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

Lazaros C Polymenakos, Chulaka Gunasekara, IBM
Walter S. Lasecki, Jonathan K. Kummerfeld, University of Michigan

Building on the success of DSTC 7 Track 1 (NOESIS: Noetic End-to-End Response Selection Challenge), we propose an extension of the task, incorporating new elements that are vital for the creation of a deployed task-oriented dialogue system. Specifically, we add three new dimensions to the challenge: (1) conversations with more than 2 participants, (2) predicting whether a dialogue has solved the problem yet, and (3) handling multiple simultaneous conversations. Each of these adds an exciting new dimension and brings the task closer to the creation of systems able to handle the complexity of real-world conversation.

Topic areas: End-to-end dialog systems, Dialog state tracking, Natural language generation for dialog systems, Dialogue disentanglement.

## Task and Evaluation

The core idea for the task follows DSTC 7 track 1, but combines aspects of several subtasks. The structure is as follows:

> Input: A conversational context and 100 utterances that could be the next message (either 99 or 100 will be incorrect).
>
> Output: A ranking of the 100 utterances and the option that the correct answer is not present.

We will also provide a set of external knowledge sources (e.g. manual pages, AskUbuntu forum discussion, and Quora answers). Unlike in DSTC 7, we will have conversations with more than 2 participants in the Ubuntu data. Starting from this basis there will be several task variations:

| **Task Name** | **Dataset(s)** | **Description** | **Evaluation** |
|---|---|---|---|
| 1. Base | Ubuntu and Advising | As described above. | Mean reciprocal rank and recall @ N for various values of N. The final metric will be the average of MRR and Recall @ 10 (as it was in DSTC 7 Track 1). |
| 2. Multi-conversation | Ubuntu | We will use data directly from the IRC channel, which contains multiple entangled conversations. This makes the task more difficult, but also means systems are tested in a setting closer to the real world. We will also provide participants with one additional piece of information - which user is uttering the message they are predicting. | |
| 3. Success | Advising | In this task, participants predict where in a dialogue the problem is solved (if at all). | Accuracy |

| | | | |
|---|---|---|---|
| 4. Disentanglement | Ubuntu | Participants are given a section of the chat logs and need to identify a set of conversations contained within that section (ie., clusters of messages that form coherent conversations). This is a useful, but not required, ability for task 2 and this subtask will allow us to measure performance on it in isolation. | Precision, recall, and F-score over complete threads and several clustering metrics (Variation of Information, Adjusted Rand Index, and Adjusted Mutual Information). |

## Data

We will provide two task-oriented datasets:

- Ubuntu, multi-participant chat conversations from the `#UBUNTU` IRC channel, in which participants discuss technical issues they are having with the Ubuntu operating system.
- Advising, conversations between two University of Michigan students simulating a student and an advisor selecting courses for the next semester.

A previous version of this data was used in DSTC 7 track 1. To prepare for DSTC 8, we will add annotations of task success in the Advising data, marking either (1) the point in each dialogue at which the original problem was solved or (2) that the problem was not solved in the dialogue. No other new annotation is required.

In DSTC 7, the Ubuntu data was automatically disentangled to provide isolated two-person conversations. For our proposed task 1, we will also use conversations with more than two people. For task 2 we do not need to disentangle data, making it easy to construct an enormous dataset (and avoiding incorrectly disentangled conversations). To avoid creating a barrier to participation in the task, we will provide automatically assigned disentanglement labels for all data. Participants can either use the provided labels or develop their own disentanglement method, providing an interesting new dimension to the task.

To support the exploration of disentanglement, we will provide a set of 62,000 messages hand-labeled with disentanglement information. This set has already been developed.

Baselines - We will also provide implementations of baseline systems for each component of the task. Participants will be welcome to use any, all, or none of the baselines as part of their submission.

## Contact

Lazaros C. Polymenakos - lcpolyme@us.ibm.com
Chulaka Gunasekara - Chulaka.Gunasekara@ibm.com
Walter S. Lasecki - wlasecki@umich.edu
Jonathan K. Kummerfeld - jkummerf@umich.edu