Navigating Wanderland: Highlighting Off-Task Discussions in Classrooms

Ananya Ganesh[1](✉), Michael Alan Chang[2], Rachel Dickler[1], Michael Regan[1], Jon Cai[1], Kristin Wright-Bettner[1], James Pustejovsky[3], James Martin[1], Jeff Flanigan[4], Martha Palmer[1], and Katharina Kann[1]

[1] University of Colorado Boulder, Boulder, CO, USA
ananya.ganesh@colorado.edu
[2] University of California Berkeley, Berkeley, CA, USA
[3] Brandeis University, Waltham, MA, USA
[4] University of California Santa Cruz, Santa Cruz, CA, USA

Abstract. Off-task discussions during collaborative learning offer benefits such as alleviating boredom and strengthening social relationships, and are therefore of interest to learning scientists. However, identifying moments of off-task speech requires researchers to navigate massive amounts of conversational data, which can be laborious. We lay the groundwork for automatically identifying off-task segments in a conversation, which can then be qualitatively analyzed and coded. We focus on in-person, real-time dialog and introduce an annotation scheme that examines two facets of dialog typical to in-person classrooms: whether utterances are pertinent to the *lesson*, and whether utterances are pertinent to the *classroom*, more broadly. We then present two computational models for identifying off-task utterances.

Keywords: Collaborative learning · Classroom discourse · Off-task speech

1 Introduction

Towards supporting learning scientists in studying collaborative discourse, we hone in on the issue of *navigating* moments of collaboration that seemingly wander from the learning goals of an activity. Some previous work operates from the premise that such off-task remarks detract from collaborative effectiveness [4]. In contrast, others argue that off-topic behavior can be critical to collaborative learning because of the strengthening of the social relationships within a group [5]. Moreover, a better understanding of off-task behaviors is a key equity issue: Langer-Osuna demonstrates that collaborative wanderings often play a critical role towards elevating the voices of youth who are marginalized in the discussion because of their identity [5].

In this paper we propose a first step towards *automatically* spotlighting critical moments of interest where seemingly off-task interactions occur. Qualitative

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
N. Wang et al. (Eds.): AIED 2023, LNAI 13916, pp. 727–732, 2023.
https://doi.org/10.1007/978-3-031-36272-9_63

approaches towards analyzing interactions are usually constrained in speed and scale by a manual expert coding process [7]. Automated detection offers the opportunity to direct researchers to relevant segments of data for further examination [10]. However, prior work on computationally modeling off-task discussions [1,2,8] focuses on text-based environments where students communicate through chat messages, which can differ in quality from the spontaneous spoken dialog that qualitative researchers study. Additionally, prior work models on-task utterances from a "productivity" perspective, which may not reflect the nuances of real-time, in-person group work in classrooms.

Towards addressing these challenges, we make two main contributions in this paper. First, we introduce an **annotation scheme** suitable for in-person classroom environments, which examines two facets of collaborative dialog, which we call lesson-focused (LF) speech and classroom-focused (CF) speech. As shown in Table 1, lesson-focused speech focuses on discussions about the current learning activity, such as discussing ideas and solutions. Classroom-focused speech discusses any relevant classroom activity, including peripheral tasks such as team management. This distinction provides flexibility for modeling, such as using keywords, as well as for downstream analysis – for instance, lesson-focused speech could be examined for speech or skills related to problem-solving, while classroom-focused speech could shed light on team dynamics.

As our second contribution, we develop **two computational models** for detecting if utterances are lesson-focused, and classroom-focused respectively, which are trained on discussions from middle school science classrooms. Both models are transformer-based sequence classifiers, and perform classification at the utterance level to indicate if an utterance is focused, non-focused, or if it is undecidable given the input information. Experimenting with multiple input signals, we find that our best LF-classifier's performance is 0.59 F1, and our best CF-classifier's performance is 0.56 F1, both strongly outperforming random and majority-class baselines. Finally, we discuss the impact of design choices such as the amount of surrounding context used, in serving the needs of the user.

2 Methodology

2.1 Dataset

We use the dataset described in Southwell et al. [9], which consists of five-minute long transcripts of small-group discussions, collected from a middle-school science classroom in the United States. The subject of instruction is a curriculum unit called Sensor Immersion (SI) focused on "programmable sensor technology". Student interactions are recorded through desk-top mics, and manually transcribed and anonymized, yielding 32 transcripts with 2133 student utterances in total.

We then annotate utterances on a 3-point scale for the "lesson-focused" and "classroom-focused" facets, indicating *focused*, *non-focused*, and *undecidable* given the information available. As shown in Table 1, LF includes content such as discussing solutions, asking for help, discussing issues with the exercise, etc.,

Table 1. Examples of utterances in our classroom dataset, and the corresponding annotations for the lesson-focused facet and the classroom-focused facet.

Utterance	LF	CF
This is actually a different button	✓	✓
So maybe there's something that we put in a wrong spot or something we didn't add	✓	✓
So we're creating a greenhouse sensor?	✓	✓
What are you doing over there [partner]?	✗	✓
No, she's supposed to navigate you.	✗	✓
Teacher, their computer shut off.	✗	✓
Hold up. Aren't you in advanced math?	✗	✗
I got a tournament this weekend so I can't get quarantined	✗	✗
We should [inaudible]	Undecidable	Undecidable

whereas CF additionally includes activities like introductions, troubleshooting equipment, as well as meta-management of roles and responsibilities in groups. Each transcript is annotated by *two* annotators experienced with linguistic annotation tasks. They are provided access to the lesson plans to understand the context of discussions, and are instructed to evaluate every utterance in context of the entire transcript, by looking at past and future utterances. On the LF facet, inter-annotator agreement is 64.7% and on the CF facet, inter-annotator agreement is 71.3%. During disagreements, if only one of the annotators picks *undecidable*, we adjudicate to the other decisive label. When annotators disagree diametrically (focused and non-focused), we manually adjudicate using a third trained annotator. Diametric disagreements comprise 41% of all disagreements on the LF facet, and 42% of all disagreements on the CF facet. Due to the in-person, spontaneous nature of the interactions, our dataset contains phenomena like deictic references, fillers and interleaved conversations, which we believe make the annotation process challenging, thereby hindering high agreement.

We divide our data into three sets at the transcript level, using 18 for training, 6 for development, and 8 for testing. Of the 1254 utterances in the training set, 74% are *focused*, 22.5% are *non-focused*, and 2.6% are undecidable for the LF facet, and 86% are *focused*, 11.8% are *non-focused*, and 2% are undecidable for the CF facet.

2.2 Models

We now investigate the research question, *how well can state of the art natural language processing models identify off-task utterances?* We develop two classification models, to classify utterances on the LF, and the CF facet respectively. Since pretrained language representation models such as RoBERTa [6] perform competitively on classification tasks, we finetune them for sequence classification

Table 2. Performance on test set for both facets, on the labels focused (✓), non-focused (✗) and undecidable (*). RoBERTa models are given utterance history of window size w, RoBERTa-UL models are given utterance and gold label history.

Model	Lesson-focused				Classroom-focused			
	Macro F1	✓F1	✗F1	* F1	Macro F1	✓F1	✗F1	* F1
Random	0.29	0.41	0.36	0.10	0.29	0.43	0.33	0.11
Majority	0.25	0.74	0.00	0.00	0.26	0.78	0.00	0.00
RoBERTa ($w = 5$)	**0.59**	0.83	0.76	0.19	**0.56**	0.83	0.53	0.32
RoBERTa-UL ($w = 4$)	0.51	0.83	0.71	0.00	0.50	0.85	0.49	0.15

here. Specifically, each classifier is built by adding a separate classification head on top of the RoBERTa model, following which we train for LF or CF classification as described below. While our models are trained for predicting both facets separately, a multi-task training setup could be explored in future work.

Input: Each input instance consists of a single utterance, preceded by a speaker ID that is an anonymized unique identifier. We experiment with passing in a context history – a sequence of utterances in a window preceding the current utterance. The utterances in the history are all concatenated, with a special token to separate each utterance. We optionally provide the gold LF/CF label corresponding to each utterance in the history as input. At inference time, the label corresponding to each previous utterance in the history is *predicted.*

The number of preceding utterances passed to the model as context is a tunable hyperparameter, w. We experiment with ten window sizes, values ranging from 0 to 9, and also experiment with setting w to be the maximum length of a transcript, that is, using all previous utterances in a transcript as context.

3 Results

Table 2 shows the performance of our best performing models, along with comparisons to two baselines, Random and Majority. The random baseline randomly selects one out of the three labels, and the majority baseline always chooses the majority class, *focused.* We report class-wise and macro-averaged F1 scores.

For the lesson-focused models, Random achieves an F1 of 0.29, while Majority achieves an F1 of 0.25. Our best performing model, with an utterance history of size 5, outperforms both these baselines by a large margin, with an F1 of 0.59. However, the model that sees context as well as label history sees a drop in performance, to 0.51 F1, potentially due to noise in predicting labels (at test time) corresponding to the history. The impact of class imbalance can be seen in the class-wise F1s – high performance of 0.83 on the majority label *focused*, and very low performance of 0.19 on *undecidable.*

In the classroom-focused models, similarly, the RoBERTa models outperform the baselines by a strong margin, and the model that sees only the utterance context achieves an F1 of 0.56, outperforming the model with context+label history. We again observe a very high performance on the *focused* label, at 0.85 F1. In comparison to the LF models, performance on the *non-focused* or non-CF labels is lower, with the best model having an F1 of 0.53. One potential direction for improving performance in future is making use of multimodal signals, such as gestures and tone. As discussed in Sect. 2.1, this could help resolve instances when the utterance alone does not provide enough information (deictic references, short ambiguous utterances), which are challenging even for humans.

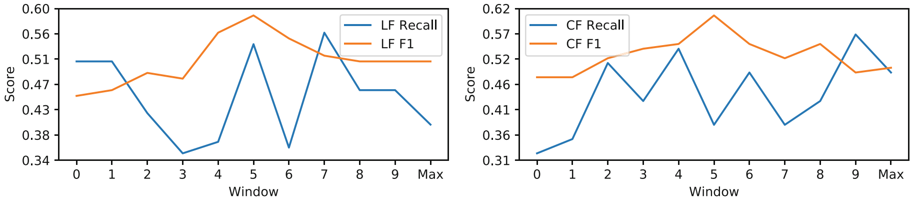

Fig. 1. Effect of context window size on recall

Recall Performance: An important indicator of the usefulness of our system to scaling qualitative coding of off-task discussions, is *recall* performance, which tells us how many off-task utterances were retrieved correctly by the model. As observed in Crowston et al. [3], a high recall is desirable in tools supporting qualitative research, to discover *all* relevant segments. Figure 1 shows recall performance on our development set for all window sizes.

We first note that maximum recall of 0.56 is observed at window size 7 for LF and 0.57 at window size 9 for CF. Further, recall performance does not necessarily follow the same trends as F1 score, which climbs until window size 5 before dropping off. Upon manually examining false negatives from models with differing window sizes, we find that the models with longer contexts tend to miss the *initiation* of an off-task segment, but correctly identify off-task utterances that are part of a longer sequence. The size of the context window must therefore serve as an important and modifiable design consideration for the researcher using our system, depending on the objective that needs to be optimized.

4 Conclusion

Towards assisting qualitative researchers in understanding the dynamics of collaboration between students, we propose using NLP to automatically highlight moments of wandering discussion. We present an annotation scheme that categorizes whether utterances are *lesson-focused* or *classroom-focused*, and annotate real-time spoken dialog. We then develop two transformer-based models for automatically classifying utterances accordingly. While our models outperform multiple baselines, our results show there is room for improvement, potentially using multi-modal information. In practice, we envision that qualitative researchers can filter large amounts of transcribed dialog based on our models' predictions, choosing to examine either of the two facets in greater detail. Further analysis can be facilitated in future through frameworks for researchers to query predictions, looking for specific types of interactions – for instance, off-task sequences that engage all speakers, or that result in increased subsequent on-task contributions.

Acknowledgments. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.

References

1. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: Off-task behavior in the cognitive tutor classroom: when students "game the system". In: Proceedings of the Conference on Human Factors in Computing Systems, pp. 383–390 (2004)
2. Carpenter, D., et al.: Detecting off-task behavior from student dialogue in game-based collaborative learning. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 55–66. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_5
3. Crowston, K., Allen, E.E., Heckman, R.: Using natural language processing technology for qualitative data analysis. Int. J. Soc. Res. Methodol. **15**(6), 523–543 (2012)
4. Godwin, K.E., et al.: Off-task behavior in elementary school children. Learn. Instr. **44**, 128–143 (2016)
5. Langer-Osuna, J., Gargroetzi, E., Chavez, R., Munson, J.: Rethinking loafers: understanding the productive functions of off-task talk during collaborative mathematics problem-solving. In: International Society of the Learning Sciences (2018)
6. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
7. Marathe, M., Toyama, K.: Semi-automated coding for qualitative research: a user-centered inquiry and initial prototypes. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2018)
8. Segal, A., et al.: Keeping the teacher in the loop: technologies for monitoring group learning in real-time. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 64–76. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_6
9. Southwell, R., Pugh, S., Perkoff, E., Clevenger, C., Bush, J., D'Mello, S.: Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. In: Proceedings of the 15th International Conference on Educational Data Mining. International Educational Data Mining Society (2022)
10. Worsley, M.: Multimodal learning analytics for the qualitative researcher. In: International Society of the Learning Sciences, Inc. [ISLS] (2018)