Model	Embedding Average Score	Embedding Greedy Score	Embedding Extrema Score
Seq2seq	0.4733 +/- 0.00010	0.4728 +/- 0.00005	0.3067 +/- 0.00009
Seq2seq (last utte)	0.5969 +/- 0.00009	0.4751 +/- 0.00005	0.3185 +/- 0.00007
VHRED	0.5151 +/- 0.00012	0.9053 +/- 0.00011	0.3252 +/- 0.00011
GSN No-speaker (1-iter)	0.7224 +/- 0.00004	1.0209 +/- 0.00008	0.5624 +/- 0.00005
GSN No-speaker (2-iter)	0.7318 +/- 0.00005	1.1199 +/- 0.00010	0.5796 +/- 0.00005
GSN No-speaker (3-iter)	0.7244 +/- 0.00004	0.9495 +/- 0.00006	0.5630 +/- 0.00004
GSN W-speaker (1-iter)	0.7133 +/- 0.00005	1.1193 +/- 0.00011	0.5860 +/- 0.00005
GSN W-speaker (2-iter)	0.7695 +/- 0.00004	1.0400 +/- 0.00006	0.6508 +/- 0.00005
GSN W-speaker (3-iter)	0.7473 +/- 0.00006	1.1610 +/- 0.00012	0.5890 +/- 0.00006

Table 1: Experimental results measured using three embedding-based metrics. Each score is followed by its standard deviation.

A Detailed Results in Embedding-based Metrics for Automated Evaluation

Experimental results measured using three embedding-based metrics under different settings and different methods with sequential data and graph data are given in Table 1. 'Seq2seq (last utte)' model is trained by only using the last utterance before the final response as the input (all utterances before it are ignored). 'n-iter' means the results are obtained after n iterations. 'No-speaker' is our GSN model without speaker information flow while 'W-speaker' has it.

B Case Study

Utterance 1: my memory card for my phone won't allow me to delete files or put files on it, it says "read only file system".	**SESSION**:
	utterance 1: i'm running on amd athlon 64 4000.
Human: have you tried—cd filepath ls -l?	**utterance 2:** then it will support 64bit if you wish to use it.
No-speaker (1-iter): please don't flood, use url to paste, don't use enter as punctuation.	**utterance 3:** i understand that, but how i make .txt files not unknown?
No-speaker (3-iter): what is the output of: uname -r?	**utterance 4:** should i use it? whats the difference?
W-speaker (1-iter): does it show up in the terminal?	**RESPONSES**:
	No-speaker (3-iter): i'm not sure what you're looking for.
W-speaker (3-iter): you can mount it with the mount command.	**W-speaker (1-iter)**: in a terminal, type "sudo apt-get update"
Seq2seq: try that. i'm not sure how to do it.	**W-speaker (3-iter)**: you need to install it from source.
VHRED: that's what i'm looking for.	**Seq2seq**: i'm not sure what you ' re trying to do.
	VHRED: you ' ll need to find out what you ' re doing .

Table 2: Sample responses generated by human, GSN, and baselines. All the responses in column 1 are for utterance 1. Column 2 is a session with the graph structure in Figure 5.

Table 2 presents some sample responses generated by our model and the baselines. From the first example, we see that the responses generated by our model are more consistent with the given utterance. The utterance in the first example (on the left in Table 2) is asking for help, not offering an answer. However, VHRED mistakenly takes it as a useful answer and gives a response "that's what i'm looking for", which may not be appropriate. The seq2seq model provides a meaningless utterance "try that. i'm not sure how to do it", which addresses nothing. In contrast to these methods, our model tries to give some useful suggestions, e.g., "you can mount it with the mount command." and "what is the output of: uname -r?", etc, which we believe are more effective.

In the second column (Table 2), based on utterances 1, 2 and 3, the speaker tried to run a 64-bit application but the application failed with an error message ".txt files unknown". In this case, our models generate meaningful responses from the training corpus: giving a suggestion to update the system (W-speaker 1-iter); advising the speaker to try another way to install the software (W-speaker 3-iter). Thus, our models (with the speaker information flow) generate responses relevant to the entire session and give substantive advice for the raised questions. These responses are more appropriate and preferred from the human perspective.