
Automatic Feedback Generation for Dialog-Based Language Tutors Using Transformer Models and Active Learning

Katherine Stasaski^{1*} and Vikram Ramanarayanan^{2,3}

¹UC Berkeley; work performed while interning at Educational Testing Service

²Educational Testing Service R&D

³University of California, San Francisco

<katie_stasaski@berkeley.edu, vramanarayanan@ets.org>

Abstract

We aim to provide non-native English learners with natural language feedback on the pragmatic appropriateness of their dialogic speech via a human-in-the-loop feedback generation model. We fine-tune a large, pre-trained transformer model on a small hand-crafted dataset of feedback paraphrases formulated from a scoring rubric. We then utilize an active learning pipeline with expert annotators to correct the model’s feedback. We find that human-rated quality and unigram diversity of generated feedback increases over time, indicating that the model improves *and* produces more diverse responses over each successive active learning iteration. Our results indicate the potential for active learning to improve targeted feedback generation at scale for language learning learners.

1 Introduction

Providing high-quality natural-language feedback to learners completing dialog exercises in an automated manner is a difficult task with the potential to greatly improve language learning experiences. Good feedback should be targeted to the learner’s response, interpretable (i.e., uses language the learner can easily understand), and actionable for the next exercise(s) the learner completes.

The goal of our work is to automatically provide language learners feedback on the pragmatic appropriateness of their responses throughout a dialog exercise, in which they converse with an automated interlocutor to practice English speaking skills. Pragmatic appropriateness includes features about the type of language a speaker uses (e.g. polite, respectful, language in the appropriate professional register) as well as linguistic strategies, such as hedging their requests.

Prior work in educational applications of machine learning have focused extensively on automated scoring of learner-completed dialogs [15]. While these systems are accurate at providing scores, they fail to provide personalized, actionable feedback to the learners. In an ideal world, one-on-one feedback would be given to every learner by an expert teacher; however, this process does not scale. Our proposed task bridges the gap between cheap and accurate but impersonal automated scoring and expensive one-on-one expert-generated feedback.

Large pre-trained transformer-based language models have made great strides in many areas of NLP [3, 23]. These models are trained on large amounts of text data, allowing them to learn syntax and vocabulary, and can later be fine-tuned on specialized, task-specific domain data. We make use of one of these models, XLNet [23], and fine-tune it for the purpose of generating natural language feedback on pragmatic appropriateness for dialogic text.

Our approach begins with a small dataset of hand-crafted pragmatic appropriateness feedback generated from an expert-created scoring rubric. From there, we fine-tune XLNet to take in the input of transcribed learner dialogs and produce the corresponding feedback. We run inference on a held-out data partition and select a subset of feedback for subsequent annotation using model uncertainty. We have human annotators correct the model-generated feedback and feed the corrections back into the training set for more fine-tuning.

We examine the following research questions:

- RQ1: Does the active learning process help the model improve over time?
- RQ2: Was the uncertainty-informed sample-selection strategy effective?

For RQ1, human-assigned quality ratings of model output improve as more active learning rounds are completed, indicating the model improves over time. Further, we obtain this improvement consistently across proficiency score levels (ranging from 1 to 4), allowing for higher-quality feedback to be generated over each active learning round for a variety of input dialogs, irrespective of the learner’s proficiency level. Semantic similarity between model- and human-generated feedback, as measured by the BERT Score [24] on a held-out, human-created test set of feedback does not improve from round-to-round; however, the diversity of responses generated does increase. We find that for RQ2, higher model confidence, measured via posterior probability of generated feedback, did not correspond to higher annotator-ranked quality.

Overall, our results indicate promise in using a human-in-the-loop system to improve a model’s feedback for learner dialogic responses. At the end of the active learning process, the majority of generated feedback by the model is labeled by annotators to be *useful*, providing preliminary evidence for future work to investigate learner interaction with the feedback-generation system.

Note that the work we present here is just one step in the direction of being able to provide targeted, useful feedback to students at scale. Providing such personalized feedback is a vast research area; this paper does not answer the broader research question of how to generate the best possible feedback a system could give a student from a pedagogical perspective. Rather, given XLNet without architecture modifications and an initial small, hand-created dataset, we are interested in how to utilize human corrections to produce the best possible feedback. The resulting feedback may not be useful for learners immediately, and future work should explore learner responses to feedback. Future work should also explore the broader area of using specialized neural models or modifications to XLNet to provide different types of feedback, such as suggesting specific modifications to learner’s phrases. Our work provides the basis for this to be done in the future.

2 Related Work

Large attention-based neural models have greatly improved word representations, leading to an improvement on a variety of NLP tasks [3, 11]. These models are trained on large corpora, allowing them to learn implicit representations of syntax and semantics. We aim to draw on these models by fine-tuning them on a specialized dataset of learner dialog transcripts and corresponding feedback.

Human-in-the-loop systems have been deployed to train dialog conversation agents, which converse with a human directly. For instance, Li et al. [11] train a question answering dialog agent using two rewards: a traditional numeric reward and natural language feedback for the model. This approach trains a dialog agent as the end goal and receives the reward as conversations are happening, compared to our model, which is aimed at providing feedback on an input dialog.

Additionally, past active learning research has explored the selection of limited samples to annotate from a dataset. Past approaches include those which propose selecting a subset of the dataset based on model uncertainty [9, 2] as well as including diversity constraints to incentivize coverage of annotated samples [18, 22]. We choose an uncertainty-based approach, as we hypothesize the model’s confidence correlates to the feedback quality.

In the intersection of NLP and educational dialogic speech, both neural and regression models have been relatively successful at predicting the score of a learner’s dialog [15], essays [17, 1], and short answer questions [16, 7]. Transformer models have been applied to score essays as well [12]. While this application is beneficial for evaluation of learner exercises, the score does not provide the specific, actionable feedback that natural language feedback can.

3 Dataset

We work with the Request Boss dataset [14, 13], which consists of non-native English speaker dialogs with an automated interlocutor. Participants are told to interact with their boss, who is played by the automated system. Participants have 2 goals throughout the conversation: (1) to schedule a meeting with their boss, and (2) to have their boss review their presentation slides before the meeting. The automated system guides them through the conversation, and their dialogic speech is recorded. For the purpose of our experiments, we work with a manually transcribed version of this dataset.

3.1 Statistics

We filter the dataset to only those with a valid score and non-empty dialog transcription, excluding conversations where technical issues prevented learner dialog from being transcribed. The resulting dataset includes 1,767 conversations. Each conversation is triple-scored by experts with extensive experience scoring standardized assessments of spoken English. The scores range between 1 and 4 along 13 distinct dimensions, covering the broad constructs of linguistic control, topic development, and interaction [13]. For our purposes, we focus on the *pragmatic appropriateness* sub-construct – see Appendix A for the scoring rubric used. We partition the dataset into 40% initial training data, 20% active learning data, 20% development data, and 20% testing data.

3.2 Artificial Data Creation

In order to create an initial training dataset to fine-tune the model, we construct a set of feedback which we pair to each dialog, D in the initial training set partition.

We begin by constructing an initial dataset of sentences based on the official scoring rubric used by annotators to score learner dialogs [13]. The rubric includes potential feedback for each score s (ranging from 1-4); see Appendix A. For each of the sentences, a paraphrase model is used to generate diverse rewordings [20, 4]². Because of the adverse effect of displaying an incomprehensible piece of feedback to a language-learning learner, we hand-select the grammatically correct paraphrases out of a set of 10 that are generated. This forms a set of responses R . We represent each (disjoint) subset of R characterized by a score s as R^s . In other words, $R = R^1 \cup R^2 \cup R^3 \cup R^4$.

To construct the training set, we first look up the score s for each conversation $c \in D_{train}$. Given s , we randomly sample 3 responses from R^s and pair with the original conversation c . This forms the initial training set, D_{train} . We follow the same process for the development dataset.

One limitation with this dataset creation method is that feedback in R^s might not apply to any conversation c which is assigned a score of s . Good feedback is as specific to a learner dialog as possible, so a given feedback $\in R^s$ might address an improvement that might not apply to c . However, as the active learning process should improve the generated feedback, we chose this approach to bootstrap an initial training dataset from scratch.

4 Methods

We propose a human-in-the-loop system to fine-tune a text generation model to provide pragmatic appropriateness feedback, visualized in Figure 1. We pre-train our model on the artificially-constructed dataset and complete 3 active learning loops to improve the model’s feedback with human corrections.

4.1 Model Specifics

The generation model, XLNet [23], is implemented in Huggingface’s transformers library [21] with a language modeling head for sequential generation. We utilize the XLNet tokenizer and the pre-trained *xlnet-base-cased* model, which has 12 layers and 12 heads. We choose to work with the *xlnet-base-cased* model because it allows us to utilize the pre-trained model weights and train on a smaller training dataset. We train each active learning iteration for 3 epochs, using the Adam optimizer [5] with an initial learning rate of $1e-5$, decay learning rate based on the development set performance, and use a batch size of 16 (given the small size of our dataset). We use a max input

²<https://github.com/laoli2046/paraphraser-1>

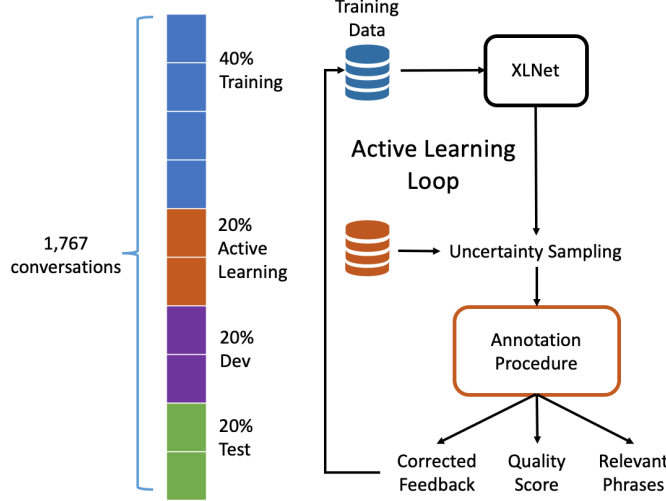


Figure 1: Active learning process diagram. The Active Learning Loop happens 3 total times.

length of 311 and a max output length of 73. We choose these values by tokenizing the artificial training dataset and choosing the maximum lengths of the learner’s utterances and scoring rubric, respectively. The training input is a concatenation of all of the learner’s dialog utterances.

4.2 Uncertainty Sampling

After training the model, we run inference over the remaining examples in the active learning partition of the dataset. We utilize the posterior probability of each generated sequence, normalized by the length of the generated sequence, to prioritize which samples we should collect corrections for. For each generated feedback sequence, $p(seq)$ is defined as:

$$p(seq) = \frac{1}{len(seq)} \sum_{t \in seq} \log(p_{model}(t|t_{-1}, t_{-2}, \dots, t_0))$$

Similar to past work which emphasizes uncertainty sampling [9, 2], we hypothesize that the least-probable generated feedback samples are those which are likely to be of poor quality and require human correction. We select the lowest 100-scored samples for round 1 and round 2 for annotation; for round 3, we choose to annotate all of the remaining 147 samples. For round 1, we randomly select 5 examples to be annotated by all annotators, to examine diversity in corrections given.

4.3 Annotation Procedure

For each active learning iteration, we show the model’s generated feedback to a set of annotators. Annotation consists of three components, also represented pictorially in Figure 1: scoring the model’s feedback, correcting the model’s feedback, and making note of relevant key words and phrases from the learner’s dialog that contribute to the feedback.

We first ask annotators to rate the quality of the model’s feedback in terms of its potential usefulness to the language learner on a 3-point Likert scale, where -1 corresponds to “not useful,” 0 corresponds to “unclear if useful or not,” and 1 corresponds to “useful.”

Next, we instruct annotators to correct the model’s feedback, producing *corrected feedback*. We ask annotators to alter the fewest words possible, as we hypothesize the model will produce better updates if the corrections are most tied to the model’s original feedback. As we are working with a relatively small dataset, we desire every sample to produce targeted model updates.

The final component of annotation includes selecting *relevant phrases* from the learner’s conversation that relate to the *corrected feedback*. These key phrases may indicate components of the learner’s response that should be improved or were particularly appropriate. The collection of these key phrases

also allows us to examine the extent to which the model attends to these key words/phrases, and therefore contribute to model interpretability.

In addition to correcting the model, we have one annotator construct a test set of size 104, where feedback and key phrases are generated from scratch without being biased by model generations.

We utilize 7 human annotators to complete annotations, as providing feedback on pragmatic appropriateness is a specialized task. However, future work can explore the feasibility of employing crowdworkers to provide this feedback at a larger scale.

4.4 Model Updates

Pairing each conversation with the newly-collected *corrected feedback*, we add the samples to the training set for retraining. We upsample each conversation in the most-recent active learning round by 3 and for all other active learning rounds by 2. At the end of training, we have completed 3 total active learning cycles, resulting in 4 models for comparison. Due to timing constraints, for round 2, we only add 80 of the possible 100 annotations to the training set. However, we include all 100 annotator quality ratings and *corrected outputs* in the analysis.

5 Results and Discussion

Our results are focused on the following two research questions:

- RQ1: Does the active learning process help the model improve over time?
- RQ2: Was the uncertainty-informed sample-selection strategy effective?

5.1 RQ1: Model Improvement

This section examines how the model’s feedback improves over multiple active learning iterations both on the active learning and test partitions of the dataset.

5.1.1 Active Learning Set Results

The most reliable measure of model improvement is the human-rated quality scores the annotators complete for each round. We measure the proportion of -1, 0, and 1 quality ratings given to the model’s feedback for each round. Results can be seen in Figure 2 (left). There is a shift as rounds increase to higher scores overall, indicating quality increases with more active learning rounds. More notably, the most prevalent score from Round 3 was 1, compared to -1 for Rounds 1 and 2. Additionally, the majority of Round 3 feedback was rated as useful.

Furthermore, we examine the quality ratings for each ground truth score of the input learner dialog (ranging from 1 to 4). Providing quality feedback to conversations scoring 1-3 are arguably more important, as these conversations have room for improvement. Figure 2 (right) shows the quality ratings broken down by gold dialog score. We note that Rounds 2 and 3 do not include dialogs scoring 1. While this category contained the fewest samples, it is potentially interesting that uncertainty sampling selected all for annotation in the first round. Comparing input scores of 2-4, we note consistent improvement among all score categories for Rounds 2 and 3. This is desirable, as we do not want performance gains to disproportionately affect one group, especially those which can most be improved by feedback. Future work could explore a sampling strategy informed by input scores.

We additionally quantify the number of edits made by the annotators, measured by the Levenshtein edit distance at the word-level with a penalty of 1 for substitutions [8]. We normalize the edit distance by the length of the model’s generated sentence, as to not unnecessarily penalize longer generated sentences. Results can be seen in Figure 3, which measures the proportion of samples falling into a given category of edit distance. We note a slight leftward shift as rounds increase, indicating fewer edits need to be made as the model improves.

To summarize, by Round 3, the majority of feedback generated by the model was rated as useful, which provides evidence that our active learning process can train a useful feedback generation model. However, an important factor to note is that the conversations were chosen with uncertainty sampling at each stage of generation. Thus, there may be an additional factor of the sampled set changing at each round which could influence quality scores. Future work should examine this more explicitly.

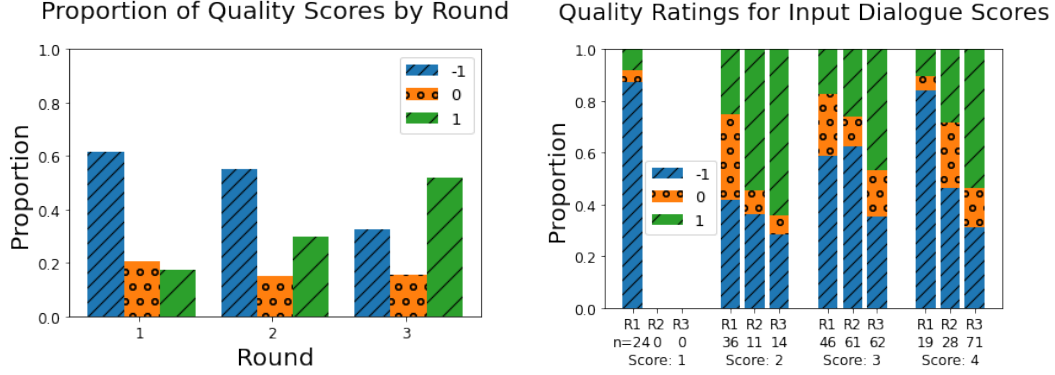


Figure 2: Left: Annotator-scored quality ratings over rounds. As rounds increase, quality ratings shift to be more positive, indicating model improvement. Right: Quality ratings segmented by ground truth score of original learner dialog. Bars are grouped by input dialog score, with rounds progressing from left to right within a group. The number of dialogs in a category is noted under the round indicator.

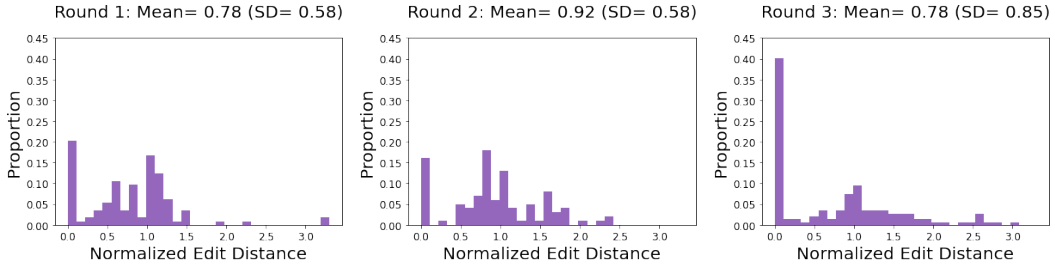


Figure 3: Edit distance normalized by input length of annotated samples; rounds progress from left to right. Lower scores indicate fewer changes made by annotators.

5.1.2 Test Set Results

As the active learning set changes from round-to-round, we provide additional analysis using the human-created test set. In particular, we report two semantic similarity metrics based on word embedding comparison: Word Mover Distance (WMD) [6] and BERT Score [24]. WMD uses context-independent embeddings, while BERT Score uses contextualized BERT embeddings. Appendix B includes selected model-generated feedback for two dialogs that received different proficiency scores.

In addition to the four models trained via active learning, we train a baseline model which does not make use of active learning. We train this model on the combination of all annotator *corrected outputs*, without artificial data. This model has the same parameter specifications as the active learning model and compares the effect of using artificial data and updating the model incrementally.

We report the WMD [6] of the model-generated output compared to the test set in Figure 4 (left) and BERT score [24] in Figure 5 (left). As the standard BERT score might not include nuanced representations for our domain-specific language, we additionally report BERT score results using the XLNet representation of a model trained to predict only the human-generated feedback. Figure 6 (left) includes these results; we note they are much higher than the generic BERT scores.

For all three automated metrics, we note the lack of significant increase in performance as more active learning rounds are completed. Additionally, the baseline model provides similar performance for all metrics. However, automated quality metrics like WMD or BERT score capture a limited component of what we desire from a feedback system. If, for instance, the model produced vague feedback, an automated quality score might assign a high score. But, the lack of diversity in model’s output is undesirable. Thus, we incorporate unigram diversity [10] into our quality metrics. Figures 4, 5, and 6 all include a version of the graph with the original values scaled down by each model’s unigram diversity scores, seen on the right. We note that combining each quality metric with diversity results in a better model at the end of training, surpassing baseline performance as well. However, we would

expect unigram diversity to increase with the collection of more data; future work should examine the usefulness of this diversity increase as it relates to accuracy and specificity of feedback.

For the diversity-scaled domain-specific BERT Score distributions, we examine potential distributional differences among gold learner dialog scores. Ideally, each model would generate similarly-scored feedback for all 4 potential input learner scores. The test set contains 5, 15, 52, and 32 dialogs scoring 1-4, respectively. We perform a Welch's t-test [19] comparing BERT Score distributions for each pair of input dialog scores for each of the 5 models and find no significant difference. This is promising, as it demonstrates that at each point along the active learning pipeline and in our baseline approach, the input score category does not have a significant impact on this metric.

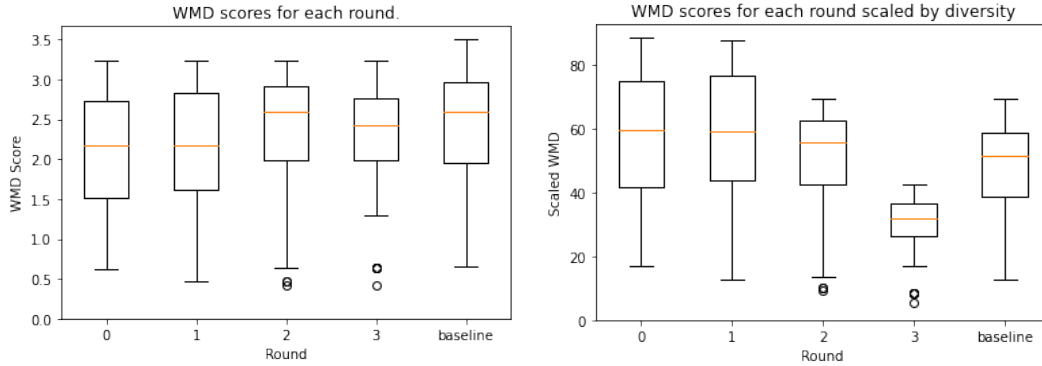


Figure 4: WMD on test set (left: unscaled, right: scaled by diversity); lower is better.

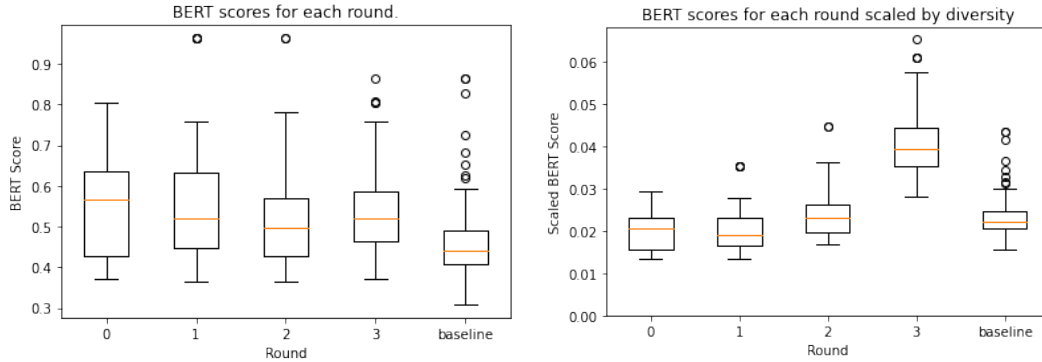


Figure 5: BERT Scores on test set (left: unscaled, right: scaled by diversity); higher is better

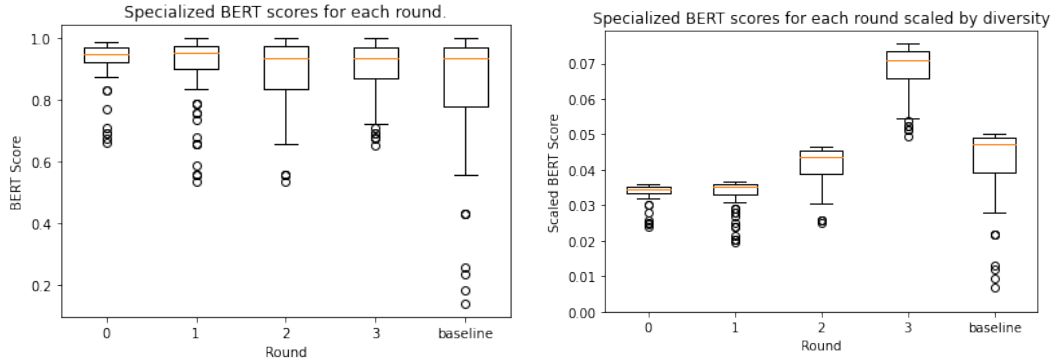


Figure 6: BERT Scores using domain-specific XLNet on test set (left: unscaled, right: scaled by diversity); higher is better.

5.2 RQ2: Uncertainty-Informed Sampling

We additionally examine our assumption behind uncertainty-based sampling, i.e., that lower model confidence corresponds to lower quality feedback. We plot the corresponding normalized posterior probability confidence scores for each of the 3 quality score categories for each round in Figure 7. We note that higher model confidence does not correspond to higher quality rating as hypothesized. One caveat to note is rounds 1 and 2 include sampling 100 pieces of feedback from the active learning set, which is decreasing in size over time. Therefore, these results should be interpreted with the knowledge that over time, the set we sample from decreases.

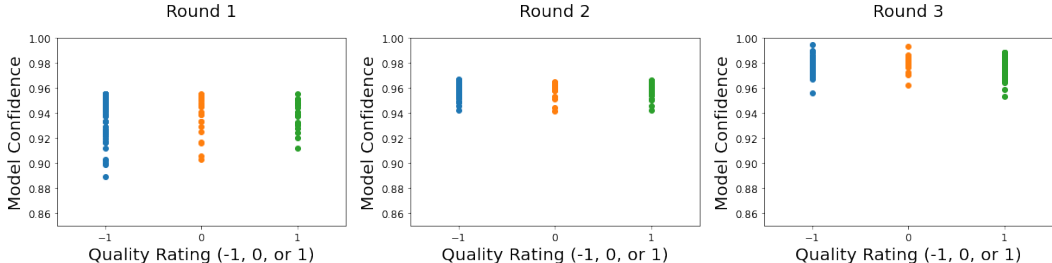


Figure 7: Confidence score distributions corresponding to each of 3 quality scores.

6 Future work and Limitations

While our results were promising and show that feedback quality and unigram diversity improves over multiple rounds of active learning, a more extensive study should be conducted to test components of the active learning process (e.g. model type, sampling method, annotation procedure). Diversity-informed sampling should be explored, as our results indicate the model is confident about poor feedback and this work has not investigated whether uncertainty-based sampling is more useful than random sampling. Additional experimentation should also examine the tradeoff between collecting more data at each round of active learning and allowing the model to perform more total updates. While expert annotators allowed us to have confidence that the model corrections and ratings were of high quality, this process does not scale to allow for this experimentation. Future work can explore the feasibility of utilizing crowdworkers to provide scalability for extensive experimentation.

Future work should explore a deeper error analysis of the rankings and edits made by the annotators. Annotators can label the model’s feedback as not useful for a variety of reasons, e.g. the feedback being unclear, not specific enough, or not applicable. Additionally, examining the types of edits being made and specific words or phrases being replaced or added could yield additional insight.

We collected key phrases that annotators marked as relevant to their corrected feedback. Future work can explore potential alignment between model attention and these key phrases as well as the effect of biasing our model to attend to these phrases. Additional future work could involve a second model to predict these phrases, to potentially include an additional type of feedback for learners.

7 Conclusion

We show that a transformer model can be trained via active learning to automatically generate feedback on the pragmatic appropriateness of dialogic text. Our preliminary results are promising; human quality ratings and unigram diversity scores increase with active learning rounds. Future work can experiment with the impact of showing this targeted feedback to language learners, to bridge the gap between vague, automated feedback generation and expensive, personalized one-to-one tutoring.

8 Acknowledgements

We thank Sozen Ozkan, Patrick Houghton, Zydrune Mladineo, Ian Blood and Melissa Fowler for help with the annotation process. We also thank the ETS NLP and Speech Research group, the Hearst Lab Research group at UC Berkeley, and the 3 anonymous reviewers for their helpful feedback.

References

- [1] Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).
- [2] Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI’05, page 746–751. AAAI Press.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- [4] Jonathan Mallinson John Wieting and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of Empirical Methods in Natural Language Processing*.
- [5] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- [6] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML’15, page 957–966. JMLR.org.
- [7] Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.
- [8] V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics. Doklady*, 10:707–710.
- [9] David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148 – 156. Morgan Kaufmann, San Francisco (CA).
- [10] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. 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.
- [11] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc’Aurelio Ranzato, and Jason Weston. 2017. Dialogue learning with human-in-the-loop. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- [12] Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA - Online. Association for Computational Linguistics.
- [13] Vikram Ramanarayanan. 2020. Design and Development of a Human-Machine Dialog Corpus for the Automated Assessment of Conversational English Proficiency. In *Proc. Interspeech 2020*, pages 419–423.
- [14] Vikram Ramanarayanan, Patrick L. Lange, Keelan Evanini, Hillary R. Molloy, and David Suendermann-Oeft. 2017. Human and automated scoring of fluency, pronunciation and intonation during human-machine spoken dialog interactions. In *Proc. Interspeech 2017*, pages 1711–1715.

- [15] Vikram Ramanarayanan, Matthew Mulholland, and Yao Qian. 2019. Scoring interactional aspects of human-machine dialog for language learning and assessment using text features. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 103–109, Stockholm, Sweden. Association for Computational Linguistics.
- [16] Jana Zuheir Sukkarieh and John Blackmore. 2009. C-rater: Automatic content scoring for short constructed responses. In *Twenty-Second International FLAIRS Conference*.
- [17] Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.
- [18] G. Wang, J. Hwang, C. Rose, and F. Wallace. 2017. Uncertainty sampling based active learning with diversity constraint by sparse selection. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6.
- [19] B. L. WELCH. 1947. THE GENERALIZATION OF ‘STUDENT’S’ PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. *Biometrika*, 34(1-2):28–35.
- [20] John Wieting and Kevin Gimpel. 2017. Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *arXiv preprint arXiv:1711.05732*.
- [21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- [22] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G. Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *Int. J. Comput. Vision*, 113(2):113–127.
- [23] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.
- [24] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Scoring Rubric

Table 1 contains the language from the scoring rubric that we utilize to create our artificial training dataset.

Score	Rubric
4	Consistently reacts to the dialog system in a positive and respectful way according to the norms for North American English, including acknowledging questions, hedging statements, recognition of the value of the interlocutor’s opinions, use of positive intonation, and recognition of power differentials, where applicable. Makes consistent use of the appropriate academic/professional register.
3	Generally acknowledges the automated interlocutor’s ideas. May occasionally convey a more aggressive stance, such as by direct challenges to interlocutor statements/questions or less positive intonation.
2	Shows some ability to establish positive and respectful rapport with social expressions or gesture. But, politeness features may be lacking (e.g., unhedged statements, less positive intonation or expression) or contributions may communicate an overly aggressive stance.
1	Little use of hedging or other strategies to soften the impact of statements or counter arguments. Tone/gesture may convey negativity or impatience.

Table 1: Pragmatic appropriateness scoring rubric, used to create initial feedback dataset.

B Selected Model Feedback

Tables 2 and 3 shows model and human responses for a given conversation (with ground truth scores of as 3 and 4, respectively) and were hand-selected as they exhibited interesting phenomena. Table 2 demonstrates the same feedback from a semantic perspective (responding in a positive way) being expressed consistently throughout active learning rounds; however, the way this is expressed changes. All models, however, don’t capture the human feedback which points out a less respectful phrase. Table 3 demonstrates the Round 1 and 2 models providing negative feedback when the human-generated feedback is positive. By Round 3, the model has returned to positive feedback.

Original Conversation	System: Hello? User: hey lisa i was just wondering if you could you know agree to have a meeting with me System: Yeah, sure I’m available on Friday at 12. Does that work for you? User: yes okay i just want to you know go through some of my presentation slides before the meeting System: Sure, no problem. Send them over. User: will do thank you very much System: Okay, see you Friday then. User: see you on friday
Human	You responded to Lisa in a positive way but you may have used a less respectful phrase.
Round 0	We heard you answer Lisa in a respectful and positive way.
Round 1	You responded to Lisa in a positive way.
Round 2	We heard you answer Lisa in a respectful and positive way.
Round 3	You might have acknowledged questions, hedged statements, recognized the value of Lisa’s opinions, used positive intonation, and recognized power differentials.
Baseline	You responded to Lisa in a positive way

Table 2: Hand-selected model responses and corresponding conversation (ground truth score of 3).

Original Conversation	System: Hello? User: hello System: Hi, how's it going? What can I do for you? User: i am good how are you um i was wondering if i can schedule a meeting with you System: Yeah, sure I'm available on Friday at 12. Does that work for you? User: yes friday works for me um i was wondering if you can um review my slides before the meeting System: Sure, no problem. Send them over. User: okay sounds good ill see you friday then oh
Human	Your tone was appropriate.
Round 0	You responded to Lisa in a positive and respectful way.
Round 1	You may have conveyed an aggressive stance.
Round 2	You may have directly challenged Lisa's statements.
Round 3	You used appropriate professional language.
Baseline	You responded to Lisa in a positive way

Table 3: Hand-selected model responses and corresponding conversation (ground truth score of 4).