More Diverse Dialogue Datasets via Diversity-Informed Data Collection

Katherine Stasaski[1], Grace Hui Yang[2], and Marti A. Hearst[1]
[1]UC Berkeley
[2]Georgetown University
[1]{katie_stasaski, hearst}@berkeley.edu
[2]huiyang@cs.georgetown.edu

Abstract

Automated generation of conversational dialogue using modern neural architectures has made notable advances. However, these models are known to have a drawback of often producing uninteresting, predictable responses; this is known as the diversity problem. We introduce a new strategy to address this problem, called Diversity-Informed Data Collection.

Unlike prior approaches, which modify model architectures to solve the problem, this method uses dynamically computed corpus-level statistics to determine which conversational participants to collect data from.

Diversity-Informed Data Collection produces significantly more diverse data than baseline data collection methods, and better results on two downstream tasks: emotion classification and dialogue generation. This method is generalizable and can be used with other corpus-level metrics.

1 Introduction

It is well-documented that neural dialogue models struggle with generating engaging, relevant responses (Li et al., 2016a) and often produce banal responses such as "Yeah." While this may be an appropriate response to a chitchat conversation, to keep a human participant engaged, diversity of responses is important. Diverse models vary the language used and the content referenced, and the generated utterances differ from the most typical conversation responses some proportion of the time. A model which only generates "Yeah," "No," and "I don't know" is not diverse and is not be engaging to converse with.

Past work has improved model diversity with innovation on model architectures and decoding strategies (Li et al., 2016a; Baheti et al., 2018; Li et al., 2017; Shao et al., 2017; Cao and Clark, 2017; Serban et al., 2017; Zhao et al., 2017). We build upon this work to propose a novel method to collect and determine more diverse data to train these models with. Our method can be used in conjunction with existing generation-specific model innovations.

Some prior work on data collection processes has prioritized diversity. For instance, Rashkin et al. (2019) prompts crowdworkers to choose an underused emotion class to generate dialogue. This work encourages coverage of emotion classes, but does not consider the likelihood that some crowdworkers are better at producing certain types of data than others.

This paper introduces Diversity-Informed Data Collection (DIDC), a new strategy for creating a dataset of conversational utterances via selecting which participants' data to include in the collection. The strategy progressively builds up a more diverse sub-corpus from an existing larger collection. The main idea is to grow the sub-corpus by adding conversations sequentially and to assess the contribution of a new participant's utterances to the diversity of the entire sub-corpus. This strategy is also applicable to on-the-fly collection of new datasets via crowdworking or similar methods. We implement DIDC with three diversity metrics: Outlier, Entropy, and Mean-IDF.

Diversity-Informed Data Collection also provides a new method for finding an upper bound on a current corpus's diversity via a Corpus-Wide Oracle which has access to information about which utterances are most diverse across the corpus.

Prior work has not used corpus-level statistics to enhance the diversity of the collected data. Instead, when collecting data with crowdworkers, researchers have sought more diverse responses by altering the task (Kang et al., 2018) or by altering the stimulus (Larson et al., 2019). Prior work that trains neural dialogue models has not made use of subsets of existing datasets that exhibit properties

of diversity.

Our experiments show this strategy yields significantly more diverse data than baseline collection processes. It also yields better, more diverse model output on two downstream tasks. Additionally, this method can be implemented for other metrics which are defined relative to the corpus.

2 Related Work

Past work in neural dialogue generation investigates how to improve diversity in conversational responses. Additionally, past work in crowdsourcing data collection has explored optimizing crowdsourcing data collection processes.

2.1 Diverse Neural Dialogue Generation

Improving model diversity is an important goal in dialogue generation (Li et al., 2016a), with several related works proposing architecture and training improvements to increase diversity.

Decoding methods to increase model diversity include Li et al. (2016a) which proposes maximizing mutual information between the source sentence and response rather than maximizing likelihood. Other approaches have focused on beam search and incentivizing diverse beams, by adding similarity constraints at decoding (Baheti et al., 2018), penalizing items on the beam that are similar and reranking resulting items (Li et al., 2016b), or penalizing words which have already been generated in a current beam (Li et al., 2017). Shao et al. (2017) uses attention over already-generated words at decode time and beam reranking. Adding a temperature parameter to sharpen the decoder's distribution has also been studied (Cao and Clark, 2017).

Neural architecture improvements have also been explored, such as conditioning on a latent variable at decode time (Serban et al., 2017; Zhao et al., 2017) or a multi-headed attention mechanism which aims to capture different parts of the context (Tao et al., 2018). Zhang et al. (2018) explore the use of Generative Adversarial Networks to incentivize diversity. These more diverse models and decoding methods can be used in conjunction with Diversity-Informed Data Collection, since it attempts to improve the data that neural models are trained on in an earlier part of the model pipeline.

2.2 Crowdsourcing

Related work in crowdsourcing has approached the optimization problem of how to assign crowdworkers to different tasks.

2.2.1 Crowdworker Task Assignment

Basu Roy et al. (2015) formulates the problem of matching crowdworkers to tasks depending on skill levels for a set of concepts, pay rates, and HIT acceptance ratio. Follow-up work extends to collaborative crowdwork, where crowdworkers need to work together (Rahman et al., 2015). Assadi et al. (2015) pursue a similar task assignment setup.

Additional work has attempted to automatically evaluate crowdworker quality of task performance and use the results to assign crowdworkers to new tasks on-the-fly (Fan et al., 2015). Further investigations have explored more adaptive assignment of tasks in real-time based on the likelihood that a participant will continually complete tasks (Kobren et al., 2015). Relatedly, Kumai et al. (2018) design a task allocation to minimize the stress of workers and maximize the resulting quality in terms of balanced skill performance.

2.2.2 Label Distribution Prediction

An additional area related to our work is crowdworker label distribution prediction. Liu et al. (2019) has a crowdworking labeling task and trains models to predict the 50-label crowdworker distribution from 5-10 labels. Yang et al. (2018) aim to predict diversity in crowdworker answers to questions about an image to determine how many crowdworker responses are required to capture this diversity.

2.2.3 Dynamic Crowdworking Tasks

Lin et al. (2018) tackle the task of employing crowdworkers to generate or label minority class examples to feed an active-learning model. They deploy a multi-armed bandit to choose crowdworking tasks based on how cheaply a minority-class example can be generated using the technique. Our approach, by contrast, adapts a distributional constraint across the entire collection. Zhou et al. (2018) explores the related task of changing crowdworker team instruction prompts.

2.2.4 Diverse Crowdworking

Data collection approaches to incentivize diverse crowdworker output have also been studied. For instance, in EmpatheticDialogues (Rashkin et al., 2019) crowdworkers are conditioned to generate a response and an emotion (such as "afraid" or "proud") associated with it. If workers do not generate text with certain emotions, they are prompted

to select only from the underused labels. This is an example of trying to get better class coverage, but does not compare crowdworker output to the entire corpus of collected responses.

Past work has also examined how the particular crowdworking task affects the diversity of crowdworker output. Kang et al. (2018) compare two crowdsourcing tasks for use in a downstream goal-oriented dialogue system and examine resulting data diversity. While Kang et al. (2018) focus on choosing a *task* which produces diverse utterances, our work focuses on choosing a *participant population* which produces diverse data compared to data which has already been collected.

Building on Kang et al. (2018), and perhaps most similar to our work is Larson et al. (2019), which tackles the problem of detecting outlier paraphrases generated by crowdworkers. To obtain multiple ways of expressing similar intent (such as opening a bank account), crowdworkers are asked to paraphrase sentences. After a round of paraphrase collection, the most diverse (the outlier) paraphrases are identified and placed back onto the crowdsourcing platform for another round of data collection.

Our method is similarly aimed at increasing diversity of collected data. However, our method adapts the participant population for a set of tasks, which can be used in addition to an approach like Larson et al. (2019) which adapts the stimulus the population works on.

3 Diversity-Informed Data Collection

We propose a method, **Diversity-Informed Data Collection**, which progressively builds up a corpus, and while doing so, identifies which conversation participants produce more diverse utterances compared to the rest of the in-progress corpus. More formally, our task is to progressively build a sub-corpus, sub_c, of a given size from a larger, pre-collected corpus, c, where utterances are tied to IDs of specific participants.

Our approach is aimed at building a diverse sub-corpus sub_c. Our approach chooses which *population* of participants to collect data from for a given round. This population changes dynamically depending on calculated participant's diversity scores.

When utilizing a human-created, pre-existing corpus, we assume responses of the dataset are well-formed and of acceptable quality. With this assumption, we can maximize diversity scores without worrying that quality will be sacrificed for this diversity. However, when using this approach to collect data on-the-fly, additional quality controls may be necessary to ensure diverse data does not come at the cost of quality.

We assess two experimental conditions: Simulated Data Collection and Corpus-Wide Oracle Upper-Bound. Simulated Data Collection is set up to mimic crowdsourcing data collection processes leveraging a large pre-collected corpus, while Corpus-Wide Oracle Upper-Bound gathers an maximally diverse sub-corpus of utterances.

3.1 Corpus

For all experiments, we utilize the pre-collected EmpatheticDialogues corpus (Rashkin et al., 2019). We experiment with this corpus because it has crowdworker IDs associated with each utterance, which allows us to experiment with varying the participant population. Future work should conduct further experimentation to examine this approach's adaptability to other chitchat and goal-oriented datasets.

The corpus has a large number of utterances (100,000) over 25,000 conversations. Each conversation is centered around a situation (such as getting a promotion at work) and is associated with one of 32 emotions, such as anger, excitement, or guilt. Each conversation takes place between two crowdworkers and is an average of 4.3 turns. There are 810 unique crowdworkers in this dataset, each completing an average of 132 utterances each across an average of 61 conversations.

Our task is to select sub_c of size 10,000 from the larger EmpatheticDialogues corpus, c. We choose 10,000 as it is a sufficient number of utterances to train downstream models but still a small proportion (10%) of the original dataset, allowing examination of differences between sub-corpora. Implementation utilizes Cornell Convokit (Chang et al., 2019).

3.2 Simulated Data Collection

We simulate real-time crowdsourcing using a large, pre-collected corpus, c. This allows for running multiple trials, each time selecting sub_c and examining significance of different diversity metrics and participant selection conditions.

We simulate collecting data on-the-fly using an artificially-constructed environment (formally described in Algorithm 1), which completes multiple rounds of data collection until the progressively built sub-corpus $size(sub_c)$ is the desired size. The

Algorithm 1: Data collection simulation environment. $ComputeDiversity$ depends on the diversity metric (Table 2), and $EvalParticipants$ depends on the participant selection approach (Table 1).

```
function GatherData(Corpus c)
sub_c = ϵ
subCorpusSize = 10,000
numConvosToCollect = 2
population = []
numParticipants = 10
while size(sub_c) < subCorpusSize do
    while size(population < numParticipants) do
        p = Sample from c.Participants
        population.append(p)
        c.Participants.remove(p)
    end
    participantDiversities = []
    for Participant p in population do
        div_p = 0
        numUtts = 0
        for i in numConvosToCollect do
            convo = sample from p.Convos
            for utt in convo do
                div_p += ComputeDiversity(utt, sub_c)
                numUtts += 1
                sub_c.append(utt)
            end
            p.Convos.remove(convo)
        end
        div_p /= numUtts
        participantDiversities.append(div_p)
    end
    // Which participants kept for next round based on diversity scores.
    toKeep = EvalParticipants(participantDiversities)
    // Which participants still have data.
    remaining = p in population where len(p.convos) ≥ numConvosToCollect
    population = (toKeep ∩ remaining)
end
```

procedure assumes a fixed number of conversation participants in each round to gather data from (set to 10 for our experiments). We collect 2 conversations from each participant, chosen to allow the algorithm to recover from a participant with low diversity utterances while not judging a participant on just one conversation.

Given a participant's conversation, the diversity of an utterance in that conversation is stated in Equation 1:

$$div_{utt} = ComputeDiversity(utt, sub_c) \quad (1)$$

where ComputeDiversity depends on the diversity metric examined. We obtain a diversity score for each participant p's set of utterances ($utts_p$) by averaging these diversity values:

$$div_p = \frac{1}{size(utts_p)} \sum_{utt \in utts_p} div_{utt} \quad (2)$$

At the end of each round of data collection, utt_p is added to sub_c for each participant. Additionally, the algorithm determines which subset of the participant population is retained for the next round based on a Participant Population Selection strategy.

Our algorithm is greedy, since the order participants are added to the simulation and the order in which conversations are sampled both affect the participant's likelihood to be retained for an additional round. However, crowdworker data collection itself is usually a greedy approach, with crowdworkers being assigned to tasks in the order they arrive and being allowed to complete many tasks until the dataset has been collected.

3.2.1 Participant Population Selection

We experiment with three conditions to determine which sub-set of *current participants* (participants which were involved in the most recent round of data collection) should be retained for the next round of data collection, summarized in Table 1.

Diverse Population: After collecting conversations from current participants, we choose to retain the most-diverse 70% of participants.

Above Mean Population: Any participant whose diversity average falls above the mean diversity average of sub_c is retained in the pool of participants.

Random Population: We compare to a special random baseline, where at each iteration we retain a random 70% of the participant population, to directly compare to the 70% of crowdworkers

Condition	Description
Diverse Population	Calculates each participant's average relative diversity for current data collection round. We retain the 70% most-diverse participants of the current round.
Above Mean Population	Calculates each participant's average relative diversity for current data collection round. Retains the participants whose diversity scores fall above the sub-corpus's mean diversity.
Random Population	Retains a random 70% of participants.
Corpus-Wide Oracle	Uses a Corpus-Wide Oracle which ranks utterances' diversities in relation to the large dataset, c. Selects the most diverse utterances from these values independent of conversations.

Table 1: Participant Population Selection conditions for Simulated Data Collection. The first three conditions are used in conjunction with Algorithm 1, while the last condition provides an upper-bound for diversity by utilizing a Corpus-Wide Oracle to determine the known most-diverse utterances.

Metric	Description
Outlier	Euclidean distance between utterance embedding and average embedding for all utterances in the sub-corpus (Larson et al., 2019)
Entropy	Entropy of utterance under a trigram language model trained on sub-corpus.
Mean IDF	Mean IDF value (Baeza-Yates et al., 1999) for words in utterance compared to the rest of the corpus.

Table 2: Diversity metrics considered for data collection.

retained in Diverse Population. We structure Random Population to collect data from roughly the same number of participants as Diverse Population, to examine differences between the resulting sub_c due to the the *selection* of which participants to retain for another round of data collection.

3.2.2 Diversity Metrics

We experiment with three diversity metrics (Outlier, Entropy, and Mean IDF), summarized in Table 2. For all metrics, a new utterance utt is compared to the sub-corpus sub_c.

The same utterance can have different diversity values depending on the utterances in sub_c. When augmenting pre-collected data, this allows for the collection of new utterances which are *relatively* diverse.

Outlier: The embedding-based Outlier metric was proposed by Larson et al. (2019). Each utterance is encoded using a Universal Sentence Encoder (USE), which creates a sentence embedding by averaging word embeddings and passing the representation through a feedforward neural network, originally trained in a multi-task setting with supervised and unsupervised NLP tasks (Cer et al., 2018).

An embedding of an utterance is created via: $E_{utt} = USE(utt)$. A *mean corpus vector* is computed by averaging all of sub_c's utterance's vectors:

$$E_{sub_c} = \frac{1}{size(sub_c)} \sum_{u \in sub_c} USE(u) \quad (3)$$

The diversity metric is the Euclidean distance between each new utterance and the mean corpus vector, or:

$$\sqrt{\sum_i (E_{u_i} - E_{sub_{c_i}})^2} \quad (4)$$

where i is a dimension in Embedding E.

Utterances which are farther from the mean corpus vector are given a higher diversity score. For Simulated Data Collection, the mean corpus vector shifts as data is collected. Therefore, depending on which utterances are already added in the sub-corpus, outlier values will change for a given utterance.

Entropy: The Entropy score is determined by a non-neural trigram language model with smoothing for unseen words. The diversity score is given by:

$$-\frac{1}{|x \in Trigram(utt)|} \sum_{\substack{x \in \\ Trigram(utt)}} p(x)\ log\ p(x) \quad (5)$$

The language model is only trained on utterances in the sub-corpus.

Mean IDF: This metric calculates the mean IDF value for each word in the utterance (Baeza-Yates et al., 1999). IDF is calculated by treating each utterance in the corpus as a document. For a given utterance utt_p and sub-corpus sub_c, Mean IDF is calculated via:

$$\frac{1}{|utt_p|} \sum_{w \in utt_p} log\left(\frac{|\{sub_c\}|}{|\{utt|w \in utt\}|}\right) \quad (6)$$

where $\{sub_c\}$ is the set of all utterances in the sub_c. The IDF of a word w in utt is the number of utterances in sub_c divided by the number of utterances containing w on a log scale.

In addition to evaluating the robustness of our approaches, multiple diversity metrics are chosen with different conceptual types of diversity in mind. Outlier uses Universal Sentence Encoder embeddings which capture content (Cer et al., 2018). Entropy considers the probability of short phrases and can capture word combination diversity. Mean IDF considers the rarity of words being used for vocabulary diversity. Depending on the downstream application for a dialogue agent, the utility of these diversity measures may vary.

3.3 Corpus-Wide Oracle Upper Bound

To provide an Upper Bound for the diversity of a sub-corpus sub_c, we create a Corpus-Wide Oracle which knows the value of each utterance's diversity compared to the entire corpus c. For each $utt \in c$, we compute diversity according to the methods in Table 2, where $sub_c = c$. For example, for Outlier, the mean corpus vector is

$$\frac{1}{size(c)} \sum_{x \in c} USE(x) \quad (7)$$

which captures utterances from the entire corpus c. We calculate a Corpus-Wide Oracle diversity score, div_{oracle}, for each utterance in c for each diversity metric.

The Corpus-Wide Oracle is used to construct sub_c of any size consisting of the most diverse utterances. This sub-corpus can be used to compare against other collection methods, such as those in Simulated Data Collection, or as a way to enhance an existing collection by selecting out the most diverse utterances.

After the Corpus-Wide Oracle ranks each utterance by diversity, we select the utterances with the top 10,000 diversity values to form sub_c. This serves as a use-case for collecting the maximally-diverse corpus for a given diversity metric.

However, the Corpus-Wide Oracle might not be the *best* 10,000 utterances to collect for a sub-corpus. The Corpus-Wide Oracle selects the utterances with the most diversity compared to the whole corpus, but this might be too much diversity without enough context since the Simulated Data Collection methods add entire conversations (not utterances in isolation) to sub_c.

4 Evaluation

We evaluate the collected corpora both in terms of how diverse each sub-corpus is as well as performance on two downstream tasks: conversation emotion classification and dialogue generation.

4.1 Overall Diversity

The first evaluation aims to answer the question of if our methods produce more diverse sub-corpora than the Random Population baseline. We examine the hypothesis that using a collection method with knowledge of diversity will result in sub_c that is significantly more diverse. For each data collection method, we compare the diversity of the sub-corpus to Random Population. Because diversity values are relative to sub_c, diversity of sub_c is measured via div_{oracle} values.

Table 3 shows the resulting div_{oracle} values for datasets collected using our methods. Each value is the average of 100 trials, in which each trial collects a 10,000 utterance sub-corpus, sub_c.

Significance results for all experiments use a two-sided t-test compared to the Random Population baseline. Both Diverse Population and Above Mean Population produce datasets which contain statistically significantly ($p < 0.001$) more diverse data compared to the Random Population baseline. The Corpus-Wide Oracle method produces the most diverse results overall, as expected as it is a collection of the top 10,000 most diverse utterances. Running Diversity-Informed Data Collection to collect datasets of size 5,000 produced similarly significant differences.

We also examine the average number of participants out of the 810 total in c that are included for each method. Note in Table 3 the difference in Average Number of Participants from Random Population and Diverse Population to Above Mean Population and Corpus-Wide Oracle. Even though Above Mean Population is more diverse than Di-

	Condition	Mean Score	Avg. #Part
Outlier	Random Population	0.974	257.4
	Diverse Population	**0.979***	262.1
	Above Mean Population	0.978*	516.9
	Corpus-Wide Oracle	**1.035***	539.0
Entropy	Random Population	−5.350	257.2
	Diverse Population	−5.320*	259.1
	Above Mean Population	**−5.294***	359.1
	Corpus-Wide Oracle	**−4.261***	481.0
Mean IDF	Random Population	5.455	256.2
	Diverse Population	**5.659***	257.7
	Above Mean Population	5.613*	357.5
	Corpus-Wide Oracle	**7.783***	546.0

Table 3: Results for diversity scores for each method of collecting corpora, by metric (Outlier, Entropy, and Mean IDF). Higher scores are better for all metrics. Also shown are the average number of participants (Avg. #Part) included out of a possible 810. * indicates statistical significance compared to the Random Population baseline ($p < 0.001$).

	Condition	SVM	FastText
Outlier	Random Population	0.224	0.050
	Diverse Population	**0.234***	0.052
	Above Mean Population	0.229	**0.077***
	Corpus-Wide Oracle	0.100*	0.057*
Entropy	Random Population	0.218	0.052
	Diverse Population	0.212†	0.049
	Above Mean Population	**0.254***	0.065*
	Corpus-Wide Oracle	0.134*	**0.102***
Mean IDF	Random Population	0.220	0.052
	Diverse Population	0.236*	0.052
	Above Mean Population	**0.257***	0.064*
	Corpus-Wide Oracle	0.131*	**0.065***

Table 4: Results for downstream classification accuracy averaged over 5-fold cross-validation over 10 trials: higher is better. The task is classification of emotions from a set of 32 possible given the text of dialogue responses in sub_c. † and * indicate $p<0.05$ and 0.001 respectively compared to Random Population.

verse Population for Entropy, it comes at the cost of more participants. Across all three diversity metrics, Above Mean Population requires about 100–200 additional participants than Diverse Population and Random Population. In an online setting where the cost to train new crowdworkers is high, the tradeoff between number of participants and diversity of content may be worth considering.

4.2 Classification

To examine the quality of the resulting sub_c's, we turn to downstream task evaluation. We first examine the task of classifying a conversation's emotions from utterance text. Following Larson et al. (2019)'s justification, we would expect more diverse sub_c to result in higher classification accuracies, because more diverse responses should cover more variation in how people express emotions in conversation.

4.2.1 Classification Method

We follow the methodology of Larson et al. (2019) who propose evaluating the diversity of goal-oriented intent paraphrases. For their use case, classification models predict the intents from the paraphrase. For our case, each conversation in the EmpatheticDialogues corpus is associated with an emotion, such as anger or guilt. There are 32 such emotions throughout the corpus. The classification task is to predict which of the 32 emotions is expressed from a given utterance. Following Larson et al. (2019), we use two classification models:

- Bag-of-Words SVM
- FastText classifier

Bag-of-Words SVM is an SVM using TF-IDF word features for prediction. The FastText classifier uses a neural classification model on top of fastText sentence embeddings (Joulin et al., 2017). The sub-corpora we collect using the different methods serve as the datasets to train these classification models.

4.2.2 Classification Results

Classification task results are summarized in Table 4. Reported scores are averaged 5-fold cross-validation and averaged over 10 runs of datasets collected from each method.

While most conditions show Diverse Population significantly outperforms Random Population, it performs worse than Random Population with Entropy SVM and Entropy FastText and performs the same in Mean IDF FastText. Above Mean Population, on the other hand, outperforms the Random Population baseline on all conditions. This could potentially be due to the larger number of participants included in Above Mean Population. Surprisingly, Corpus-Wide Oracle does not perform the best in each category. We conjecture that too many diverse responses do not allow a classifica-

tion model to learn common characteristics.

4.3 Generation

Because the ultimate goal of collecting more diverse dialogue data is generating more diverse text, we evaluate diversity of neural text generation models trained on resulting corpora.

4.3.1 Generation Method

Our task is to generate the next utterance in a dialogue, where the data collection processes collect utterances for sub_c. To train generation models, the input is the most recent parent utterance for each utt in sub_c, and utt is the target sentence to generate. When utt is the starting utterance in a conversation, the input is the situation associated with the conversation (such as planning a vacation).

We train Sequence-to-Sequence models (Sutskever et al., 2014) with a 2-layer bidirectional encoder, hidden size 500, word vector size 64, Adam optimizer (Kingma and Ba, 2014), learning rate 0.001, trained for 3000 steps with batch size 32. Models are implemented using OpenNMT (Klein et al., 2017). We opt to use a standard model as it has fewer parameters to learn from smaller sub-corpora. We use the same parameter settings for all trained models.

4.3.2 Generation Results

Generation task results are summarized in Table 5. We report on both mean and median length of model responses. Distinct-1 and Distinct-2 measure the proportion of unigrams and bigrams respectively in the set of model responses which are unique (Li et al., 2016a). We also report diversity of the generated responses calculated by the metrics used in sub_c collection (see Table 2).

Our method results in models which produce more diverse output compared to baseline Random Population data collection. Interestingly, Diverse Population and Above Mean Population split the win on producing more diverse outputs. Corpus-Wide Oracle diversity results are sometimes lower and overall shorter in length than other methods; a potential reason is this condition only samples utterances, not conversations.

Responses from the model trained on each sub_c are evaluated with all 3 diversity metrics, to examine potential interactions. Collecting sub_c with Entropy results in higher Mean IDF (and vice versa) compared to Random Population. Collecting sub_c with Outlier results in slightly lower Mean IDF (and vice versa) for Diverse Population and Above Mean Population compared to Random Population. There is not a consistent signal between Outlier and Entropy. Future work can further examine the relationships among these diversity metrics.

5 Discussion

Diversity Considerations: Compared to a random baseline, Diversity-Informed Data Collection results in more diverse data than Random Population, which is shown to be more effective on downstream tasks. Future work can explore the effect of simultaneously optimizing multiple desirable measurements of diversity.

However, we acknowledge that maximum diversity might not be what is desired and does not always result in the best downstream task performance, as indicated by the low Corpus-Wide Oracle downstream task performance. While we have not examined the tradeoff between diversity and quality, this can be explored in future work.

Generalizability: Diversity-Informed Data Collection is generalizable to metrics other than diversity. Concretely, DIDC should be used when a desired metric (1) can compare one sample (or set of samples) to the in-progress dataset and (2) has variation among the participant population.

Additionally, Diversity-Informed Data Collection can be applied to areas outside of dialogue data collection. For instance, DIDC could apply to collecting data with different emotions or sentiment. Another extension is to a specialized application domain, such as collecting dialogues for educational tutoring purposes, where our method could be used to collect more data from students who generate text consistent with certain types of misconceptions.

Crowdworking Deployment: We evaluated on simulated crowdworking data by leveraging an existing corpus. This choice stems from the desire to test multiple runs of methods in a controlled environment, to reliably determine significance, and to work with data with an assumed level of quality. That said, our approach can be applied to real crowdworking tasks. Data can be gathered from several participants in parallel, where crowdworkers are added and offered new tasks or assigned qualifications based on their diversity.

If our method is deployed in paid crowdworking tasks, Diverse Population might be more cost-effective. In this particular investigation, we find

	Condition	Mean Length	Median Length	D-1	D-2	Outlier	Entropy	Mean IDF
Outlier	Random Population	7.6	**7**	0.114	0.296	0.981	–3.088	5.504
	Diverse Population	**9.7**	**7**	0.110	0.279	0.989*	–3.354*	5.297§
	Above Mean Population	8.1	**7**	0.063	0.169	0.960*	–3.083	5.067*
	Corpus-Wide Oracle	3.8	4	**0.204**	**0.448**	**1.042***	**–2.968***	**6.789***
Entropy	Random Population	**8.8**	**8**	0.101	0.265	0.981	–3.281	5.263
	Diverse Population	7.7	7	**0.122**	**0.317**	0.978	–3.197§	5.411†
	Above Mean Population	6.6	6	0.092	0.226	0.982	–3.057*	5.474*
	Corpus-Wide Oracle	4.9	5	0.112	0.316	**0.985**§	**–2.935***	**5.781***
Mean IDF	Random Population	6.1	6	0.120	0.294	0.988	–3.036	5.526
	Diverse Population	6.7	6	0.131	0.322	0.986	–2.955§	5.797§
	Above Mean Population	**7.2**	**7**	0.071	0.187	0.976*	–2.937*	5.655
	Corpus-Wide Oracle	3.4	3	**0.214**	**0.449**	**1.008***	**–2.421***	**8.327***

Table 5: Downstream model generation results; higher numbers are better for all metrics. †, §, and * indicate $p<0.05$, 0.01, and 0.001 respectively. As Distinct-1 and Distinct-2 are summary statistics, we did not test significance.

Diverse Population requires 100-200 fewer participants than Above Mean Population to create a dataset. Due to the time required to train new participants, there is a tradeoff between training a new worker and collecting more data form current participants.

Caution should be taken in using this method on-the-fly without a quality check. Standard quality control methods (e.g., crowdworker qualifications, manual examination, crowdworker verification) should be deployed for from-scratch data collection.

Crowdworker Fairness: Another important consideration for a live deployment is the crowdworker's perspective of fairness. Because some crowdworkers are retained for more data collection than others, communicating this possibility to crowdworkers is essential (Brawley and Pury, 2016). Crowdworking best practices involve disclosing which quality metrics are being used to workers to set clear expectations (Bederson and Quinn, 2011). Additionally, combining our method with a method which alters the task crowdworkers complete (Kang et al., 2018) as opposed to restricting the crowdworking population could be a way to balance fairness with crowdworkers. Different task and population combinations could allow for all crowdworkers to participate in more tasks.

6 Conclusion

We propose a method, Diversity-Informed Data Collection, which leverages this to produce more diverse datasets than the standard approach, and which performs better on downstream tasks. We define diversity of an utterance compared to the other utterances in a corpus. This allows for measurement of the impact of adding each utterance to the corpus. Working under the same assumption that a subset of participants produce diverse data compared to the corpus, our method can be extended to other diversity measures and can be modified to work with other corpus-level metrics.

Acknowledgements

This work was supported by an AWS Machine Learning Research Award, an NVIDIA Corporation GPU grant, a UC Berkeley Chancellor's Fellowship, a National Science Foundation (NSF) Graduate Research Fellowship (DGE 1752814) and an NSF CAREER Award (IIS-1453721). We thank the three anonymous reviewers for their helpful comments. We additionally thank Cathy Chen, David Gaddy, Daniel Fried, Lucy Li, and Nate Weinman for their helpful feedback.

References

Sepehr Assadi, Justin Hsu, and Shahin Jabbari. 2015. Online assignment of heterogeneous tasks in crowdsourcing markets. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California, USA*, pages 12–21. AAAI Press.

Ricardo A. Baeza-Yates, Berthier Ribeiro-Neto, et al.

1999. *Modern Information Retrieval*, chapter 3, Modeling. ACM press New York, USA.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980, Brussels, Belgium. Association for Computational Linguistics.

Senjuti Basu Roy, Ioanna Lykourentzou, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das. 2015. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal—The International Journal on Very Large Data Bases*, 24(4):467–491.

Benjamin B. Bederson and Alexander J. Quinn. 2011. Web workers unite! addressing challenges of online laborers. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 97–106, New York, NY, USA. Association for Computing Machinery.

Alice M. Brawley and Cynthia L.S. Pury. 2016. Work experiences on mturk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior*, 54:531 – 546.

Kris Cao and Stephen Clark. 2017. Latent variable dialogue models and their diversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 182–187, Valencia, Spain. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2019. Convokit: The cornell conversational analysis toolkit.

Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. 2015. Icrowd: An adaptive crowdsourcing framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, page 1015–1030, New York, NY, USA. Association for Computing Machinery.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Yiping Kang, Yunqi Zhang, Jonathan K. Kummerfeld, Lingjia Tang, and Jason Mars. 2018. Data collection for dialogue system: A startup perspective. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 33–40, New Orleans - Louisiana. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Ari Kobren, Chun How Tan, Panagiotis Ipeirotis, and Evgeniy Gabrilovich. 2015. Getting more for less: Optimized crowdsourcing with dynamic tasks and goals. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 592–602, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Katsumi Kumai, Masaki Matsubara, Yuhki Shiraishi, Daisuke Wakatsuki, Jianwei Zhang, Takeaki Shionome, Hiroyuki Kitagawa, and Atsuyuki Morishima. 2018. Skill-and-stress-aware assignment of crowd-worker groups to task streams. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018*, pages 88–97.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. 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 517–527, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Christopher H. Lin, Mausam, and Daniel S. Weld. 2018. Active learning with unbalanced classes and example-generation queries. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2018, Zürich, Switzerland, July 5-8, 2018*, pages 98–107. AAAI Press.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. 2019. Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 1111–1120, New York, NY, USA. ACM.

H. Rahman, S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das. 2015. Task assignment optimization in collaborative crowdsourcing. In *2015 IEEE International Conference on Data Mining*, pages 949–954.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3295–3301. AAAI Press.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating long and diverse responses with neural conversation models. *CoRR*, abs/1701.03185.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4418–4424. International Joint Conferences on Artificial Intelligence Organization.

Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual question answer diversity. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 1815–1825, Red Hook, NY, USA. Curran Associates Inc.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Sharon Zhou, Melissa Valentine, and Michael S. Bernstein. 2018. In search of the dream team: Temporally constrained multi-armed bandits for identifying effective team structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 108:1–108:13, New York, NY, USA. ACM.