Starting from "Zero": An Incremental Zero-shot Learning Approach for Assessing Peer Feedback Comments

Qinjin Jia[1], **Yupeng Cao**[2], and **Edward F. Gehringer**[1]

[1]Department of Computer Science, North Carolina State University, Raleigh, NC, USA
{qjia3,efg}@ncsu.edu
[2]Department of Electrical & Computer Eng., Stevens Institute of Tech., Hoboken, NJ, USA
{ycao33}@stevens.edu

Abstract

Peer assessment is an effective and efficient pedagogical strategy for delivering feedback to learners. Asking students to provide quality feedback, which contains suggestions and mentions problems, can promote metacognition by reviewers and better assist reviewees in revising their work. Thus, various supervised machine learning algorithms have been proposed to detect quality feedback. However, all these powerful algorithms have the same Achilles' heel: the reliance on sufficient historical data. In other words, collecting adequate peer feedback for training a supervised algorithm can take several semesters before the model can be deployed to a new class. In this paper, we present a new paradigm, called incremental zero-shot learning (IZSL), to tackle the problem of lacking sufficient historical data. Our results show that the method can achieve acceptable "cold-start" performance without needing any domain data, and it outperforms BERT when trained on the same data collected incrementally.

1 Introduction

Peer assessment is a process whereby students assess other students' assignments by writing review comments against a set of assessment criteria provided by the instructor. This pedagogical strategy has been extensively applied across various academic fields and has demonstrated its effectiveness over the past decades (Double et al., 2020). Furthermore, peer assessment serves as a crucial tool for delivering necessary feedback in massive open online courses (MOOCs), as this assessment strategy allows MOOCs to scale up the feedback process while minimizing ongoing support costs.

Nevertheless, the benefits of peer assessment can only be achieved with quality peer feedback (Ashton and Davies, 2015; Van Zundert et al., 2010). Course staff can manually review the credibility of each submitted feedback, but this is very inefficient. Hence, there has been a surge of interest in automating the assessment of feedback quality by machine-learning algorithms. These algorithms typically assess quality by determining whether the feedback comprises certain features (e.g., contains "suggestion" and "problem" statements) (Nelson and Schunn, 2009). If those characteristics are not present in the submitted reviews, the peer-assessment system could suggest that the reviewer revise the feedback to add the missing features.

Although these machine-learning algorithms for assessing feedback quality are very effective, they all have the same Achilles' heel: dependence on enough domain-specific peer-feedback data. That is, for each new discipline, it takes several school terms to collect sufficient data before the model can be applied. Thus, a desideratum of peer-assessment platforms is an effective quality-assessment model that does not require domain-specific historical data in *"cold-start"* condition (i.e., no domain data is available for training). Additionally, this model should be capable of using incrementally collected data to progressively improve its performance.

In this paper, we present an approach, named *Incremental Zero-shot Learning* (IZSL), for addressing lack of historical data in automated feedback-quality evaluation. The core idea of the method is to treat the problem of detecting quality feedback as a natural language inference (NLI) task and utilize the pre-trained BART-based NLI model (Yin et al., 2019) to assess feedback quality. Our results show that IZSL can achieve acceptable performance in the "cold-start" condition on different datasets, and IZSL can substantially outperform BERT (Devlin et al., 2018) after training on the same incrementally collected data.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes datasets. Section 4 elaborates on our IZSL method for assessing feedback quality. Section 5 presents experimental results. Section 6 concludes the paper and provides some discussion about future work.

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 46 - 50

Peer-Feedback Comments	Sugg.	Prob.
No model tests have been added. Basic controller tests generated by scaffold and devise are available.	0	1
The design is written great. It will be better to explain more about the pattern used.	1	0
A little short. Make the conclusion more powerful and mention how you would address it as a teacher.	1	1

Table 1: Sample data. The first two samples are from *CS-Peer-Feedback*. The last sample is from *Ed-Peer-Feedback*. "Sugg." and "Prob." indicate whether the comment provides suggestions and mentions problems, respectively.

2 Related Work

2.1 Automated Peer-Feedback Assessment

Automated peer-feedback assessment is defined as a task of automatically analyzing peer-feedback comments written by students and highlighting low-quality comments that need to be revised. The goal of the task is to improve the overall quality of peer feedback and consequently improve students' learning. As the first step towards building an effective automated peer-feedback assessment system, Cho (2008) pioneered various machine-learning methods to classify peer-feedback units.

Subsequent work typically focused on designing more sophisticated features or using deep-learning algorithms to improve the performance. For example, Xiong and Litman (2011) designed features to represent feedback by combining generic linguistic features and specialized features. Ramachandran et al. (2017) utilized word-order graphs to represent review texts to assess the quality of feedback. Xiao et al. (2020b) leveraged various deep-learning approaches to detect whether the peer-feedback comments contain problem statements.

After that, researchers have noticed and tried to address the problem of lacking training data for new curricula. For instance, Xiao et al. (2020a) attempted to reduce the need for domain-specific data by applying transfer-learning and active-learning techniques. Jia et al. (2021) proposed to leverage multi-task learning to alleviate the problem. Despite the fact that these techniques can considerably reduce the need for historical data, none of them can help when we do not have any domain data.

2.2 Zero-shot Learning

Traditionally, zero-shot learning most often refers to the task of training a classifier on one set of labels and then evaluating it on a different set of labels that the classifier has never seen before (Wang et al., 2019). With the emergence of the pre-training and fine-tuning paradigm, "zero-shot learning" has been generalized to refer to the situation where a pre-trained language model is used to predict for a downstream task that it was not even fine-tuned on.

Yin et al. (2019) proposed to use a pre-trained NLI model as an out-of-the-box zero-shot text classifier and achieved promising results. A major advantage of this method over other zero-shot learning methods (e.g., Schick and Schütze, 2020) is that NLI-based zero-shot learning does not need access to task-specific hand-crafted prompt sentences.

3 Dataset

We captured data from Expertiza. In this system, learners can submit their work and write feedback comments on peers' submissions based on a set of rubric prompts. For example, each reviewer might be asked to provide a comment for the criterion, "Does the design incorporate all of the functionality required?" In this paper, the terms "feedback comments," "review comments," and "peer feedback" are used interchangeably to mean the textual responses to criterion in the rubric.

We obtained two datasets from the aforementioned peer-review platform for this study. The first dataset, *CS-Peer-Feedback*, is derived from a graduate-level object-oriented development course. This dataset consists of 12,053 data points and is mildly imbalanced. The second dataset, *Ed-Peer-Feedback*, comes from a graduate-level education course. The dataset contains 172 data points and is also mildly skewed. Some sample peer-feedback comments are displayed in Table 1.

All feedback comments have been manually annotated by a fluent English speaker who is familiar with the course context. To measure the reliability of the labels, we randomly sampled 100 comments from each dataset and asked a second annotator to annotate them. We measured the inter-annotator agreement on each set of 100 randomly selected samples using Cohen's κ coefficient. The average κ scores for the CS-Peer-Feedback dataset and the Ed-Peer-Feedback data were 0.88 and 0.85, respectively. These scores suggest that the annotations are reliable (Cohen's $\kappa > 0.81$ (McHugh, 2012)).

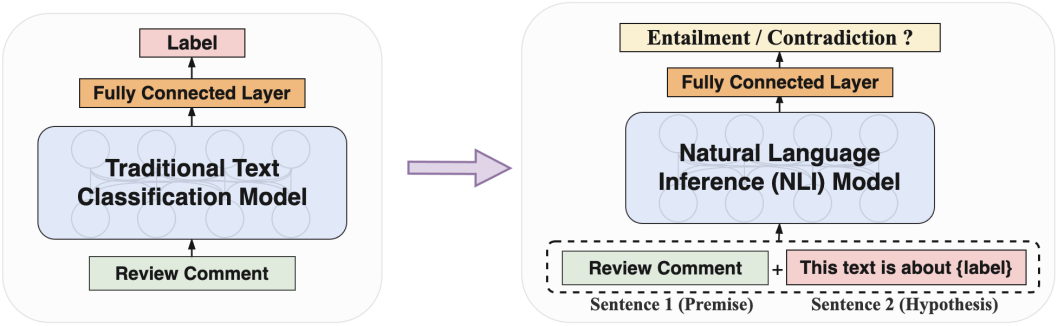

Figure 1: The key idea of IZSL is to convert the problem of evaluating peer-feedback comments into an NLI problem. The left part shows the traditional text classification setting for assessing feedback comments. The right part shows the NLI setting that treats the peer feedback as the premise and uses the label to formulate the hypothesis.

4 Methodology

4.1 Problem Formulation

We formulate the task of evaluating peer-feedback comments as follows: suppose that during the t-th semester of a class, we could collect a dataset $D^t = (X^t, Y^t)$ consisting of N^t data samples, where $X^t = \{x_1^t, x_2^t, \ldots, x_{N^t}^t\}$ denotes a set of N^t feedback comments collected in the t-th semester, and Y^t denotes corresponding labels indicating whether the feedback provides suggestions and/or mentions problems. In practice, these annotations can be obtained by, e.g., asking reviewees to determine if the received feedback contains the features. Additionally, it is worth noting that the labels Y^t can only be used for training after they are collected, i.e., after the t-th semester. **In the "cold-start" condition** (i.e., without any historical data, in the 0-th semester), the task of IZSL is to craft a classifier $\mathcal{F}_{IZSL}$ that can effectively make predictions for feedback comments X^0 without using any domain data to train the model. **In the incremental learning phase** (i.e., $t > 0$), we would have historical data $D^{<t}$, where $D^{<t}$ means $(D^0, D^1, ..., D^{t-1})$ (the data we collected in the first $(t-1)$ semesters). The task of IZSL in this phase is to update the classifier $\mathcal{F}_{IZSL}$ using all historical data $D^{<t}$ and to predict more accurately the labels for peer-feedback comments X^t.

4.2 Incremental Zero-shot Learning

We now describe our IZSL approach for classifying feedback comments. As shown in Figure 1, the overall idea of IZSL is to convert a text classification problem into a natural language inference (NLI) problem. NLI is the task of determining whether, given a premise, a hypothesis is true (entailment) or false (contradiction). We typically treat the text to be classified (i.e., feedback comments) as the premise, and construct the hypothesis from the class name of the label, "This text is about {label}," where "{label}" can be "suggestions" or "problems". If the NLI model tells us that the premise is likely to entail the hypothesis, we can conclude that the label is associated with the input feedback comment and vice versa.

We use BART (Lewis et al., 2019) to craft the NLI model and initialize all parameters with the "bart-large-mnli" checkpoint[1] (Yin et al., 2019), which is pretrained on the multi-genre NLI (MNLI) dataset. **In the "cold-start" condition**, using the pretrained weights makes us have an out-of-the-box NLI model for assessing feedback quality for any curriculum without needing historical data. This is not possible for traditional text classification models, since they need domain data to tune the output fully-connected layer. Then, **in the incremental learning phase**, we use incrementally collected data to further fine-tune the NLI model.

4.3 Baseline Classification Method

Although traditional text classification models cannot be applied in the "cold-start" condition, a BERT-based classifier is implemented to compare the performance of IZSL in the incremental learning phase. We build the classifier by stacking a dense layer on top of BERT. The parameters of BERT are initialized using a pretained checkpoint,[2] and the weights of the dense layer are randomly initialized using the uniform distribution. Then, we fine-tune the model utilizing the same incremental acquisition data as when fine-tuning IZSL.

[1] https://huggingface.co/facebook/bart-large-mnli
[2] https://huggingface.co/bert-base-uncased

Data	Model	Suggestions				Problems			
		F_1	P	R	AUC	F_1	P	R	AUC
0	IZSL	**61.2**±2.4	**70.1**±2.8	**59.6**±1.7	**73.5**±3.2	**60.9**±2.4	**63.1**±2.5	**61.2**±2.2	**70.3**±2.3
50	BERT	63.0±16.0	64.8±8.7	68.7±22.2	82.6±10.6	63.4±5.7	69.9±3.2	66.2±2.5	79.9±3.6
	IZSL	**91.5**±1.3	**90.0**±2.3	**94.5**±2.7	**97.5**±1.1	**84.8**±2.0	**85.4**±1.5	**85.0**±2.2	**93.4**±1.0
100	BERT	65.4±14.1	63.6±15.7	69.6±14.1	85.7±2.3	69.5±9.9	73.7±9.0	71.2±8.4	81.0±10.1
	IZSL	**92.3**±1.5	**92.5**±2.5	**92.2**±2.7	**97.1**±1.3	**87.0**±2.8	**87.7**±2.6	**87.0**±2.8	**94.1**±1.4
250	BERT	77.9±7.4	76.2±7.0	82.2±7.6	90.9±4.8	83.1±7.1	83.1±7.0	83.6±7.0	90.0±6.1
	IZSL	**93.5**±1.5	**92.8**±2.9	**94.4**±0.9	**97.9**±0.6	**87.9**±0.8	**87.8**±0.9	**88.4**±1.0	**94.4**±0.7
500	BERT	81.1±7.2	79.2±8.2	84.6±4.4	93.0±4.1	87.3±1.2	87.4±1.2	87.3±1.4	93.5±0.9
	IZSL	**93.5**±0.8	**92.8**±1.3	**94.2**±1.2	**98.2**±1.0	**89.1**±1.0	**89.1**±1.1	**89.1**±1.0	**94.9**±0.8
750	BERT	90.7±0.5	90.4±1.5	91.0±1.2	97.6±0.4	87.2±5.9	86.6±7.8	88.2±3.2	94.3±1.4
	IZSL	**93.7**±1.9	**92.7**±3.2	**94.9**±0.6	**98.2**±0.5	**90.2**±0.4	**90.2**±0.6	**90.3**±0.4	**95.6**±0.2
1000	BERT	91.7±1.0	90.5±1.1	93.2±2.0	98.1±0.9	88.8±1.0	88.8±0.9	88.9±1.1	94.6±0.6
	IZSL	**93.8**±0.9	**92.7**±1.4	**94.9**±0.9	**98.2**±0.4	**90.4**±1.3	**90.2**±1.2	**90.7**±1.6	**95.9**±1.3

Table 2: Performance evaluation of BERT (baseline) and IZSL on *CS-Peer-Feedback*. The first column is the number of training samples used. The best results in each setting are marked in bold. Confidence interval = 95% .

Data	Model	Suggestions				Problems			
		F_1	P	R	AUC	F_1	P	R	AUC
0	IZSL	**60.5**±2.0	**67.1**±3.2	**59.6**±1.6	**68.8**±0.7	**57.2**±2.3	**57.5**±2.2	**59.6**±3.0	**64.4**±2.3
50	BERT	52.1±14.7	51.3±18.0	56.7±10.9	69.6±9.0	56.3±11.7	59.8±20.4	56.9±7.6	67.4±7.6
	IZSL	**78.1**±3.3	**76.9**±2.7	**82.0**±6.8	**87.8**±1.4	**81.7**±2.7	**84.2**±3.4	**80.4**±4.7	**94.2**±1.0
100	BERT	68.7±14.8	77.6±11.7	68.5±15.6	80.3±9.2	62.0±14.6	66.3±21.8	64.3±10.5	75.8±15.0
	IZSL	**82.2**±1.8	**80.7**±1.8	**86.1**±4.2	**90.8**±1.3	**84.3**±3.8	**87.5**±3.6	**82.5**±5.0	**93.4**±2.2

Table 3: Performance evaluation of BERT (baseline) and IZSL on *Ed-Peer-Feedback* with 95% confidence interval.

5 Evaluation

5.1 Experimental Setup

Training and Optimization Details. We train our models on eight NVIDIA RTX6000 GPUs (24GB each) with a total batch size of 8, a learning rate of 2e-5/3e-5/5e-5, epochs of 2/3, and the Adam optimizer (Kingma and Ba, 2014).

Handling the Imbalanced Datasets. To alleviate the problem of class imbalance, we employ a cost-sensitive approach. Specifically, we weight the cross-entropy loss function based on the frequency of each class in the training set.

5.2 Results and Discussion

The evaluation results are shown in Tables 2 and 3. The first row (i.e., for "Data" = 0) of each table shows the performance of IZSL when we do not have any historical data. Then, the following rows of each table compare the results of IZSL and BERT when trained with incrementally collected data. **In the "cold-start" phase**, the F_1 scores for the labels "Suggestions" and "Problems" on the CS-Peer-Feedback dataset are 61.2 and 60.9, respectively. On the Ed-Peer-Feedback dataset, the F_1 scores for these two labels are 60.5 and 57.2, respectively. The results suggest that IZSL can achieve acceptable "cold-start" performance on data from different disciplines, considering that it does not use any domain data. However, it is worth noting that the performance of the IZSL model varies on datasets from different domains. It is still unclear how we can estimate the "cold-start" performance of IZSL on a particular dataset. We leave this research question to future studies. **In the incremental learning phase**, we surprisingly find that the F_1 scores of IZSL quickly jump to over 91.5 and 84.8 on the CS-Peer-Feedback dataset after training with only dozens of training samples, and we make a similar finding on the Ed-Peer-Feedback dataset. Our hypothesis for IZSL to perform better than BERT in "low-data" settings is that NLI-based classification models have better generalization ability than traditional classification methods. However, this hypothesis needs to be further tested by extensive experiments. By examining the following rows of the tables, the results clearly show that IZSL can consistently outperform BERT on all metrics across all settings, and the confidence intervals suggest that the performance of IZSL is more stable. **To summarize**, IZSL can achieve acceptable "cold-start" performance and consistently outperform the BERT model in the incremental learning phase, especially when we only have dozens of incrementally collected data points.

6 Conclusion and Future Work

The quality of peer feedback plays a vital role in peer assessment. However, lacking historical data for new curricula is a persistent problem. Our work proposes a novel method for assessing feedback quality by converting it into an NLI problem. The approach can potentially be generalized to other pedagogical tasks. Future plans include investigating how to improve "cold-start" performance.

References

Scott Ashton and Randall S Davies. 2015. Using scaffolded rubrics to improve peer assessment in a mooc writing course. *Distance education*, 36(3):312–334.

Kwangsu Cho. 2008. Machine classification of peer comments in physics. In *Educational Data Mining 2008*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kit S Double, Joshua A McGrane, and Therese N Hopfenbeck. 2020. The impact of peer assessment on academic performance: A meta-analysis of control group studies.

Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F Gehringer. 2021. All-in-one: Multi-task learning bert models for evaluating peer assessments. *arXiv preprint arXiv:2110.03895*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Melissa M Nelson and Christian D Schunn. 2009. The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, 37(4):375–401.

Lakshmi Ramachandran, Edward F Gehringer, and Ravi K Yadav. 2017. Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education*, 27(3):534–581.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Marjo Van Zundert, Dominique Sluijsmans, and Jeroen Van Merriënboer. 2010. Effective peer assessment processes: Research findings and future directions. *Learning and instruction*, 20(4):270–279.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Yunkai Xiao, Gabriel Zingle, Qinjin Jia, Shoaib Akbar, Yang Song, Muyao Dong, Li Qi, and Edward Gehringer. 2020a. Problem detection in peer assessments between subjects by effective transfer learning and active learning. *International Educational Data Mining Society*.

Yunkai Xiao, Gabriel Zingle, Qinjin Jia, Harsh R Shah, Yi Zhang, Tianyi Li, Mohsin Karovaliya, Weixiang Zhao, Yang Song, Jie Ji, et al. 2020b. Detecting problem statements in peer assessments. *arXiv preprint arXiv:2006.04532*.

Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.