NOESIS II: Predicting Responses, Identifying Success, and Managing Complexity in Task-Oriented Dialogue

Chulaka Gunasekara,[1] **Jonathan K. Kummerfeld,**[2] **Luis Lastras**[1] and **Walter S. Lasecki**[2]
IBM Research AI, USA[1] University of Michigan, USA[2]
{chulaka.gunasekara,lastasl}@ibm.com {jkummerf,wlasecki}@umich.edu

Abstract

Real-world conversation often involves more than two participants and complex conversation structures, but most datasets for dialogue research simplify the task to make it more tractable. This shared task built on prior tasks for goal-oriented dialogue, moving towards more realistic settings. Seventeen teams participated in the primary task, predicting the next utterance in a multi-party conversation, and several teams participated in supplementary tasks. All of the datasets have been publicly released, providing a standard benchmark for future work in this space.

1 Introduction

Dialogue research at DSTC has focused on two-party conversations, which leaves out a range of challenges, e.g., determining who an utterance is directed at. At the same time, there has been growing interest in the development of systems that engage with people online in group settings, such as Internet Relay Chat (IRC), Slack, Twitter, and Reddit. Fortunately, these platforms also provide an exciting source of data for studying dialogue and developing systems. Previous tasks in the Dialogue System Technology Challenges have considered a range of tasks based on collected human-human dialogue.

We introduce a new dataset and extend a prior dataset to support challenging new dialogue problems. Our new data is an improved disentanglement of the `#Ubuntu` IRC channel, based on the method proposed by Kummerfeld et al. (2019).[1] Unlike prior work, we include conversations with any number of speakers. We also extend the Advising dataset from DSTC 7 track 1 (Gunasekara et al. 2019), adding annotations of when students accept or reject an advisor's class suggestion.

Using these datasets, we ran a primary task and three supplementary tasks. The primary task was next utterance selection given a partial conversation and one hundred options for the next utterance. The first supplementary task changed the context to be raw chat logs, containing a mixture of conversations and only part of the current one. This is the first time the true natural setting of these conversations has been considered (rather than after extracting isolated conversations). The second supplementary task explored the prediction of success in the Advising dialogue. Finally, we included a disentanglement task, in which the input is a raw chat log and the output is a set of independent conversations.

This task is a direct extension of DSTC 7 track 1 (Gunasekara et al. 2019). That task considered the same settings (Ubuntu and Advising), but focused on next utterance selection in 1-1 conversations. Our primary task was designed based on that experience, using the same parameters for many aspects of the set up while adding new challenges.

Seventeen teams participated in the task. All teams submitted output for the primary track for Ubuntu, ten participated in the primary track for Advising, three for the first two supplementary tasks, and one team for the final supplementary task. Many teams improved substantially over the baseline system provided. The best approach varied across the two datasets, though both were built with BERT (Devlin et al. 2019). The data, baselines, and evaluation metrics have been released[2], enabling future work to accurately compare with the results from this challenge.

2 Task

The primary task was next-utterance selection. In this problem, each example consists of a partial dialogue and a set of options for what the next utterance is in the dialogue. Participants must rank the potential messages plus the possibility that the true next message is not in the set. We followed the configuration from DSTC 7 track 1, with one hundred options for the next message. In 20% of cases, the true next message is not in the set. Participants were permitted to use external knowledge sources in their system. Those that were used are outlined in the description of submissions below.

2.1 In-Channel Selection

The first supplementary task was a variant of the main task in which the conversation context was not a complete prefix

[1]This data is **not** the same as the Lowe et al. (2015) corpus or the DSTC 7 Track 1 data, but it is based on the same raw chat logs.

[2]https://github.com/dstc8-track2/NOESIS-II/

Time	Speaker	Message
12:30	s_0	how can i boost microphone volume? The volume is toooooo low
12:30	s_1	s_0 , look for a microphone boost in alsamixer
12:30	s_2	s_0 : type 'alsamixer' into terminal
12:31	s_0	how the heck do i use alsamixer? :P what is microphone ?
12:32	s_0	how do i choose volume on input s_2 ?
12:33	s_2	s_0 : arrow keys up and down
12:33	s_0	s_2 , yes i understand that. But wich one of those things am i supposed to choose ?
12:33	s_2	s_0 : you wanted input, right?
12:34	s_0	s_2 , yes. But i there is no way i can turn that up. :S
12:34	s_2	s_0 : press tab to go over to capture, then turn it up
12:34	s_0	aha :) thanks

Speaker	Message
Student	Hello!
Advisor	Hi!
Student	I am currently trying to figure out what courses to take next semester.
Student	Could you suggest any?
Advisor	Let me see. Give me a minute to go over your transcript
Advisor	Can you tell me what your preferences are?
Student	Of course! I am interested in Computer Science, video game design is something that has always been interesting for me.
Advisor	Eecs 280 I should a prerequisite for most computer science classes, including game design
Student*	Okay yeah I will take that course. Do you know of any other prerequisites for game design?
Advisor	Eecs 281 is also necessary, and unfortunately you can't take both 280 and 281 in the same semester.
Advisor	You should take Eecs 203 as that is also a prerequisite for most Eecs classes
Student	Okay thanks for the info! Are both EECS 203 and EECS 280 project based?
Advisor	280 is all project based and 203 is not, but don't let that fool you. Many students say 203 is harder than 280
Student	Oh wow okay so do you think that taking them both in the same semester will be manageable?
Advisor	If you have a good grasp of probability and combinations it I should perfectly manageable

Figure 1: Examples of data in NOESIS II track: new dialogues from Ubuntu (top) and prior dialogues from Advising (bottom). The utterance marked with a * in the Advising dialogue is a point at which the student accepts a suggestion (used in the second supplementary task).

of a single conversation, but instead a section of chat from the raw Ubuntu IRC channel. The raw chat often contained multiple conversations, including cases where speakers participate in multiple conversations simultaneously. To reduce ambiguity about which conversation the next message is part of, we provided the identity of the speaker.

2.2 Advising Success

The second supplementary task considered identification of task success in the Advising data. Specifically, we provided a partial conversation and participants had to identify utterances that indicated the student had accepted or rejected the advisor's suggestion. Cases were also included in which no utterance accepting or rejecting the suggestion was present.

2.3 Disentanglement

The final supplementary task considered the process used to extract conversations from chat logs. We provided raw sections of the logs as input and requested sets of messages as output, where each set corresponded to a distinct conversation.

Property	Advising	Ubuntu
Dialogues	700	496,469
Average Number of Speakers	2	2.6
Utterances / Dialogue	18.6	7.2
Tokens / Utterance	9.8	11.4
Utterances / Unique utterances	4.4	1.2
Tokens / Unique tokens	10.5	44.1

Table 1: Comparison of the two data sources (based on training, development, and test data). Tokens are identified by splitting on whitespace.

3 Data

We considered two domains, using the same data sources as in DSTC 7 track 1. Both were task oriented, but one was much broader in scope and had more data (Ubuntu) while the other was smaller and more focused (Advising). Figure 1 shows examples of the data used to prepare the task and Table 1 shows stats about the conversations. A training set with answers was provided to participants to use as they wished.

Time	Speaker	Message
20:56	s_1	s_0 : I suggest you read the link ubottu sent you to understand what each digit represents
20:58	s_0	k thanks
21:00	s_0	http://www.onlineconversion.com/...
21:00	s_0	this really helps! :) now i get the 4, 2, 1 deal :)
21:05	**s_6***	hello! how to restore after rm? I delete my directory with projects, oh :
21:06	s_7	s_6 : Write a good eulogy and hold a memorial service.
21:06	s_7	s_6 : There's no way to, AFAIK
21:06	s_8	unless you made backups
21:07	s_2	!recover
21:07	ubottu	Some tools to recover lost data are listed and ...
21:14	**s_6***	s_8 : strange, crontab -l gave me this: 0 5 * * 1 tar -zcf /var/backups/home.tgz /home/, but no backups, uh :
21:15	s_8	no such file in /home? s_6 ?

Figure 2: The laat 12 lines of an example from the development set for supplementary task 1. Participants are also told that the next speaker is s_6 (who is marked in bold and with a * above).

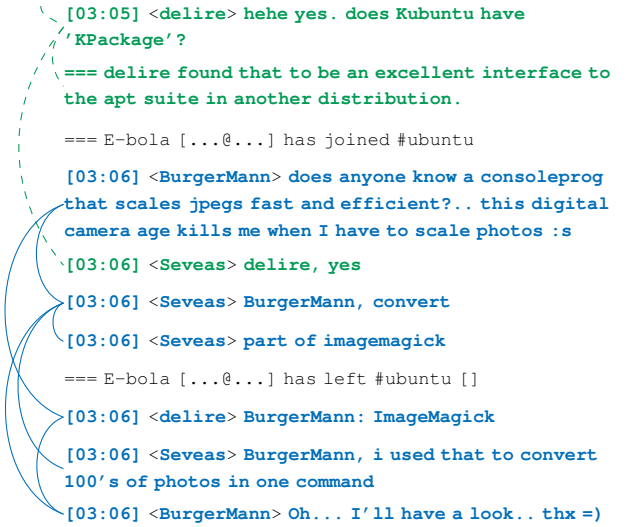

Figure 3: Example instance for the third supplementary task, with colours used to indicate conversations and curved lines indicating reply-to relations between messages. Participants only receive the time, speaker, and message as input. They are evaluated on the conversations only (not reply-to relations).

For the evaluation period, inputs for the test set were provided. The answers for the test set were not released until after the challenge was complete.

3.1 Ubuntu

A new set of disentangled Ubuntu IRC dialogues was provided for this challenge based on recent work (Kummerfeld et al. 2019). These were derived from the raw Ubuntu logs directly, not from any prior corpus. The dataset consisted of multi-party conversations extracted from the Ubuntu IRC channel.[3] A typical dialogue starts with a question that was asked by one participant, and then other participants respond with either an answer or follow-up questions that then lead to a back-and-forth conversation. In this challenge, the context of each dialogue contains at least three messages between the participants. The next turn in the conversation is guaranteed to be from one of the participants who has spoken so far. We pre-processed the data to identify references to speakers, as shown in Figure 1.

For the first supplementary task, we use raw samples from the channel, with pre-processing for speaker identification. For the third supplementary task, we used data from Kummerfeld et al. (2019).[4] Unlike the main task data, we did not pre-process the data for the disentanglement task, i.e., speakers appeared with their usernames as they did in the original channel.

The test data for each task was chosen so that it did not overlap with any other sets. For example, the test data for the main task came from a portion of the IRC log that was not used for training or testing in any other subtask. This was done to avoid information leakage across tasks and data.

Data Split We randomly split the conversations into training, development, and test sets. The development set had 4,827 conversations, the test set had 5,529 conversations, and the training set had the rest.

For the first supplementary task there were 112,262 instances for training, 9,565 for development, and 9,027 for testing. For the third supplementary task, we use the training, development and test split from Kummerfeld et al. (2019). That includes 67,463 training messages with links, 2,500 for development, and 5,000 for the test set.

3.2 Advising

This dataset contains two party dialogues that simulate a discussion between a student and an academic advisor. The purpose of the dialogues is to guide the student to pick courses that fit not only their curriculum, but also personal preferences about time, difficulty, areas of interest, etc. The conversations used were the same as those used in DSTC 7 track 1 (Gunasekara et al. 2019). They were collected by having students at the University of Michigan act as the two roles using provided personas. Structured information in the form of a database of course information was provided, as well as the personas (though at test time only information available to the advisor was provided, i.e. not the explicit student preferences). The data also includes paraphrases of the sentences and of the target responses.

Data Split The training, development, and test sets were the same as in DSTC 7 track 1. The development and test sets are based on 100 raw conversations, each paraphrased five times and then cut off at different points. The training set

[3] https://irclogs.ubuntu.com/

[4] The test data from that work was not released until after the deadline for DSTC 8 submissions.

Team	Approach
1	Similarity matching between context and candidates with TF-IDF and Cosine Similarity on words and subwords.
3	Domain-specific model trained on raw data and fine tuned on the training dataset.
4	Similarity matching with BERT, using a threshold for when to return no answer.
5	An ensemble combining a cross-encoder with full self-attention and a poly-encoder using separate self-attention.
6	BERT with a data augmentation method.
7	RoBERTa with an extra self-attention layer to capture the relative importance of utterances by a speaker.
8	Multilingual universal sentence encoder as part of the dual encoder pipeline with upsampling to balance the data.
9	An ensemble of ESIM and Dual Encoder-based models with BERT and GloVe, using a theshold for predicting no answer.
10	BERT.
11	BERT with course information concatenated for Advising.
12	An ensemble of BERT models with (1) the cross-encoder and poly-encoder, (2) augmented data, and (3) a re-ranking process.
13	An ensemble of ESIM and BERT with a gradient boosting classifier.
14	BERT with a randomly wired network trained with binary classification.
15	RoBERTa with augmented training data using raw dialogues and binary classification as the training method.
16	A combination of the Dual-LSTM encoder and USE for Ubuntu. For Advising, a similarity based approach to match with training instances.
17	BERT based similarity matching, using a threshold to identify no answer cases.

Table 2: Summary of approaches used in submissions. One team did not provide a description.

is based on 500 conversations, also paraphrased five times, but then remixed many times.

For the second supplementary task, we use the same data split. Instances were annotated by one of the authors.

4 Results

4.1 Participants

We provided baselines for task 1 and task 4. The task 1 baseline was a slightly modified form of the encoder-decoder baseline provided in track one of DSTC 7. The task 4 baselines was from Kummerfeld et al. (2019), a feedforward neural network that uses GloVe embeddings and structural features to represent the conversation.

Table 2 summarizes the approaches used by all participants. One clear trend was a switch from the ESIM (Chen et al. 2017) model used by participants in DSTC 7 to BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019). While many participants used a version of BERT, there was still a broad range of results. This indicates the importance of elements like defining the loss, data augmentation, text segmentation, and so on in achieving strong results.

4.2 Evaluation Metrics

A range of metrics were considered. The main task and the first supplementary task followed DSTC 7 track 1, using the mean of (a) recall at 10 and (b) mean reciprocal rank (MRR). The second supplementary task used accuracy, measured as whether each example was correct or not. The final supplementary task used precision, recall, and F-score over complete conversations and several clustering metrics (Variation of Information (Meila 2007), Adjusted Rand Index, and Adjusted Mutual Information). These treat each message as an item and conversations as clusters.

4.3 Discussion

Main Task Performance varied significantly on the main task, with the best teams scoring far higher than the reference baseline provided. As in DSTC 7, the Advising data proved harder. Unlike in DSTC 7, the best approach varied across the datasets, with the best approach on Ubuntu coming second on Advising and the best approach on Advising coming 5th on Ubuntu.

In-Channel Selection As expected, shifting to the more realistic setting of the raw channel led to lower performance. The size of the drop did vary, from 0.035 to 0.128 in score.

Team	Recall@ 1	5	10	MRR	Score
Baseline	0.212	0.421	0.565	0.325	0.445
1	0.245	0.382	0.457	0.319	0.388
2	0.626	0.857	0.931	0.729	0.830
3	0.649	0.904	0.949	0.760	0.855
4	0.550	0.863	0.927	0.683	0.805
5	0.663	0.943	0.974	0.786	0.880
6	0.552	0.763	0.835	0.651	0.743
7	0.620	0.708	0.710	0.658	0.684
8	0.212	0.451	0.572	0.327	0.449
9	0.536	0.855	0.911	0.680	0.796
10	0.593	0.872	0.922	0.713	0.818
11	0.462	0.633	0.694	0.539	0.617
12	0.720	0.948	0.977	0.819	0.898
13	0.669	0.922	0.961	0.778	0.869
14	0.634	0.738	0.759	0.680	0.720
15	**0.761**	**0.958**	**0.979**	**0.848**	**0.913**
16	0.238	0.484	0.608	0.356	0.482
17	0.550	0.863	0.927	0.683	0.805

Table 3: Results for the main task (Ubuntu data).

Team	Recall@ 1	5	10	MRR	Score
Baseline	0.222	0.493	0.622	0.355	0.489
2	0.242	0.522	0.680	0.368	0.524
3	0.224	0.526	0.676	0.374	0.525
4	0.140	0.370	0.508	0.263	0.385
6	0.230	0.510	0.612	0.365	0.488
9	0.202	0.458	0.600	0.329	0.464
11	0.192	0.342	0.434	0.271	0.352
13	0.254	0.560	0.690	0.401	0.545
15	0.306	0.632	0.762	0.455	0.609
16	0.010	0.048	0.092	0.048	0.070
17	**0.564**	**0.806**	**0.878**	**0.677**	**0.777**

Table 4: Results for the main task (Advising data).

Team	Recall@ 1	5	10	MRR	Score
3	0.505	0.755	0.834	0.621	0.727
13	0.596	0.847	0.904	0.707	0.806
15	**0.706**	**0.916**	**0.957**	**0.799**	**0.878**

Table 5: Results for the in-channel next-utterance selection task (Ubuntu).

Team	Exact Match	Precision	Recall	F1
3	80.0	**83.2**	**80.2**	**81.7**
13	66.2	70.7	66.2	68.4
15	**80.8**	**83.2**	**80.2**	**81.7**

Table 6: Results for the suggestion acceptance task (Advising).

Team	P	R	F	VI	Rand	AMI
Baseline	36.3	39.7	38.0	0.915	0.650	0.837
3	**44.3**	**49.6**	**46.8**	**0.933**	**0.752**	**0.865**

Table 7: Results for the conversation disentanglement task (Ubuntu).

This suggests that the complications introduced in the raw setting are real, but surmountable.

Advising Success All three teams that attempted this new task performed well. Task success could be a good signal for training dialogue systems with reinforcement learning, and so these results are an encouraging sign that automated training via interaction with people may be feasible (with success detection as the reward).

Two teams had almost exactly the same results (3 and 15). Investigating these further, we found several differences in the patterns of errors in their output. Team 15 tended to predict "No Decision Yet" more often, achieving higher recall and lower precision that team 3 on that category. The trend was reversed for "Accept", with team 3 predicting it more often and achieving higher recall and lower precision. For identifying "Reject", the results were extremely similar.

Disentanglement Only one team attempted this supplementary task, but they achieved strong performance, improving over the baseline by 7.8 F_1. This is still far from perfect performance, indicating that this problem remains an open challenge.

5 Conclusion

This task introduced a new collection of dialogues from the Ubuntu IRC channel and several new task variations. Seventeen teams attempted part of the challenge, producing systems that significantly improved over the provided baseline. This task also complements DSTC 7 task 1, extending to new types of challenges in dialogue.

Acknowledgements

This material is based in part upon work supported by IBM under contract 4915012629. Any opinions, findings, conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of IBM.

References

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1657–1668.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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), 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Gunasekara, C.; Kummerfeld, J. K.; Polymenakos, L.; ; and Lasecki, W. S. 2019. Dstc7 task 1: Noetic end-to-end response selection. In *7th Edition of the Dialog System Technology Challenges at AAAI 2019*.

Kummerfeld, J. K.; Gouravajhala, S. R.; Peper, J. J.; Athreya, V.; Gunasekara, C.; Ganhotra, J.; Patel, S. S.; Polymenakos, L.; and Lasecki, W. S. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3846–3856.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints* arXiv:1907.11692.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 285–294.

Meila, M. 2007. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis* 98(5):873–895.