Selective In-Context Data Augmentation for Intent Detection using Pointwise V-Information

Yen-Ting Lin* **Alexandros Papangelis†** **Seokhwan Kim†** **Sungjin Lee†**
Devamanyu Hazarika† **Mahdi Namazifar†** **Di Jin†** **Yang Liu†** **Dilek Hakkani-Tur†**
*National Taiwan University
†Amazon Alexa AI
ytl@ieee.org {papangea,seokhwk,sungjinl,dvhaz,mahdinam,djinamzn,yangliud,hakkanit}@amazon.com

Abstract

This work focuses on in-context data augmentation for intent detection. Having found that augmentation via in-context prompting of large pre-trained language models (PLMs) alone does not improve performance, we introduce a novel approach based on PLMs and pointwise V-information (PVI), a metric that can measure the usefulness of a datapoint for training a model. Our method first fine-tunes a PLM on a small seed of training data and then synthesizes new datapoints – utterances that correspond to given intents. It then employs intent-aware filtering, based on PVI, to remove datapoints that are not helpful to the downstream intent classifier. Our method is thus able to leverage the expressive power of large language models to produce diverse training data. Empirical results demonstrate that our method can produce synthetic training data that achieve state-of-the-art performance on three challenging intent detection datasets under few-shot settings (1.28% absolute improvement in 5-shot and 1.18% absolute in 10-shot, on average) and perform on par with the state-of-the-art in full-shot settings (within 0.01% absolute, on average).

1 Introduction

Intent detection, defined as the identification of a user's intent given an utterance, is a fundamental element in task-oriented dialogue systems, usually occurring within the Natural Language Understanding (NLU) component. One of the practical challenges of training and deploying NLU modules is data scarcity, due to various reasons, such as under-represented languages, privacy and ethical concerns, or simply the cost of collecting and annotating sufficiently large amounts of data for new intents. Consequently, accurately identifying intents in limited-resource scenarios has drawn attention from the community (Papangelis et al., 2021; Mehri and Eric, 2021; Zhang et al., 2021b, for example).

There are three main families of approaches that address the challenge of limited data for intent detection: data augmentation (Peng et al., 2021; Li et al., 2021), focusing on generating high-quality synthetic training and evaluation data; few-shot learning (Zhang et al., 2020, 2021b), focusing on creating learning algorithms that can cope with limited amounts of data; and transfer learning (Namazifar et al., 2021), focusing on learning algorithms that can generalize across domains (therefore not requiring in-domain data). In this work, we follow the data augmentation approach, which is a general method that attempts to augment a human-authored dataset with a large set of synthetically-generated instances. Most recent work has suggested using Pre-trained Language Models (PLMs) for data augmentations under various setups, e.g., (Peng et al., 2021), showing great improvements in performance. However, simply generating a large number of synthetic data points is not enough; we need to consider the quality of each data point, i.e., how beneficial it would be to the model's performance if that synthetic data point is added to the training set. This is an important issue since the model might learn to overfit to synthetic datapoints (which may be low quality, represent specific use cases, etc.) and thus under-perform on real data.

In this work, we propose to apply Pointwise $\mathcal{V}$-Information (PVI) (Ethayarajh et al., 2022) for data augmentation, in a way that leverages a PLM to generate synthetic examples that are relevant and beneficial for training the downstream model, which in our case is an intent classifier. Our contributions are as follows:

- We propose a novel filtering method based on PVI (Ethayarajh et al., 2022) to filter out examples that are not relevant or helpful to the desired intent.
- We conduct experiments on three challenging intent detection datasets and show that our

* Work done during internship at Amazon Alexa AI

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1463–1476
May 2-6, 2023

method achieves state-of-the-art performance.

- We conduct an in-depth study and present a comprehensive analysis of the factors that influence performance, including ablation studies and comparisons with alternative methods.

The rest of the paper is organized as follows: In Section 2 we present relevant work and in Section 3 we introduce our method. In sections 4 and 5 we discuss training details, experiments, and results. In section 6, we present our analysis and discuss alternative approaches we investigated. In section 7 we conclude, and in the following sections we discuss limitations and ethical considerations.

2 Related Work

Intent Detection Intent detection is the task of identifying the user's intent by mapping the user's natural language utterance into one of several predefined classes (Hemphill et al., 1990; Coucke et al., 2018). It is a critical component in the pipeline of task-oriented dialogue systems, as it is used to determine the user's goal and to trigger an appropriate system action (Raux et al., 2005; Young et al., 2013). Several datasets have been proposed to evaluate the performance of intent detection models (Casanueva et al., 2020; Liu et al., 2019a; Larson et al., 2019, for some recent examples). With the availability of such datasets, intent detection has been extensively studied in the literature. Recently, pre-trained language models (e.g., BERT (Devlin et al., 2019)) have been shown to be effective in intent detection (Bunk et al., 2020; Zhang et al., 2020, 2021a,b; Mehri and Eric, 2021).

Data Augmentation Data augmentation is a widely-used technique to address the problem of data scarcity. Paraphrasing the data is one of the ways frequently used for augmentation and can produce more diverse synthetic text with different word choices and sentence structures while preserving the meaning of the original text. Paraphrasing methods have been shown to be effective in many natural language processing tasks (Gupta et al., 2018; Edunov et al., 2018; Iyyer et al., 2018; Wei and Zou, 2019; Cai et al., 2020; Okur et al., 2022; Panda et al., 2021; Jolly et al., 2020). However, such methods often fail to generate more challenging and semantically diverse sentences that are important for the robustness of the downstream models.

Recently, conditional generation – using a PLM to produce text conditioned on some label – has become the dominant paradigm of data augmentation (Bowman et al., 2016; Kumar et al., 2019; Anaby-Tavor et al., 2020; Kumar et al., 2020; Yang et al., 2020a; Lee et al., 2021). This is usually achieved by fine-tuning a language model to produce the original text given the label.

In the field of intent detection, previous work has proposed using data augmentation techniques to generate synthetic training data (Sahu et al., 2022; Papangelis et al., 2021). Sahu et al. (2022) also used PLMs to generate augmented examples, but they require human effort for labeling. This is a challenging task since it is expensive to annotate large amounts of data.

Our approach involves data valuation, similar to the concepts of Ghorbani and Zou (2019); Mindermann et al. (2022). However, our approach differs from such previous work in two key ways. First, Ghorbani and Zou (2019) only evaluated the quality of the training set after training them, whereas we evaluate the synthetic examples before training the task model. Second, Mindermann et al. (2022) selected points that minimize the loss on a holdout set, whereas we select synthetic examples that are reasonably challenging to the task model. Our approach aims to address the problem of data scarcity by evaluating the synthetic examples generated by PLMs and selecting the most valuable examples to augment the training data.

In-context Learning Large language models such as GPT-3 (Brown et al., 2020) and OPT (Zhang et al., 2022) have shown to be able to perform many natural language processing tasks with in-context learning. In this paradigm, the model is provided with a few exemplars based on which it performs the respective task.

In-context learning is a promising solution for few-shot learning. Because of the effectiveness in few-shot performance, in-context learning has been applied to a wide range of NLP tasks. For dialogue tasks, in-context learning has been applied to intent classification (Yu et al., 2021), semantic parsing (Shin and Durme, 2022), and dialogue state tracking (Hu et al., 2022).

However, PLMs require a large amount of computational resources and the limitation on input length restricts the application of PLMs to intent detection tasks with large numbers of intents (e.g., 150 intents in CLINC (Larson et al., 2019)), where

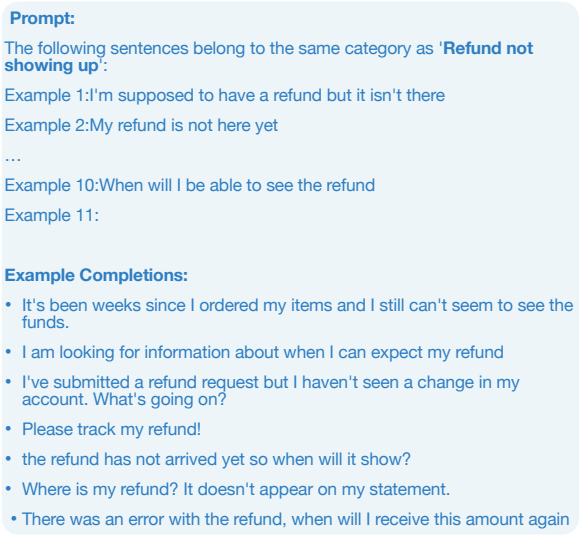

Figure 1: An example of the prompt used to generate synthetic examples. The intent class is *refund not showing up*. Completions are generated by a pre-trained language model via sampling. Note that 5-shot experiments only use 5 examples from the training set.

we cannot fit examples for each intent in the input. One solution would be to call the model multiple times, each time with a subset of the possible intents. This would lead to increased inference time and may also impact performance. Consequently, Yoo et al. (2021); Sahu et al. (2022) leveraged in-context learning and PLMs to generate synthetic examples for intent detection, instead of directly deploying the PLM. However, they did not consider the quality of the generated examples, which may lead to the model overfitting on examples that are not relevant to the desired intent.

3 In-Context Data Augmentation

In the following section, we describe our proposed two-stage method for data augmentation, which we refer to as In-Context Data Augmentation (**ICDA**). The overall procedure is summarized in Algorithm 1. We apply ICDA to the task of few-shot intent detection, which involves classifying a user utterance x into an intent label $y \in Y$. ICDA aims to generate synthetic examples x' such that they would belong to a given intent y.

3.1 Synthesizing Examples

The core idea is to use a large pre-trained language model such as GPT-3 (Brown et al., 2020) or OPT (Zhang et al., 2022) to generate synthetic data in the context of the training set. In particular, for each intent class, we create a natural language context (prompt) that contains the intent class name, a set of real training examples under the same intent class, and an incomplete example. For instance, the prompt for the intent class *refund_not_showing_up* is shown in Figure 1. We feed the prompt to the language model and obtain a set of synthetic examples as outputs. In this work, we use OPT-66B (Zhang et al., 2022) as the language model to generate a set of examples for each intent class. We adopt typical decoding with $\tau = 0.9$ (Meister et al., 2022) and set repetition penalty to 1.1 following Keskar et al. (2019) to generate the synthetic examples.[1] Due to the fine-grained nature of intents, and the sampling-based generation aiming to produce a set of diverse datapoints, we expect some of the generated utterances to not match the given intent.

Note that our method leverages PLMs in a way that is orthogonal to the intent detection model. Unlike other methods that use the same model to directly predict the intent class of a user utterance, we use a PLM to generate synthetic training instances. These instances are then used to augment the actual training data and train a smaller intent detection model. This approach leverages the power of PLMs while preserving the independence of the intent detection model design.

3.2 PVI Filtering

As mentioned above, given the stochastic nature of synthetic data generation, we expect some of the synthetic utterances not to match the given intent. To address this phenomenon, we filter generated instances and retain only those that are relevant and helpful to the desired intent classes.

Specifically, we apply Pointwise V-Information (Ethayarajh et al., 2022) - an idea originally suggested for understanding how difficult a dataset is - as a filter to discard unhelpful datapoints. PVI of an utterance x with respect to its corresponding intent class, y, is defined as:

$$\text{PVI}(x \to y) = -\log_2 g^*[\emptyset](y) + \log_2 g'[x](y)$$

where, in this work, g' and g^* are the intent detection models finetuned with and without the input x, respectively. $\emptyset$ is a special token that is used to indicate the absence of an input utterance.

Intuitively, PVI measures the amount of information that the input x provides to the intent detection

[1] Implementation details are available from https://huggingface.co/docs/transformers/main_classes/text_generation

Algorithm 1: In-Context Data Augmentation with PVI Filtering

Input: Task Model $\mathcal{V}$, Language Model $\mathcal{PLM}$, Data Multiplier m, PVI Threshold Function ϵ

Output: Task Model g

Data: Seed Data $\mathcal{D}_{\text{train}} = \{(\text{input } x_i, \text{gold label } y_i)\}_{i=1}^{n}$

$g' \leftarrow$ Finetune $\mathcal{V}$ on $\mathcal{D}_{\text{train}}$

$\emptyset \leftarrow$ empty string

$g^* \leftarrow$ Finetune $\mathcal{V}$ on $\{(\emptyset, y_i)|(x_i, y_i) \in \mathcal{D}_{\text{train}}\}$

$\mathcal{D}_{\text{synthetic}} \leftarrow \text{Prompt}(\mathcal{PLM}, \mathcal{D}_{\text{train}})$

for $(x_i, y_i) \in \mathcal{D}_{\text{synthetic}}$ **do**

  $\text{PVI}(x_i \rightarrow y_i) \leftarrow -\log_2 g^*[\emptyset](y_i) + \log_2 g'[x_i](y_i)$

$\mathcal{D}_{\text{synthetic}} \leftarrow \{(x_i, y_i)|(x_i, y_i) \in \mathcal{D}_{\text{synthetic}} \ \& \ \text{PVI}(x_i \rightarrow y_i) > \epsilon(y_i)\}$

$g \leftarrow$ Finetune $\mathcal{V}$ on $\mathcal{D}_{\text{train}} \bigcup \mathcal{D}_{\text{synthetic}}$

model (compared to the absence of meaningful input). A high PVI value indicates that the input x provides a lot of information to the model, and thus is more likely to be helpful when training the model to classify instances of the intent class y. On the contrary, a low PVI value indicates that the input x provides little information to the model, and thus is likely to be irrelevant to the intent class y (Ethayarajh et al., 2022).

We set a threshold ϵ (tunable parameter) to determine which x are retained and conduct experiments to study the effect of the threshold in Section 6. Algorithm 1 defines ϵ as a function of y to allow flexibility in its definition: either a fixed threshold for all intent classes, or a different threshold per intent class.

4 Experimental Setup

4.1 Datasets

To evaluate the effectiveness of our approach in intent detection in cases where we have a large number of often semantically similar intent labels, we chose the BANKING (Casanueva et al., 2020), HWU (Liu et al., 2019a), and CLINC (Larson et al., 2019) datasets and compare with recent state-of-the-art baselines. BANKING comprises 13,083 utterances in a single banking domain and 77 intents. HWU includes 25,716 utterances with 64 intents across 21 domains. CLINC contains 23,700 utterances with 150 intents across 20 domains.

	Full-shot mult.	Few-shot mult.
XS	-	1x
S	1x	4x
M	2x	16x
L	4x	64x
XL	-	128x

Table 1: To assess the impact of the synthetic data size on performance, we experiment with several data multipliers (*synthetic data size = source data size x mult.*).

4.2 Training

In our experiments, we use RoBERTa-LARGE (Liu et al., 2019b) as the intent detection model $\mathcal{V}$ in Algorithm 1. We use OPT-66B[2] (Zhang et al., 2022) as the language model $\mathcal{PLM}$ to generate synthetic examples and set the data multiplier m to be 128[3]. We set the PVI threshold function ϵ to be the average PVI under each intent class in the validation set, where the PVI is computed using the same models as in Algorithm 1. We train RoBERTa-LARGE for 40 epochs with a batch size of 16, a learning rate of $1e-5$, and the AdamW optimizer (Loshchilov and Hutter, 2019). We use the HuggingFace Transformers library (Wolf et al., 2020) for all experiments.

4.3 Baseline Models

We compare our proposed method with the following baselines:

RoBERTa-BASE + Classifier is a baseline that uses RoBERTa-BASE (Liu et al., 2019b) with a linear classifier on top (Zhang et al., 2020).

USE is a universal sentence encoder pre-trained on 16 languages supporting multiple down-stream tasks (Yang et al., 2020b).

CONVERT is an intent detection model finetuned from dual encoder models, which is pre-trained on (input, response) pairs from Reddit (Henderson et al., 2020).

CONVBERT fine-tunes BERT on a large open-domain dialogue corpus with 700 million conversations (Mehri et al., 2020) .

CONVBERT + Combined is an intent detection model based on CONVBERT, with example-driven training based on similarity matching and observers for transformer attentions. It also conducts task-adaptive self-supervised learning with masked language modeling (MLM) on the intent detection

[2]We used p3dn.24xlarge AWS EC2 instances for our experiments.

[3]This means that we generate m times the available training data, e.g. (5 x 77) x m in the 5-shot BANKING case.

datasets. Here, “Combined" represents the best MLM+Example+Observers setting in the referenced paper (Mehri and Eric, 2021).

DNNC (Discriminative Nearest-Neighbor Classification) is a discriminative nearest-neighbor model, which finds the best-matched example from the training set through similarity matching. The model conducts data augmentation during training and boosts performance by pre-training on three natural language inference tasks (Zhang et al., 2020).

CPFT (Contrastive Pre-training and Fine-Tuning) is the current state-of-the-art in few-shot intent detection on the selected datasets. It is pre-trained on multiple intent detection datasets in a self-supervised contrastive manner and then fine-tuned with supervised contrastive learning (Zhang et al., 2021b).

5 Experimental Results

We conduct experiments on three benchmark datasets to validate the effectiveness of our proposed method. We first use OPT-66B to generate augmentation examples and then apply our method to enhance a RoBERTa-Large model trained on three datasets. We repeat all experiments with 5 random seeds and report the average performance in Full-shot and Few-shot settings. To investigate the effect of the synthetic data size, we experiment with a variety of multipliers (see Table 1 for notations). Results are shown in Table 2.

Full-shot settings. In this setting, we use the entire training set for each domain. The proposed method achieves the best performance on BANKING and comparable results on HWU and CLINC. In particular, on BANKING, we improve the CONVBERT + Combined baseline (Mehri and Eric, 2021) by 0.59% (absolute) and the RoBERTa-Large baseline by 0.72% (absolute). Compared with the CONVBERT + Combined, which is pre-trained on intent detection datasets in a self-supervised fashion and adds examples-driven training and specific model architectural design, our method achieves similar results with much simpler model design. Furthermore, our method is orthogonal to model architectures and can be integrated with any other approach for further improvement.

We also find that ICDA improves the performance of the RoBERTa-Large model on HWU and CLINC. This highlights the effectiveness of our method for enhancing intent detection models. Moreover, state-of-the-art performance on BANKING with the proposed method and RoBERTa-Large shows that our method is capable of generating high-quality augmentation examples to enhance the RoBERTa-Large model on the most fine-grained intent detection task.

Few-shot settings. In this setting we only use a small number of instances (datapoints) per class. We evaluate our method in both 5-shot and 10-shot settings and compare it with several strong baselines. Our proposed method outperforms all baselines on all datasets in both 5-shot and 10-shot settings. ICDA-M achieves the best performance in 5-shot settings on BANKING dataset and ICDA-XL achieves the best performance on HWU and CLINC datasets in 5-shot settings and on all datasets in 10-shot settings. All configurations of our method significantly improve the performance of a RoBERTa-Large model trained on any of the three datasets. Compared with CPFT (Zhang et al., 2021b), which utilizes contrastive learning for few-shot intent detection with extra data, our method achieves better performance without any additional human-annotated data. This showcases the advantage of our method for few-shot intent detection.

We also observe that our method consistently improves the performance of the baseline model as the number of synthetic datapoints increases from XS to XL. This indicates that the generated instances from our method can gradually cover more and more information of real instances and are capable of providing more useful information for model training.

6 Analysis and Discussion

In this section, we analyze the performance of ICDA and other approaches we tried. We first identify several factors that affect performance, and then present evidence that ICDA works by transferring knowledge from the pretrained generator to the task model. We then discuss a data-relabelling experiment and an experiment using uncertainty measures or data cartography (Swayamdipta et al., 2020) as filters.

6.1 Factors that Affect ICDA Performance

ICDA is effective at various training sizes. Throughout this work, we conduct experiments with different seed data sizes[4] to study the effect of

[4]By *seed data*, we mean data taken from each dataset, i.e. not synthetic data produced by ICDA.

Model	BANKING 5	BANKING 10	BANKING Full	HWU 5	HWU 10	HWU Full	CLINC 5	CLINC 10	CLINC Full
RoBERTa-Base + Classifier	74.04	84.27	-	75.56	82.90	-	87.99	91.55	-
USE	76.29	84.23	92.81	77.79	83.75	91.25	87.82	90.85	95.06
CONVERT	75.32	83.32	93.01	76.95	82.65	91.24	89.22	92.62	97.16
USE+CONVERT	77.75	85.19	93.36	80.01	85.83	92.62	90.49	93.26	97.16
CONVBERT	-	83.63	92.95	-	83.77	90.43	-	92.10	97.07
+ MLM	-	83.99	93.44	-	84.52	92.38	-	92.75	97.11
+ MLM + Example	-	84.09	94.06	-	83.44	92.47	-	92.35	97.11
+ Combined	-	85.95	93.83	-	86.28	**93.03**	-	93.97	**97.31**
DNNC	80.40	86.71	-	80.46	84.72	-	91.02	93.76	-
CPFT	80.86	87.20	-	82.03	87.13	-	92.34	94.18	-
RoBERTa-Large + Classifier	78.99	86.08	93.70	74.44	84.11	92.13	89.89	93.56	96.80
+ **ICDA**-XS	80.29	86.72	-	81.32	85.59	-	91.16	93.71	-
+ **ICDA**-S	81.95	87.37	93.66	81.97	86.25	92.33	91.22	93.98	96.97
+ **ICDA**-M	**84.01***	88.64	93.73	81.84	87.36	92.12	91.93	94.71	97.06
+ **ICDA**-L	83.90	89.12	**94.42***	81.97	86.94	92.57	92.41	94.73	97.12
+ **ICDA**-XL	83.90	**89.79***	-	**82.45***	**87.41***	-	**92.62***	**94.84***	-

Table 2: Intent Detection Accuracy (in %) in few-/full-shot settings with augmented data from OPT-66B. Numbers in bold are the best results and numbers with * are statistically significant by t-test ($p < 0.05$) compared to the baselines (**5 / 10** examples per intent).

	Model	BANKING	HWU	CLINC
	RoBERTa-Large	86.08	84.11	93.56
	All	84.19	84.57	94.24
	All w/ relabeling	87.05	85.22	93.02
PVI	Global Low PVI	73.99	69.61	85.42
PVI	Global High PVI	87.38	86.27	94.27
PVI	Per-Intent Low PVI	76.49	71.84	89.33
PVI	Per-Intent High PVI	**88.64**	**87.36**	**94.71**

Table 3: Intent Detection Accuracy (in %) for RoBERTa-Large model in 10-shot settings with ICDA-M synthetic instances from OPT-66B. Numbers in bold are statistically significant by t-test ($p < 0.05$). "All" represents using all synthetic data without PVI filtering. and "All w/ relabeling" represents using "All" and an oracle intent classifier to relabel the synthetic data.

training size. By looking at the results in Table 2, we observe that our proposed method consistently improves the accuracy of the downstream model in all training sizes. Also, as the training size decreases, we see that the ICDA improvement increases significantly. For example, on BANKING, the improvement goes from 0.72% in the full shot setting to 5.02% as the training size decreases to 5-shot. This indicates that ICDA is more effective when we have few training data available.

PVI filtering threshold. To study the effect of the threshold function ϵ, we conduct experiments with two different threshold functions: *Global*, and *Per-Intent*. *Global* means that the PVI threshold is the same for all intent classes, which is the average PVI value in the validation set. *Per-Intent* means that the PVI threshold is different for each intent class, which is the average PVI value under each intent class in the validation set. As a sanity check, we also conduct experiments using synthetic instances with PVI values lower than the threshold (*Low PVI)* as opposed to the normal (*High PVI)* instances.

We show the results in Table 3 (bottom half), where we see that *Per-Intent High PVI* filtering performs the best. Compared to using all synthetic training data without filtering (referred to as *All*), we see that *High PVI* filtering in general helps in improving accuracy. In BANKING, for example, when PVI filtering is applied with *Per-Intent High PVI*, the accuracy is 88.64% with 10-shot training size, which is significantly better than the result without PVI filtering (84.19%) – the same holds for the other two datasets. For the *Low PVI* conditions, we observe that performance drops significantly. This indicates that the model overfits on those examples that are not relevant to the desired intent. We discuss the *All w/ relabelling* condition in Section 6.3.

In Figure 2, we plot the F1 score against the PVI score of the test set instances grouped by intent, showing that some classes are harder than others, further supporting why we need a threshold per class rather than a global one.

6.2 Why Does ICDA Work?

PVI filtering discards mislabeled examples. We believe that the success of ICDA is because of not only the high diversity of the synthetic instances produced by the generator, but also the fact that PVI filtering effectively discards digressed instances. To verify this hypothesis,

Data	Prompt Label	Generated Sentence	PVI
BANKING	Refund not showing up (PVI Threshold: 5.79)	I didn't see my refund appear on my statement.†	**6.10**
		Where did your refund end up at? Please send it back immediately since I had a return and then refunded your purchase in full and I want it all returned back to my credit card.†	**5.81**
		Can we please cancel my return & resend my goods again	3.97
		Please confirm what is the reason for delay in payment for my purchase?	-3.86
HWU	alarm query (PVI Threshold: 4.28)	show me all the alarms and remind me about them†	**5.69**
		i want to be notified when alarm goes off.†	**4.36**
		how do i delete or disable alarms.	3.18
		list all the events on this date	-5.13
CLINC	accept reservation (PVI Threshold: 6.53)	does hanover steakhouse take reservations†	**6.74**
		are there any restaurants that take reservations for dinner in philadelphia†	**6.58**
		how many days prior is required for making reservations	5.39
		what time does bibiana's in greenwich open	−4.31

Table 4: Synthetic examples generated from OPT-66B. † indicates the sentences that belong to the same intent as the prompt label from our manual assessment; and bold denotes the PVI values over the threshold for given label.

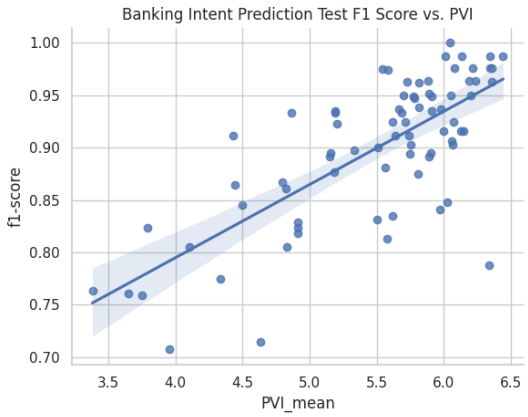

Figure 2: Intent Detection F1 score per intent class (circle) of the BANKING test set, justifying why we need a PVI threshold per intent.

Data	Split	D-1 ↑	D-2 ↑	Self-BLEU ↓	PPL ↓
Bank.	Test	-	-	-	12.14
	10-shot	0.15	0.54	0.24	**17.34**
	ICDA	**0.21**	**0.66**	**0.11**	21.33
HWU	Test	-	-	-	14.84
	10-shot	0.25	0.71	0.07	**26.97**
	ICDA	**0.30**	**0.78**	**0.03**	28.52
CLINC	Test	-	-	-	14.77
	10-shot	0.15	0.49	0.28	**34.23**
	ICDA	**0.20**	**0.60**	**0.17**	37.34

Table 5: Quantitative metrics of fluency and diversity of real and synthetic utterances in 10-shot settings as measured with distinct-1 (D-1), distinct-2 (D-2), self-BLEU, and perplexity.

we randomly sample several synthetic instances from the OPT-66B generator and manually assess if each instance follows the same intent as the prompt label. We show some examples in Table 4. We observe that instances that are relevant to the desired intent are assigned high PVI values, and instances that are not relevant to the desired intent are assigned low PVI values. This further indicates that the per-intent threshold function provides an effective indicator of relevance. For example, in the BANKING dataset, most relevant instances have PVI values greater than 5.79, and most non-relevant instances have PVI values less than 5.79. This indicates that PVI filtering is an effective method for discarding mislabeled data points.

ICDA produces fluent and diverse utterances. We hypothesize that our proposed method is effective because it introduces more fluent and diverse utterances. We therefore compare synthetic data under the 10-shot XS condition (i.e., we generate 10 synthetic datapoints) with the original 10-shot datapoints taken from the training data. Then we use a GPT2 model trained on the test set of each benchmark dataset to calculate the perplexity of the generated utterances. We also use the same synthetic set to calculate the distinct-1, distinct-2, self-BLEU, and perplexity (PPL) metrics. We report the results in Table 5 and observe that our proposed method generates more diverse utterances as shown by distinct-1, distinct-2, and self-BLEU. This indicates that our proposed method harnesses the generation power of the OPT-66B generator. Additionally, the perplexity of synthetic utterances is slightly higher than the human-annotated training set. These results suggest that our proposed method generates more diverse utterances, which can help the task model to learn a better representation.

6.3 Data Relabelling

Following Sahu et al. (2022), we wanted to see if it is effective to use the available data to train an intent classifier and then use it to relabel the synthetic data. Intuitively, such a method would correct mistakes in the generation process. To test the feasibility of this approach, we train an oracle classifier using the entire training data of each dataset and use this as an upper bound. The results are shown in Table 3 ("All w/ relabeling"), where we see that while promising, this approach underperforms ICDA.

7 Conclusion

We introduced **I**n-**C**ontext **D**ata **A**ugmentation, a novel data augmentation framework to generate synthetic training data, preserving quality and diversity. We demonstrate that ICDA is effective on multiple intent detection benchmarks, with state-of-the-art few-shot performance. Our analysis shows that ICDA tends to perform better in low-resource settings and that our PVI filtering strategy is important for performance. Future work includes applying ICDA to other conversational understanding tasks such as slot filling and dialogue state tracking, and incorporating other filtering or data selection strategies for further performance gains.

Limitations

In this section we take BANKING as a case study to motivate PVI and discuss some of the limitations of our approach. Figure 3 shows how much we gain (or lose) in F1 score when we use a custom threshold for each class vs. a fixed threshold. While most classes benefit, there are clearly many that show performance degradation. Another limitation is the size of the model we use to generate synthetic instances (OPT-66B); in general the larger the model is, the better the generated data is.

Ethical Considerations

As with any work involving PLMs (or *foundation models*), due to the data and training methods, there is inherent risk of generating biased, toxic, harmful, or otherwise unwanted output. Regarding our work in particular, as we show in Figure 3, the model's performance on some of the classes can degrade. More analysis needs to be done before deploying our approach, since it is unclear whether it will introduce a bias towards certain types of classes.

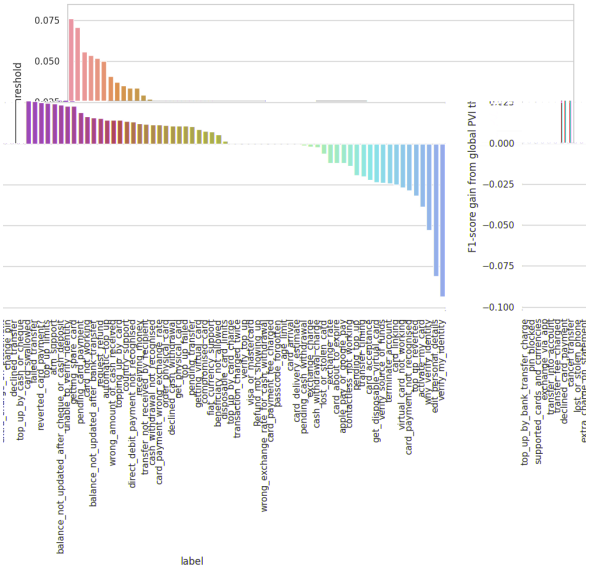

Figure 3: This figure shows the difference in Intent Detection F1 score for each intent, if we have a PVI threshold per-class VS having a fixed PVI threshold. See larger figure in Appendix.

References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7383–7390. AAAI Press.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21. ACL.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. DIET: lightweight language understanding for dialogue systems. *CoRR*, abs/2004.09936.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 746–751. AAAI Press / The MIT Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with *V*-usable information. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Amirata Ghorbani and James Y. Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5149–5156. AAAI Press.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, USA, June 24-27, 1990*. Morgan Kaufmann.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. *CoRR*, abs/2203.08568.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1875–1885. Association for Computational Linguistics.

Shailza Jolly, Tobias Falke, Caglar Tirkaz, and Daniil Sorokin. 2020. Data-efficient paraphrase generation to bootstrap intent classification and slot labeling for new features in task-oriented dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 10–20, Online. International Committee on Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *CoRR*, abs/2003.02245.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the 2nd Workshop on*

Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 1–10, Hong Kong, China. Association for Computational Linguistics.

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.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *CoRR*, abs/2102.01335.

Shiyang Li, Semih Yavuz, Kazuma Hashimoto, Jia Li, Tong Niu, Nazneen Fatema Rajani, Xifeng Yan, Yingbo Zhou, and Caiming Xiong. 2021. Coco: Controllable counterfactuals for evaluating dialogue state trackers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 165–183. Springer.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tong Luo, Kurt Kramer, Dmitry B. Goldgof, Lawrence O. Hall, Scott Samson, Andrew Remsen, and Thomas Hopkins. 2004. Active learning to recognize multiple types of plankton. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004*, pages 478–481. IEEE Computer Society.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2979–2992, Online. Association for Computational Linguistics.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tür. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *CoRR*, abs/2009.13570.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *CoRR*, abs/2202.00666.

Sören Mindermann, Jan Markus Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. 2022. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15630–15649. PMLR.

Mahdi Namazifar, Alexandros Papangelis, Gökhan Tür, and Dilek Hakkani-Tür. 2021. Language model is all you need: Natural language understanding as question answering. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7803–7807. IEEE.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4114–4125, Marseille, France. European Language Resources Association.

Subhadarshi Panda, Caglar Tirkaz, Tobias Falke, and Patrick Lehnen. 2021. Multilingual paraphrase generation for bootstrapping new features in task-oriented dialog systems. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 30–39, Online. Association for Computational Linguistics.

Alexandros Papangelis, Karthik Gopalakrishnan, Aishwarya Padmakumar, Seokhwan Kim, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Generative conversational networks. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–120, Singapore and Online. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2021. Data augmentation for spoken language understanding via pretrained language models. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 1219–1223. ISCA.

Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskénazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 885–888. ISCA.

Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 441–448. Morgan Kaufmann.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *Advances in Intelligent Data Analysis, 4th International Conference, IDA 2001, Cascais, Portugal, September 13-15, 2001, Proceedings*, volume 2189 of *Lecture Notes in Computer Science*, pages 309–318. Springer.

Greg Schohn and David Cohn. 2000. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 839–846. Morgan Kaufmann.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5417–5425. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020a. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020b. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE*, 101(5):1160–1179.

Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. Few-shot intent classification and slot filling with retrieved examples. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 734–749, Online. Association for Computational Linguistics.

Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert Y.S. Lam. 2021a. Effectiveness of pre-training for few-shot intent classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1114–1120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021b. Few-shot intent

detection via contrastive pre-training and fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1906–1912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

A Data Cartography and Uncertainty

Apart from relabelling, we investigated two additional approaches to rank synthetic instances as easy or hard to classify. We used data cartography (Swayamdipta et al., 2020) and classification uncertainty to guide our filtering.

Data cartography classifies the training data in four categories: Easy-to-learn, Low-Correctness, Ambiguous, Hard-to-Learn using training dynamics (i.e. the model's confidence in the true class, and the variability of this confidence across epochs).

For uncertainty modeling, we assign uncertainty scores to each training instance in a *cross-validation* manner. We first split the training set into 5 folds, hold one fold out as validation, and predict on the validation with the classifier trained on the remaining 4 folds. We tried the following uncertainty measures: Contrastive Active Learning (AL) (Margatina et al., 2021), Least Confidence (Culotta and McCallum, 2005), Prediction Entropy (Schohn and Cohn, 2000; Roy and McCallum, 2001), and Breaking Ties (Scheffer et al., 2001; Luo et al., 2004).

We conducted experiments using the above approaches to select data that amounts to one third of the total training data in BANKING (i.e., we select the top 33% hardest examples, etc.). As an additional baseline, we include a random filter, i.e., a randomly sampled 33% portion of BANKING. Table 6 shows the results, where we see that the performance actually degrades when compared to using the entirety of the data. We experimented with a few more variations in the filtering thresholds but no combination improved performance and we do not report those results here. See Figures 5 and 6 in the Appendix B for a visualization of the BANKING data map.

		100% Train	92.89
		Random	89.50
33% Train	Uncertainty	Contrastive AL	88.54
		Least Confidence	89.08
		Breaking Ties	89.20
		Prediction Entropy	89.23
	Cartography	Easy to Learn	90.44
		Ambiguous	90.94
		Low Correctness	91.00
		Hard to Learn	91.26

Table 6: Intent Detection Accuracy (in %) for ConvBERT model, trained on different selections of BANKING77 under full-shot settings.

B Figures

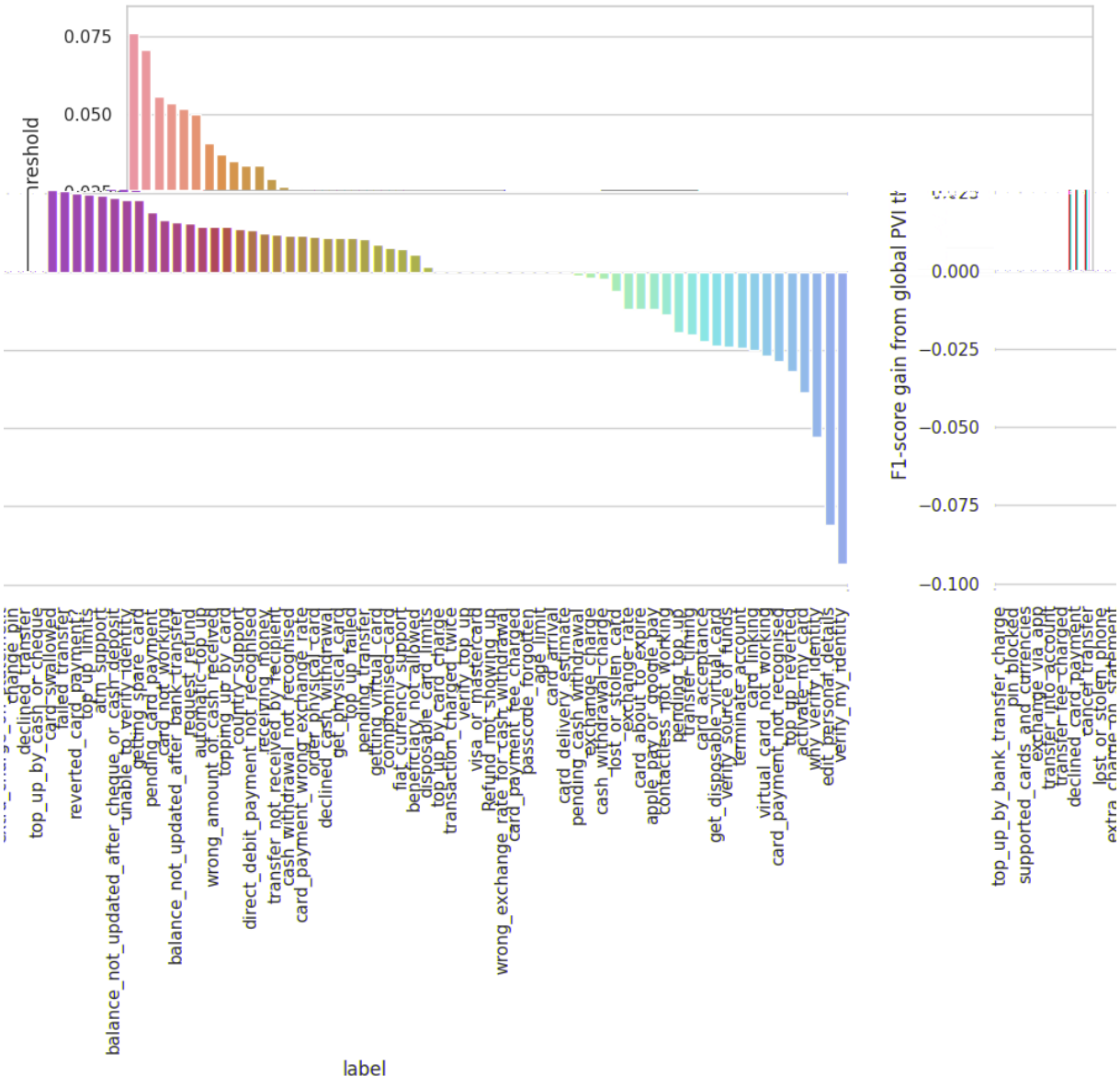

Figure 4: This figure shows the difference in F1 score for each intent, if we have a PVI threshold per-class VS having a fixed PVI threshold (Enlarged Figure 3).

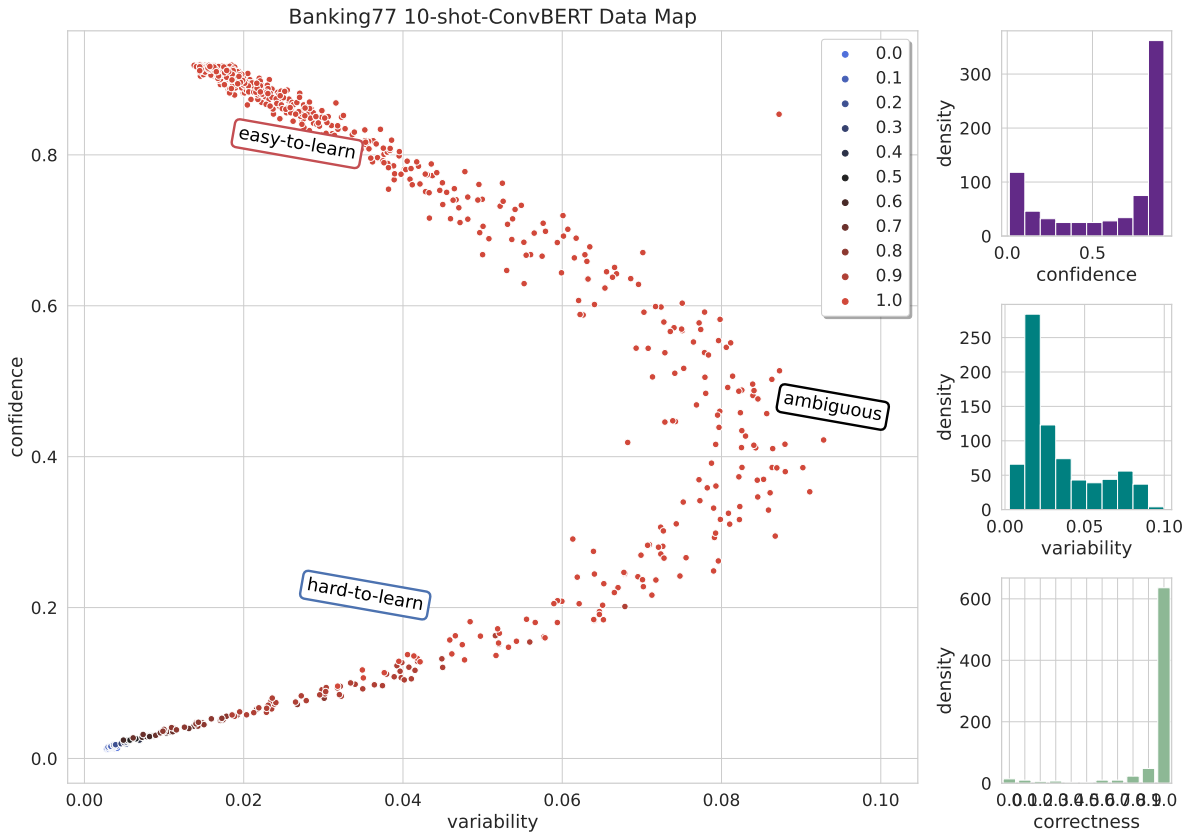

Figure 5: Data map for BANKING (10-shot).

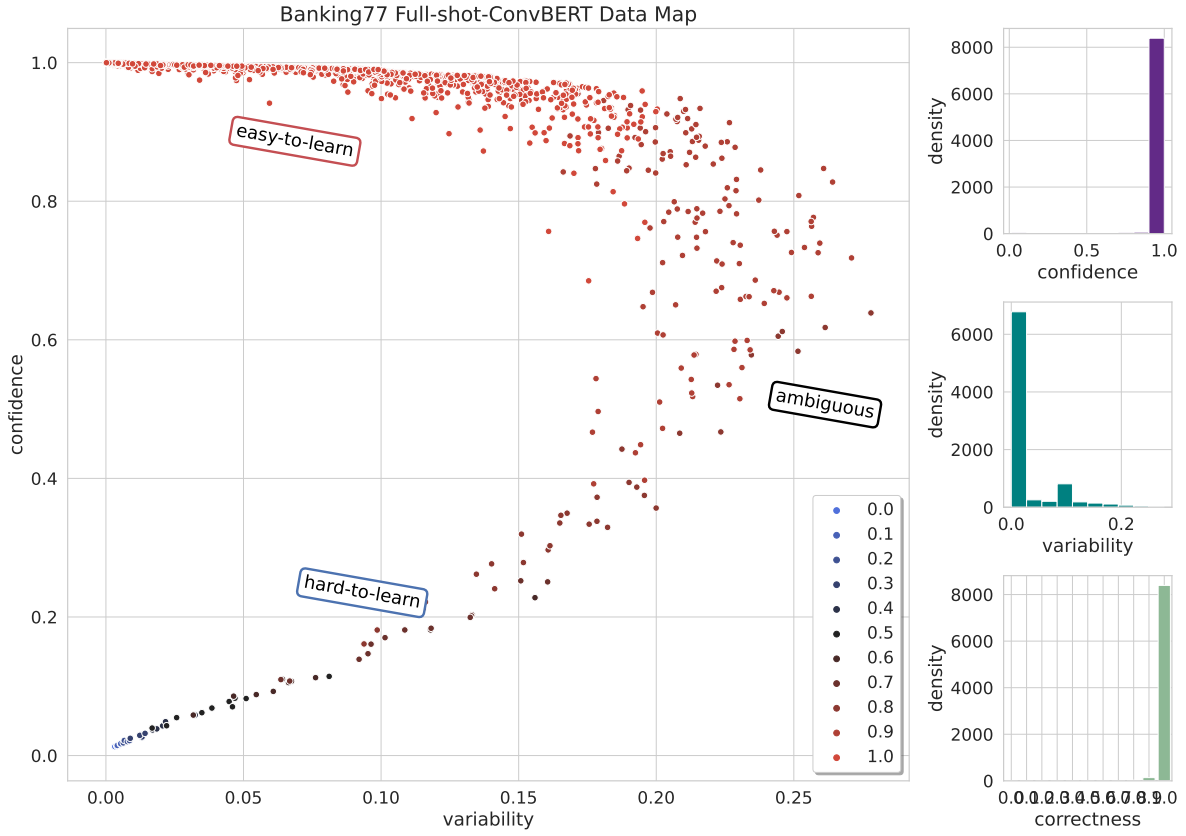

Figure 6: Data map for BANKING (full-shot).