Improving Classroom Dialogue Act Recognition from Limited Labeled Data with Self-Supervised Contrastive Learning Classifiers

Vikram Kumaran **Jonathan Rowe** **Bradford Mott**
North Carolina State University
{vkumara, jprowe, bwmott}@ncsu.edu

Snigdha Chaturvedi
UNC Chapel Hill
snigdha@cs.unc.edu

James Lester
North Carolina State University
lester@ncsu.edu

Abstract

Recognizing classroom dialogue acts has significant promise for yielding insight into teaching, student learning, and classroom dynamics. However, obtaining K-12 classroom dialogue data with labels is a significant challenge, and therefore, developing data-efficient methods for classroom dialogue act recognition is essential. This work addresses the challenge of classroom dialogue act recognition from limited labeled data using a contrastive learning-based self-supervised approach (SSCon). SSCon uses two independent models that iteratively improve each other's performance by increasing the accuracy of dialogue act recognition and minimizing the embedding distance between the same dialogue acts. We evaluate the approach on three complementary dialogue act recognition datasets: the TalkMoves dataset (annotated K-12 mathematics lesson transcripts), the DailyDialog dataset (multi-turn daily conversation dialogues), and the Dialogue State Tracking Challenge 2 (DSTC2) dataset (restaurant reservation dialogues). Results indicate that our self-supervised contrastive learning-based model outperforms competitive baseline models when trained with limited examples per dialogue act. Furthermore, SSCon outperforms other few-shot models that require considerably more labeled data [1].

1 Introduction

Dialogue analysis offers significant potential for improving our understanding of classroom learning and teaching by modeling dialogue between students and teachers. Studies of classroom dialogue can provide deep insight into how students learn most effectively and engage with each other and with teachers (Mercer et al., 2019; Mercer, 2010; Resnick et al., 2010; Hmelo-Silver, 2004). A long-standing goal in analyzing classroom dialogue is to understand how student-student and student-teacher dialogues lead to better student learning outcomes (Wendel and Konert, 2016). This work addresses the problem of dialogue act recognition in K-12 classroom dialogues.

Dialogue act recognition has garnered considerable attention and is useful for many tasks such as dialogue generation and understanding (Chen et al., 2022; Lin et al., 2021; Goo and Chen, 2018). Recent efforts in dialogue act recognition are built on large-scale pre-trained language models (Qin et al., 2021, 2020; Wang et al., 2020; Raheja and Tetreault, 2019; Chen et al., 2018). These models demonstrate high performance on standard datasets but require substantial labeled training data and, in some cases, combine other corroborative labels such as sentiment. Finding labeled public datasets of K-12 classroom dialogues is challenging for several reasons. First, there are concerns about participants' privacy and security. Second, researchers often develop individualized coding schemes specific to their design framework, and research context (Mercer, 2010; Song et al., 2019; Hao et al., 2020; Song et al., 2021). Therefore, even when labeled datasets are available, the assortment of coding schemes makes it challenging to cross-train across datasets. Third, classroom dialogue utterances are usually specific to a given subject matter, making generic labeled dialogue datasets less useful as auxiliary data. It is, therefore, essential to develop the capability to build dialogue act recognition models from limited labeled training data.

Our research addresses the lack of large-scale labeled classroom dialogue datasets by using a self-supervised contrastive learning-based model (SSCon) trained using limited labeled data. SSCon uses contrastive learning to transform the dialogue utterance representation into a new embedding space where identical dialogue acts cluster together, and distinct dialogue acts are pushed further apart. The system iteratively improves per-

[1] Our code is available at https://gitlab.com/vkumara/SSCon

Findings of the Association for Computational Linguistics: ACL 2023, pages 10978–10992
July 9-14, 2023

formance, as the contrastive learning step benefits self-supervision, even when presented with limited labeled data. Experiments show that SSCon outperforms competitive baselines with just tens of labeled examples per dialog act in both a K-12 mathematics classroom dataset and an everyday conversation dataset, DailyDialog. Our key contributions are the following:

- We propose a novel self-supervised contrastive learning dialogue act recognizer.
- We test our model on multiple datasets in distinctly different domains under label-scarce settings. Our experiments show that our model outperforms strong baselines.
- We illustrate with an ablation study why our model outperforms the baseline.

2 Contrastive Learning Model

2.1 Problem Definition

A dialogue $D = (u_1, u_2, \cdots, u_N)$ in a dialogue dataset consists of a sequence of N utterances, and a set of dialogue acts A. Dialogue act recognition (DAR) is defined as a classification problem, that involves recognizing the dialogue act $da_i = DAR(u_i|u_{i-1}, \cdots, u_{i-m})$ for utterance u_i, given its context, a set of previous m utterances, where $da_i \subseteq A$. The task is formalized as a multi-class classification problem and sometimes a multi-label classification problem, depending on the coding scheme used.

2.2 Approach and Assumptions

While our approach is primarily evaluated on the multi-class classification problem, we also test our model on the multi-label dataset DSTC2 (Henderson et al., 2014), with promising results.

Since we test a few-shot learning scenario, we start with a small set of labeled utterance examples I and the remaining training set U of unlabeled utterances. We also have a set of labeled utterances T set aside as a validation set.

2.3 SSCon Overview

In this section, we describe the architecture of our model, the self-supervised contrastive learning (SSCon) based multi-class classifier, shown in Figure 1. Our model operates in multiple stages. We begin by finetuning a large pre-trained transformer-based language model (PLM) trained on large publicly available dialogue datasets using our domain-specific dialogue dataset. We use the finetuned PLM to generate dialogue embeddings for our model. In Stage 1, we use the finetuned PLM to encode the utterance and its dialogue history. We also use sentence-BERT (Reimers and Gurevych, 2019) to create a latent representation of each utterance. A classifier built on the latent representation makes an initial soft prediction of the dialogue act. We distill the initial high-confidence predictions as soft labels for Stage 2. In Stage 2, the soft labels from Stage 1 train an encoder using contrastive learning. It translates the latent representations from Stage 1 into a vector in a different encoding space where identical dialogue acts cluster together while distinct dialogue acts are separated. In Stage 3, we pass the utterance representations through the encoder trained in Stage 2 to get the embeddings, which will be the input to classify the dialogue acts. The high probability soft labels from the Stage 3 classifier are sent back to Stage 1 as input to self-supervise the model in the next iteration.

2.4 Pretraining

For our pretraining stage, we finetune DialoGPT (Zhang et al., 2020), a dialogue PLM built on GPT2 (Radford et al.) and trained on 147M Reddit and other online conversations, licensed under MIT License. Given an utterance u_i and its context of a set of previous m utterances, the dialogue PLM is finetuned to maximize the conditional probability for the subsequent utterance, $P(u_{i+1}|u_i, u_{i-1}, \cdots, u_{i-m})$. For pretraining, we create a dataset consisting of target utterances and its context consisting of preceding m utterances. SSCon uses the finetuned PLM's hidden state as the representation of the utterance u_i given its dialogue context $(u_{i-1} \cdots u_{i-m})$.

2.5 Stage 1: Context Classifier

The Stage 1 classifier concatenates the pretrained dialogue PLM's last layer embedding $H_i \in R^{768}$ with the sentence-BERT embedding $S_i \in R^{384}$ of the utterance u_i and predicts the dialog act, y_i^{stage1} for the utterance,

$$y_i^{stage1} = CL_{stage1}(X_i, \Theta) \quad (1)$$

$$X_i = [H_i; S_i] \quad (2)$$

$$H_i = DialoGPT(u_i, u_{i-1}, \cdots, u_{i-m}) \quad (3)$$

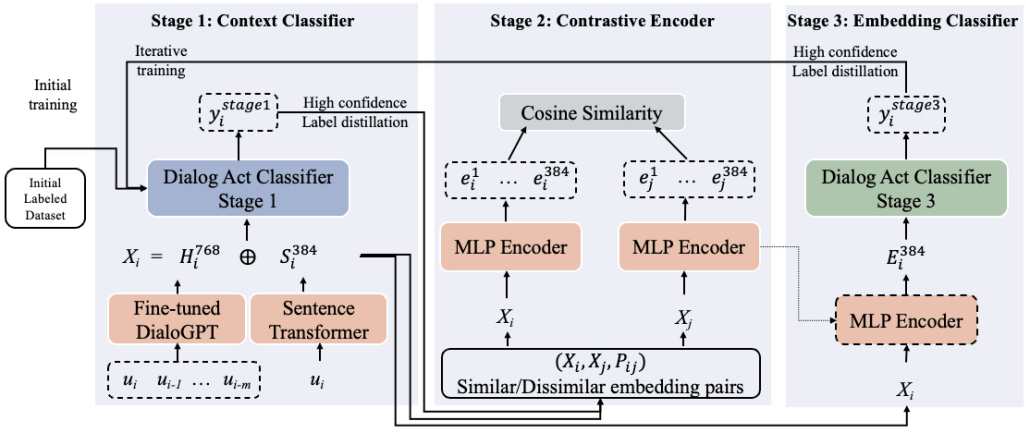

Figure 1: Self-supervised contrastive learning (SSCon) multiclass classifier

$$S_i = SentenceBERT(u_i) \quad (4)$$

Where y_i^{stage1} is the predicted dialogue act class for the last utterance u_i and CL_{stage1} is the Stage 1 multi-class classifier with trainable weights Θ. The concatenated vector X_i represents the utterance u_i as an independent sentence and in the context of its dialogue $(u_{i-1}, \cdots, u_{i-m})$. The output of Stage 1 is a dialogue act label (and the corresponding prediction probability) for each utterance in the dataset. For the choice of classifier, we explored both MLP classifiers (Haykin, 1994) and XGBoost (Chen and Guestrin, 2016); however, any reasonable classifier can be utilized. We chose the XGBoost classifier as it was faster to train and run without any impact on the performance of our model. However, for one of the datasets (DSTC2), where an utterance could have multiple labels, we used an MLP-based multi-label classifier.

In the first iteration of the self-supervision process, the training data used for the classifier is limited to the number of available labeled samples, usually between ten to a hundred examples per class. In subsequent iterations, the number of samples increases based on the soft labels from earlier iterations.

2.6 Stage 2: Contrastive Encoder

The prediction from the Stage 1 classifier, CL_{stage1}, is used as (soft) ground truth for Stage 2. Specifically, we use the dialogue act labels with the highest confidence in terms of prediction probability as new soft labels along with the initial labeled examples used in Stage 1. Including high confidence, soft-labeled samples increases the effective size of the training set with each iteration. Using this data, we adopt a contrastive training approach. This training process creates a network that can encode the finetuned PLM latent representation into a space where utterances with the identical dialogue act class are close together while utterances with distinct dialogue act labels are farther apart. The labeled samples are paired to generate positive and negative triplets $P = (X_i, X_j, P_{ij})$ where X_i (Equation 2) is the concatenated encoding defined in the previous section for a given utterance u_i, and P_{ij} is 1 if both the utterances map to the same dialogue act or -1 if they map to different dialogue acts. In the case of multi-label utterances, P_{ij} is the cosine distance between the one-hot encoding of the dialogue act vectors of the two utterances. The encoder is a five-layer MLP that transforms the $X_i \in R^{1152}$ concatenated latent representation into an $E_i \in R^{384}$ encoding. We use a twin encoder network to train on the positive/negative triplets P using cosine similarity between output embeddings as the similarity score. Given a triplet (X_i, X_j, P_{ij}) the network trains,

$$E_i = B_{encoder}(X_i) \quad (5)$$

$$E_j = B_{encoder}(X_j) \quad (6)$$

$$l_{ij} = cosine\_similarity(E_i, E_j) \quad (7)$$

$$loss_{ij} = MSE(l_{ij}, P_{ij}) \quad (8)$$

The trained encoder transforms each utterance, into a new encoding, $E_i = B_{encoder}(X_i)$, where $E_i \in R^{384}$ is the generated embedding of the MLP encoder ($B_{encoder}$).

2.7 Stage 3: Embedding Classifier

In this final stage of our model, we use the encoder from the Contrastive Encoder network that was trained in the previous stage to convert each X_i

into the embedding E_i used as input to the Stage 3 classifier.

$$y_i^{stage3} = CL_{stage3}(E_i, \Phi) \quad (9)$$

where y_i^{stage3} is the dialogue act labels for the utterance u_i and CL_{stage3} is a multi-class classifier with trainable weights Φ. Like Stage 1, we use an XGBoost classifier. We use the dialogue act labels with the highest confidence in terms of prediction probability from Stage 1 as the soft label for training the classifier.

The training starts with a small set of labeled utterance examples L, and the remaining training set U of unlabeled utterances. During every iteration of the self-supervision process, the model labels the unlabeled U utterances in our training set. We filter out the utterances labeled with low confidence (low prediction probability of CL_{stage3}). The distilled silver-labeled instances are moved from U to L, along with the initial labeled examples. The updated L is the training set for Stage 1, starting the next iteration of the self-supervision process.

3 Experimental Setup

3.1 Implementation Details

The pretraining stage involves fine-tuning the DialoGPT model with utterance data from a given dialogue dataset. The training set uses nine previous utterances as context. We finetune a HuggingFace pre-trained "dialoGPT-small" base model on a single GPU for four epochs. The version of DialoGPT we used is a 12-layer transformer. We use the hidden state vector H_i of the end-of-sentence tag in the 12$^{\text{th}}$ layer of the transformer as the embedding representing the last utterance in the sequence.

Stage 1 of the model uses a dialog act classifier. We implemented an MLP-based classifier and an XGBoost classifier. There was no difference in performance between the two classifiers, so we picked XGBoost as it is a standard baseline. We use the hidden state, H_i ,of the last layer as the embedding to represent the dialogue context of the utterance. For the sentence-BERT embedding S_i of the utterance we used a pre-trained network ("all-MiniLM-L6-v2") that is trained on roughly 1B sentence pairs (Reimers and Gurevych, 2019). The input to the model is the concatenated vector of H_i and S_i. After running the model for various distillation thresholds, a threshold of 0.85 was used for high-probability soft-label distillation based on a simple grid search on the threshold parameter as discussed in the appendix.

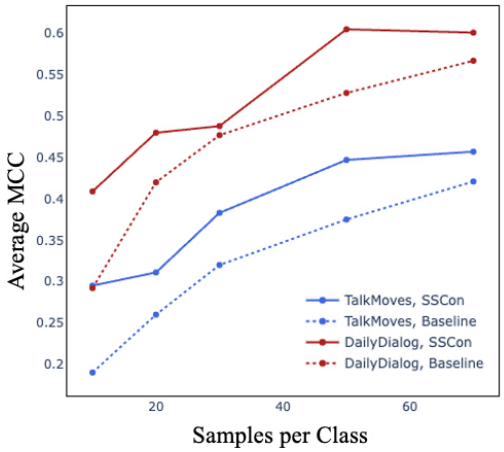

Figure 2: Improvement of performance over a baseline XGBoost classifier by our SSCon classifier.

Stage 2 of the model is a contrastive encoder network. Each encoder network in the twin network is a 5-layer MLP with a 20% dropout between the layers and an output embedding vector of dimension 384. About 1.2 million similarity pairs are used to train the twin network for 4–6 epochs on a single GPU. Stage 3 is an XGBoost classifier head on top of the encoder trained in Stage 2. The iterative self-supervision process continues until we show no improvement in the validation data. We run 5–10 iterations. In our experiments, we use a validation set to determine when to stop the iterations. We report the results on the test set.

3.2 Datasets

We use three datasets in our experiments: the TalkMoves, the DailyDialog, and the Dialogue State Tracking Challenge 2 datasets. The TalkMoves dataset (Suresh et al., 2022a) consists of 567 human-annotated class video transcripts of K-12 mathematics lessons between teachers and students. A human-transcribed dataset consists of 174,186 teacher utterances and 59,874 student utterances. The dialogue act labels in the dataset have an inter-rater agreement score above 90% for all labels. The dialogue acts for student utterances include 'relating to another student', 'asking for more information', 'making a claim', and 'providing evidence'. The dialogue acts for teacher utterances include: 'keeping everyone together', 'getting students to relate', 'restating', 'revoicing', 'pressing for accuracy', and 'pressing for a reason'. For the TalkMoves dataset, we train our model on student

utterances.

DailyDialog is a multi-turn dialogue dataset (Li et al., 2017) consisting of everyday conversations. The dataset consists of 13,118 conversations between multiple people. Four categories of dialogue acts are coded in the dataset. They are 'inform' (45%), 'questions' (29%), 'directives' (17%), and 'commissive' (9%).

To evaluate our approach on multi-label data set, we train on the DSTC2 dataset (Henderson et al., 2014). This dataset contains dialogues between crowdsourced workers and automated dialogue systems in the restaurant reservations domain, with 1000 train and test dialogues and about 21 dialogue acts.

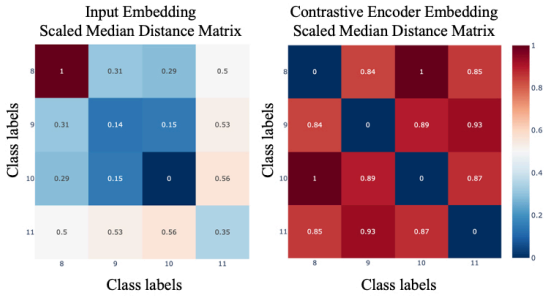

Figure 3: Heatmaps show the median cosine distance between embedding examples (TalkMoves dataset) for each dialog act type (8–relating to another student; 9-Asking for more info; 10-making a claim; 11-providing evidence/reasoning). The left heatmap corresponds to the input embeddings. The right heatmap corresponds to the Contrastive Encoder embeddings

3.3 Baselines

We run our experiments on different datasets described above. The TalkMoves dataset does not yet have many published baselines for dialogue act recognition, so we compare our approach against multiple baseline models. One is a baseline XGBoost classifier using the same utterance representation input as our model. This classifier is the same as our Stage 1 classifier. The second baseline is a self-supervised XGBoost classifier similar to SSCon without the contrastive learning in Stage 2. For the third baseline, we use an embedding prototype distance-based classifier. The average sentence embedding vector for labeled examples from each class represents a prototype for each label. We then measure the cosine distance of every utterance's sentence embedding in our test set against each class-prototype embedding. An utterance is assigned the class label of the closest prototype in the embedding space. To validate SSCon against current state-of-the-art dialogue act recognition models, we compare our results on the DailyDialog dataset against a few-shot learning model and a model that uses all available labeled training data. The state-of-the-art Co-GAT model (Qin et al., 2021) uses all training data available, including related sentiment labels on the utterance. Trained on very limited data, we do not expect our model to beat Co-GAT but use its performance as an upper bound. ProtoSeq (Guibon et al., 2021) is a sequential prototypical network trained to work in a few-shot fashion. However, they do use all training data to train their network. We use this model for comparison purposes as our approach is a type of few-shot learner using a small number of samples.

We also compare SSCon to standard XGBoost as we do for the TalkMoves dataset. We also compare SSCon dialogue act recognition results on DSTC2, a multi-label problem, against the self-supervised student-teacher approach by (Mi et al., 2021a).

3.4 Evaluation Metrics

The various baselines are compared with SSCon using four metrics. The Matthews Correlation Coefficient (MCC) for multi-class classification has a range of -1 to 1 and handles imbalanced datasets well. In some cases, the baselines only report the F1 (macro) score, the arithmetic mean of individual class F1 scores giving equal weight to all classes. We also report macro-averaged precision and recall.

4 Results

The first dataset we consider is the TalkMoves dataset. To evaluate the models in a few-shot learning scenario, we experiment using 10-70 labeled instances per dialogue act type (less than 1% of the overall data). Table 1 shows our results. SSCon shows an improvement of about 7-10% over the Utterance Embedding XGBoost baseline and the Prototype Distance Classifier when trained on 70 labeled examples. When SSCon is compared with the self-supervised XGBoost classifier without contrastive learning, it shows an improvement of 1% for the 70 labeled examples case but an improvement of more than 60% for the 10 labeled example case. This difference in performance suggests that contrastive learning is more beneficial when working with fewer labeled examples. Besides the three baselines, we also report the performance of a pub-

Dataset	Model Type	Labeled Examples per Class	MCC	F1 (macro)	Precision (macro)	Recall (macro)
TalkMoves Student	SSCon Classifier	10	0.295±0.042	0.412±0.044	0.454±0.034	0.512±0.016
		30	0.383±0.035	0.495±0.033	0.492±0.032	0.588±0.012
		70	**0.457±0.015**	**0.566±0.015**	**0.534±0.014**	**0.652±0.011**
	Utterance Embedding XGBoost	70	0.421±0.025	0.507±0.023	0.502±0.019	0.615±0.015
	Self Supervised XGBoost without Contrastive Learning	10	0.183±0.047	0.327±0.037	0.361±0.024	0.397±0.031
		70	0.451±0.021	0.514±0.021	0.535±0.015	0.609±0.016
	Prototype Distance Classifier	70	0.369±0.015	0.463±0.022	0.457±0.009	0.558±0.011
	Transformer for Seq Classifier (Suresh et al., 2022b)	All training	0.6716	0.7312	-	-
		70% training	0.588±0.005	0.653±0.005	-	-
Daily Dialog	SSCon Classifier	10	0.409±0.065	0.471±0.037	0.550±0.064	0.510±0.044
		30	0.488±0.046	0.564±0.043	0.608±0.013	0.619±0.039
		70	**0.601±0.022**	**0.660±0.018**	**0.663±0.009**	**0.687±0.015**
	Utterance Embedding XGBoost	70	0.567±0.013	0.638±0.008	0.639±0.008	0.671±0.009
	ProtoSeq (Guibon et al., 2021)	Few-shot	0.392±0.023	0.352±0.030	-	-
	Co-GAT (Qin et al., 2021)	All training	-	0.794	0.81	0.781
DSTC2	SSCon Classifier	1%		**0.293±0.025**	0.322±0.041	0.293±0.007
		10%		**0.447±0.008**	0.456±0.024	0.449±0.011
	ToD-BERT-ST (Mi et al., 2021b)	1%		0.285±0.040	-	-
		10%		0.405±0.090	-	-

Table 1: Comparison of results for the Daily Dialog, TalkMoves and DSTC2 dataset against baselines.

lished model that uses 100% (21K labeled samples) and 70% of the labeled training data. SSCon works in a label-scarce scenario, so our results are lower than the model using the complete labeled data set. Figure 2 shows the impact of labeled example counts on the overall performance of the SSCon classifier and an XGBoost classifier. While our iterative approach increases the performance on average by almost 50% over the baseline for a small training size (10 labels per class), it is about 9% for a larger labeled set size (70 labels per class). We notice that the overall performance improvement using SSCon is more with small labeled sets, and the improvement tapers as the labeled examples count increases.

The second dataset we consider is the DailyDialog dataset with four possible dialogue acts. Like the TalkMoves dataset, we consider 10 to 70 labeled examples per class. Table 1 shows that our performance, with just 70 instances per class (0.6% of 21521 labeled examples), is within 16% of the state-of-the-art dialogue act recognition model (Qin et al., 2021) that uses 100% of all the training data labels plus auxiliary sentiment labels. We also show that we outperform the other few-shot learning model, ProtoSeq (Guibon et al., 2021), by a significant margin even though they train using the entire labeled dataset. We investigate the change in performance of SSCon with increasing size of the labeled training set (Figure 2). As before, the performance improvement of SSCon over a standard XGBoost classifier is more significant for smaller label sets and less so as the number of samples increases. With ten samples per class, SSCon performance improvement over the baseline classifier is almost 41%. With seventy labels per class, the improvement is only about 9%.

We also compare our results on the DSTC2 dataset on restaurant inquiries. This dataset differs from the other two as each utterance can have multiple dialogue act labels. SSCon performs better on the DSTC2 dataset than the baseline self-supervising student-teacher model ToD-BERT-ST (Mi et al., 2021a) by 3 to 10%, depending on the size of the labeled training dataset. The ToD-BERT-ST model uses a data augmentation approach, while we use a contrastive learning approach. Both these approaches are complementary, and future work should explore using them together.

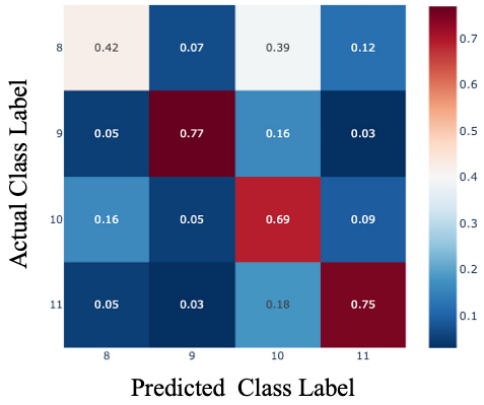

Figure 4: Confusion matrix with counts normalized by actual label counts (TalkMoves Dataset) for the SSCon classifier (70-samples per class training set). Dialog acts: 8–relating to another student; 9-Asking for more info; 10-making a claim; 11-providing evidence/reasoning.

Classifier	**LastUtterance**(u_i)
"*relating to another*" predicted as "*making a claim*"	"Reciprocal" "I don't agree, I measured them' "by five"
"*relating to another*" predicted as "*relating to another*"	"He didn't show his work" "You don't know that' "The first one"
"*making a claim*" predicted as "*making a claim*"	"Two fifths" "Y intercept' "power of three"

Table 2: : Examples of true positive and false negative for a couple of the TalkMoves dialog acts

5 SSCon Component Analysis

5.1 Clustering in Embedding Space

The fundamental intuition behind our approach is that contrastive learning brings similar utterances close to each other and dissimilar utterances further apart in the embedding space. Figure 3 shows the spatial clustering of labels in the embedding space of the trained Contrastive Encoder for one iteration with the TalkMoves dataset. Each square in the heatmap corresponds to the median cosine similarity distance between individual dialogue act class examples. To compute this distance, we take utterances from each dialogue act class and calculate the median cosine distance between their embeddings. Each cell in the heatmap shows a scaled distance, with the blue color corresponding to closer embeddings and red corresponding to the embeddings being farther away. The heatmap on the left corresponds to the distance matrix between embeddings provided as input to Stage 2. The heatmap to the right corresponds to the distance matrix for embeddings that the Contrastive Encoder has transformed. Please note that the figure uses scaled distance values, and we show four student dialog act classes.

When looking at the heatmap on the left for the input feature space, instances belonging to the dialogue act "relating to another student" (first row, label 8) are closer to the instances belonging to the dialogue act, "making a claim" (third row, label 10) than among themselves. After being transformed by the Contrastive Encoder (heatmap on the right), the smallest distances for each dialogue act fall on the diagonal (blue), as one expects with clustered labels. The diagonal corresponds to utterances with the same label. The non-diagonal terms, which are utterances with different labels, are pulled farther apart (red).

5.2 Classifier Performance by Dialogue Act

Figure 4 shows a confusion matrix with counts normalized by the ground truth label counts. In the TalkMoves dataset, the classifier struggles the most with the dialogue act corresponding to "relating to another student" (label 8). It is not able to clearly distinguish between "relating to another student" and "making a claim" (label 10). Label-scarce training suggests we only cover a limited variety of examples for each dialogue act class. In Table 2, we show some examples of correct and incorrect predictions. The first row shows utterances of the type "relating to another" mislabeled as "making a claim". The other two rows are true positive examples. We can see that the example utterances are hard to distinguish between "making a claim" or "relating to another" even for a human. The difference is in the context of the dialogue, and with limited training data, such distinctions are hard to make and might lead to overfitting.

5.3 Labeled Sample Selection

The initial set of labeled samples, in essence, drives the final performance of SSCon. Figure 5 shows the performance distribution of the baseline, XGBoost, for multiple runs for different sample sizes. Two trends are evident as the number of samples increases for every dataset. The first trend is that the performance improves as more samples are available for training. The improvement in performance with sample size is an expected trend as more examples indicate more information for the classifier to model. The second trend is that the variance

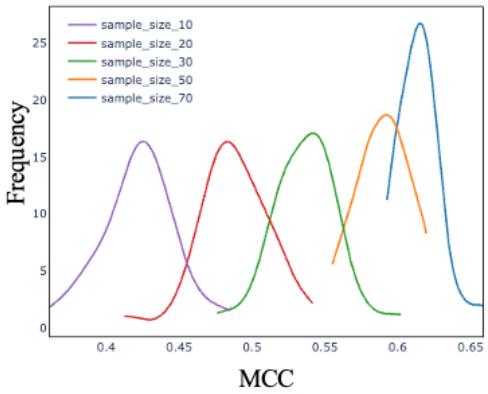

Figure 5: Drift in performance as we increase the number of labeled samples (TalkMoves dataset).

in performance decreases as the number of samples increases. One can assume a certain level of variability in the labeled training sets data's pattern of utterances for each dialogue act class. The dataset can only capture some of the variances in the data when the number of labeled examples is small. Hence, the classifier's performance can vary considerably depending on included examples for smaller training sets.

6 Related Work

Kalchbrenner and Blunsom (2013) proposed a hierarchical network architecture that combines CNN and RNN models to capture discourse structure. Subsequent work on dialogue act recognition (DAR) followed a similar approach using hierarchical architectures combining CNN and LSTM (Lee and Dernoncourt, 2016; Ji et al., 2016; Liu et al., 2017). More recently, researchers enhanced the model architecture by adding a CRF layer for classification (Kumar et al., 2018; Chen et al., 2018). Raheja and Tetreault (2019) build on earlier work that solves DAR as a sequential labeling problem using deep hierarchical networks by adding context-aware self-attention with promising results on standard benchmark datasets. More recent work has built off publicly available large pre-trained language models, significantly reducing the required training data. Qin et al. (2020) showed that combining associated labels such as sentiment with DAR can improve performance. However, all the above architectures require significant training data to achieve peak performance. Our model (SSCon) works with limited training data as it builds on dialogue context captured by a finetuned pre-trained language model. We use an iterative self-supervised training approach combined with a contrastive learning step to accommodate the lack of large labeled training datasets.

The idea of using pre-trained models to learn a few examples has been shown to be successful for natural language processing (Miller et al., 2000; Fei-Fei et al., 2006; Brown et al., 2020), and specifically, task-oriented dialogue systems (Liu et al., 2021b; Wu et al., 2020). Using few-shot learning, Mi et al. (2022) show they can improve dialogue state tracking, intent recognition, and natural language generation with limited labeled data using talk-specific instructions through prompts. Guibon et al. (2021) propose a prototypical network for sequence labeling on conversational data. While their network is trained to support few-shot DAR, they still require significant data for episodic training. The approaches mentioned above are not suitable for label-scarce situations. We show that SSCon outperforms their model's performance on the benchmark dataset with limited labeled data.

Pretrained language models have been used for classifying texts in the K-12 math education context. Shen et al. (2021) applied a BERT-based model to classify knowledge components in descriptive math texts. Loginova and Benoit (2022) employed an LSTM model to predict math problem difficulty trained on a question-answer dataset. These models work on descriptive text samples, a distinct use case from classroom dialogues. Okur et al. (2022) developed a speech dialogue system using MathBERT for natural language understanding, trained on significant pre-labeled data, differing from our limited labeled data scenario.

Self-supervision is a viable approach for DAR with limited training data. Mi et al. (2021a) use a teacher-student model iterative approach to improve performance. They use a novel text augmentation technique that adds to each iteration's training data. We show that SSCon improves upon their results on standard datasets. Our model uses contrastive learning (Hadsell et al., 2006) to cluster the different classes. A contemporary work by Tunstall et al. (2022) uses a similar contrastive learning-based approach for few-shot text classification. Liu et al.'s (2021a) trans-encoder combines two learning paradigms, cross, and bi-encoder, in a joined iterative framework to build state-of-the-art sentence similarity models. Our self-supervised contrastive learning-based (SSCon) based multiclass classifier also combines two learning paradigms

like the trans-encoder by Liu et al. (2021a). An embedding-based classifier, and a twin network, trained to leverage different training goals, help each other improve. One model uses the whole dialogue context captured in a hidden layer embedding to train a classifier. At the same time, the other is a twin neural network taking the contrastive representation learning approach, clustering same dialogue acts and separating distinct dialogue acts within the embedding space.

7 Conclusion

Classroom dialogue analysis can yield significant insight into student learning. However, collecting and coding classroom dialogue datasets is very labor-intensive. To address this problem, we introduce a novel self-supervised contrastive learning approach that can automate a portion of this process and make it more efficient even when limited labeled is available. We show that our approach improves on other methods that work on limited labeled datasets. The results also show that our approach can match and exceed the performance of some models trained on the fully labeled dataset.

Limitations

Selecting a representative set of examples to label becomes essential when working with limited labeled data. In this work, we use uniform sampling for our results, which might not be the best approach. We discuss these limitations in more detail in the appendix.

While we evaluate our model on a multi-label dataset (DSTC2) and show improvement over standard baselines, the effectiveness of our approach on such problems needs more investigation.

Ethical Considerations

While our algorithm is primarily a tool for improving classifier performance in label-scarce settings and uses publicly available, anonymized datasets, we acknowledge the potential ethical implications it may carry. Despite the neutral nature of our tool, it could unintentionally propagate or amplify biases favoring certain styles of communication. If used in real-time settings or without proper checks in place, it could inadvertently alter the natural dynamics of the classroom as teachers or students might modify their behavior based on how they believe the classifier categorizes their utterances. Like any other, we recognize that our tool could make mistakes or be misused by over-relying on quantitative aspects over qualitative aspects of instruction. Therefore, real-world application requires continuous vigilance and open dialogue with practitioners and stakeholders to ensure its use benefits teaching and learning.

Acknowledgements

This work is supported by the National Science Foundation under award DRL-2112635. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly supervised data augmentation through prompting for dialogue understanding. *arXiv preprint arXiv:2210.14169*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234.

Li Fei-Fei, Robert Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Chih-Wen Goo and Yun-Nung Chen. 2018. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 735–742. IEEE.

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, and Chloé Clavel. 2021. Few-shot emotion recognition in conversation with sequential prototypical networks. In *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Tianyong Hao, Xieling Chen, and Yu Song. 2020. A topic-based bibliometric analysis of two decades of research on the application of technology in classroom dialogue. *Journal of educational computing research*, 58(7):1311–1341.

Simon Haykin. 1994. *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Cindy E Hmelo-Silver. 2004. Problem-based learning: What and how do students learn? *Educational psychology review*, 16(3):235–266.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of NAACL-HLT*, pages 332–342.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality*.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of NAACL-HLT*, pages 515–520.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Zhaojiang Lin, Bing Liu, Seungwhan Moon, Paul Crook, Zhenpeng Zhou, Zhiguang Wang, Zhou Yu, Andrea Madotto, Eunjoon Cho, and Rajen Subba. 2021. Leveraging slot descriptions for zero-shot cross-domain dialogue StateTracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5640–5648, Online. Association for Computational Linguistics.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2021a. Trans-encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. In *International Conference on Learning Representations*.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021b. Robustness testing of language understanding in task-oriented dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2467–2480.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Ekaterina Loginova and Dries Benoit. 2022. Structural information in mathematical formulas for exercise difficulty prediction: a comparison of NLP representations. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 101–106, Seattle, Washington. Association for Computational Linguistics.

Neil Mercer. 2010. The analysis of classroom talk: Methods and methodologies. *British journal of educational psychology*, 80(1):1–14.

Neil Mercer, Sara Hennessy, and Paul Warwick. 2019. Dialogue, thinking together and digital technology in the classroom: Some educational implications of a continuing line of inquiry. *International Journal of Educational Research*, 97:187–199.

Fei Mi, Yasheng Wang, and Yitong Li. 2022. Cins: Comprehensive instruction for few-shot learning in task-oriented dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11076–11084.

Fei Mi, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings. 2021a. Self-training improves pre-training for few-shot learning in task-oriented dialog systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1887–1898.

Fei Mi, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings. 2021b. Self-training improves pre-training for few-shot learning in task-oriented dialog systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1887–1898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik G Miller, Nicholas E Matsakis, and Paul A Viola. 2000. Learning from one example through shared

densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 464–471. IEEE.

Eda Okur, Saurav Sahay, Roddy Fuentes Alba, and Lama Nachman. 2022. End-to-end evaluation of a spoken dialogue system for learning basic mathematics. In *Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)*, pages 51–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Libo Qin, Wanxiang Che, Yangming Li, Mingheng Ni, and Ting Liu. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8665–8672.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13709–13717.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. 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 3727–3733.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Lauren B Resnick, Sarah Michaels, and Catherine O'Connor. 2010. How (well structured) talk builds the mind. *Innovations in educational psychology: Perspectives on learning, teaching and human development*, pages 163–194.

Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean McGrew, and Dongwon Lee. 2021. Classifying math knowledge components via task-adaptive pre-trained bert. In *Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part I 22*, pages 408–419. Springer.

Yu Song, Tianyong Hao, Zhinan Liu, and Zixin Lan. 2019. A systematic review of frameworks for coding towards classroom dialogue. In *International Symposium on Emerging Technologies for Education*, pages 226–236. Springer.

Yu Song, Shunwei Lei, Tianyong Hao, Zixin Lan, and Ying Ding. 2021. Automatic classification of semantic content of classroom dialogue. *Journal of Educational Computing Research*, 59(3):496–521.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022a. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022b. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Dong Wang, Ziran Li, Haitao Zheng, and Ying Shen. 2020. Integrating user history into heterogeneous graph for dialogue act recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4211–4221.

Viktor Wendel and Johannes Konert. 2016. Multiplayer serious games. In *Serious Games*, pages 211–241. Springer.

Chien-Sheng Wu, Steven CH Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

A Implementation Details

In the paper, we describe the iterative approach of our model, where we alternate between a contrastive representation learning model and a standard classification model to improve performance.

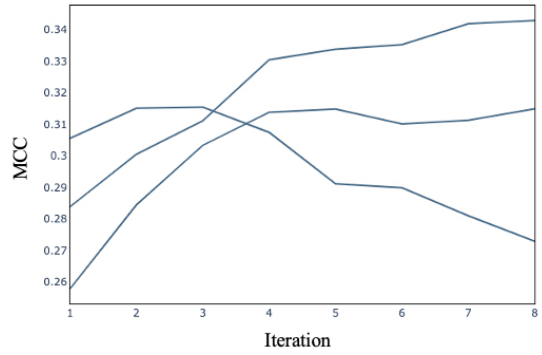

Figure 6: The model's performance changes with each iteration for multiple runs. Each run starts with different labeled data sets of 10 examples per class (TalkMoves dataset).

Two aspects determine the decision to stop the iteration. Figure 6 shows the model's performance in subsequent iterations for the TalkMoves dataset, starting from a label set of size ten examples per dialog act. A held-out dataset was used to calibrate the model's performance. One can see that the performance improvement is non-monotonic. In most cases, stopping after just one or two iterations is necessary to get maximum performance. The system is trying to improve based on minimal information in the labeled set, so as the size of the label set is small, there is a potential for over-fitting to the information contained in the label set. Another implementation detail is the selection of threshold confidence levels to distill soft labels as we expand the training set of the next model at the end of an iteration. Figure 7 shows the performance of our model for various prediction probability distillation thresholds. As mentioned earlier in the main paper and observed in Figure 7, 85% is the distillation threshold with the best performance distribution.

B Sample Selection by Active Learning

We experimented with using active learning techniques to pick labeled examples (Ren et al., 2021). We first trained a model with just ten initial labeled examples per class. We used the model to find a new set of ten samples for each class with the least confidence and added them to the labeled set using human labels. Figure 8 compares our active learning with uniform random selection approaches as a baseline. We can see that the random approach has much more variance for every label set size. However, we also note that the best-performing model was a random set for every label size. This difference in performance might be because our ap-

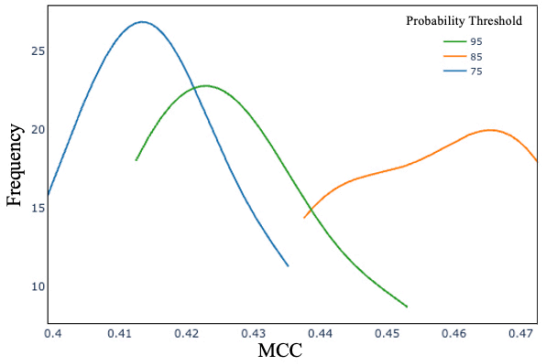

Figure 7: The figure shows our model's performance distribution for different soft label distillation thresholds. The runs were for 70 labeled samples per class as the initial training set from the TalkMoves dataset.

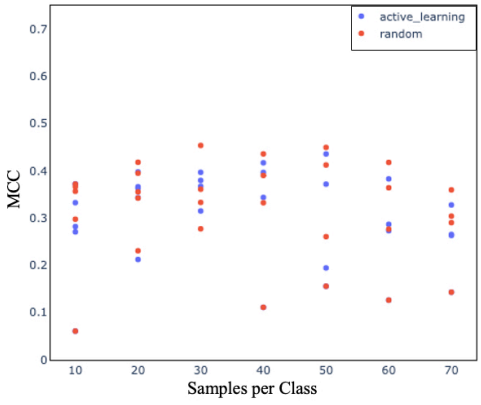

Figure 8: Comparison between active learning selection vs uniform random selection of labeled samples.

proach for active sample selection was naive. When we add a new sample based on low prediction confidence, there is no guarantee that we will improve the model. The new sample might be predicted with low confidence either because our model has limited information to decide or because the sample itself is an outlier. If it is the former adding the sample might help, but if it is the latter, adding the sample increases the number of outliers in our training set. As our training data is limited, outliers in the input data disrupt overall performance. We plan to experiment with more sophisticated active learning techniques for sample selection in the future.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
It is in a separate section after section 7 (Conclusion)

☑ A2. Did you discuss any potential risks of your work?
It is in the Ethical Considerations section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2, 3.2

☑ B1. Did you cite the creators of artifacts you used?
2,3.2

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
2.4, 3.2

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3.2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3.2

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.