain-OOS dataset under 5-shot setting. Darker colors indicate overlaps.

CLINC-Single-Domain-OOS	K	Train	Dev.	Test
In-scope	10	500	500	500
ID-OOS	-	-	400	350
OOD-OOS	-	-	200	1000
BANKING77-OOS	K	Train	Dev.	Test
In-scope	50	5905	1506	2000
ID-OOS	-	-	530	1080
OOD-OOS	-	-	200	1000

Table 1: Statistics of CLINC-Single-Domain-OOS and BANKING77-OOS dataset.

4 Empirical Study

4.1 Experimental Setting

We implement all the models following public code from Zhang et al. (2020), based on the HuggingFace Transformers library (Wolf et al., 2019) for the easy reproduction of experiments. For each component related to the five pre-trained models, we use their base configurations, e.g., `roberta-base` for RoBERTa. We take grid search on the hyper-parameters for each experiment and train the model ten times for each hyper-parameter set to select the best threshold δ (introduced in Section 2) on the development set. We apply the best model and threshold to the test set. We mainly conduct the experiments in 5-shot, e.g., five training examples per in-scope intent, and 10-shot; we also report partial results in the full-shot scenario.

4.2 Overall Results

Table 2 shows the results of few-shot intent detection on the test set for 5-shot and 10-shot settings. In both settings, the in-scope accuracy of ID-OOS examples tends to be lower than that of OOD-OOS examples, and the gap becomes larger for OOS recall and precision. It is interesting to see that ToD-BERT, which is pre-trained on several task-oriented dialog datasets, does not perform well in our scenario. The results indicate that the pre-trained models are much less robust on the ID-OOS intent detection. Compared with the results on the two single domains of the CLINC-Single-Domain-OOS dataset, we can find that the performances become much worse on the larger fine-grained BANKING77-OOS dataset. Especially the in-scope accuracy and OOS precision are pretty low, even with more training examples. This finding encourages more attention to be put on fine-grained intent detection with OOS examples.

4.3 Analysis and Discussions

One key to the OOS detection is a clear separation between in-scope and OOS examples in terms of the model confidence score (Zhang et al., 2020). Figure 1 illustrates the differences in confidence score distributions. The confidence scores of ID-OOS examples are close or mixed with the scores of in-scope intents, and are higher than the OOD-OOS examples, showing that separating ID-OOS examples is much harder than separating OOD-OOS examples.

Among comparisons of the pre-trained models, ALBERT performs worst, and RoBERTa performs better than other models in general since the confidence score received by in-scope examples is higher than that received by the OOS examples. Figure 2 also shows similar results. We conjecture that pre-trained models with more data, better architecture and objectives, etc., are relatively more robust to OOD-OOS and ID-OOS examples than the others. Comparing the RoBERTa 5-shot and full-shot confidence distributions, the ID-OOS confidence scores are improved, indicating over-confidence to separate semantically-related intents (i.e., ID-OOS examples).

Next, we inspect what ID-OOS examples are misclassified, and we take RoBERTa as an example as it performs better than other models in general. Figure 3 shows the confusion matrices of RoBERTa w.r.t. the 'banking' domains, under full-shot setting. We can see that the model is extremely likely to confuse ID-OOS intents with particular in-scope intents. We expect this is from our ID-OOS design, and the trend is consistent across evaluated models.

Now one question arises: *what causes the model's mistakes?* One presumable source is the

5-shot		In-scope accuracy			OOS recall			OOS precision		
		Banking	Credit cards	BANKING77-OOS	Banking	Credit cards	BANKING77-OOS	Banking	Credit cards	BANKING77-OOS
ID-OOS	ALBERT	54.1 ± 6.9	55.5 ± 8.1	20.3 ± 2.4	86.3 ± 8.1	75.9 ± 11.2	89.5 ± 1.5	57.9 ± 3.3	55.8 ± 4.3	39.8 ± 0.7
	BERT	75.2 ± 2.9	74.1 ± 4.6	25.4 ± 3.6	81.8 ± 10.5	76.5 ± 9.7	90.9 ± 0.6	70.8 ± 2.5	68.1 ± 3.2	41.3 ± 1.4
	ELECTRA	64.8 ± 4.8	71.0 ± 7.3	30.9 ± 2.3	89.4 ± 4.3	75.8 ± 6.1	87.5 ± 2.4	65.1 ± 3.0	67.1 ± 4.8	43.0 ± 0.8
	RoBERTa	83.8 ± 1.7	64.5 ± 5.6	43.0 ± 2.9	78.4 ± 6.2	86.8 ± 5.4	83.1 ± 4.3	78.6 ± 1.5	63.3 ± 3.4	46.3 ± 1.9
	ToD-BERT	75.1 ± 2.3	67.4 ± 4.2	35.5 ± 1.5	75.8 ± 9.5	72.3 ± 3.4	82.7 ± 1.8	69.4 ± 3.6	61.3 ± 2.3	43.8 ± 0.1
OOD-OOS	ALBERT	63.1 ± 5.7	55.5 ± 8.1	20.3 ± 2.4	85.3 ± 3.4	92.5 ± 4.0	97.3 ± 2.5	83.4 ± 1.7	81.5 ± 3.1	39.9 ± 1.3
	BERT	75.2 ± 2.9	74.1 ± 4.6	39.0 ± 3.1	93.4 ± 3.7	95.5 ± 2.7	94.1 ± 1.6	88.8 ± 1.4	88.4 ± 1.9	49.0 ± 1.8
	ELECTRA	75.5 ± 4.0	71.0 ± 7.3	39.1 ± 2.7	87.3 ± 4.3	87.6 ± 4.2	93.1 ± 4.3	88.8 ± 2.1	87.0 ± 2.7	48.7 ± 1.1
	RoBERTa	83.8 ± 1.7	81.2 ± 4.0	62.1 ± 2.9	97.0 ± 0.9	96.7 ± 1.4	93.9 ± 1.4	92.9 ± 0.6	91.4 ± 1.8	68.7 ± 2.2
	ToD-BERT	83.0 ± 1.6	75.8 ± 5.0	52.9 ± 1.5	91.9 ± 1.0	96.7 ± 0.9	88.4 ± 1.7	92.8 ± 0.6	89.6 ± 2.1	66.0 ± 1.2
10-shot										
ID-OOS	ALBERT	77.8 ± 2.7	66.7 ± 7.8	27.3 ± 3.4	77.6 ± 13.0	79.8 ± 6.4	87.6 ± 1.3	72.2 ± 2.9	64.0 ± 4.1	42.4 ± 1.3
	BERT	77.5 ± 1.7	80.3 ± 3.7	52.5 ± 1.7	87.5 ± 9.2	74.5 ± 6.9	77.3 ± 3.2	73.8 ± 1.7	73.1 ± 3.3	50.8 ± 1.1
	ELECTRA	79.5 ± 2.9	78.0 ± 2.5	40.1 ± 2.7	85.2 ± 9.1	86.5 ± 5.8	84.0 ± 1.7	75.4 ± 2.7	73.3 ± 2.9	46.1 ± 1.1
	RoBERTa	76.6 ± 0.9	81.0 ± 5.5	59.7 ± 1.2	86.4 ± 6.3	83.9 ± 6.9	79.1 ± 1.7	72.7 ± 1.5	75.8 ± 5.2	55.8 ± 1.1
	ToD-BERT	80.7 ± 2.5	80.6 ± 0.9	54.3 ± 1.8	79.5 ± 6.1	70.2 ± 5.9	76.9 ± 2.7	75.4 ± 1.4	71.9 ± 2.6	52.1 ± 1.2
OOD-OOS	ALBERT	77.8 ± 2.7	66.7 ± 7.8	30.5 ± 6.5	90.6 ± 4.0	95.0 ± 3.4	92.7 ± 6.3	89.8 ± 1.0	85.7 ± 2.7	47.1 ± 1.9
	BERT	77.5 ± 1.7	90.1 ± 1.9	64.2 ± 0.5	96.8 ± 1.2	91.1 ± 4.4	91.4 ± 3.2	90.0 ± 0.7	95.5 ± 1.1	68.9 ± 1.0
	ELECTRA	79.5 ± 2.9	88.6 ± 2.1	40.1 ± 2.7	94.8 ± 1.7	89.1 ± 2.2	97.6 ± 1.0	90.7 ± 1.2	94.2 ± 1.1	47.9 ± 1.4
	RoBERTa	89.2 ± 1.3	87.5 ± 3.3	70.3 ± 0.3	95.6 ± 1.0	94.6 ± 2.4	94.0 ± 0.8	95.4 ± 0.5	94.0 ± 1.4	73.3 ± 1.5
	ToD-BERT	86.5 ± 2.6	86.5 ± 0.6	60.6 ± 1.8	96.0 ± 0.5	96.4 ± 0.5	94.9 ± 0.9	94.2 ± 1.2	93.7 ± 0.3	63.3 ± 0.9

Table 2: Testing results on the "Banking" and "Credit cards" domains in CLINC-Single-Domain-OOS and BANKING77-OOS datasets. Note that as the best δ is selected based on $(A_{\text{in}} + R_{\text{oos}})$, the in-scope accuracy could be different in the scenarios of OOD-OOS and ID-OOS (see Figure 2).

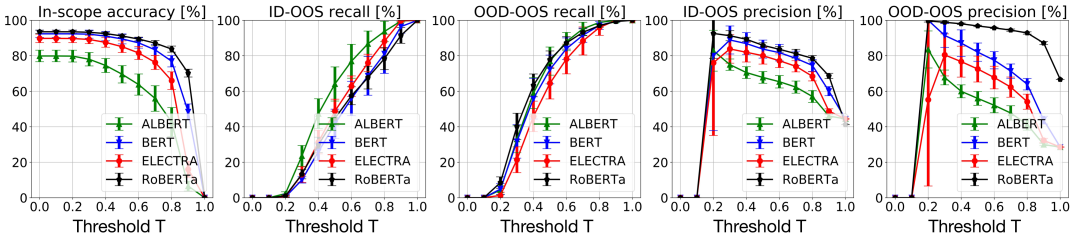

Figure 2: Results on the "Banking" domain in CLINC-Single-Domain-OOS dataset (Dev. set) under 5-shot setting.

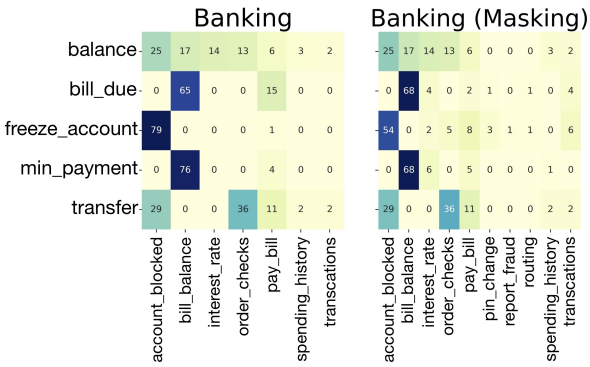

Figure 3: Full-shot confusion matrices on the development set with and without masking ("Banking", RoBERTa). Vertical axis: ID-OOS; horizontal axis: in-scope (only predicted intents considered).

keyword overlap. We checked unigram overlap, after removing stop words, for the intent pairs with the three darkest colors in 'banking' based on Figure 3. We then masked top-5 overlapped unigrams from the corresponding intent examples in the development set using the *mask* token in the RoBERTa masked language model pretraining and conducted the same evaluation.[2] Figure 3 shows that most of the confusing intent pairs are still misclassified even without the keyword overlap. Table 3 shows two intent pairs with the overlapped

[2]We did not mask the top-10 or top-15 overlapped unigrams, as there are already many masked tokens in the user utterance when setting the threshold to 5. See also Table 3.

Intent pair	bill_due & bill_balance
Unigram overlap	bill (60), pay (9), need (9), know (8), due (7)
Masked ID-OOS example	i [mask] to [mask] what day i [mask] to [mask] my water [mask] → bill_balance (confidence: **0.84**)
Intent pair	improve_credit_score & credit_score
Unigram overlap	credit (99), score (76), tell (7), want (3), like (3)
Masked ID-OOS example	i'd [mask] to make my [mask] [mask] better → credit_limit_change (confidence: **0.86**)

Table 3: Examples investigated for the unigram overlap analysis. The overlap frequency is also presented.

words and their masked ID-OOS examples. It is surprising that the examples show counterintuitive results. That is, even with the aggressive masking, the model still tends to assign high confidence scores to some other in-scope intents. We also adopted state-of-the-art methods with contrastive learning on few-shot text classification (Gao et al., 2021) and intent detection (Zhang et al., 2021). However, we did not achieve promising improvements on OOD-OOS and ID-OOS detection, and we leave more explorations to future work.

5 Conclusion

We have investigated the robustness of pre-trained Transformers in few-shot intent detection with OOS samples. Our results on two new constructed datasets show that pre-trained models are not robust on ID-OOS examples. Both the OOS detection tasks are challenging in the scenario of fine-grained intent detection. Our work encourages more attention to be put on the above findings.

References

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Paulo Cavalin, Victor Henrique Alves Ribeiro, Ana Appel, and Claudio Pinhanez. 2020. Improving out-of-scope detection in intent classification by using embeddings of the word graph space of the classes. In *EMNLP*, pages 3952–3961.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and trends® in information retrieval*, 13(2-3):127–298.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *EMNLP*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained Transformers Improve Out-of-Distribution Robustness. In *ACL*, pages 2744–2751.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*.

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*, pages 1311–1316.

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. *arXiv preprint arXiv:1907.11692*.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: Deep Open Classification of Text Documents. In *EMNLP*, pages 2911–2916.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. ToD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogues. *EMNLP*.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. 2021. Unsupervised out-of-domain detection via pre-trained transformers. In *ACL*, pages 1052–1061.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. In *ACL*, pages 3521–3532.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and S Yu Philip. 2021. Few-shot intent detection via contrastive pre-training and fine-tuning. In *EMNLP*, pages 1906–1912.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, S Yu Philip, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *EMNLP*, pages 5064–5082.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2019. Out-of-domain Detection for Natural Language Understanding in Dialog Systems. *arXiv preprint arXiv:1909.03862*.