SILVER: Self Data Augmentation for Out-of-Scope Detection in Dialogues

Chunpeng Ma and **Takuya Makino**
Megagon Labs, Tokyo, Recruit Co., LTD.
{ma.chunpeng,makino}@megagon.ai

Abstract

Detecting out-of-scope (OOS) utterances is crucial in task-oriented dialogue systems, but obtaining enough annotated OOS dialogues to train a binary classifier directly is difficult in practice. Existing data augmentation methods generate OOS dialogues automatically, but their performance usually depends on an external corpus. Herein we propose SILVER, a **self** data augmentation method that does not use external data. It improves the accuracy of OOS detection (false positive rate: 90.5% → 47.4%). Furthermore, SILVER successfully generates high-quality **in-domain** (IND) OOS dialogues in terms of naturalness (percentage: 8% → 68%) and OOS correctness (percentage: 74% → 88%), as evaluated by human workers.

1 Introduction

Task-oriented dialogue systems [2, 3] require human operators to deal with intentions that are beyond their capacities, raising the issue of out-of-scope (OOS) detection.

Due to the lack of OOS annotations in open-world settings, previous research usually detects OOS samples **indirectly** resorting to in-scope (INS) samples. Recently, data augmentation methods [4, 5] have made it possible to detect OOS **directly** using a binary classifier. One such method is GOLD [6]. It uses simple rules to replace utterances in known OOS dialogues with sentences selected from a large pool. However, the dependence on external corpora prevented the realization of GOLD's full potential.

We propose a method called **S**elf **I**terative OOS **L**abeling **v**ia **E**nsembling T**r**ees (**SILVER**), overcoming issues of GOLD. It build pools from training data, detects OOS using an ensemble of decision trees [7], and generates OOS dialogues iteratively. Figure 1 compares dialogues generated by GOLD and SILVER.

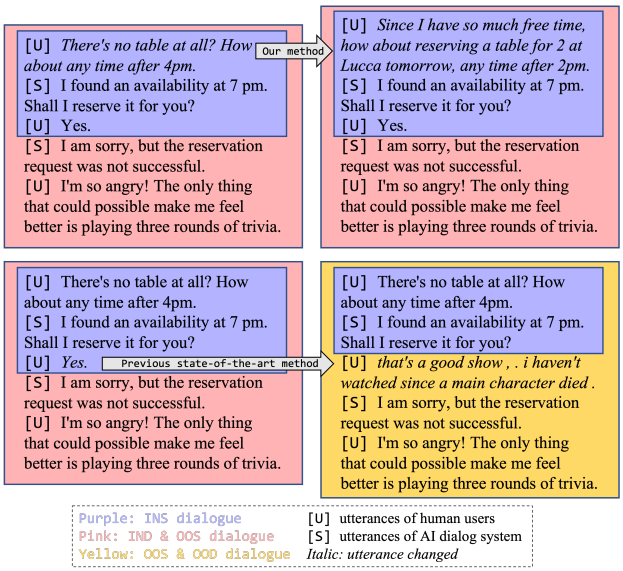

Figure 1 Comparison of GOLD and SILVER. To automatically generate an OOS dialogue, SILVER replaces the first utterance with an IND utterance, while GOLD replaces the third utterance with an OOD utterance, making the dialogue become OOD and incomprehensible.

2 GOLD: Generating Out-of-scope Labels with Data Augmentation

GOLD [6] is the data augmentation method most closely related to this work. Given a small set of annotated OOS dialogues (1% of the size of INS), GOLD replaces utterances with sentences selected from an external pool to generate new OOS dialogues. Selected sentences should be in the neighborhood of the original utterances. Then GOLD filters the generated OOS, and combines predictions of different methods via majority voting. Filtered OOS dialogues are concatenated with the original annotated OOS dialogues and are used to train a binary classifier.

GOLD has a practical appeal. Labor-intensive data col-

This is a shortened version of the paper [1] published in the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics.

lection and annotation of OOS are unnecessary, and the data augmentation method is orthogonal to the classification improvements. Both advantages extend its applicability to real scenarios. However, dependence on external corpora limits its performance.

3 SILVER: Methodology

Figure 2 outlines the framework of SILVER. First, we sample a small set of dialogues from the training data. These dialogues are known OOS. Then candidates are generated by randomly choosing one utterance from seed OOS dialogues and swapping it with an utterance extracted from INS (ref. §3.1). After generating numerous candidates, an ensemble classifier is used for election (ref. §3.2). Selected dialogues are concatenated to seed OOS samples, increasing the number of available OOS dialogues. Iterating this process several times provides sufficient data to train a binary classifier for OOS detection (ref. §3.3).

3.1 Self candidate generation

Candidates are generated by swapping utterances in seed OOS dialogues with those in the pool. To achieve this, two questions must be answered.

How should the utterance pool be built? All utterances of INS dialogues in the training data are used to build the utterance pool because we aim to generate candidates without using external corpora. Furthermore, for a task-oriented dialogue system, we assume that utterances from the user and system are in different clusters. Hence, two pools are built: (1) one for system utterances and (2) one for user utterances.

How should an appropriate utterance be selected? Two criteria are considered to determine appropriate utterances: (1) high similarity to the original utterance and (2) high divergence between each other. Figure 3 illustrates their trade-off.

(1) **High similarity**: By selecting utterances similar to the original one, the naturalness of the original OOS dialogue is kept. This means that the blue points in Figure 3, which were selected by GOLD, will not be selected by SILVER.

(2) **High divergence**: Dialogues generated by simply modifying some words in the original utterances do not improve the classification performance. We hope the generated dialogues differ from each other. This means that selecting the black points in Figure 3 should be avoided.

Therefore, only **appropriate** utterances in the sense of high similarity and high divergence (i.e., green points in Figure 3) should be selected. In practice, utterances are selected from the set $\mathcal{N}(16) - \mathcal{N}(4)$, where $\mathcal{N}(k)$ is the set of k-nearest utterances from the original utterance.

3.2 Tree ensemble

SILVER classifies OOS candidates via the gradient tree boosting algorithm [8]. Feature sets consist of three parts.

(1) Probability-based feature. An intent classifier is trained as the supporting model. Then, given a dialogue d, the supporting model outputs the probability distribution over all intent labels: $[p_1(d), \ldots, p_l(d)]$, where l is the number of possible intent labels, and $\forall i \in [1, l], 0 \leq p_i(d) \leq 1$. Based on this probability distribution, the probability-based feature is calculated as:

$$X_{prob}(d) = [\sigma^{-1}(p_1(d)), \ldots, \sigma^{-1}(p_l(d))], \quad (1)$$

where $\sigma(\cdot)$ is the standard logistic function.

(2) Distance-based feature. Given a dialogue d, the distance-based feature is calculated as below:

$$\begin{aligned} X_{dist}(d) = [&\mathrm{Dist}(h_{\mathrm{BERT}}(d), \overline{h_{\mathrm{BERT}}(\mathcal{D}_1)}), \\ &\ldots, \mathrm{Dist}(h_{\mathrm{BERT}}(d), \overline{h_{\mathrm{BERT}}(\mathcal{D}_l)})], \end{aligned} \quad (2)$$

where $\mathrm{Dist}(\cdot, \cdot)$ is the cosine distance between two vectors, and $h_{\mathrm{BERT}}(d)$ is the representation of the last hidden layer given input d. $\mathcal{D}_i$ is the collection of all dialogues with intent label i, and $\overline{h_{\mathrm{BERT}}(\mathcal{D}_i)}$ is their average.

(3) Ensemble-based feature. This is the average of the output probability distributions of three different runs by randomly dropping out different nodes of the baseline intent classifier, which is given as:

$$X_{drop}(d) = \left[\frac{1}{3}\sum_{k=1}^{3}\sigma^{-1}(p_1^k(d)), \ldots, \frac{1}{3}\sum_{k=1}^{3}\sigma^{-1}(p_l^k(d))\right], \quad (3)$$

where $p_i^k(d)$ is the probability of intent label i at the k-th run given dialogue d, after dropping out some nodes of the neural network. The dropout percentage is 10%.

The final feature set is the concatenation of X_{prob}, X_{dist} and X_{drop}. It is trained on **sampled** training data. Thus, no extra annotation is needed.

3.3 Iterative data augmentation

SILVER generates sufficient data in an iterative manner. After each iteration, newly generated dialogues are aggre-

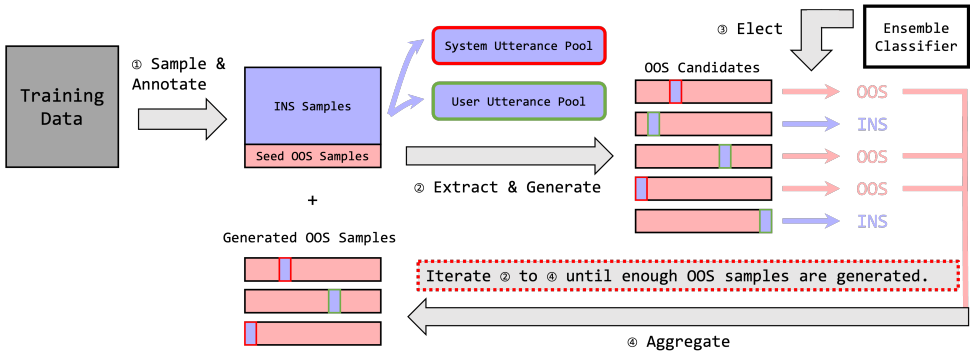

Figure 2 Framework of SILVER.

Modules			STAR				FLOW				ROSTD			
Pool	Elect	Iter.	AUROC↑	AUPR↑	FPR@.95↓	FPR@.90↓	AUROC↑	AUPR↑	FPR@.95↓	FPR@.90↓	AUROC↑	AUPR↑	FPR@.95↓	FPR@.90↓
Ext.	Rnd.	✗	0.7827	0.3618	90.5%	77.8%	0.6692	0.1503	89.1%	80.3%	0.9918	0.9224	1.96%	1.28%
Ext.	MV	✗	0.8456	0.4501	75.4%	59.7%	0.7111	0.1789	83.7%	76.4%	0.9967	0.9613	0.30%	0.30%
Ext.	TE	✗	0.8632	0.4721	63.9%	48.4%	0.7287	0.2183	79.2%	72.1%	0.9985	0.9805	0.15%	0.13%
Ext.	TE	✓	0.8858	**0.4906**	56.9%	38.3%	0.7373	0.2299	76.9%	69.3%	**0.9991**	**0.9910**	**0.09%**	**0.09%**
Int.	Rnd.	✗	0.7843	0.2623	82.7%	74.3%	0.7825	0.2608	79.7%	72.4%	0.8594	0.2967	31.7%	15.0%
Int.	MV	✗	0.8363	0.3618	71.5%	49.3%	0.8030	0.2995	71.3%	60.6%	0.9966	0.9744	0.25%	0.09%
Int.	TE	✗	0.8643	0.3952	60.7%	40.4%	0.8215	0.3368	56.9%	45.1%	0.9971	0.9680	0.22%	0.13%
Int.	TE	✓	**0.8992**	0.4212	**47.4%**	**33.9%**	**0.8319**	**0.3379**	**55.0%**	**43.9%**	0.9971	0.9680	0.22%	0.13%
	GOLD		0.8683	0.4450	56.0%	40.9%	0.8022	0.3243	60.6%	49.5%	0.9990	0.9933	0.17%	0.09%

Table 1 Results of OOS detection. Column "Pool" means whether **external** (Ext.) data (PersonaChat [9]) or **internal** (Int.) data (i.e., training data) is used to generate utterance pool. Column "Elect" gives different election methods: random selection (Rnd.), majority voting (MV), or tree ensemble (TE). Column "Iter." indicates whether dialogues are generated iteratively (✓) or not (✗). Therefore, Int. + TE + ✓means all components of SILVER are applied. **Best** and runner-up of different configurations are denoted by **bold** and underlined texts, respectively. Last line is copied from [6].

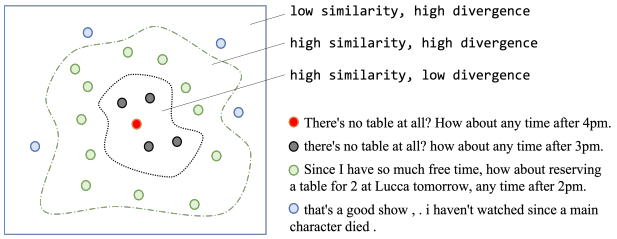

Figure 3 Trade-off between similarity and divergence when selecting appropriate utterances from an utterance pool.

gated and considered known OOS dialogues. Then these are used to generate more dialogues in the next iteration. This iterative process generates **high-quality** dialogues with **high efficiency**.

High quality. The candidate list is kept small, and contains only appropriate (i.e., high similarity & divergence) dialogues, which are rarely INS dialogues. When combined with a powerful ensemble classifier, the generated dialogues have a satisfactory quality.

High efficiency. Because INS dialogues rarely exist in the candidate list, many generated dialogues remain after election. Consequently, the number of available OOS dialogues increases rapidly, reaching the target number in only a few iterations.

4 Experiments

4.1 Datasets and configurations

We conducted experiments on STAR [10], FLOW [11] and ROSTD [12] data.

The supporting model was a classifier finetuned on the task of intent classification, consisting of a pretrained BERT model[1)], with two feed-forward layers above. The model inputs were the first 256 words of the dialogues. The model was optimized using the Adam algorithm [13].

We forced the size of seed OOS dialogues to be 1% of INS, and the target number of generated OOS dialogues was 24 times the seed size.

Experiment results were evaluated using the following metrics: (1) AUROC, area under receiver operating characteristic curve, (2) AUPR, area under precision-recall curve, and (3) FPR@θ, false positive rate with threshold θ.

1) https://huggingface.co/bert-base-uncased

Method	OOS	Naturalness
GOLD	74%	8%
SILVER	88%	68%

Table 2 Human evaluation results of 50 dialogues generated by GOLD and SILVER. Numbers are the percentages of real OOS/natural dialogues.

	# Unique utterances in generated data	#Unique utterances in utterance pool
GOLD	4,153	93,472
SILVER	5,289	32,320

Table 3 Numbers of unique utterances.

4.2 Results on OOS detection

Table 1 shows the key experiment results of SILVER for OOS detection. To verify the effectiveness of the three key components of SILVER corresponding to §3.1 to §3.3, we modified or removed some of these components. We can see that except for special cases (e.g., length-1 dialogues in ROSTD), combining all three components of SILVER achieves the best performance for OOS detection.

4.3 Evaluation of data quality

Next, we analyzed the quality of the generated data. It should be emphasized that the quality of generated dialogues is evaluated **intrinsically** not extrinsically. Specifically, we focus on evaluating (1) the quality of the generated dialogue **itself** and (2) the generated data **as a whole**. Herein the gain of the classification performance contributed by generated data is **not** considered. Although **intrinsic** high-quality does not necessarily contribute to extrinsic tasks directly, it is indispensable in practice.

To evaluate the quality of generated dialogue **itself**, we evaluate whether the generated dialogues are (1) OOS and (2) natural. The second row of Table 2 shows the human evaluation results of dialogues generated by SILVER. Compared with the first row, SILVER outperformed GOLD on both evaluation metrics.

To evaluate the quality of generated data **as a whole**, we compared the generated data of GOLD and SILVER with the original data, resulted in the following observations.

SILVER-generated data has a larger diversity. It is possible that one utterance was selected twice during generation. This reduces the diversity of the generated data. Table 3 shows the numbers of unique utterances in the generated data and the utterance pools. Although SILVER utilized a much smaller pool (built on training data), the generated data contains more unique utterances, indicating a larger diversity.

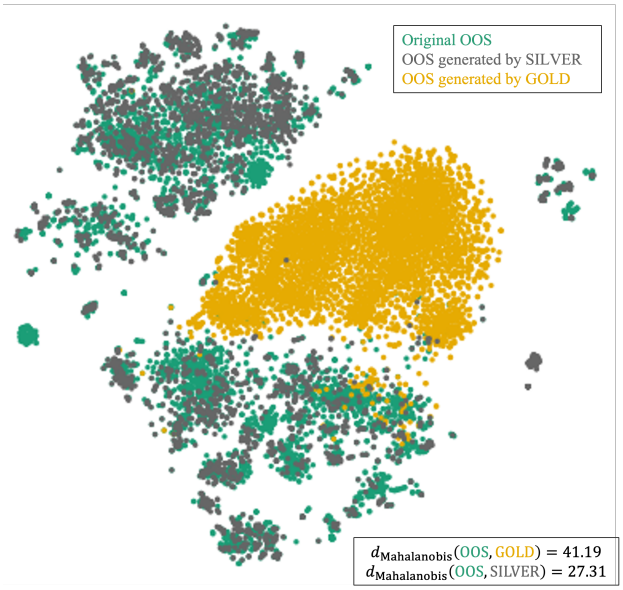

Figure 4 Visualization of the original and generated OOS dialogues.

SILVER generated IND OOS data. An advantage of SILVER is its ability to generate INS OOS dialogues. We calculated the representations of the original OOS dialogues and the dialogues generated by GOLD or SILVER using vanilla RoBERTa [14]. Figure 4 shows the 2-dim t-SNE visualization [15] of these representations along with the average Mahalanobis distances between clusters. The OOS dialogues generated by GOLD differed from the original ones, indicating that these dialogues are OOD. In contrast, the overlap between the original OOS and SILVER-generated dialogues is large, implying that SILVER generates IND data.

5 Conclusion

We proposed SILVER to generate OOS dialogues without using external data. The components in SILVER are designed to overcome issues and realize the full potential of state-of-the-art augmentation methods. Using only training data, SILVER successfully generated high-quality IND OOS dialogues, which not only contributed to the improved performance of **extrinsic** tasks such as OOS detection, but are also natural enough **intrinsically**, indicating the potential for future applications.

References

[1] Chunpeng Ma and Takuya Makino. Silver: Self data augmentation for out-of-scope detection in dialogues. In **Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics**, pp. 26–38, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.

[2] Pawe l Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[3] Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. SalesBot: Transitioning from chit-chat to task-oriented dialogues. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6143–6158, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[4] Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1268–1283, Online, November 2020. Association for Computational Linguistics.

[5] Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. Data augmentation and learned layer aggregation for improved multilingual language understanding in dialogue. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2017–2033, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Derek Chen and Zhou Yu. GOLD: Improving out-of-scope detection in dialogues using data augmentation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 429–442, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. **Advances in neural information processing systems**, Vol. 12, , 1999.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**, pp. 785–794, 2016.

[9] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[10] Johannes EM Mosig, Shikib Mehri, and Thomas Kober. Star: A schema-guided dialog dataset for transfer learning. **arXiv preprint arXiv:2010.11853**, 2020.

[11] Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. Task-Oriented Dialogue as Dataflow Synthesis. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 556–571, 09 2020.

[12] Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 7764–7771, Apr. 2020.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of machine learning research**, Vol. 9, No. 11, 2008.

A Details of Datasets

Throughout this paper, we conducted experiments on three datasets: STAR [10], FLOW [11] and ROSTD [12]. Statistics of each dataset is shown in Table 4. All these datasets consist of task-oriented dialogues. Each dialogue consists of one or several utterances between human and chatbot/human. All utterances are in English.

split	STAR	FLOW	ROSTD
train	22,051/1,248	60,119/4,499	30,521/3,200
dev	2,751/178	3,239/228	4,181/453
test	2,708/168	3,227/239	8,621/937

Table 4 Numbers of INS/OOS dialogues in each dataset.

All these datasets follow the MIT license. Copyrights belong to their creators. Our use of these datasets was consistent with their intended use, i.e., for the research on dialogues of natural languages. All datasets are sufficiently anonymized to make identification of individuals impossible. We randomly sampled 100 dialogues, and asked human workers to check these dialogues. We found that these dialogues do **not** contain any information that names or uniquely identifies individual people, and do **not** contain offensive content.

B Details of Experiments

We have reported key configurations in Section 4.1. In this section, we report more details of experiments to reproduce reported results.

All experiments were conducted on Google Cloud Platform.[2] The instance used for experiments contains one GPU (Nvidia T4).

For data augmentation, we implement the tree ensemble module using XGBoost library.[3] Grid search is used for searching the best hyperparameters for tree ensemble. For STAR dataset, if we use all three types of features, it takes about 70 minutes for tree ensemble.

After data augmentation, we train a binary classifier to detect OOS dialogues. The classifier consists of a `bert-base-uncased` model (109 million parameters), and two feed-forward layers (231 thousand parameters). We resort to libraries (e.g., `pytorchlightning`,[4] `transformers`,[5] etc.) to simplify implementation. For STAR dataset, it takes about 10 minutes for each epoch. We stop training after 13 epochs, and select the model with the largest AUROC on the development data as the final model.

For evaluation, we resort to `scikit-learn` library.[6] Specifically, we use `roc_auc_score` and `average_precision_score` functions to calculate AUROC and AUPR, respectively. Calculation of FPR@θ is implemented by ourselves, by simply combining `roc_curve` function with binary search.

C Details of Human Annotation

To evaluate the quality of automatically generated dialogues, we randomly sampled 50 dialogues and ask human workers to check them manually. Human evaluation results have been reported in Table 2. In this section, we report more details of human annotation.

For OOS correctness, we gave annotators the following instruction.

> Is this dialogue really out-of-scope? For example, the chatbot can only deal with hotel reservation, but the customer asks today's weather. Another example is that the customer becomes angry because the chatbot cannot understand his/her intention.

For naturalness, we gave annotators the following instruction.

> Does the replaced utterance make the whole dialogue strange? Specifically, if the dialogue remains to be natural after replacing by the new utterance, then label this dialogue as "natural," otherwise, label this dialogue as "unnatural."

All annotators are full-time employees affiliated in the same team as the authors. They all have high levels of English proficiency, and are able to annotate dialogues correctly. Annotation was done in an in-house environment, and all dialogues are used only for the purpose of research. After annotation, no ethics issues were reported.

2) https://console.cloud.google.com/
3) https://xgboost.ai
4) https://www.pytorchlightning.ai
5) https://github.com/huggingface/transformers
6) https://scikit-learn.org/