A Dataset of Topic-Oriented Human-to-Chatbot Dialogues

Varvara Logacheva Mikhail Burtsev Valentin Malykh
Vadim Poluliakh Alexander Rudnicky Iulian Serban
Ryan Lowe Shrimai Prabhumoye Alan W Black
Yoshua Bengio

February 21, 2018

Abstract

This document contains the description of dataset collected during the first round of Conversational Intelligence Challenge (ConvAI) which took place in July 2017. During this evaluation round we collected over 2,500 dialogues from 10 chatbots and 500 volunteers. Here we provide the analysis of dataset statistics and outline some possible improvements for future data collection experiments.

0 Disclaimer

Dialogues collected in this dataset can contain strong words and insults. Views and opinions expressed by chatbots as well as human volunteers who participated in data collection do not necessarily reflect the position of authors.

1 Introduction

The development of dialogue systems is hampered by the inability to evaluate them automatically. This problem is particularly crucial for non-goal-oriented dialogue systems (chatbots). In contrast to goal-oriented dialogue systems, chatbots do not have any formal criterion of successful conversation. Their quality is based on user experience, it cannot be easily formalised.

A recently suggested solution of this problem is to train a model to predict user rating of dialogue. Such model should be trained on real user scores. However, the existing models cannot perform well enough to replace human scores. One of the main obstacles to good quality is the insufficient training data. While there exist many datasets of human-to-human conversations [], human-to-bot conversations with quality labellings are scarce.

The First Conversational Intelligence Challenge (ConvAI) aimed at evaluating the performance of chatbot systems. Human volunteers conversed with chatbots and evaluated them. As a byproduct of this evaluation we acquired a dataset of human-to-bot conversations labelled for quality. This data can be used to train a metric for evaluating dialogue systems. Moreover, it can be used in the development of chatbots themselves: it contains the information on the quality of utterances and entire dialogues, that can guide a dialogue system in search of better answers.

We describe the statistics of collected data, analyse its properties and outline some possible improvements for future data collection experiments.

2 Data collection

The aim of the competition was to establish a task for evaluating non-goal-oriented dialogue systems. Such systems do not have any particular goal in conversation. In order to fully test their capabilities and make the task more formal we specified a constraint on the topic of conversations: a chatbot and a human volunteer should discuss an excerpt from a Wikipedia article that we provide. These texts were taken from the SQuAD dataset [Rajpurkar et al., 2016]. The peers are encouraged (but not strictly required) to discuss this text.

We created a framework for collecting conversations of humans and bots which operates on Telegram and Facebook messaging service. When a user starts a conversation, the framework randomly assigns her a bot or another user, so the user does not know if s/he is talking to a bot or a human.

During the conversation user can evaluate the quality of her peer's answers. Below every peer's utterance a user is shown two buttons: "thumbs up" and "thumbs down" to indicate whether the answer was appropriate or inappropriate, respectively. This evaluation is not compulsory, user can continue the conversation without giving scores to peer's utterances.

After the conversation is finished, user is asked to evaluate the whole dialogue along three dimensions: overall quality of dialogue, breadth of dialogue (how thoroughly peers discussed the topic suggested in the opening text) and engagement of peer. All three parameter are given scores from 1 (bad) to 5 (good).

3 Dataset

3.1 Statistics of dialogues

The dataset contains the total of 4,750 dialogues. These include 4,224 human-to-bot dialogues and 526 human-to-human conversations. The average number of utterances per dialogue is 10.5 and the average utterance length is 7.1 words.

The statistics of dialogues are summarised in table 1. These statistics are computed for the whole dataset. When computing the number of words per utterance we excluded 20 dialogues which contained utterances of over 100 words — these were cases when users were copying and pasting dialogue *context* or typing in other senseless answers.

We give joint statistics for all dialogues as well as separate figures for human-to-human and human-to-bot dialogues, as these turn out to be different in some respects. Two humans usually have longer and more diverse (in terms of the number of used words) conversations than a human and a bot. These facts are apparently related: if a peer uses richer vocabulary, s/he is better at capturing his/her partner's attention for a longer time. On the other hand, utterances themselves are shorter in human-to-human dialogues. This probably suggests that humans try to be more explicit when talking to machine so that it understands them better.

However, some of these dialogues contain zero utterances — this means that a user finished a dialogue without saying anything. Also, there are non-empty dialogues where all utterances come from one user. The distribution of dialogue lengths is shown in figure 1. It can be seen that over 700 dialogues contain 0 to 5 utterances, and dialogues of over 40 utterances are extremely rare.

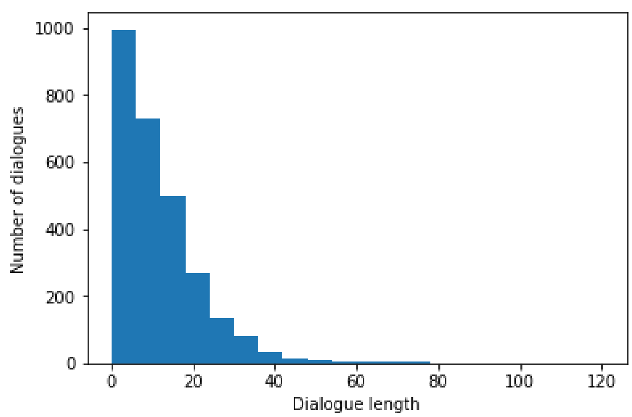

Figure 1: Distribution of dialogue lengths in utterances.

The statistics of dialogues are summarised in table 1. We give joint statistics for all dialogues as well as separate figures for human-to-human and human-to-bot dialogues, as these turn out to be different in many respects. Two humans usually have longer and more diverse (in terms of the number of used words) conversations than a human and a bot. These facts are apparently related: if a peer uses richer vocabulary, s/he is better at capturing his/her partner's attention for a longer time.

On the other hand, utterances themselves are shorter in human-to-human dialogues. Figure 2 shows that humans in general generate shorter utterances. This is apparently explained by the fact that some bots use retrieval approaches, i.e. select their answers from a database. It contains meaningful and grammatically correct sentences which are usually relatively long. On the other hand, users can output extremely short answers (e.g. "?", "O!", ":)" etc).

	All dialogues	Human-to-bot	Human-to-human
Total number of dialogues	2,778	2,337	441
Empty dialogues	119 (4.3%)	102 (4.4%)	17 (3.9%)
One-sided dialogues*	560 (20.2%)	520 (22.3%)	40 (9.1%)
Long dialogues**	1719 (61.9%)	1409 (60.3%)	310 (70.3%)
Utterances per dialogue	10.95	10.73	12.13
Words per utterance	7.31	7.48	6.51
Characters per utterance	32.54	33.60	27.56
Unique words per dialogue	45.72	45.00	49.58

Table 1: Dataset statistics: number of dialogues with different characteristics.
* *one-sided* dialogues are dialogues where one of users did not produce any utterances.
** *long* dialogues are dialogues consisting of at least three turns, where one turn is an utterance from one user + utterance from another user.

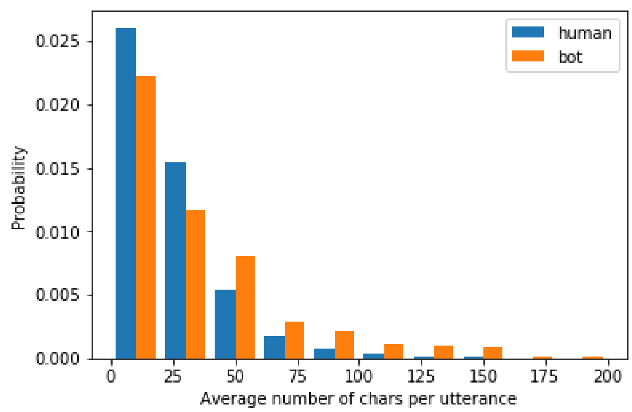

Figure 2: Distribution of utterances length in characters for humans and bots.

Filtering of dataset We are interested in getting a clean dataset of high quality. It should not have empty dialogues or dialogues which do not have any user evaluation (as those are useless for training of a dialogue evaluation metric).

We filtered the dataset according to these two parameters: we used only long dialogues (having at least 2 utterances from each users) and dialogues which had at least one utterance-level score. This left us with a half of the initially collected dialogues.

As we see, the present size of the dataset is barely suitable for training of models that can evaluate bot quality at the utterance level, because there are not enough utterance-level scores. However, all dialogues have dialogue-level scores.

	All dialogues	Human-to-bot	Human-to-human
Total	2,778	2,337	441
Long dialogues	1719 (61.9%)	1409 (60.3%)	310 (70.3%)
Long & one or more utt. rated	1356 (48.8%)	1085 (46.4 %)	271 (61.5 %)
Long & 50% or more utt. rated	1058 (38.1%)	886 (37.9%)	172 (39%)
Long & all utt. rated	603 (21.7%)	564 (24.1%)	39 (8.8%)

Table 2: Size of filtered dataset.

4 Evaluation of dialogues

4.1 Dialogue-level evaluation

After the end of a dialogue a user was asked to rate it in terms of three parameters: peer's engagement, breadth and overall quality of dialogue. Similarly to previous experiments on dialogue data collection [Lowe et al., 2017], we found that these three dialogue-level metrics are strongly correlated: Pearson r scores between any two of those metrics is 0.86 to 0.87.

The distribution of overall quality scores is plotted in figure 3. The number of bad dialogues exceeds that of others: 42% of dialogues were rated with "1".

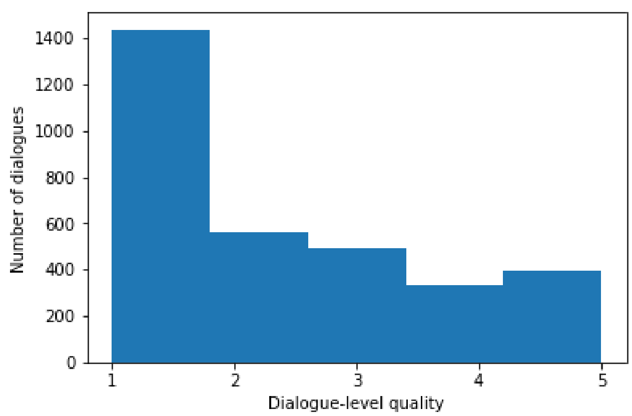

Figure 3: Distribution of dialogue-level quality scores.

Human vs bot dialogue scores As with other statistics, quality of human and bot dialogues differ significantly. Figure 4 shows the distribution of dialogue-level quality scores separately for bots and humans. As expected, humans perform much better. However, in around 30% cases participants of human-to-human dialogues still rated their peers' performance as bad. This can indicate that users produced utterances which were irrelevant to the previous dialogue. Alternatively, a user could just dislike answers from her peer, although they were grammatical and relevant. Finally, a user could just rate dialogues randomly.

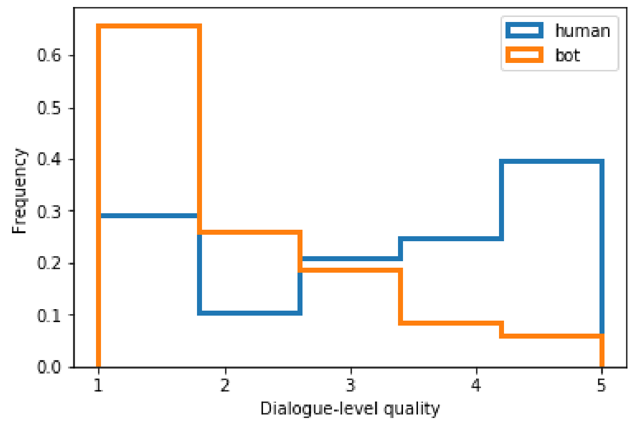

Figure 4: Distribution of dialogue-level evaluation of human and bot peers.

Reliability of scores The latter situation occurs in human-to-bot dialogues as well. We conducted manual analysis of some human-to-bot dialogues that were rated high (had a score of "5") and found out that some of these dialogues are of quite low quality and shouldn't be considered as examples of successful human-bot interaction (see an example below). This observation requires more thorough analysis. We suggest that highly-rated dialogues should be re-rated, preferably by several users.

```
   In 550 BC, Cyrus the Great, son of Mandane and Cambyses I,
took over the Median Empire, and founded the Achaemenid Empire
by unifying other city states.  The conquest of Media was a result
of what is called the Persian Revolt.  The brouhaha was initially
triggered by the actions of the Median ruler Astyages, and was
quickly spread to other provinces, as they allied with the Persians.
Later conquests under Cyrus and his successors expanded the empire
to include Lydia, Babylon, Egypt, parts of the Balkans and Eastern
Europe proper, as well as the lands to the west of the Indus and
Oxus rivers.
   User:  I will learn that the conquest of Media was a result
of what is called the Persian Revolt .  Who told you that?
   User:  Hello!
   Bot:  a result of
   User:  How old are you?
   Bot:  spread
   User:  Spread what?
```

Bot: The

An example of unsuccessful human-to-bot dialogue rated by a human with "5" for quality, breadth and peer engagement.

Multiple dialogue-level scores Note that dialogue-level evaluation was provided by all users, so human-to-human dialogues were evaluated twice. This gave us a possibility to compare the evaluation of the same dialogue by both participants. Figure 5 shows the distribution of differences between dialogue-level quality scores given by two human interlocutors. It can be seen that scores given by different users were quite consistent: in 69% cases the difference between scores does not exceed 1 (i.e. participants rated a dialogue with the same or neighbouring scores). The Spearman correlation of the scores is 0.45.

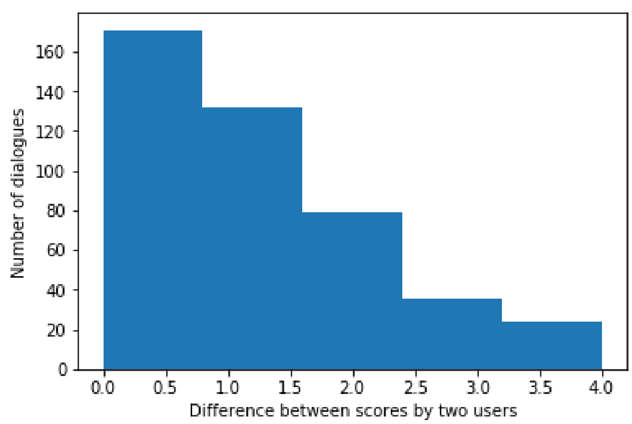

Figure 5: Distribution of differences between dialogue-level scores given to the same dialogues by two humans.

4.2 Utterance-level evaluation

As opposed to the quality of dialogues, quality of utterances was evaluated in terms of a binary scale. This task is apparently difficult to perform during the conversation: 45.6% of utterances were not rated. On the other hand, there can be a different interpretation of the absence of score: a user might not be sure whether a response was good or not. We suggest that in next data collection experiments utterance-level scores should be ternary (analogously to [Yu et al., 2016] where an utterance can be classified as "Appropriate", "Inappropriate" or "Interpretable", with the latter meaning that an utterance did not fit to the context perfectly, but could still be interpreted as an adequate answer). Another alternative would be to

Ditribution of rated/unrated utterances In order to better understand why we got so few utterance-level scores we performed analysis of scores. Our intuition is that if unrated items mean ambiguous quality, then percentage of such items should be close for all dialogues. On the other hand, if some users do not rate utterances because they find on-the-fly evaluation difficult, the distribution of ranked utterance within a dialogue will be user-dependent.

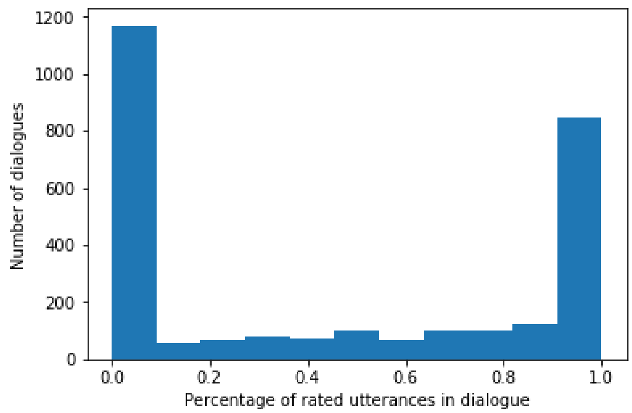

Figure 6: Distribution of dialogues by percentage of rated utterances.

We discovered that the latter hypothesis is true. We plotted the distribution of dialogues by the percentage of rated utterances in them (see figure 6). Almost 1200 dialogues (40.7%) have no rated utterances. On the other hand, half of the rest (802 or 28.8%) have all utterances rated. The rest 30% are partially rated with utterance-level scores. Interestingly, the plot in figure 6 shows a slight upward trend for these dialogues — there are more dialogues with higher percentage of rated utterances. The average percentage of rated utterances for dialogues with at least one rated utterance is 74.8%. This means that if a user rates utterances in a dialogue, s/he tries to rate all of them.

Percentage of rated utterances vs dialogue length One of the reasons of such behaviour might be the length of a dialogue. It could happen that as a conversation gets longer, user gets tired of giving ranks to items, so longer dialogues might have smaller percentage of rated utterances. In order to test this hypothesis we examined the distribution of rated utterance percentages with respect to dialogue lengths (see figure 7). It is clearly seen that the percentages are distributed evenly across dialogues with different lengths. Therefore, user's commitment to ranking dialogues holds till the end of a dialogue regardless of its length.

Varying user behaviour As it was already briefly mentioned before, the vast majority of dialogues was contributed by a relatively small number of users

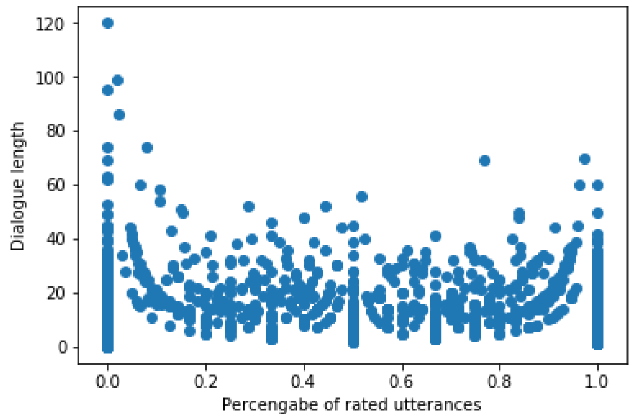

Figure 7: Relationship between the percentage of rated utterances in a dialogue and dialogue length.

(85% conversations were conducted by 29 volunteers). Therefore, we decided to examine their behaviour in terms of utterance-level scores. Figure 8 shows that almost a third of these users rated a small (10% or less) proportion of utterances. Another large group of users clusters around 50%. And only 4 users always rated the majority of utterances, i.e. have an average of 60% or more. This observation again confirms the hypothesis about difficulty of on-the-fly evaluation of dialogues.

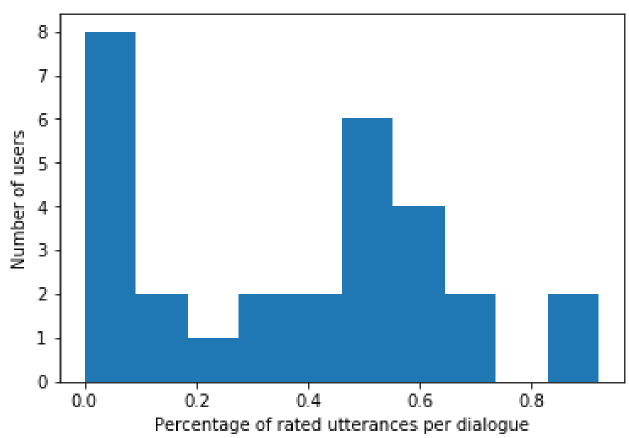

Figure 8: Distribution of active users by average percentage of rated utterances per dialogue.

Utterance-level scores The distribution of utterance-level quality scores themselves is shown in figure 9 (unrated utterances were discarded). As with dialogue-level scores, here humans perform much better, but also occasionally produce some utterances which were rated as bad by a peer: 13.5% rated user responses were considered inappropriate. Among rated bot utterances, 58.6% are inappropriate.

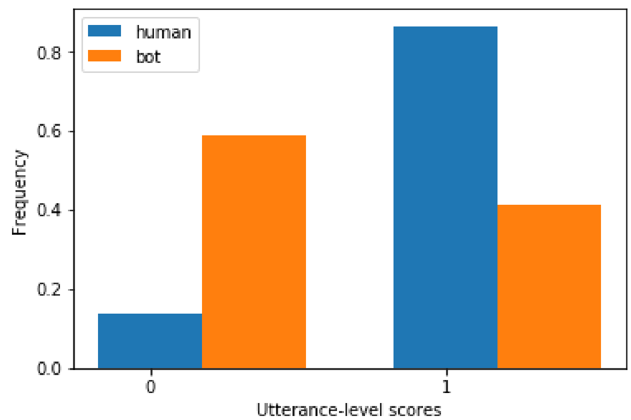

Figure 9: Distribution of utterance-level quality scores for humans and bots.

Dialogue-level vs utterance-level scores We were also interested to see if utterance-level scores matched the dialogue-level ones: if the overall dialogue is good, are individual utterances also appropriate within the dialogue? In order to check that we took an average of utterance-level and dialogue-level scores and computed their correlations (only for dialogues where at least one utterance was rated). It turns out that utterance-level and dialogue-level scores correlate quite strongly — their Pearson r score is 0.6. The plot in figure 10 shows their correspondence.

Figure 10 shows correlation of the averaged dialogue-level and utterance-level scores. However, the similar level of correlation holds for individual dialogue-level metrics: table 3 shows Pearson r score for dialogue-level quality, breadth and engagement separately.

Metric	Pearson r
Quality	0.585
Breadth	0.564
Engagement	0.550
Averaged	0.599

Table 3: Correlation of dialogue-level metrics with utterance-level user scores.

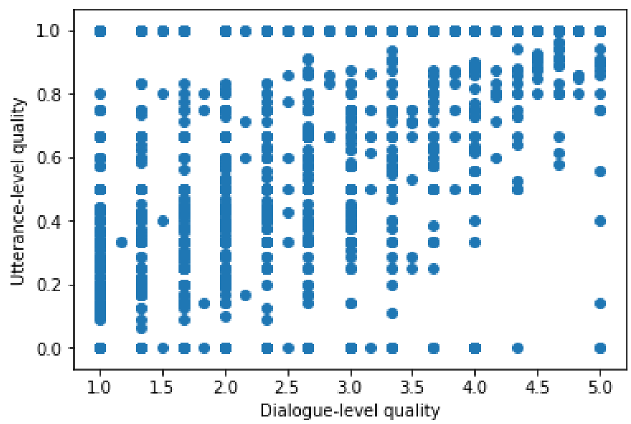

Figure 10: Relationship between dialogue-level and utterance-level quality scores.

4.3 Dialogue statistics vs user scores

Simple statistics Let us see if we can approximate real dialogue quality (i.e. user scores) with any other dialogue properties. First, we compare user scores against various quantitative parameters. We observe moderate correlation of dialogue quality the number of unique words (Pearson r score of 0.39), unique trigrams (0.35) and the number of utterances (0.31) in dialogue. That is reasonable, because a longer dialogue means that a chatbot managed to say something interesting to a user and attract his/her attention, and large number of unique tokens or ngrams in a dialogue implies a diverse conversation.

Use of dialogue context Besides that, we decided to check how useful is the context that is provided in the beginning of every conversation. We suggested that all participants of a dialogue discuss the provided paragraph of text, hence adding an implicit goal to conversation. Now we want to check if the contexts were used as conversation topics. We do that by checking if the most characteristic words of the context appeared later in the conversation.

We define the most characteristic words as words with the highest tf-idf score. This score is computed for a collection of documents (in our case a collection of paragraphs used as contexts) and is high for words which occur often in the current document and rarely in other documents — this means that these words are representative for this document. We compute tf-idf score for each word in all contexts. Then we take 15 words with the highest tf-idf score from each context and compute how many times any of these words occurs in the corresponding dialogue. This gives an indication of whether the participants discussed the topic of the context.

It turns out that almost half of dialogues does not contain any of representative words. This means that in half of cases neither users nor bots tried to

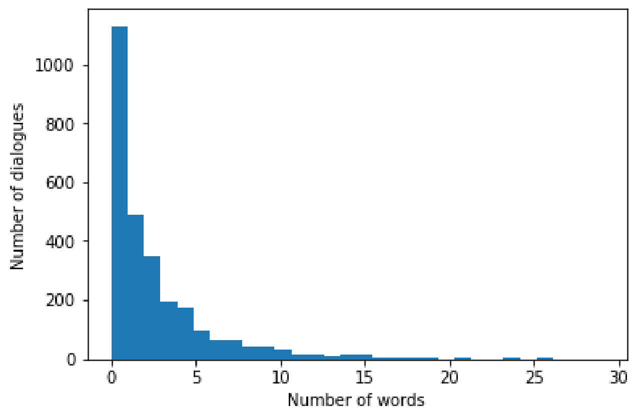

Figure 11: Number of occurrences of top-15 representative words from the context in the dialogue.

discuss the suggested paragraph. Another observation is that there is only weak correlation between the breadth of conversation and the use of representative words (0.16). The *breadth* evaluation metric was supposed to capture how good the conversation was in terms of coverage of a suggested topic. However, this weak correlation suggests that either the use of representative words does not mean that topic has been covered, or users did not understand the purpose of the *breadth* metric.

Most of correlations we report above are for dialogue-level *quality* metric. However, close correlations are observed for other metrics (see table 4). This confirms the high correlation of *quality*, *breadth* and *engagement* scores given by users.

Metric	Quality	Breadth	Engagement
# of unique words	0.408	0.364	0.425
# of unique trigrams	0.368	0.319	0.387
# of utterances	0.321	0.283	0.334
# of topic words	0.199	0.164	0.181

Table 4: Correlation of dialogue-level scores and dialogue statistics.

5 Quality of individual bots

We computed quality of individual bots at the dialogue and utterance levels by averaging all scores for a bot. Note that we did not consider unrated utterances and short dialogues (dialogues with 2 or less utterances for each participant).

During the human evaluation round we added some extra bots to the ones that participated in the competition. That was done in order to increase the diversity of bots (and to prevent users from running into the same bot all the time). We computed human scores for those additional bots as well, but they will not be included to the official ranking. In the tables below names of bots which participated in the competition are written **in bold**.

Bot name	Quality	Engagement	Breadth	**Total**
poetwannabe	2.366	2.310	2.207	2.294
DATA Siegt	2.320	2.400	1.953	2.224
bot#1337	2.295	2.219	2.094	2.203
RLLChatBot	2.228	2.244	2.024	2.165
Plastic world	2.181	2.319	1.993	2.164
poetess	2.172	2.207	2.069	2.149
kAIb	2.011	1.991	1.780	1.928
Q&A	2.000	1.833	1.833	1.889
DeepTalkHawk	1.427	1.433	1.401	1.420
PolyU	1.329	1.286	1.271	1.295

Table 5: Dialogue-level quality of bots. Bots are sorted from best to worst according to the averaged values of all metrics (**Total** column). Bots **in bold** are those which participate in the official competition.

Table 5 shows the average scores of individual dialogue-level metrics and the average of all scores given to dialogues of a bot (shown in the rightmost column). As it has already been shown, the scores are mostly bad and not very diverse: the average values for bots range from 1.3 to 2.3. The three dialogue-level metrics are strongly correlated at the system level (i.e. rankings of systems under different metrics are very close), therefore, we use the average of all dialogue-level scores to rank the bots.

The utterance-level scores produce a slightly different ranking of bots (shown in table 6). However, it shows strong correlation (Pearson r of 0.85) with dialogue-level ranking of the same systems. Here we see a larger variation: the average utterance-level scores range from 0.5 to 0.06. This ranking is not guaranteed to be fair because each bot has on average 50-70% of rated items, and the unrated ones were discarded for this evaluation. On the other hand, this holds for all bots, so they are on an equal footing.

6 Bots vs humans

We already discussed the differences between human and bot behaviour in dialogues. Here we sum up the main tendencies:

- Humans use shorter utterances in dialogue,
- Human-to-human dialogues are longer (which shows growing engagement of peers),
- Human performance in dialogue (both utterance- and dialogue-level) is generally rated high, but not exclusively high, which suggests that either

Bot name	Quality
DATA Siegt	0.512
poetwannabe	0.467
kAIb	0.453
bot#1337	0.433
RLLChatBot	0.430
poetess	0.380
Plastic world	0.372
Q&A	0.326
DeepTalkHawk	0.195
PolyU	0.061

Table 6: Utterance-level quality of bots. Bots are sorted from best to worst. Bots **in bold** are those which participate in the official competition.

human utterances or scores (or both) are not always reliable.

7 Discussion

The human evaluation round unraveled several flaws in our experimental setup. First of all, there were issues in evaluation process. We realised that many users struggled with on-the-fly evaluation of peer utterances. We suggest that utterance-level labelling should be conducted separately, after the dialogues are generated. Also, as some of volunteers suggested, they sometimes couldn't decide if an utterance was suitable or not, so the binary scale (relevant / irrelevant) should be replaced with the ternary scale (relevant / interpretable / irrelevant).

A different way of making utterance-level evaluation easier could be the reduction of the task's cognitive load by giving only one option — "irrelevant". Thus a user can explicitly mark a peer's response if s/he did not like it. The absence of answer will be regarded as a positive score.

Dialogue-level evaluation can also sometimes be inaccurate and should be conducted separately. Furthermore, dialogue-level metrics we use now are strongly correlated and also have close correlation scores with other parameters, which shows their redundancy. We suggest that in the next data collection experiments we should use one dialogue-level score, namely overall quality. The use of multiple metrics does not give new information, but increases user's cognitive load.

Another problem that we encountered is the uselessness of contexts that we provided in the beginning of dialogues. As our analysis showed, the majority of users or bots did not use it in conversations. Furthermore, volunteers complained about the big size and high complexity of texts. Therefore, we suggest that texts should be shorter and be selected from a different source with more common topics, simpler language and shorter sentences.

References

[Lowe et al., 2017] Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., and Pineau, J. (2017). Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. *Acl*, pages 1–19.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Emnlp*, (ii):2383–2392.

[Yu et al., 2016] Yu, Z., Xu, Z., Black, A. W., and Rudnicky, A. I. (2016). Chatbot Evaluation and Database Expansion via Crowdsourcing. *WOCHAT workshop*.