The Second Conversational Intelligence Challenge (ConvAI2)

Emily Dinan[1], Varvara Logacheva[2], Valentin Malykh[2], Alexander Miller[1], Kurt Shuster[1], Jack Urbanek[1], Douwe Kiela[1], Arthur Szlam[1], Iulian Serban[3], Ryan Lowe[4,1], Shrimai Prabhumoye[5], Alan W Black[5], Alexander Rudnicky[5], Jason Williams[6], Joelle Pineau[1,4], Mikhail Burtsev[2] and Jason Weston[1]

[1]Facebook AI Research
[2]Moscow Institute of Physics and Technology
[3]University of Montreal
[4]McGill University
[5]Carnegie Mellon University
[6]Microsoft Research

Abstract

We describe the setting and results of the ConvAI2 NeurIPS competition that aims to further the state-of-the-art in open-domain chatbots. Some key takeaways from the competition are: (i) pretrained Transformer variants are currently the best performing models on this task, (ii) but to improve performance on multi-turn conversations with humans, future systems must go beyond single word metrics like perplexity to measure the performance across sequences of utterances (conversations) –in terms of repetition, consistency and balance of dialogue acts (e.g. how many questions asked vs. answered).

1 Overview of the competition

The Conversational Intelligence Challenge[1] aims at finding approaches to creating high-quality dialogue agents capable of meaningful open domain conversation. Today, the progress in the field is hampered by the absence of established benchmark tasks for non-goal-oriented dialogue systems (chatbots) and solid evaluation criteria for automatic assessment of dialogue quality. The aim of this competition was therefore to establish a concrete scenario for testing chatbots that aim to engage humans, and become a standard evaluation tool in order to make such systems directly comparable, including open source datasets, evaluation code (both automatic evaluations and code to run the human evaluation on Mechanical Turk), model baselines and the winning model itself.

---

[1]http://convai.io/

1

This is the second Conversational Intelligence (ConvAI) Challenge; the previous one was conducted under the scope of NeurIPS 2017 Competitions track. Taking into account the results of the previous edition, this year we improved the task, the evaluation process, and the human conversationalists' experience. We did this by making the setup simpler for the competitors and by making the conversations more engaging for humans. We provided a dataset from the beginning, ConvAI2, a variant of the Persona-Chat dataset, whose training set consists of conversations between crowdworkers who were randomly paired and asked to act the part of a provided persona (randomly assigned, and created by another set of crowdworkers) in a conversation with their partner. The paired workers were instructed to chat naturally and to get to know each other during the conversation. This produced interesting and engaging dialogues that learning agents can try to mimic. The Persona-Chat task was designed to facilitate research into alleviating some of the issues that traditional chit-chat models face, and with the aim of making such models more consistent and engaging by endowing them with a persona [1].

Competitors' models were compared in three ways: (i) automatic evaluation metrics on a test set hidden from the competitors; (ii) evaluation on Amazon Mechanical Turk; and (iii) 'wild' live evaluation by volunteers having conversations with the bots. We declared winners in the automatic evaluation tracks, but the grand prize was awarded to the best performing system in human evaluations.

The winner in the automatic evaluation tracks by a substantial margin was the team Hugging Face, however the grand prize winner from human evaluations was Lost in Conversation (Hugging Face coming in second place, with 23 entrants in total). There are a number of key takeaways from our analysis of the results, indicating that the automatic evaluations show some correlation to human evaluations, but fail to take into account important aspects of multi-turn conversation that humans consider important, in particular the balance of dialogue acts throughout the conversation (e.g. the amount of questions asked versus answered).

1.1 Previous competitions

There have been a number of competitions on question answering (e.g. quiz bowl) which can be seen as single-turn goal-directed dialogue, as well as competitions on goal-directed dialogue involving dialogue state tracking (including 5 iterations of the DSTC challenge), e.g. for booking restaurants or tourist information. Those do not explicitly address the "chit-chat" setting of dialogue about general topics which is not goal-directed, although later DSTC challenges do address chit-chat.

The first edition of the Conversational Intelligence Challenge took place at the NeurIPS 2017 Competition track in the form of a live competition. The task was for an agent to carry out intelligent and natural conversations about specific snippets from Wikipedia articles with humans, which was not engaging for many human participants.

Ten dialogue systems participated in the 2017 competition. The majority of them

combined multiple conversational models such as question answering and chit-chat systems to make conversations more natural. The evaluation of chatbots was performed by human assessors. More than 1,500 volunteers participated and over 4,000 dialogues were collected during the competition. All of the data as well as the solutions of the winners are available via the competition repository.[2,3] The best bot had a final dialogue quality score of 2.746 compared to 3.8 for human performance. This demonstrates that current technology allows supporting dialogue on a given topic but with quality much lower than that of humans.

In contrast to the first edition, the 2018 competition focused on open domain chit-chat, rather than on encyclopedic facts. To our knowledge, no other competition has focused on a dataset like this. Importantly, we provided a large training set and validation set in a standard setup, complete with code for baseline systems for entrants to obtain clear automatic evaluation metrics to improve upon. In the 2017 ConvAI competition, no data was initially provided. Instead, data was collected by volunteers as the competition progressed, which may have led to fewer participants.

Outside of NeurIPS, the most similar competition is probably the Alexa Prize[4]. This is a competition to build a socialbot that can converse coherently and engagingly with humans on popular topics for 20 minutes. The top bots were selected by Amazon Alexa customers and the Amazon panel and competed head-to-head in front of three judges in November 2017. Alexa Prize data and models are not made public, in contrast to this competition in which data, winning models, and training code are made available. The results of the Alexa Prize mostly confirmed that ensembles are useful in such tasks and did little to drive fundamental algorithm research. Another small scale analogue is the Loebner Prize, which was given to the system that was judged to hold the most "human-like" textual conversations.[5]

2 Competition description and set-up

2.1 Data

The CONVAI2 dataset for training models is publicly available in ParlAI[6]. It is a variant of the PERSONA-CHAT task, differing only by some formatting changes and a different (new) test set. The task consists of short, chit-chat dialogues between two speakers, each of whom has an assigned profile. See Table 1 for a dialogue example. The speakers were instructed to chat naturally and get to know each other. The assigned sets of profile information, or personas, are natural and descriptive, and contain common topics of human interest that the speakers often discuss during their conversation. Each persona contains at least 5

---

[2]`http://convai.io/2017/data/`

[3]`https://github.com/DeepPavlov/convai/tree/master/2017/solutions`

[4]`https://developer.amazon.com/alexaprize`

[5]`https://en.wikipedia.org/wiki/Loebner_Prize`

[6]https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks/convai2

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Table 1: Example dialogue from the PERSONA-CHAT dataset. Person 1 is given their own persona (top left) at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation.

sentences, and is drawn from a set of over 1,155 personas (at training time), setting aside 100 never before seen personas for validation. The dataset statistics are given in Table 2.

As the original PERSONA-CHAT test set was released, we crowdsourced further data for a hidden test set unseen by the competitors for automatic evaluation. The hidden test set consisted of 100 new personas and over 1,015 dialogs.

To avoid modeling that takes advantage of trivial word overlap, additional rewritten sets of the same train and test personas were crowdsourced, with related sentences that are rephrases, generalizations or specializations, rendering the task much more challenging. For example "I just got my nails done" is revised as "I love to pamper myself on a regular basis" and "I am on a diet now" is revised as "I need to lose weight."

	# examples	# dialogues	# personas
Training set	131,438	17,878	1,155
Validation set	7,801	1,000	100
Hidden test set	6,634	1,015	100

Table 2: Statistics of the ConvAI2 dataset (based on PERSONA-CHAT).

The task aims to model normal conversation when two interlocutors first meet, and get to know each other. Their aim is to be engaging, to learn about the other's interests, discuss their own interests and find common ground. The task is technically challenging as it involves both asking and answering questions and maintaining a consistent persona. Conversing with current chit-chat models for even a short amount of time quickly exposes their weaknesses [2, 3]. Common issues with chit-chat models include: (i) the lack of a consistent personality [4] as they are typically trained over many dialogues each with different speakers, (ii) the lack of an explicit long-term memory as they are typically trained to produce an utterance given only the recent dialogue history [3], and (iii) a tendency to produce non-specific answers like "I don't know" [5]. With this task we aim to find models that address those specific issues [1].

Note that for training, competitors were allowed to use other additional training data as long as it was made public (or was already public).

2.2 Metrics

We first evaluated all submissions on a set of automatic metrics. The top 7 teams from the automatic metrics were then evaluated by humans:

- **Automatic metrics** - Perplexity, F_1, and hits@1/20. These were computed on the hidden test. As some methods are not applicable to some metrics – perplexity, for instance, is only scored for probabilistic generative models – participants could choose which metric or metrics they wanted their model to be evaluated on. We chose these metrics as they are commonly used in the literature to evaluate dialogue models. For instance, hits@k is often used to evaluate retrieval-based dialogue models [6, 7, 8] and perplexity is often used to evaluate probabilistic generative dialogue models [9, 10].

  - Perplexity — a metric of text fluency which is computed as $\frac{1}{m}\sum_{i=1}^{m} log p(w_i)$ for sentence $\mathbf{w} = w_1, w_2, ..., w_m$. This metric is computed only for probabilistic generative models.
  - F_1-score — $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. In the context of dialogue, precision is the fraction of words in the predicted response that are contained in the gold response, and recall is the fraction of words in the gold response that were in the predicted response. This can be computed for any model, retrieval-based or generative.
  - Hits@1/20 — hits@1/N is the accuracy of the next dialogue utterance when choosing between the gold response and $N - 1$ distractor responses (here, we use $N = 19$). Distractor responses are random responses from the dataset. Any model that can assign a score to a given candidate utterance can compute this metric. Such a method could then in principle be used in a retrieval model to score retrieved candidates.

- **Human evaluations** -

- **Amazon's Mechanical Turk:** Given the entrants' model code, we ran live experiments where workers on Amazon's Mechanical Turk (Turkers) chatted to a given model following instructions identical to the creation of the original dataset, but with new profiles, and then scored its performance. Performance was evaluated by asking Turkers how much they enjoyed talking to the model on a scale of 1-4 after having a conversation of 4-6 dialogue turns each. We also had Turkers verify which persona the model was using given the choice between the correct persona and a random one.
- **'Wild' Live Chat with Volunteers:** We solicited volunteers to chat to the models in a similar way to the Mechanical Turk setup, the primary difference being that in this setting, workers are unpaid. This setup was hosted through the Facebook Messenger and Telegram APIs.

2.3 Baselines and code available

Source code for baseline methods for the competition were provided in the open source system ParlAI [11][7], including training loop and evaluation code. The example models were the methods developed in [1], which we consider strong baselines. They included a retrieval-based Key-Value Memory Network, and two generative models: an LSTM-based attentive Seq2Seq model and a LSTM-based language model.

2.4 Rules

- Competitors had to provide their source code so that the hidden test set evaluation and live experiments could be computed without the team's influence, and so that the competition would have further impact as those models could be released for future research to build off them. Code could be written be in any language, but a thin python wrapper needed to be provided in order to work with our evaluation and live experiment code via ParlAI's interface.

- Each team could only submit a maximum of once per month during the automatic metrics round.

- We required that the winning systems also released their training code so that their work was reproducible (although we also encouraged that for all systems).

- Competitors were required to indicate which training sources were used to build their models, and whether (and how) ensembling was used.

- Competitors were free to augment training with other datasets as long as they were publicly released (and hence, reproducible). Hence, all entrants were expected to work on publicly available data or release the data they used to train.

---

[7]https://github.com/facebookresearch/ParlAI/tree/master/projects/convai2

2.5 Timeline

- **April 21:** Competition began: automatic metrics leaderboard, baselines, and submission instructions were posted.

- **May 9** Hackathon: We organized a non-compulsory hackathon around the competition: DeepHack.Chat. At the hackathon teams aimed to improve their systems, took part in live human evaluations, and listened to lectures from researchers in the field.

- **July 10:** 'Wild' evaluation opened. Participants were able to submit their models to be evaluated by live volunteers.

- **September 30:** Submissions for the automatic metrics round were closed. We invited the top seven teams from this round to prepare their submissions for the Mechanical Turk evaluation portion of the competition.

- **December 9:** Winner of the competition was announced at NeurIPS 2018.

2.6 Prize

The grand prize for the winner of the human evaluations was awarded $20,000 in funding for Amazon Mechanical Turk, in order to encourage further data collection for dialogue research. The winner in the automatic metrics received $5,000 in AWS compute.

3 Results and Analysis

3.1 Automatic Metrics

We had over 23 teams submit models to be evaluated for the automatic metrics. The rank of each team was determined by sorting by the minimum rank of the score in any of the three metrics (F_1, Hits@1, and Perplexity). Recall that participants could choose which metric or metrics they wanted their system to be evaluated on: few teams chose to be evaluated on all three of the metrics. The Hugging Face team performed the best in every single metric and was therefore determined to be the winner of this round. All participants and their scores on the hidden test set are shown in Table 3.

The top seven teams made it to the next round. Notably, each of these teams surpassed our baseline models in some metric. The High Five team chose not to participate in the human evaluation round, so ultimately six teams participated in the next round. Refer to Section 4 for a description of the models submitted from the top-performing teams.

3.1.1 Further Analysis and Additional Automatic Metrics

Revised Personas We also evaluated models (from the teams in the top 7) that were capable of ranking – i.e. models that were evaluated on the Hits@1 metric – on the

Team Names	Perplexity	Hits@1	F_1
1. Hugging Face	16.28	80.7	19.5
2. ADAPT Centre	31.4	-	18.39
3. Happy Minions	29.01	-	16.01
4. High Five	-	65.9	-
5. Mohd Shadab Alam	29.94	13.8	16.91
6. Lost in Conversation	-	17.1	17.77
7. Little Baby (AI 小奶娃)	-	64.8	-
8. Sweet Fish	-	45.7	-
9. 1st-contact	31.98	13.2	16.42
10. NEUROBOTICS	35.47	-	16.68
11. Cats'team	-	35.9	-
12. Sonic	33.46	-	16.67
13. Pinta	32.49	-	16.39
14. Khai Mai Alt	-	34.6	13.03
15. loopAI	-	25.6	-
16. Salty Fish	34.32	-	-
17. Team Pat	-	-	16.11
18. Tensorborne	38.24	12.0	15.94
19. Team Dialog 6	40.35	10.9	7.27
20. Roboy	-	-	15.83
21. IamNotAdele	66.47	-	13.09
22. flooders	-	-	15.47
23. Clova Xiaodong Gu	-	-	14.37
Seq2Seq + Attention Baseline	29.8	12.6	16.18
Language Model Baseline	46.0	-	15.02
KV Profile Memory Baseline	-	55.2	11.9

Table 3: Automatic Metrics Leaderboard.

"revised" test set. Recall that we crowdsourced additional rewritten sets personas as a way of measuring how much models rely on word overlap between utterances and personas for their performance, as the revised ones have little or no overlap with the original personas. The results are shown in Figure 1. The Hugging Face team performed the best on the revised task, with Little Baby close behind. The performance of the baseline Key-Value Memory Network baseline greatly deteriorated given the revised personas. Hence, we found the success of the best competitor's models as a good result, which we believe is due to their use of sufficient pretraining and regularization, among other factors.

Last Utterance (Parrot) Distractor   We also evaluated how adding a distractor candidate affected the performance of these ranking models. Namely, we added the last partner message to the list of candidates to rank. A model should only in very rare circumstances
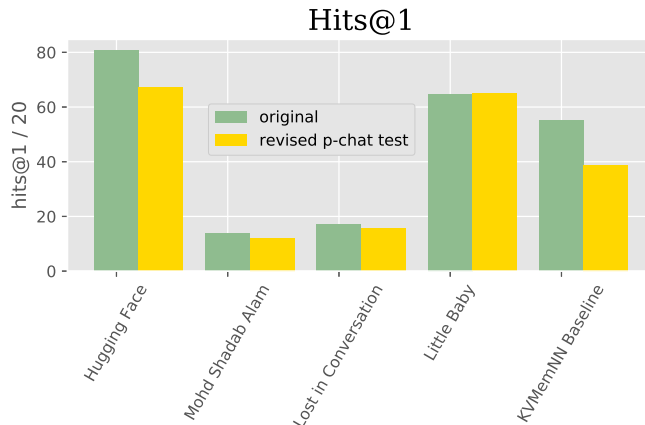
Figure 1: **Revised Test Set.** Hits@1 on the revised test set vs. on the regular test set.

parrot the speaking partner, so the Hits@1 metric should remain at a similar score with and without this distractor. See Figure 2 for the results. Most models suffered with this metric, showing they probably rely too much on word overlap with the last utterance when performing ranking (generally a response does have word overlap with the last utterance, but still it should not be a copy – this makes this a somewhat difficult function for models to learn). The Hugging Face model was the most resistant to this type of attack, but still suffered to some degree.

3.1.2 F_1 Metric Toy Baseline

During the automatic evaluation stage of the competition, we discovered that always re-plying with *"i am you to do and your is like"* would outperform the F_1 score of all the models in the competition. This toy baseline was constructed simply by picking several frequent words from the training set. Specifically, always replying with this message gives an F_1 score of 19.6 on the test set and 20.5 on the validation set (compare to Hugging Face's scores of 19.5 and 19.1 on the test and validation sets respectively). In [12], the authors showed that word overlap metrics do not correlate well with human judgment for dialogue response generation systems. This is another indication that we should reconsider using F_1 to evaluate our dialogue systems.

3.2 Human Evaluations

3.2.1 Wild Evaluation

Recall that we solicited volunteers to chat with and evaluate the top-performing models through the Facebook Messenger and Telegram APIs. This setup is closer to real-world
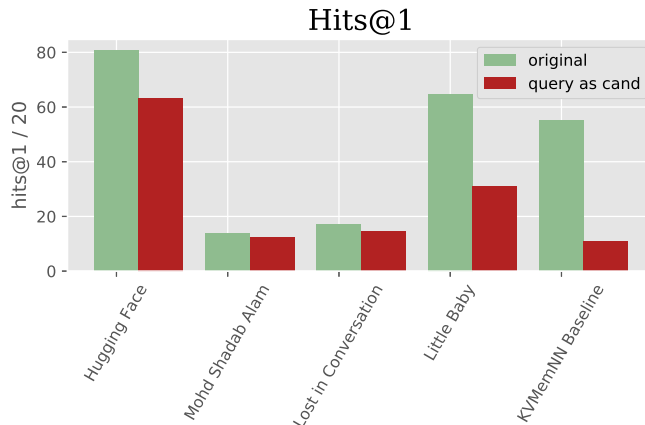
9

Figure 2: **Distractor Candidate.** Hits@1 on the test set when we add the query (last partner message) as a candidate.

chatbot applications than the Mechanical Turk evaluation set-up. As volunteers – unlike Turkers – are not paid and therefore less likely to follow the instructions closely, we expected the distribution of these results to be quite different than those from the Mechanical Turk evaluation. Two notable (but expected) differences between the results of the Mechanical Turk and "wild" evaluations are that the wild evaluation had relatively fewer good (i.e. long and consistent) dialogues, and Mechanical Turk workers tended to rate the models higher than the volunteers in the wild evaluation. A detailed description of our "wild" dataset and a comparison with the data from the first ConvAI competition will be described in detail in separate publication[8].

Given that many of the dialogues collected during the wild evaluation were either senseless, offensive, or simply not in line with the instructions, and we did not have a systematic way of distinguishing these from the "good" dialogues, we decided to discount the results from the wild evaluation altogether. See Table 4 for example conversations. The data collected from this evaluation is available in the competition repo[9]. Detecting this type of undesirable behavior remains an open (but we believe solvable) problem for evaluation of models in a live system.

3.2.2 Mechanical Turk Evaluation

Since the results of the wild evaluation were ultimately discounted, the winner of the human evaluation round – and therefore the winner of the competition – was determined by per-

---

[8]ConvAI2 dataset of non-goal-oriented human-to-bot dialogues (2019). V. Logacheva, V. Malykh, A. Litinsky, M. Burtsev

[9]`http://github.com/DeepPavlov/convai/data`

[HUMAN:]	Ok. Bye	[HUMAN:]	Hey!
[BOT:]	Bye.	[BOT:]	hello there how are you? do you like dogs?
[HUMAN:]	Bye	[HUMAN:]	Good ! I like dogs, I grew up in a farm
[HUMAN:]	Bye	[BOT:]	that is cool. i've a german shepard.
[HUMAN:]	Yo	[HUMAN:]	what is your dog's name ?
[BOT:]	What video games?.	[BOT:]	his name is pepper. do you have any pets?
[HUMAN:]	Bye	[HUMAN:]	I used to have a dog but he died.....
[BOT:]	OK.	[BOT:]	that is so sad! do you watch nhl hockey?
[HUMAN:]	Bye	[HUMAN:]	no... not a big fan of sport ?
[HUMAN:]	Bye	[BOT:]	no, i love watching it. i love the nhl
[**Score**:]	5/5	[**Score**:]	4/5

Table 4: Example of a bad conversation (left) and a good conversation (right) from the wild evaluation. The model received a score of 5 for the spam conversation, and 4 for the good conversation.

formance in the Mechanical Turk Evaluation. As announced at the NeurIPS Competition Track Workshop, the Lost in Conversation team won the competition.

The set-up of the Mechanical Turk evaluation was nearly identical to the set-up we used to collect the original PERSONA-CHAT dataset. The chat interface is shown in Figure 3. For each evaluation, we paired a human worker with a model, assigned each of them personas, and instructed the humans to chat with and get to know their partner. Dialogues were of length 4-6 turns each. Following a short conversation, we asked workers "How much did you enjoy talking to this user?" and had them answer on a scale of 1-4. We asked this question as a way of measuring how engaging the models were and used this as our primary evaluation metric. We used a single evaluation metric as is typically standard: many papers with human evaluations only report one type of metric (often quality or appropriateness) [13, 14, 15, 16]. Lastly, we tested whether the human could distinguish the persona the model was using from a random one. We crowdsourced 100 evaluations for each model. Samples conversations from some of the models are given in Appendix A.

The results are shown in Table 5. Lost in Conversation won the competition with an engagingness score of 3.11 out of 4. We attempted to reduce annotator bias in the engagingness scores by using a Bayesian calibration method recently proposed in [17]. The results from before and after calibration are given in Figure 4. The calibration did not affect the ordering of the scores, and the scores reported in the final leaderboard are post-calibration.

3.2.3 Further Analysis of Results

Length Statistics In an attempt to understand the results from the Mechanical Turk evaluations, we analyzed various word statistics on the conversation logs. We measured the average length of both the bot and human responses for each team's evaluation, as
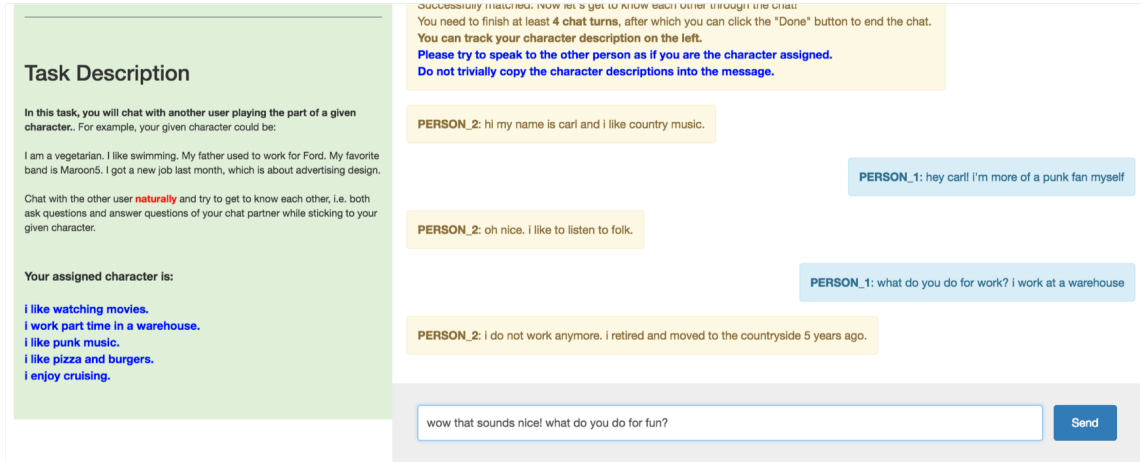
Figure 3: **Mechanical Turk Evaluation Interface.** The chat interface used for the Mechanical Turk portion of the evaluation was intentionally similar to the interface used to collect the original dataset.
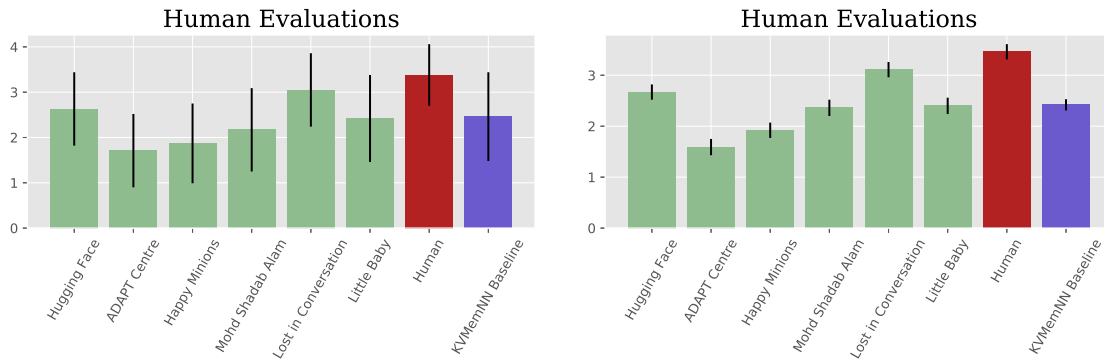

Figure 4: **Mechanical Turk Evaluation: Engagingness.** Results before (left) and after (right) Bayesian calibration. The calibration did not alter the ordering of the scores.

Team Names	Engagingness (1-4)	Persona Detection (0-1)
1. Lost in Conversation	3.11	0.9
2. Hugging Face	2.68	0.98
3. Little Baby (AI 小奶娃)	2.44	0.79
4. Mohd Shadab Alam	2.33	0.93
5. Happy Minions	1.92	0.46
6. ADAPT Centre	1.6	0.93
Human	3.48	0.96
KV Profile Memory (Baseline)	2.44	0.76

Table 5: Human Evaluation Results

Team Names	Engagingness (1-4)	# words (model)	# words (human)	# chars (model)	# chars (human)
1. Lost in Conversation	3.11	10.18	11.9	39.2	48.2
2. Hugging Face	2.67	11.5	11.9	44.4	49.2
3. Little Baby (AI 小奶娃)	2.4	11.5	11.3	51.5	47.3
4. Mohd Shadab Alam	2.36	9.5	10.2	33.8	42.5
5. Happy Minions	1.92	8.0	10.2	27.9	42.5
6. ADAPT Centre	1.59	15.1	11.8	60.0	48.0
Human	3.46	-	13.7	-	57.7

Table 6: **Average response length in Mechanical Turk logs.**

shown in Table 6. Models with higher evaluation scores tended to get longer responses from humans, which can be considered as an implicit engagement score. However, this is possibly skewed by humans mimicking the length of the bot's utterances, e.g. consider ADAPT Centre's results. We note that when humans are speaking with other humans, they have much longer utterances on average than the models do. We believe this is related to their production of more generic, less engaging utterances.

Rare Word Statistics We also looked to see how often rare words were used in the conversation logs. In Table 7, Freq1h and Freq1k indicate the frequency with which the model used words that appear fewer than 100 or 1,000 times in the training corpus. The hypothesis here is that utterances with some rare words might be less generic and hence more interesting/engaging, rendering higher human evaluation scores. The results show that humans use substantially more rare words than any of the models (aside from Little Baby, which is a retrieval rather than generative model), and the bottom three models do have lower Freq1h scores than the top three; otherwise, however, the relationship be-

Team Names	Engagingness (1-4)	Freq1h (model)	Freq1h (human)	Freq1k (model)	Freq1k (human)
1. Lost in Conversation	3.11	2.2	3.4	9.9	13.2
2. Hugging Face	2.67	2.5	4.2	9.0	15.6
3. Little Baby (AI 小奶娃)	2.4	4.9	3.7	18.3	15.6
4. Mohd Shadab Alam	2.36	1.3	3.2	9.5	14.1
5. Happy Minions	1.92	0.3	4.1	4.3	14.3
6. ADAPT Centre	1.59	1.7	3.5	8.8	15.1
Human	3.46	4.8	4.3	17.2	16.3

Table 7: **Rare word frequencies in Mechanical Turk logs.**

tween evaluation score of the models and their use of rare words is not completely clear. We suspect that is because this is just one factor among many that would need to be disentangled.

Word and Utterance Repetition Statistics We then looked at how often the models repeated themselves in conversations with humans. Table 8 shows the frequency of unigram, bigram, and trigram repeats in the model responses, as well as how often the model's responses were unique in the logs. Again, it is clear the humans repeat themselves very infrequently, but there is not a clear relationship between our proxy measures of repetition with the human evaluation scores. We suspect this is because there are more subtle instances of repeating that our proxies do not measure, and the proxies have already been optimized by many models (e.g. by doing n-gram or full utterance blocking). For example we observed instances like "i like watching horror" followed by "i love watching scary movies" occurring, but these are not captured well by our metrics. Finally, overall utterance uniqueness should ideally be close to 100% with the same utterance rarely being repeated across conversations, with humans at 99%. While Hugging Face's model was at 97%, many other models were lower, with the winner Lost in Conversation at 86%. A low uniqueness score could be problematic for a deployed system, as it might make users tire of it repeating itself. However, as our competition evaluations involve very short dialogues, this likely did not impact human evaluations.

Blind Evaluation Following the above analyses, it was still unclear why the Lost in Conversation model had a statistically significant human evaluation win over the Hugging Face model, even though the Hugging Face model performed much better in the automatic evaluations. To better understand this, we performed a blind evaluation ourselves of a random sample of the Mechanical Turk evaluation logs from these two teams, giving each conversation a score between 1 and 4 and making comments about the model's performance. The average score given to this subset of conversations is shown in Table 9. These scores

Team Names	Engagingness (1-4)	Unigram Repeats	Bigram Repeats	Trigram Repeats	Unique Responses
1. Lost in Conversation	3.11	2.11	5.6	2.67	86%
2. Hugging Face	2.67	1.49	5.04	0.6	97%
3. Little Baby (AI 小奶娃)	2.4	2.53	2.69	1.43	91%
4. Mohd Shadab Alam	2.36	3.48	11.34	7.06	83%
5. Happy Minions	1.92	1.62	6.56	3.81	53%
6. ADAPT Centre	1.59	6.74	11.53	1.44	98%
Human	3.46	1.83	2.47	0.51	99%

Table 8: **Repeats in Mechanical Turk logs.**

	Hugging Face	Lost in Conversation
Turker	2.8	3.29
Blind Annotator 1	2.47	2.78
Blind Annotator 2	2	2.71

Table 9: **Blind Evaluation Results.** Average engagingness score (1-4) for the randomly sampled subset of conversations.

show a clear example of annotator bias: Annotator 2 is more lenient than Annotator 1, and both are considerably harsher with their judgments than the average Turker. (The latter may be partially attributed to the fact that Turkers are paid to give ratings and therefore more likely to be generous with them.) Despite this, all annotators are in agreement about which model was better.

Asking questions   Reading through the comments made by the blind annotators afterwards, we noticed that while both models suffered from errors involving repetition, consistency or being "boring" at times, a common complaint about the Hugging Face model was that it "asked too many questions." In order to determine to what extent this was true, we analyzed the Mechanical Turk logs and measured how often each model response began with a question word (like "who," "what," "when," "where," "why," or "how") and how often the response contained a question mark.

The results are given in Figure 5. It is clear that the Hugging Face model is indeed a large outlier. Notably, you can see that in the 100 conversations it had, it began a response with a question word 107 times whereas humans only did this 12 times. When the model asks too many questions it can make the conversation feel disjointed, especially if the questions do not relate to the previous conversation. Friendly chit-chat requires a delicate balance of question-asking and question-answering. The tentative conclusion
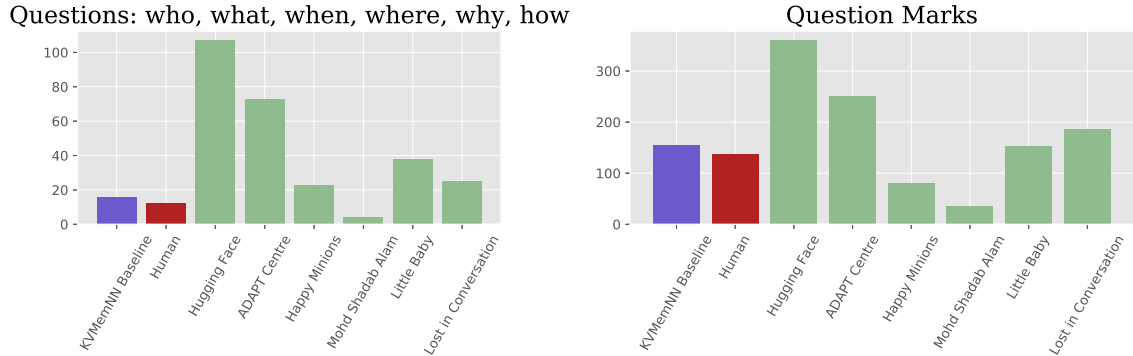
Figure 5: **How often did the models ask questions?** We measured (on the left) how often the models began their response with "who," "what," "when," "where," "why," or "how," as well as (on the right) how often the models' responses contained at least one question mark as an estimate for how often the models asked questions when conversing with humans.

that we draw here is that the tendency to ask too many questions negatively affected the human evaluation results for the Hugging Face model. Future work should consider how we can automatically evaluate this type of conversation-level performance rather than just utterance-level performance.

Persona Detection   Lastly, looking at the persona detection scores from the Mechanical Turk evaluation in Table 5, we note that most models did relatively well in this metric (with the exception of the Happy Minions model). Recall that this score is the percentage of the time that the annotaters were able to to distinguish the model's persona from a random one. We often observed models repeating the persona sentences almost verbatim, which might lead to a high persona detection score but a low engagingness score. Training models to use the persona to create engaging responses rather than simply copying it remains an open problem.

4   Participating Models

We include a short summary of the model types used for some of the top competitors in Table 10. Some of the authors of these models plan to write detailed papers describing their models[18]. Please also refer to the slides at the website written by the model's authors[10]. The winner's (Lost in Conversation's) code is also publicly available[11].

---

[10]http://convai.io/NeurIPSParticipantSlides.pptx

[11]https://github.com/atselousov/transformer_chatbot

Team Names	Model Summary
Lost in Conversation	Generative Transformer based on OpenAI GPT. Trained on PERSONA-CHAT (original+revised), DailyDialog and Reddit comments.
Hugging Face	Pretrained generative Transformer (Billion Words + CoNLL 2012) with transfer to PERSONA-CHAT [18].
Little Baby	Profile-Encoded Multi-Turn Response Selection via Multi-Grained Deep Match Network. Modification of [19]: better model + data augmentation via translation.
Mohd Shadab Alam	Seq2Seq + Highway model. Glove + language model vector. Transfer learning strategy for Seq2Seq tasks.
ADAPT Centre	Bi-directional Attentive LSTM. Pretrained via GloVe embeddings + Switchboard, Open Subtitles.

Table 10: **Brief model descriptions of some of the top competitors.**

5 Conclusions and Future Work

There are several important takeaways from the 2018 competition. We conclude that pretrained Transformer variants are currently the best performing models on this task, but future systems must go beyond single word metrics like perplexity to measure the performance across sequences of utterances (conversations), as these metrics fail to capture multi-turn aspects of human evaluation. The latter has implications both for the quality of the automatic metrics evaluating our models, and the quality of our model training objectives.

Models The best models in the competition were variants of the generative Transformer architecture. Those models have rather high capacity and thus cannot be trained on ConvAI2 (PERSONA-CHAT) data alone, but must be either pretrained or multitasked with additional large datasets. One can use dialogue datasets to pretrain, but it seems as though the system still works well with language modeling datasets that are not explicitly dialogue (e.g. the Billion Words corpus). Many other tweaks to the base models were tried, such as trying to optimize the automatic metrics directly, but without direct ablations with human evaluation it is difficult to state here the effects of all these components.

Retrieval models fared a little worse than generative models in the human evaluations, although we are unsure if this is true in general, or because no very strong retrieval model was proposed. With a Transformer-based retrieval model it is possible to get Hits@1 in excess of 80% but no such method was tried by a competitor (see Table 3, Hugging Face used a two-head Transformer model, but opted to generate rather than retrieve). In our opinion, looking at the outputs from the generative systems in the competition, they still fall short of the most interesting and engaging comments of humans (which sometimes

retrieval models choose); however, the generic responses from generative models are often low-risk or "safe" responses, which may give them higher scores. A retrieve and refine approach (combining generative and retrieval methods) is another possibility that was not explored in the competition [20].

Finally, better sentence representations are being developed all the time. This competition was run before the BERT model [21] was released which has been shown to improve many NLP tasks. Hence, we expect these models to improve on ConvAI2 as well. In fact, some improved results have already been reported since the competition took place [22].

Automatic vs. Human Evaluation It remains an open problem to find the best automatic evaluation metrics for dialogue. There is not enough data from the competition to measure correlation between the automatic metrics we tried and human evaluations in depth. Clearly a randomly initialized model has poor values for all of these metrics, whereas training to optimize any of them will improve human evaluations. The problem is more whether the finer-grained differentiation of relatively similar models can be automatically measured. We believe each automatic metric evaluates at least some aspects of what humans consider a "good" model but misses other aspects. As such, optimizing only one of these metrics can fail to address important issues. For example, optimizing per-word perplexity fails to address the search strategy of a model when generating a full utterance, e.g. it is not affected by beam search choices. Optimizing Hits@1 is a per-utterance metric that fails to address the full conversational flow (as the gold dialogue history between two humans is used for that metric, not what the model previously said). Some models optimize F_1 and do well, however it also has major issues (see Section 3.1.2). Further, it is very hard to compare retrieval and generative models other than by human evaluation.

Nevertheless, we find the use of automatic metrics important for several reasons. If we desire to be able to train our models offline at least initially (which we believe we do) then we need an offline training objective, which typically relates to offline metrics. Hence, if we understand how human evaluations relate to offline metrics, not only will we understand the dialogue task better, but we will know how to perform such offline training. Additionally, for our competition it would have been very difficult to filter models for the human evaluation stage without the use of automatic metrics.

Towards Multi-turn Evaluation We thus believe we are still missing some key offline metrics, but have hope that they are possible to find. We identified that the current metrics fail to measure the multi-turn aspects of human evaluation, in particular in terms of repetition, consistency and balance of dialogue acts. Even the best competitors' models often failed to be self-consistent across a few dialogue turns, which we believe was at least partly responsible for lowering their evaluation score. For example, "i am a professional runner. you? i love running" followed by "i'm not very athletic" or "i work as a snowboard instructor" followed by "i work for a food company" are both unlikely continuations of a

18

conversation. Even if they happen infrequently, these problems are particularly jarring for a human speaking partner when they do happen.

In a related problem, we observed the models asking questions that are already answered, e.g. one model asks "what do you do for a living?" even though the human earlier stated "i work on computers" resulting in the human replying "I just told you silly".

One possible solution to these problems is the use of dialogue natural language inference (NLI) [23], a new task that has been proposed that evaluates exactly these problems. It works by providing pairs of utterances as input, and the task is then to predict if the pair entail, are neutral or contradict. This is exciting because it can allow us to both (i) fix our model's consistency problems by training on this new task and (ii) evaluate to what extent our model's consistency problems are fixed using the evaluation set.

Finally, in Section 3.2.3 we identified that models that do not balance question asking with answering over multiple turns might can cause human evaluations to suffer. Given this information, it may be possible to construct new metrics that measure these kind of balances so that we can optimize them (to look more similar to human data, for instance). A recent paper, published after the competition took place, studies methods of generative control for optimizing such multi-turn attributes with promising results [24].

Towards more complex tasks Going forward, even if we can completely solve the ConvAI2 (Persona-Chat) task (i.e. reach human performance), it is still only a meet-and-greet task involving getting to know someone for a few dialogue turns, with shallow topics and quick context switches. Clearly many aspects of an intelligent agent are not evaluated by this task, such as the use of long-term memory or in-depth knowledge and deeper reasoning. For example, in Table 1 "Game of Thrones" is mentioned, but a model imitating this conversation would not really be required to know anything more about the show, as in ConvAI2 speakers tend to shallowly discuss each other's interest without lingering on a topic for too long. Subsequent competitions could explore this issue further. Such a study is feasible as several new datasets are being released to explore such a setting, in particular the Wizard of Wikipedia task involves using knowledge from Wikipedia to discuss open-domain topics [25]. The DSTC7 competition[12] also recently addressed this topic, however the evaluation was not multi-turn.

6 Acknowledgements

---

[12]http://workshop.colips.org/dstc7/

References

[1] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[2] Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*, 2016.

[3] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[4] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[5] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[6] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 285–294. The Association for Computer Linguistics, 2015.

[7] Wenchao Du and Alan W. Black. Data augmentation for neural online chats response selection. In Aleksandr Chuklin, Jeff Dalton, Julia Kiseleva, Alexey Borisov, and Mikhail Burtsev, editors, *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 52–58. Association for Computational Linguistics, 2018.

[8] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. Multi-turn response selection for chatbots with deep attention matching network. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1118–1127. Association for Computational Linguistics, 2018.

[9] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing, 2017.

[10] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3776–3784. AAAI Press, 2016.

[11] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.

[12] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *CoRR*, abs/1603.08023, 2016.

[13] Oriol Vinyals and Quoc V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015.

[14] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In Su et al. [26], pages 1192–1202.

[15] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics, 2016.

[16] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Su et al. [26], pages 2122–2132.

[17] Ilya Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. Importance of a search strategy in neural dialogue modelling. *CoRR*, abs/1811.00907, 2018.

[18] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, 2019.

[19] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*, 2016.

[20] Jason Weston, Emily Dinan, and Alexander H Miller. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*, 2018.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Real-time inference in multi-sentence tasks with deep pretrained transformers. *arXiv preprint arXiv:1905.01969*, 2019.

[23] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*, 2018.

[24] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*, 2019.

[25] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

[26] Jian Su, Xavier Carreras, and Kevin Duh, editors. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. The Association for Computational Linguistics, 2016.

A   Example Dialogues

Example dialogues for some of the top models are given in Figures 6, 7, 8, 9, 10, and 11.

[BOT:]	hello, how are you today?
[HUMAN:]	I'm great! I just got back from skydiving! I love doing that!
[BOT:]	that sounds like fun! i love to paint.
[HUMAN:]	Awesome! What do you like to paint?
[BOT:]	i love to draw portraits of my favorite cars.
[HUMAN:]	I'd love to see them! I live in Arizona, where do you live?
[BOT:]	i live in california. i love the beach!
[HUMAN:]	I've been to California! I love to shop at their Barnes and Noble store!
[BOT:]	i love shopping at kroger!
[HUMAN:]	Kroger is great! I wish they sold pet food for my pet lizard though!

[**Score**:]   4/4

Figure 6: Lost in Conversation: example Mechanical Turk conversation.

[HUMAN:]	Hello there, I like rock climbing do you have any hobbies?
[BOT:]	i like nature. do you? i like nature.
[HUMAN:]	I do love nature, especially when going on walks
[BOT:]	what is your favorite rock? mine is rock.
[HUMAN:]	Well I like rock music, does that count?
[BOT:]	i think it does. what do you do for fun?
[HUMAN:]	Well I work at a restaurant that I love, and rock climbing
[BOT:]	that sounds fun! i'm into nature.
[HUMAN:]	Cool, so am I
[BOT:]	do you have any pets? do you have pets?

[**Score**:]   2/4

Figure 7: Hugging Face: example Mechanical Turk conversation.

[BOT:]	whats its name, she good off leash?
[HUMAN:]	I don't actually have any dogs.
[BOT:]	its name is noodle she does not need a leash i talk to her in my sleep
[HUMAN:]	Wow! Does she talk back
[BOT:]	why do you not believe her
[HUMAN:]	I don't even know what she said.
[BOT:]	after my workouts i drink a protein shake instead coffee, do not like coffee
[HUMAN:]	Alright, that's cool.

[**Score**:]   1/4

Figure 8: Little Baby: example Mechanical Turk conversation.

[BOT:]	hi how are you today?
[HUMAN:]	Hello how are you?
[BOT:]	i'm good. just got done reading a book.
[HUMAN:]	Which book did you read?
[BOT:]	i don't read much, i am more into reading
[HUMAN:]	Where do you live?
[BOT:]	i live in the us. i love to read
[HUMAN:]	I live in Alaska
[BOT:]	i live in the us, i love to read
[HUMAN:]	Do you like to read?

[*Score*:]   1/4

Figure 9: Mohd Shadab Alam: example Mechanical Turk conversation.

[BOT:]	i love to read
[HUMAN:]	Nice, I'm more of a numbers guy myself. I use to be an accountant.
[BOT:]	well, i am not sure what you mean.
[HUMAN:]	I worked with numbers. I recently moved away to a beach house because I got tired of it
[BOT:]	i am sorry to hear that. do you have any hobbies?
[HUMAN:]	Just hanging out with my dog, I guess. Trying to start a new business.
[BOT:]	i am not sure what you mean.
[HUMAN:]	Exactly, it's hard to find meaning
[BOT:]	yeah, i guess. i am not sure what that means.
[HUMAN:]	I do work, for pay, occasionally, as a business person.

[*Score*:]   1/4

Figure 10: Happy Minions: example Mechanical Turk conversation.

[HUMAN:]	Hello, have you ever gone scuba diving?
[BOT:]	no i do not have a job yesterday. do you have any hobbies
[HUMAN:]	I love going scuba diving in hawaii.
[BOT:]	i love scuba diving yesterday. do you like to swim? i do
[HUMAN:]	It's ok, I prefer to play with my pet snakes.
[BOT:]	i love to swim yesterday. i like to go to the beach with my job
[HUMAN:]	I just got divorced so maybe I should go to the beach.
[BOT:]	you should go with your job yesterday. do you go to a lot of the beach
[HUMAN:]	Only when I visit hawaii.
[BOT:]	i love the beach so i have to go with my new job
[HUMAN:]	I traveled the Americas playing guitar

[*Score*:]   2/4

Figure 11: ADAPT Centre: example Mechanical Turk conversation.