Benchmarking the Covariate Shift Robustness of Open-world Intent Classification Approaches

Sopan Khosla
AWS AI Labs, Amazon
sopankh@amazon.com

Rashmi Gangadharaiah
AWS AI Labs, Amazon
rgangad@amazon.com

Abstract

Task-oriented dialog systems deployed in real-world applications are often challenged by out-of-distribution queries. These systems should not only reliably detect utterances with unsupported intents (*semantic shift*), but also generalize to *covariate shift* (supported intents from unseen distributions). However, none of the existing benchmarks for open-world intent classification focus on the second aspect, thus only performing a partial evaluation of intent detection techniques. In this work, we propose two new datasets (CLINC14-COV and HWU12-COV) that include utterances useful for evaluating the robustness of open-world models to covariate shift. Along with the i.i.d. test set, both datasets contain a new cov-test set that, along with out-of-scope utterances, contains in-scope utterances sampled from different distributions not seen during training. This setting better mimics the challenges faced in real-world applications. Evaluating several open-world classifiers on the new datasets reveals that models that perform well on the test set struggle to generalize to the cov-test. Our datasets fill an important gap in the field, offering a more realistic evaluation scenario for intent classification in task-oriented dialog systems.

1 Introduction

Open-world classification has been extensively studied in both NLP and CV. Reliably refraining from prediction on samples from out-of-scope labels is of utmost value (Zhang et al., 2021), especially to ensure safety (e.g. autonomous driving) and high quality performance of ML models in production environments. Yang et al. (2021) term this as *semantic shift* detection.

With the advent of voice/text-based task-oriented dialog assistants, it is important to distinguish between supported and unsupported intents to ensure that the classifier does not return garbage when it is barraged with queries from intents it has not been trained on. Several state-of-the-art datasets have been proposed to evaluate the performance of open-world classifiers for intent detection. For example, CLINC (Larson et al., 2019), ROSTD (Schuster et al., 2019; Gangal et al., 2020), HWU64 (Liu et al., 2021), etc.

However, to the best our knowledge, none of the existing benchmarks for intent classification incorporate another fundamental aspect of inference in production. Not only should an open-world classifier reliably handle *semantic shift*, it should also generalize (or be robust) to inference-time *covariate shift* where $P_{train}(y|x) = P_{test}(y|x)$ but $P_{train}(x) \neq P_{test}(x)$ (Shimodaira, 2000; Moreno-Torres et al., 2012; Yang et al., 2021; Wang et al., 2022). In industrial settings, it is common practice for ML systems to be trained on some amount of synthetic data. In general, for most real-world applications the production distribution is often unknown. The classifier, however, is still expected to output correct predictions regardless of this potential shift from what it has observed during training.

In this work, we propose two new English (Bender, 2011) benchmarks, CLINC14-COV and HWU12-COV, that fill this gap by focusing on both semantic and covariate shift to evaluate the performance of intent classifiers. We leverage existing state-of-the-art intent classification datasets to specifically design a test set (cov-test) that, along with out-of-scope utterances, contains in-domain queries generated from a different distribution to the training set. The latter are collected by identifying *equivalence clusters* across different state-of-the-art intent classification datasets. Elements within an equivalence cluster contain intent classes that, despite being sourced from different datasets, share the same underlying intent. In total, CLINC14-COV cov-test contains 420 queries across 14 intents, while the cov-test split in HWU12-COV has 1080 queries across 12 intents.

We evaluate a range of open-world intent classifiers and out-of-scope detection techniques on our

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 14–23
November 20–23, 2022. ©2022 Association for Computational Linguistics

#	Equivalence Clusters
1	rostd:alarm/set_alarm, hwu64:alarm_set, massive:alarm_set, clinc:alarm
2	rostd:alarm/cancel_alarm, hwu64:alarm_remove, massive:alarm_remove
3	rostd:alarm/show_alarms, hwu64:alarm_query, massive:alarm_query
4	rostd:weather/find, hwu64:weather_query, snips:GetWeather, massive:weather_query, clinc:weather
5	hwu64:calendar_query, massive:calendar_query, clinc:calendar
6	hwu64:cooking_recipe, massive:cooking_recipe, clinc:recipe
7	hwu64:datetime_query, massive:datetime_query, clinc:time, clinc:date
8	hwu64:general_repeat, massive:general_repeat, clinc:repeat
9	hwu64:qa_definition, massive:qa_definition, clinc:definition
10	hwu64:takeaway_order, massive:takeaway_order,, clinc:order
11	hwu64:transport_traffic, massive:transport_traffic, clinc:traffic
12	rostd:reminder/show_reminders, clinc:reminder
13	snips:PlayMusic, hwu64:play_music, massive:play_music, clinc:play_music
14	snips:BookRestaurant, clinc:restaurant_reservation
15	snips:AddToPlaylist, clinc:update_playlist
16	banking:declined_card_payment, clinc:card_declined

Table 1: Equivalence clusters (<*dataset:intent*>). Elements within a cluster represent labels that, despite being from different source datasets, share the same underlying intent.

datasets. Our experiments show that all methods perform relatively poorly on the new cov-test sets. In the full-setting, we find a drop in performance of more than 10 absolute F1 and Accuracy points from test to cov-test. We observe a smaller drop for few-shot classification suggesting that such a setting might lead to more robust intent classifiers. We also analyse the affect of covariate shift with and without semantic shift, and find that not only does the existence of both phenomena better mimic production scenarios, it also results in a more challenging setting for classifiers. Our results show that the current models are less reliable when exposed to queries with covariate shift, especially in the open-world setting. We hope that the new datasets will enable future work to fill this gap in the research and development of dialog systems.[1]

2 Dataset

We introduce two new datasets that contain utterances to evaluate the robustness of intent-classifiers to both covariate shift and semantic shift.

2.1 In-Scope Data Collection

To collect in-scope utterances, we leverage the existing state-of-the-art intent-classification datasets including HWU64 (Liu et al., 2021), MASSIVE (FitzGerald et al., 2022), CLINC (Larson et al., 2019), ROSTD (Schuster et al., 2019), SNIPS (Coucke et al., 2018), BANKING (Casanueva et al., 2020) as our starting points.

Dataset	TRAIN	VAL	TEST	COV-TEST
CLINC14-COV	1400	280 (100)	420 (1000)	420 (1000)
HWU12-COV	5055	815 (100)	1028 (1000)	1080 (1000)

Table 2: Data Statistics for our proposed benchmarks – #ID (#OOS) utterances in each split. COV-TEST depicts the newly introduced test set with covariate shift.

Equivalence Clusters. We manually go through the different intents and corresponding utterances in the above-mentioned datasets and define Equivalence Clusters (ECs) as clusters of labels across these datasets that represent similar underlying intents. Overall, we identify 16 such clusters (as shown in Table 1). The nature of these ECs gives rise to a natural covariate shift. Each element in the cluster comes from a different dataset and therefore can be safely assumed to be generated from a dissimilar underlying distribution (examples utterances shown in Table 3). We leverage this property to create our two new benchmarks.

CLINC14-COV. To collect this dataset, we consider the equivalence clusters that contain atleast one CLINC intent. We leverage the CLINC intents in 14 such clusters to build the in-domain training, development, and test set. Rest of the elements in those 14 clusters are used to populate the cov-test set. For example, from Cluster 1, *clinc:alarm* utterances are make up the train/dev/test; whereas utterances from *rostd:alarm/set_alarm, hwu64:alarm_set* are used for cov-test. Finally, to ensure balance among the in-domain classes in CLINC14-COV cov-test, we randomly sample 30 utterances for each intent.[2]

HWU12-COV. We first take the 12 ECs that contain a *hwu64* or *massive* intent. Then, we populate the train/dev/test/cov-test splits using the same procedure as discussed for CLINC14-COV. So, for Cluster 6, *hwu64/massive:cooking_recipe* are considered i.i.d., whereas *clinc:recipe* queries are added to the cov-test set. For cov-test, we randomly sample 90 utterances for each of the 12 intent classes.

2.2 Out-of-Scope Data Collection

We use the existing CLINC OOS samples as out-of-scope data for different splits of our benchmarks.

[1] https://github.com/sopankhosla/cov_shift_intent_datasets

[2] i.i.d. test set also contains 30 utterances per intent class.

EC#	Utterance	Source Dataset : Label
1 (alarm_set)	tomorrow i would like an alarm for 9 tomorrow please add an alarm called "fitness" set an alarm for two hours from now please wake me up after 2 hours	clinc:alarm rostd:alarm/set_alarm hwu64:alarm_set rostd:alarm/set_alarm
4 (weather)	what will the weather be like in samoa at 6 pm Are we expecting snow this week? how does the weather feel should i take my raincoat with me now	snips:GetWeather rostd:weather/find clinc:weather hwu64:weather_query
13 (play_music)	play a song for me play my women of rock playlist please select the first song in my itunes library next play justin bieber's sorry	clinc:play_music snips:PlayMusic hwu64:play_music massive:play_music
16 (card_declined)	The payment for the card did not go through My card payment has been declined how come my credit card isn't working i could not buy food using my card when i was in vietnam	banking:declined_card_payment banking:declined_card_payment clinc:card_declined clinc:card_declined

Table 3: Example utterances from different equivalence clusters (EC).

OOS samples in CLINC test set are also used for the new cov-test. We refer the reader to Larson et al. (2019) for more details. Table 2 provides details on statistics for both benchmark datasets.

3 Benchmark Evaluation

We evaluate a range of open-world intent classification approaches on the new benchmarks.

Unsupervised OOS Detection. The term unsupervised here refers to the absence of OOS samples during training. For this setup, we consider approaches that leverage a confidence-score to distinguish between in-domain and out-of-scope instances. Confidence scores can be calculated using logits like **M**aximum **S**oftmax **P**robability (Hendrycks et al., 2020; Hsu et al., 2020) or distance-based statistics like **Maha**lanobis distance and **Cosine** similarity (Zhou et al., 2021). In addition, we show results for **KNN-C** (Zhou et al., 2022) that uses cosine distance to arrive at a local outlier factor score, and **ADB** (Zhang et al., 2021) that learns adaptive spherical decision boundaries.

Pseudo k+1 OOS Detection strategies focus on generating synthetic out-of-scope samples. These synthetic samples are then included in the training regime of the open-world classifier in a k+1 multi-class classification setup, with k ID and 1 (pseudo) OOS classes. For our experiments, we consider recent algorithms like **ODIST** (Shu et al., 2021) and **DCLOOS** (Zhan et al., 2021) under this umbrella.

4 Experimental Setup

Evaluation Metrics. In line with Shu et al. (2017); Lin and Xu (2019); Khosla and Gangadharaiah (2022), we evaluate the perfromance of the various approaches on accuracy (Acc) and macro F1-score on known classes ($F1_{In}$), open class ($F1_{Out}$), and all classes combined ($F1_{All}$).

Hyperparameters. For a fair comparison, we use the *bert-base-uncased* encoder from HuggingFace for classification with most of the default hyperparameters.[3] We experiment with training batch sizes $\{32, 64, 128\}$. Model with batch size 64 performs the best across all datasets. The learning rate for ID classifier training is set to 2e-5. For ADB, KNN-C, ODIST, and DCLOOS, we use the default hyperparameters in their released code. [4,5]

Threshold Selection. For MSP, Maha, and Cosine we follow Khosla and Gangadharaiah (2022) and extract a random subset from the validation data (VAL-HOLD) for threshold selection. The in-domain classifier is not exposed to this random subset for development. For threshold tuning, we maximize $Acc_{in} + Recall_{out}$ on VAL-HOLD. For other methods, we follow their released source code.

5 Results and Analysis

Here, we present the results of our experiments.

Full Setting Open-world Classification. Table 4 shows the results on full setting open-world classification.[6] The compared state-of-the-art methods see a significant drop in performance from i.i.d. test set to cov-test on both benchmark datasets.

On CLINC14-COV, the models consistently lose 8-12 Acc points, and 18-20 $F1_{All}$ points, a large

[2] Each result is an average of 10 runs with different seeds.

[3] https://huggingface.co/bert-base-uncased

[4] roberta-base results are present in the Appendix.

[5] All experiments are run on a Tesla V100 16GB GPU.

[6] We report the averaged scores on 10 random seeds and the std. dev. values for brevity.

	Performance on TEST				COV-TEST			
	$F1_{All}$	$F1_{In}$	$F1_{Out}$	Acc	$F1_{All}$	$F1_{In}$	$F1_{Out}$	Acc
CLINC14-COV								
MSP	88.5	88.2	93.3	91.2	72.9	71.7	89.0	84.2
Cosine	91.4	91.0	96.1	94.4	71.6	70.2	90.8	86.0
Maha	92.4	92.1	96.7	95.2	**73.2**	**71.9**	**91.4**	**86.9**
ADB	89.1	88.7	94.2	92.1	71.9	70.9	85.8	80.6
KNN-C	**92.5**	**92.2**	96.6	95.2	71.2	69.8	91.0	86.3
ODIST	90.5	90.1	95.6	93.8	72.3	71.0	90.6	85.7
DCLOOS	**92.5**	**92.2**	**97.0**	**95.7**	59.7	57.6	89.3	83.2
HWU12-COV								
MSP	86.9	87.5	80.4	83.1	80.4	80.7	76.6	78.7
Cosine	92.0	92.1	90.4	90.6	**85.4**	**85.3**	**86.6**	**86.6**
Maha	92.3	92.4	90.8	91.1	84.5	84.3	**86.6**	86.5
ADB	88.8	89.1	85.3	86.6	84.7	84.9	82.6	83.7
KNN-C	92.1	92.3	89.8	90.3	84.0	83.9	85.7	85.7
ODIST	90.0	90.1	88.1	88.7	83.2	83.1	84.2	84.3
DCLOOS	**93.6**	**93.7**	**92.9**	**93.0**	80.5	80.1	85.2	84.1

Table 4: Full-setting open-world classification results. Although the compared state-of-the-art methods perform well on i.i.d. test, they struggle to generalize to the new cov-test.

part of which can be attributed to their poor performance on *in-domain* classes (F_{in}) on this set. On HWU12-COV, the drop is smaller yet still significant, with performance ($F1_{All}, Acc$) going down from low 90s to mid 80s. Similar to CLINC14-COV, we see large differences (around 10 points) between the $F1_{in}$ scores on test vs cov-test sets.

Unsupervised vs Pseudo k+1 OOS Detection. We also observe that the unsupervised open-world classification algorithms seem to suffer slightly smaller drops in $F1_{in}$ from test to cov-test as compared to their pseudo k+1 counterparts (Table 4). Although DCLOOS achieves a very high $F1$ on CLINC14-COV i.i.d. test ($F1_{in} = 92.2\%, F1_{All} = 92.5\%$), its performance on cov-test is substantially impacted ($F1_{in} = 57.6\%, F1_{All} = 59.7\%$). This holds true for HWU12-COV as well where DCLOOS scores the lowest $F1_{in}, F1_{All}$ on cov-test amongst all systems studied in this work. Compare this to unsupervised approaches like Cosine and Maha that achieve the highest F_1 scores on cov-test while remaining competitive on i.i.d. test.

Few-shot Classification. Next, we study the impact of covariate-shift in the few-shot setting. Table 5 shows the results for 5, 10-shot classification.

Expectedly, the performance in the few-shot setting on i.i.d. test is lower than what was achieved in the full-setting. This difference is larger on HWU12-COV as compared to CLINC14-COV. It is interesting to see, however, that the drop in Acc and F_1 from test to cov-test is lower than that in the full-setting. The gap is almost non-existent for HWU12-COV. This seems to indicate that the few-shot setting might be more robust to covariate shift as the models do not overfit on the training data.

	Performance on TEST				COV-TEST			
	$F1_{All}$	$F1_{In}$	$F1_{Out}$	Acc	$F1_{All}$	$F1_{In}$	$F1_{Out}$	Acc
CLINC14-COV (5-Shot)								
MSP	79.7	79.1	88.0	84.3	65.8	64.4	85.0	78.6
Cosine	82.2	81.4	92.3	89.0	66.9	65.3	88.9	82.6
Maha	**84.2**	**83.5**	**93.3**	**90.2**	**68.5**	**67.0**	**89.9**	**83.8**
ADB	80.4	79.5	93.1	89.8	57.0	54.7	88.3	81.6
CLINC14-COV (10-Shot)								
MSP	83.6	83.0	91.3	88.1	70.0	68.7	87.8	82.1
Cosine	85.6	85.0	93.6	90.8	69.7	68.2	89.8	84.3
Maha	**86.5**	**86.0**	93.7	90.9	**71.5**	**70.2**	**90.0**	**84.4**
ADB	84.0	83.3	**93.8**	**91.0**	64.8	63.0	89.5	83.8
HWU12-COV (5-Shot)								
MSP	69.2	69.1	70.7	69.4	71.9	72.0	71.7	71.5
Cosine	70.7	70.4	74.4	71.9	72.5	72.3	74.9	73.1
Maha	**73.8**	**73.1**	**81.8**	**77.1**	**73.7**	**73.1**	**81.1**	**77.2**
ADB	63.8	62.8	76.7	70.7	59.4	58.0	75.7	69.9
HWU12-COV (10-Shot)								
MSP	77.3	77.4	75.6	76.3	80.0	80.2	76.5	78.2
Cosine	80.9	80.7	83.6	81.6	81.9	81.7	84.0	82.9
Maha	**83.2**	**82.9**	**86.5**	**84.2**	**82.2**	**81.9**	**85.8**	**84.2**
ADB	77.2	76.8	81.9	78.9	75.2	74.6	81.8	79.3

Table 5: Few-shot classification results for unsupervised open-world classification. The drop in performance from test to cov-test seems to be smaller than that observed in the full-setting. We note that this gap is almost non-existent on HWU12-COV.

Covariate Shift in Open-world Setting. Finally, we also discuss the differential impact of covariate shift in the absence and presence of semantic shift.

On CLINC14-COV (Figure 1), Maha (seed 0) is extremely accurate in its predictions about the ID classes in the presence of semantic shift. But, it classifies some OOS samples incorrectly (Fig. 1 *left*). For covariate shift, we find that in the absence of any semantic shift, the model is robust enough for most intents (*middle*). However, when both phenomena occur together, as is the case in the newly proposed cov-test, model's outputs go awry and it considers several of the ID samples to be OOS (*right*). For example, in the closed-world setting (no open-intent), the model only misclassifies 2 *date* samples from cov-test (*middle*). However, this number goes up to 24 when covariate shift is introduced in the open-world setting (*right*). This seems to be a result of the model's reduced confidence on cov-set ID utterances, ultimately lowering their score below the OOS detection threshold. Open-world classification methods end up introducing tighter conceptual boundaries around each ID class as compared to their closed-world counterparts thus making it easier to confuse ID examples with covariate shift as OOS. We observe a similar trend for other methods, but exclude those results for brevity.

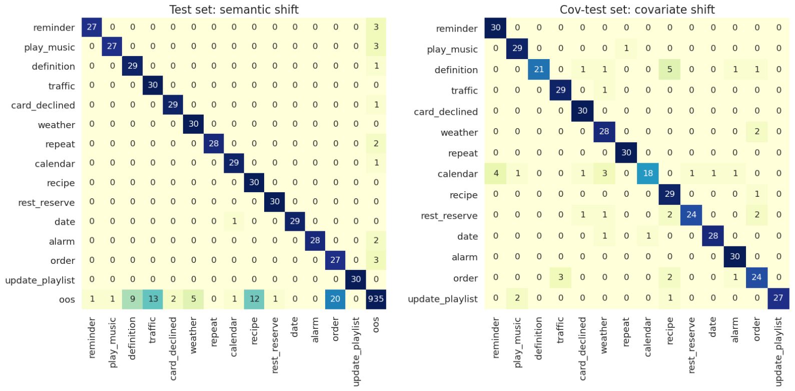

Figure 1: Confusion matrix for Maha (seed 0) on CLINC14-COV with and without covariate and semantic shifts. Covariate shift by itself (middle) does not seem to significantly affect classifier's performance. However, when present along with semantic shift (right), often the case in real-world scenarios, it adversely affects the prediction accuracy (e.g. on intents like *date, calendar*).

6 Related Work

Intent Classification Corpora. There are several available state-of-the-art datasets to evaluate intent classification models (e.g. Larson et al. (2019); Lee et al. (2019); Liu et al. (2021), (Liu et al., 2021; Xu et al., 2015; Casanueva et al., 2020)). While some of these corpora also contain out-of-scope utterances in their test sets, none of them include non i.i.d. in-scope samples. Our new CLINC14-COV and HWU12-COV fill this gap by incorporating such samples in a new cov-test set to evaluate the robustness of intent classification models to both covariate shift and semantic shift.

Evaluating model robustness. Prior works have proposed datasets with adversarial examples to evaluate model robustness. Jia and Liang (2017) show that inserting text can confuse QA systems. Ribeiro et al. (2020) propose a behavioral checklist, an automated test data modification framework to probe model robustness on sentiment analysis and machine comprehension. Whereas, works like Peng et al. (2021); Krone et al. (2021) show that models trained on clean data often struggle to generalize to noisier inputs (e.g. spelling errors, speech disfluencies). In this work, we propose challenge sets that evaluate model robustness to covariate shift. These new benchmarks complement prior art by introducing a new dimension for probing robustness of open-world intent classification systems.

Larson et al. (2020) used crowdsourcing to generate paraphrases of test samples tabooing the use of certain key words. They showed that models trained on the standard datasets struggled on these samples. Although similar in motivation, our benchmark creation approach differ from theirs. Instead of manual paraphrasing, we extract distributionally shifted examples from the equivalent intent classes in the existing state-of-the-art datasets.

Equivalence Clusters. Our notion of *equivalence clusters* is similar to the notion of *collisions* proposed concurrently in Larson and Leach (2022). They introduce the task of intent collision detection when updating the intent classification dataset to incorporate more intents, and show that model performance suffers if new data does not take colliding intents into consideration. On the other hand, we use semantically similar intents in our equivalence clusters to create a challenging test set that evaluates model robustness to covariate shift.

7 Conclusion

In this work, we propose two new benchmark datasets to evaluate open-world intent-classification techniques on their robustness to covariate shift. We leverage previously proposed intent-detection datasets to construct equivalence clusters whose elements represent intent labels that come from different datasets but refer to the same underlying intent class. The nature of these clusters results in a natural covariate shift, as utterances corresponding to each element can be assumed to be generated from a different distribution. These benchmarks test models in the presence of both semantic and covariate shift, a setting that better mimics the challenges faced in real-world production scenarios. We evaluate a range of state-of-the-art open-world classification techniques on our datasets and find that despite their superior performance on i.i.d. test data, they fail to generalize on the covariance test samples. We believe that our datasets and analysis will lead to developing more robust systems for task-oriented dialog.

References

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

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.

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*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7764–7771.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Sopan Khosla and Rashmi Gangadharaiah. 2022. Evaluating the practical utility of confidence-score based techniques for unsupervised open-world classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 18–23.

Jason Krone, Sailik Sengupta, and Saab Mansour. 2021. On the robustness of goal-oriented dialogue systems to real-world noise. In *ICLR 2021 Workshop on Robust and Reliable Machine Learning in the Real World*.

Stefan Larson and Kevin Leach. 2022. Redwood: Using collision detection to grow a large-scale intent classification dataset. *arXiv preprint arXiv:2204.05483*.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K Kummerfeld. 2020. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8097–8106.

Sungjin Lee, Hannes Schulz, Adam Atkinson, Jianfeng Gao, Kaheer Suleman, Layla El Asri, Mahmoud Adada, Minlie Huang, Shikhar Sharma, Wendy Tay, and Xiujun Li. 2019. Multi-domain task-completion dialog challenge. In *Dialog System Technology Challenges 8*.

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.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 165–183. Springer.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021. Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4418–4429.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task 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 3795–3805.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Lei Shu, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2021. Odist: Open world classification via distributionally shifted instances. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3751–3756.

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.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*.

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3521–3532.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14374–14382.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. KNN-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141, Dublin, Ireland. Association for Computational Linguistics.

Appendix

A Label Distribution

In Table A1, we show the label distribution for CLINC14-COV and HWU12-COV. The new cov-test sets contain a uniform distribution for each in-domain intent class, with the 30 and 90 new utterances per intent for CLINC14-COV and HWU12-COV respectively. Train, dev, and test splits for HWU12-COV are not balanced. Please note that this is a property of the original HWU64 corpus.

Dataset	Labels (#train, #dev, #test, #cov-test)
CLINC14-COV	reminder (100, 20, 30, 30), play_music (100, 20, 30, 30), definition (100, 20, 30, 30), traffic (100, 20, 30, 30), card_declined (100, 20, 30, 30), weather (100, 20, 30, 30), repeat (100, 20, 30, 30), calendar (100, 20, 30, 30), recipe (100, 20, 30, 30), restaurant_reserve (100, 20, 30, 30), date (100, 20, 30, 30), alarm (100, 20, 30, 30), order (100, 20, 30, 30), update_playlist (100, 20, 30, 30)
HWU12-COV	alarm_query (288, 36, 53, 90), cooking_recipe (361, 59, 91, 90), qa_definition (425, 71, 76, 90), alarm_remove (174, 24, 32, 90), weather_query (728, 143, 175, 90), play_music (794, 141, 195, 90), datetime_query (501, 81, 107, 90), transport_traffic (272, 38, 34, 90), calendar_query (724, 119, 145, 90), takeaway_order (290, 38, 41, 90), alarm_set (341, 47, 60, 90), general_repeat (157, 18, 19, 90)

Table A1: Label distribution: CLINC14-COV & HWU12-COV.

B Extended Results

The main paper shows results for open-intent classification methods built on top of *bert-base-uncased* encoder. Here, for completion, we also provide the scores achieved by methods which leverage *roberta-base* instead. Table A2 contains results for full-setting and 5,10-shot settings on both new benchmarks. Similar to the trends seen for *bert-base-uncased*, we find that there is a significant drop in F1 and Accuracy from test to cov-test in full setting. For $F1_{All}$ this gap is more than 15 absolute points on CLINC14-COV, and about 10 points on HWU12-COV. In the few-shot scenario, we note that this gap is smaller, and almost non-existent for HWU12-COV(5,10-Shot). Overall, roberta models yield slightly higher scores on both test and cov-test as compared to their bert counterparts.

C Confusion matrices

In Figures A1 and A2, we show the confusion matrices for Maha (seed 0) model on CLINC14 and HWU12 respectively. The four plots depict model's confusion in the presence and absence of covariate and semantic shift. We find that for both datasets, introducing both phenomena together results in the most difficult setting, with CLINC14-COV cov-test being harder than HWU12-COV cov-test.

	Performance on TEST				**COV-TEST**			
	$F1_{All}$	$F1_{In}$	$F1_{Out}$	Acc	$F1_{All}$	$F1_{In}$	$F1_{Out}$	Acc
				CLINC14-COV (Full)				
MSP	90.0	89.6	94.4	92.4	73.1	71.9	89.5	84.7
Energy	90.0	89.7	94.1	92.2	73.2	72.1	89.4	84.6
Cosine	**92.5**	**92.3**	**96.2**	**94.6**	73.2	71.9	**90.6**	**86.0**
Maha	**92.5**	92.2	**96.2**	**94.6**	73.3	72.1	90.5	85.7
ADB	88.8	88.5	93.2	91.0	**73.4**	**72.5**	85.7	80.8
				CLINC14-COV (5-Shot)				
MSP	81.6	81.0	90.1	86.5	66.9	65.5	86.8	80.5
Energy	80.1	79.5	87.9	84.5	66.6	65.2	85.4	78.9
Cosine	86.8	86.2	94.3	91.8	**71.5**	**70.1**	90.8	85.6
Maha	**87.6**	**87.1**	**95.1**	**92.7**	70.9	69.4	**91.3**	**85.9**
ADB	86.1	85.5	94.5	92.0	66.9	65.2	90.0	84.5
				CLINC14-COV (10-Shot)				
MSP	84.2	83.7	91.0	87.9	68.4	67.0	87.1	81.2
Energy	84.0	83.4	91.6	88.5	70.1	68.8	88.3	82.6
Cosine	88.7	88.3	**95.0**	92.8	71.7	70.4	90.6	85.6
Maha	**89.0**	**88.5**	**95.0**	**92.9**	**72.6**	**71.3**	**90.8**	**85.9**
ADB	86.0	85.6	92.5	89.7	70.8	69.5	88.4	83.0
				HWU12-COV (Full)				
MSP	89.4	89.9	83.9	85.7	81.6	81.8	79.8	81.1
Energy	89.9	90.4	85.0	86.5	80.6	80.7	80.2	80.9
Cosine	93.4	93.5	92.3	92.2	83.5	**83.2**	**87.2**	**86.6**
Maha	**93.8**	**93.8**	**92.9**	**92.8**	82.8	82.5	87.1	86.3
ADB	89.5	89.8	86.8	87.9	**84.4**	**84.5**	84.1	84.7
				HWU12-COV (5-Shot)				
MSP	73.3	73.5	71.8	72.6	73.5	73.6	71.5	72.5
Energy	72.9	73.3	68.3	71.5	73.9	74.3	69.3	72.3
Cosine	79.7	79.2	86.2	**83.0**	**78.8**	**78.2**	**86.6**	**82.8**
Maha	**79.9**	**79.3**	**86.5**	**83.0**	77.5	76.8	85.8	81.9
ADB	78.2	77.7	83.5	80.7	73.8	73.0	83.0	79.7
				HWU12-COV (10-Shot)				
MSP	79.0	79.2	76.7	77.7	79.7	80.0	75.6	77.4
Energy	79.2	79.6	75.4	77.3	80.1	80.5	75.0	77.6
Cosine	84.8	84.6	87.0	85.6	**83.7**	**83.5**	86.8	**85.6**
Maha	**85.3**	**85.1**	**88.2**	**86.6**	83.2	82.9	**87.3**	85.4
ADB	82.4	82.3	83.9	82.6	81.2	80.9	83.8	83.0

Table A2: Full-setting and few-shot classification results for unsupervised open-world classification (*roberta-base*).

D Example Predictions

In Table A3, we provide Maha (seed 2) model's predictions on (atmost) five randomly sampled utterances from test and cov-test of CLINC14-COV. As shown, the utterances that are incorrectly classified for intent classes like *definition, alarm, card_declined, rest_reserve*, we do not find linguistic expressions that frequently occur in the correctly classified subset. For example, for *card_declined*, incorrectly classified queries consistently lack any explicit mention of "card". Similarly, for *definition*, most correctly classified utterances use words like "define", "mean" to depict their intent. Whereas, incorrect ones use phrases like "tell me". We observe that the linguistic differences between test and cov-test of *rest_reserve* are more subtle.

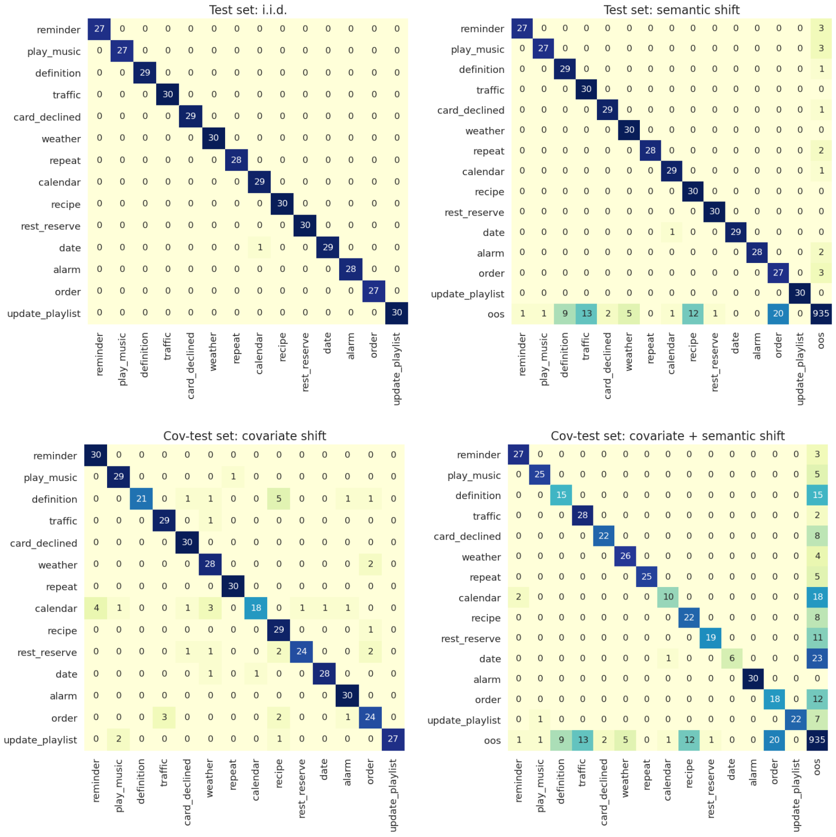

Figure A1: Confusion matrix for Maha (seed 0) on CLINC14-COV with and without covariate and semantic shifts. Covariate shift by itself (bottom left) does not significantly affect classifier's performance. However, when present along with semantic shift (bottom right), often the case in real-world cases, it adversely affects the prediction accuracy (e.g. on intents like *order, calendar*).

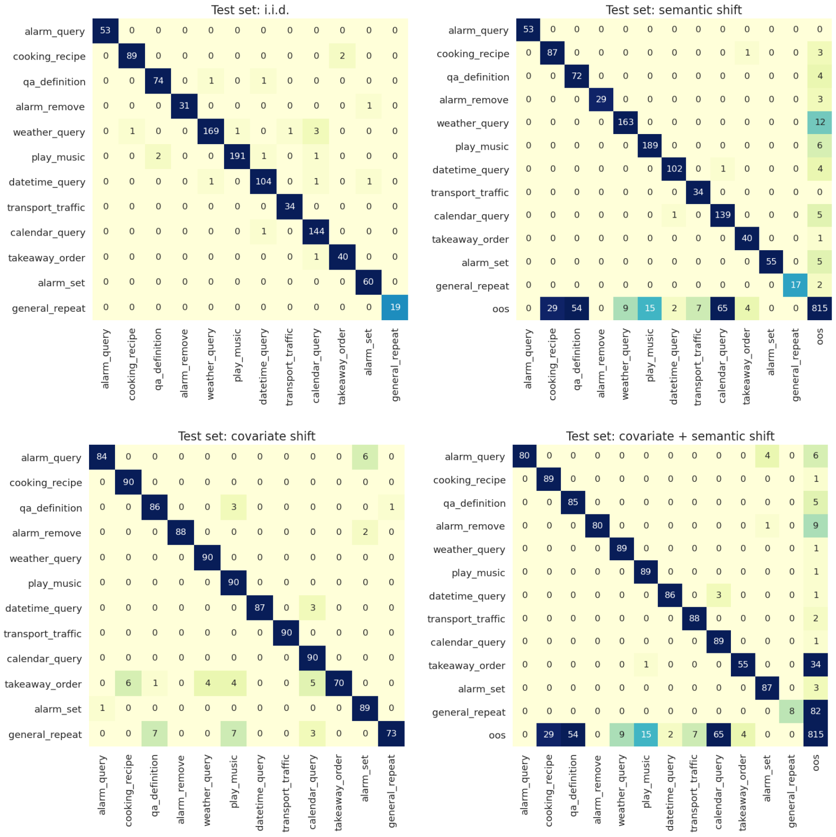

Figure A2: Confusion matrix for Maha (seed 0) on HWU12 with and without covariate and semantic shifts.

	TEST	COV-TEST
definition	what does amicable mean ✔	what are the definitions of orange ✔
	i'd like to know what bitcoin means ✔	what is photosynthesis ✔
	what's the definition of remunerative ✔	spell and define oscillate ✔
	define antebellum ✔	define framework ✔
	can you tell me what dendrofilous means ✔	what is the definition of the word perpetual ✔
	i heard some woman say she was going to yerd me, what's that mean ✗	tell me all about hurricane ✗
		what is computer ✗
		tell me about morel mushrooms ✗
		what is a mango ✗
		give me the description about smartphone ✗
alarm	i need an alarm set now ✔	Reset the alarm for the beginning of the movie tonight ✔
	i'd love to set an alarm ✔	set alarm for 8 am ✔
	set the alarm now ✔	Set a daily alarm for 17h00 ✔
	i would like to have an alarm set for me ✔	Set alarm for 6 am, Mon-Fri ✔
	i need an alarm ✔	please ring the wake up alarm at eight am next saturday ✔
	i need to up by noon ✗	
	can you wake me up at noon ✗	
card_declined	why did i get rejected on my card ✔	My card was not accepted. ✔
	why was my card not accepted ✔	Why has my card payment been declined? ✔
	i was in thailand and i could not use my card to buy snacks ✔	I couldn't pay with card in a shop ✔
	why was my card not working at target ✔	I was trying to purchase something at the store today and my card has been declined. Why has this happened? ✔
	can you tell me why my card got declined ✔	My card payment did not complete. ✔
	how come i got declined ✗	You have declined my payment. ✗
		Why was my Payment declined ✗
		Why are you declining my payment? Everything was fine. ✗
		Why did it decline my payment? ✗
		My latest payment was declined, I was told everything was back to working order. What happened? ✗
rest_reserve	i need a table for two at the havana at nine ✔	make a reservation in a popular sicilian bar place nearby for me only tomorrow ✔
	get me a table for five at itta bena at three ✔	book me a reservation for a party of 3 at a pub in northern mariana islands ✔
	could you reserve table for 3 at carlos jr under the name adam at 4 ✔	book a reservation for an oyster bar ✔
	reserve table for 5 at red robin under the name sara at 3 ✔	table for 8 at a popular food court ✔
	are there any open reservations at outback tonight ✔	i d like a table for midday at the unseen bean ✔
		i want to book a restaurant for my father in law and i in buckner a year from now ✗
		book a table for nine people in svalbard and jan mayen ✗
		i want to book a jewish restaurant in gambia ✗
		book a table at a fried chicken restaurant ✗
		find a restaurant in fm that servec quiche ✗

Table A3: CLINC14-COV: Five random correctly and incorrectly classified examples (Maha; seed 2) across four intent classes in TEST and COV-TEST.