#	#Turn	Utterance	Pred.	Human
1	1	@user ur prolly tired now , arent u ? wanna sleep , don't cha ? (:	4.67	5.00
	2	@user atm jus chillin nd chattin nd listening to music :P haha xxx	3.70	2.00
	3	@user x factor tickets came so im siva happy dancing :) might do some fanfic xx wbu ? xx	3.26	1.00
	4	@user goood :) do u want to be in my 1d fanbook ? xx	3.07	2.00
	5	@user : o gosh ppl these days :/ ... ok tell me everything 2moz p . s send me the link so i can send it to et :) xx dont stress it :)	2.89	1.00
	6	@user yer im fine thanks xxx	3.18	3.00
	7	@user okay , thanks xxx	3.42	1.00
	8	@user :) you okay ? xxx	3.26	1.00
	9	@user yeah im good thanks :) you ? xx	3.11	4.00
2	0	if i could i would hurt you ... a lot ! ! i'm so sick of you blaming me for everything that goes wrong ! !	4.86	5.00
	1	@user awh ... is everything ok sweetie ?	4.72	5.00
	2	@user thanks ... it is ... i just can't wait to see how it is ...	3.72	3.00
	3	@user wait , what ?	3.51	5.00
	4	@user i have a new phone . i have to get it .	2.54	3.00
	5	@user how much do you pay ?	2.20	3.00
	6	@user pay attention	2.65	1.00
	7	@user i would have to do that	1.82	1.00
	8	@user you could of been in my head .	2.27	1.00
	9	@user yeah , i don't know .	3.03	1.00
3	0	hashtag @user @user @user @user @user @user @user @user	4.47	5.00
	1	@user thanks for the follow . do you have at bat on your phone yet ?	4.95	5.00
	2	@user there's a lot going on ... but i think is good to say what u think here and at the boards too . specially there !	3.55	2.00
	3	@user pretty much just as well as the current macbook pros do ...	2.42	1.00
	4	@user how well does it run photoshop / illustrator ?	2.67	2.00
	5	@user this is like sixth sense . can i have your games consoles ?	3.00	1.00
	6	@user probably , but i don't know if it'd be appropriate in this case . i try only to use our funds for things i know i need to know .	2.15	1.00
	7	@user well we can scan thru a few joints i'm workin on , or take the traditional route n pen one down	2.09	1.00
	8	@user i learned gordon's rhythm patterning in methods classes , not much orff	2.21	1.00
	9	@user its like a long story with sequence of events lmao	1.76	1.00

Table 3: Randomly sampled output. The conversation is sampled at random and *AutoJudge* rates each turn.

A Training Details

Model Training. For all models, we used a bidirectional LSTM to encode the turns, and a unidirectional LSTM for both the context encoder and decoder. We specify the number of units for the LSTMs to 500, 1000, 1000 for the turn-encoder, context-encoder and decoder respectively. We use the pretrained 300 dimensional FastText embeddings (Mikolov et al., 2018), which we refine during the training. In order to avoid too large vocabularies, we limit the vocabulary size to 20k distinct tokens. The generative models are trained to minimize the reconstruction error. For the VHRED and MrRNN, we refer to the original papers for the loss function formulation. The Dual Encoder is trained to minimize a contrastive loss function $log\sigma(c^T r_{Ture}) + \sum_{n \in N} log\sigma(-c^T r_n)$, where c is the context encoding, r_{True} is the correct response encoding and N is a set of negative samples. For each training sample we sampled 10 negative examples uniformly at random from the training set. All models are optimized using the *Adam* optimizer (Kingma and Ba, 2014), with a $lr = 0.001$ and a batch size of 80.

AutoJudge Training We trained *AutoJudge* using the pre-trained VHRED model to encode the context and the response. During the training only the matrix M gets optimized. We also experimented with non-linear transformation on these encodings, which did not yield any improvements. Similar to (Lowe et al., 2017), we use $\alpha = 0.01$ and $\beta = 32$. *AutoJudge* is optimized using *Adam* optimizer (Kingma and Ba, 2014), with a $lr = 0.001$ and a batch size of 512.

B Reinforcement Learning

For reinforcement learning, we use the pre-trained HRED system as our initial policy. We apply Policy Gradient as described above. We experimented with various episode batch sizes (1, 10, 100, 1000), i.e. in sample n episodes at once to reduce variance. However, it had no impact on the performance. We also experimented with different formulations, i.e. using Advantage Actor Critics in order to reduce the variance.

In Table 1, we show the rolling average return over the course of 100 episodes. We used a batch size of 100 and we used the standard Policy Gradient formulation. The reward oscillates, which is due to finding new local maxima. He maximal observed reward is at 37 after 80 episodes. However, the generated dialogues are all empty, i.e. the dialogue system always returns the "end-of-sequence" token right away.

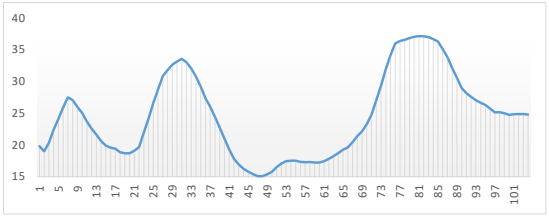

Figure 1: Data distributions for both the overall data and for the systems.