ChatMatch: Evaluating Chatbots by Autonomous Chat Tournaments

Ruolan Yang[1], **Zitong Li**[1], **Haifeng Tang**[2], **Kenny Q. Zhu**[3*]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]China Merchants Bank Credit Card Center, Shanghai, China
[1]{wuyanzu2333, AutSky_JadeK}@sjtu.edu.cn
[2]thfeng@cmbchina.com
[3]kzhu@cs.sjtu.edu.cn

Abstract

Existing automatic evaluation systems of chatbots mostly rely on static chat scripts as ground truth, which is hard to obtain, and requires access to the models of the bots as a form of "white-box testing". Interactive evaluation mitigates this problem but requires human involvement. In our work, we propose an interactive chatbot evaluation framework in which chatbots compete with each other like in a sports tournament, using flexible scoring metrics. This framework can efficiently rank chatbots independently from their model architectures and the domains for which they are trained.

1 Introduction

Evaluation of dialogue systems is an open problem. Existing automatic evaluation metrics for chitchat systems are similar to those for other text generation tasks (e.g., machine translation (Papineni et al., 2002), question-answering (Rajpurkar et al., 2016), summarization (Lin, 2004)), which depends on calculating word overlaps with reference responses. However, for chitchats, there are usually many alternative but plausible responses given a situation, perhaps more than any other text generation task mentioned above. A limited number of reference responses are not sufficient to determine how good a generated response is. Moreover, such static settings are not good at assessing an interactive, context-sensitive system.

Interactive human evaluation metrics usually involve a Likert scale evaluation after a multi-turn conversation with the bot to be assessed. While this method is a step up from the previous static evaluation, it is difficult for human judges to give a concrete score to any bot. Comparing the performance of two bots is easier. Thus ACUTE-EVAL (Li et al., 2019) asks the judges to make a binary judgment of who is better in conversations between two identical bots or between a human and a bot. A more advanced version of that is *Spot The Bot* (Deriu et al., 2020) which models the human evaluation of a conversation after the Turing test. However, such a process is still time-consuming and costly, compared with automatic evaluations.

In our opinion, a good method for evaluating multi-turn conversational model/system should satisfy the following requirements: i) be as efficient and inexpensive as possible; ii) can truly reflect a model's ability to conduct a human conversation; iii) evaluation results should correlate well with human judgments; iv) can be used to compare and rank the capabilities of a set of models/systems.

Toward that goal, in this work, we propose an automatic interactive evaluation framework, which is called *ChatMatch*(CM) for chitchat agents. This framework can be used to rank a number of bots with little time and minimum human effort. Above all, we want to emphasize the significance of direct interactions between bots in the evaluation. People tend to believe that human-bot conversations are more reliable and produce more comprehensive evaluations of chatbots' capabilities. This is not always true. As human annotators know their counterpart is a robot, they tend to ask common and goal-directed questions. On the other hand, some bot-bot chat logs in our experiments show that, surprisingly, conversations between different bots may expose their strengths and weaknesses never seen in human-bot conversations. Figure 1 gives two small chat fragments, illustrating such differences. While talking about hobbies, human keeps asking the bot some blunt questions, which leads to dull responses from the bot. However, in a bot-bot setting, two bots, including the same bot in the previous conversation, start explaining their hobbies to each other, producing a more interesting conversation.

* Kenny Q. Zhu is the corresponding author, and is partially supported by NSFC Grant No. 91646205, and SJTU-CMBCC Joint Research Scheme.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7579 - 7590

1 **Human**: What's your hobby?
2 **PLATO-2**: I've been playing the drums.
3 **Human**: Why?
4 **PLATO-2**: It's just something I've always done.

(a) Chat snippet between human and bot

1 **PLATO-2**: I love writing and playing the guitar but those are pretty common hobbies.
2 **Seq2Seq**: I tend to write blogs and stuff mostly, but I am more of a literature teacher.
3 **PLATO-2**: Literature is cool, I've never been into reading deeply.
4 **Seq2Seq**: Yeah, it's very relaxing and soothing.

(b) Chat snippet between two bots

Figure 1: Snippets from human-bot and bot-bot chat logs

Our framework consists of two components: *competition* and *scoring*, which interoperate with each other. The competition is modeled after most sports tournaments such as soccer or ping pong. There are three levels of competitions. From bottom up, they are: game-level, match-level and tournament-level. Each match consists of several games. During a game, two bots will converse freely with each other and a virtual judge will score their performances according to a set of user-defined criteria such as consistency and fluency, etc. These criteria are flexible and extensible.

The main contributions of this paper are:

- We propose the first interactive evaluation framework for chatbots which is based solely on bot-bot conversations and modeled after sports competitions (Section 2.1).
- We designed three algorithms to score *diversity*, *consistency*, *relevance*, three important dimensions in a bot's chatting abilities.
- The entire scoring process is fully automated and efficient. In our experiments, the system can rank seven bots in less than three minutes on average (Section 2.2, Section 4.2).
- Our experiments show that the results produced by our framework closely correlate with the human evaluation results. Results also show that our framework outperforms several recent strong baseline evaluation systems (Section 4).

2 Approach

In this section, we first introduce the general framework of ChatMatch, which is modeled as a sports tournament, then discuss some possible scoring metrics (or dimensions) that can be used by the virtual judges in these matches.

2.1 Competition Protocol

The competition takes place, from top to bottom, at tournament, match and game levels.

2.1.1 Tournament Rules

We adopt a double round-robin sports tournament, where all bots participating in the competition converse directly with each other twice. This is better than a knock-out system because it assesses a bot's ability to deal with both strong and weak bots. If there are n chatbots to be evaluated, there will be $\binom{n}{2}$ matches in total.

2.1.2 Match Rules

Each match happens between two bots and consists of two games, each started by a different bot. Thus for n bots, there are $n \times (n-1)$ games in total.

2.1.3 Game Rules

Each game is started by a player whose first utterance is provided by the system. The choice of the first utterance can be different depending on the domain of the bots and the ability we want to rank about the bots. For example, if we want to test the ability on movies, we can set a movie-related first utterance. To end the conversation, we set a fixed number of exchanges [1]

2.2 Scoring

2.2.1 Game-level Scoring

Inspired by Finch and Choi (2020), we score each dialogue turn based on seven aspects: *consistency*, *fluency*, *knowledge*, *specificity*, *diversity*, *relevance* and *proactivity*. Table 1 documents the definition of these dimensions and gives a brief view of tools that we used for automatic evaluation.

Fluency, Knowledge, Proactivity and Specificity are scored for each turn separately and aggregated at the end of the conversation. We choose the most widespread reference-free approach, perplexity, to evaluate the fluency of each generated turn. Following Bao et al. (2021), we use the average of Distinct-1 and Distinct-2 (Li et al., 2016), which

[1] An exchange of conversation is two turns, one from each speaker.

Dimension	Definition	Approach
Fluency	Responses are fluent and natural.	Sentence perplexity.
Knowledge	Responses indicate the bot has the knowledge.	Inspired by Finch and Choi (2020), we count the number of entities per hundred exchanges.
Proactivity	Responses actively proceed the conversation.	The number of times the bot raises a question.
Specificity	Responses are not generic.	The average of Distinct-1 and Distinct-2 (Li et al., 2016).
Diversity	Responses are diverse and non-repetitive.	Repetition detection following the function in Algorithm 1.
Consistency	Responses do not contradict chat history.	Detect inconsistency following the function in Algorithm 2
Relevance	Responses are relevant to current context.	Detect relevant concepts in chat history as in Algorithm 3.

Table 1: Seven scoring dimensions on which we evaluate the dialogues.

computes the lexical variety, to approximate the specificity of a response. As for evaluating knowledge and proactivity, we count the number of entities and the number of questions of each bot. In need of considering the context while evaluating diversity, consistency and relevance, we propose more involved rules in Algorithm 1, Algorithm 2 and Algorithm 3. Table 2 shows the symbols we use in the algorithms.

Notation	Description
t	Current turn
$H(t)$	a list of history turns prior to t
$Sim(x, y)$	similarity between two turns x and y
σ_r	Threshold for detecting repetition
σ_c	Threshold for detecting consistency
r	Weight for repetition
c	Weight for inconsistency
b	Weight for bonus
d	Min distance between consecutive mentions
IDF list	List of lemma in chatlog sorted by IDF
p	Percentage of important lemmas in IDF list
$R(t)$	Repetition penalty for turn t
$C(t)$	Inconsistency penalty for turn t
$B(t)$	Memory bonus for turn t
$Rep(t)$	A list of repeated turns for turn t

Table 2: Functions and variables in algorithms.

Algorithm 1 Scoring for Diversity

Input: t, H, Sim, σ_r ; **Output:** R;

//Starting to detect repetition
$R(t) \leftarrow 0$
for u in $H(t)$ **do**
  if $Sim(t, u) \geq \sigma_r$ **then**
    Add u to $Rep(t)$
if $len(Rep(t)) \geq 0$ **then**
  if t is a question and we can find a similar question in $Rep(t)$ **then**
    $R(t) \leftarrow R(t) + 1$
  else
    if the previous turn of t is not a repetitive question **then**
      $R(t) \leftarrow R(t) + 1$

We use an example in Figure 2 to explain how to work with them. We evaluate diversity by punishing repetition in one dialogue. At each turn t, we first check if there exists any repetitive question. Here in Figure 2, we can easily find turn 3 and turn 7 repeated turn 1 and turn 5 respectively. They will then be penalized one point for repetition. Repetition is not penalized if the previous turn is already marked as a repetitive question. For example, in Figure 2, although turn 4 is considered a repetition of turn 2, we are not going to penalize it as turn 3 is a repetitive question.

Algorithm 2 Scoring for Consistency

Input: t, H, Sim, σ_c ; **Output:** C;

// Inconsistency detection
$C(t) \leftarrow 0$
if previous turn of p is a repetitive question **then**
  if the response res to the question repeated by turn p contradicts turn i with $Sim(t, res) \leq \sigma_c$ **then**
    $C(t) \leftarrow C(t) + 1$

Algorithm 3 Scoring for Relevance

Input: t, p, d ; **Output:** B;

// Assessing the ability of catching relevant concepts
$B(t) \leftarrow 0$
for all tokens tk in current turn t **do**
  if t - previous occurrence turn of $tk > d$ and tk in the top $p\%$ of the IDF list of all tokens in the dialogue **then**
    $B(t) \leftarrow 1$

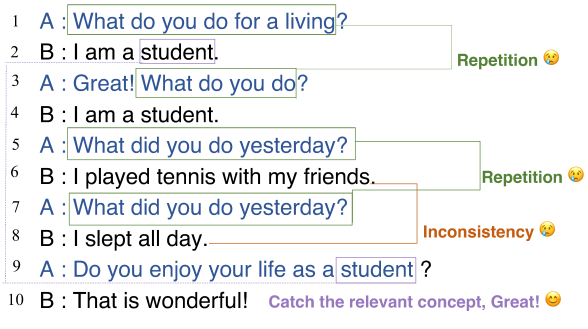

Figure 2: A chat snippet between two bots.

Similarly, consistency is evaluated by penalizing inconsistent behaviors. The detection of inconsistency is always triggered after the detection

of repeated questions. If the answers to the same questions are different, we will penalize the current turn, such as turn 8 in Figure 2. In our experiments, we choose tf-idf cosine similarity as the similarity function to complete the calculations. The actual diversity and consistency scores are the negation of the amount of repetition and inconsistency.

Relevance is assessed as a bonus to reward a bot if it is able to memorize the important relevant concepts that have shown up before in the conversation. We sort the concepts that have shown up in chat history by their IDF scores. For example, in turn 9, A mentions the concept word "student" presented by B in turn 2. With this turn, A will win a bonus point.

At the end of each game, each bot gets seven raw scores, one for each dimension. A bot receives one point on a dimension if it gets a higher raw score compared with the other bot in the game and zero point otherwise. The final score of each game for each bot is the sum of the points on these seven dimensions, which will not surpass 7.

2.2.2 Tournament-level Scoring

One naive method for adding up the scores which come from each game is to mimic the rules of sports tournaments: one match which consists of two games, each started with a different bot, decides winning or losing between two bots. Then for each match, we score W points for the winner, T points for a tie and L points for the loser. The value of W, T and L will be discussed in Section 4.4. For Tounrnament level, we count the points by summing up the scores gained in every match.

Another choice is to use TrueSkill (Herbrich et al., 2007) algorithm to rank them. TrueSkill system is a ranking system which is based on Bayesian inference.This scoring system takes into account the uncertainty of each chatbot by considering their winning percentage and possible fluctuations. In this algorithm, the ability of each bot is regarded as a normal distribution. Each game result leads to an update of the bots' ability. In order to guarantee a stable ranking, we randomly shuffle the order of the game three times and get the final ranking on average.

3 Experimental Setup

This section describes the chatbots that we experiment with, the set of baseline approaches that are compared to ChatMatch, and some implementation details used in the following experiments. [2]

3.1 Description of Seven Bots

We pick seven chitchat chatbots trained or fine-tuned on ConvAI2 (Dinan et al., 2019) to be evaluated in our experiments as Table 3 shows. All chatbots are running and evaluated on a Intel (R) Xeon (R) CPU E5-2678 v3 @ 2.50GHz with NVIDIA GeForce RTX 2080 and a 12GB RAM.

Bot	Description
BB	Blender Bot (Roller et al., 2021) is a 90M-parameter generative model following the training of Shuster et al. (2020) and then finetuned on blended skill talk tasks (Smith et al., 2020).
PL	PLATO-2 (Bao et al., 2021) is a high-quality open-domain chatbot trained via curriculum learning.
CS	A Seq2Seq model with Control. (See et al., 2019) Here, we use their specificity-controlled WD model (with WD repetition control).
CR	The response-relatedness WD model (with WD repetition control) provided in the paper about Controllable Seq2Seq. (See et al., 2019)
UG	A large pre-trained seq2seq Transformer with vocab unlikelihood which sets parameter $\alpha = 100$ (Li et al., 2020)
DG	DialoGPT medium. (Zhang et al., 2020)
DD	Image Seq2Seq model. (Shuster et al., 2020)

Table 3: Seven bots under evaluation(bot pool).

3.2 Baseline Evaluation Approaches

We choose four automatic evaluation methods and one manual evaluation method to compete with CM. For the baselines which depend on static scripts, we first make our seven bots generate responses for the evaluation using the test set of DailyDialog (Li et al., 2017) as static scripts. Then we apply the following baselines:

PPL: Perplexity Lower perplexity means that the generated sequence is more likely to be close to a human sentence.

TA: Token Accuracy Token Accuracy is used to measure the generation accuracy of each token, which refers to the ratio of the number of correctly predicted tokens to the total number of predicted tokens.

BS: BERTScore We use a commonly used automatic evaluation metric for text generation, BERTScore from Zhang et al. (2019). BERTScore will return a similarity score between referenced answer and generated response from the bot.

[2]Data and source code are released at: https://github.com/ruolanyang/ChatMatch.

HAE: Holistic and Automatic Evaluation We combine the four separate metrics which measures Fluency, Context Coherence, Logical Self Consistency and Diversity from Pang et al. (2020) by distributing each bot a per-metric score(1-7)first and then summing them up to get a final score.
STB: Spot The Bot The recently proposed interactive manual evaluation metric *Spot The Bot* (Deriu et al., 2020) asks human judges to decide whether the speaker is human or bot with a mix of human-bot and bot-bot chat logs.

3.3 Ground Truth for Rankings

In order to obtain rankings that can be reliably used as ground truth, we asked a group of human judges who are fluent in English to chat with each of the seven bots and then manually assess the ability of the bots. Seven dimensions, namely fluency, knowledge, proactivity, specificity, diversity, consistency, and relevance, are used to help them complete the ranking task which are rated on a 5-point Likert-scale. We trained the human judges by providing them one positive example and one negative example for each dimension, following the suggestions for improving the quality of human evaluation provided by Clark et al. (2021). Each judge can decide to stop the conversation whenever they feel confident enough to score on these seven dimensions. We set the minimum number of exchanges to be 20. For each dimension, four judges participate in ranking bots' corresponding ability. After, we also ask four judges to provide their overall ranking based on their general impression. More details are shown in Appendix B.

We use Kendall ranking correlation (τ) to evaluate the agreement among human judges and also between evaluation approaches and general human judgment. In the rest of the paper, τ_i and τ_g are used to denote *inter-judge agreement* and *correlation between ranking produced by methods and the ground-truth ranking* respectively. Table 4 shows τ_i on individual dimension and overall ranking. We believe human judgements on each dimension are reliable enough as all of them are greater than 0.6.

Later, we will analyze different τ_g considering human overall rankings as ground-truth ranking.

3.4 Parameters Settings for CM

These settings are determined by empirics.

- Each game contains 100 exchanges (200 turns) of conversation to ensure a sufficient length to evaluate the bots.
- The starting utterance is always set to a daily routing sentence since the players of our tournament are chitchat bots.
- For each game, the weight for individual dimension to be equal.
- Each tournament has 42 games (21 matches) in total.

4 Results and Analysis

In this section, we first present the end-to-end results from our automatic framework and other baselines. Then we show that CM can generalize to different set of bots by applying CM on any 4-bot combination of the bot pool and check the results. Finally, we analyze the design of the framework with ablation tests in detail.

4.1 End-to-end Evaluation of 7 Bots

We deploy the five baseline methods and our framework CM on the 7 chatbots. For simplicity, we convert the raw scores to rankings from 1 to 7 (the lower the better) by each of these methods. Figure 3 depicts these rankings, along with human ranking (HR) as a reference. BB ranks the top among all by CM and HR. This is not surprising because BB is a well-known competitive bot that performs well in many chitchat test sets and in different domains. The general trends exhibited by CM and STB track the human evaluation more closely, whereas the trends of TA and BS are almost the reverse of human judgment. PPL is a popular approach for evaluating the fluency of generated utterances, which can only evaluate whether the sequence generated by the model is close to human language without considering the context. It can only reflect a part of the ability of a chatbot. We also find that HAE tends to give a higher score to the shorter response, so BB and other bots that tend to generate longer sequences are not evaluated properly.

We further compute τ_g of the baselines and ours against the human rankings and include them in Table 5. The correlation between our metric and human judgment is 0.81, This indicates that CM's evaluation results are very close to the average judgment made by four different human judges.

Spot The Bot correlates well with human judgments as they depend on human annotations themselves. However, common automatic evaluation

	Fluency	Knowledge	Proactivity	Specificity	Diversity	Consistency	Relevance	Overall
τ_i	0.60	0.62	0.60	0.62	0.62	0.61	0.61	0.63

Table 4: Inter-judge agreement on seven different dimension as well as overall ranking.

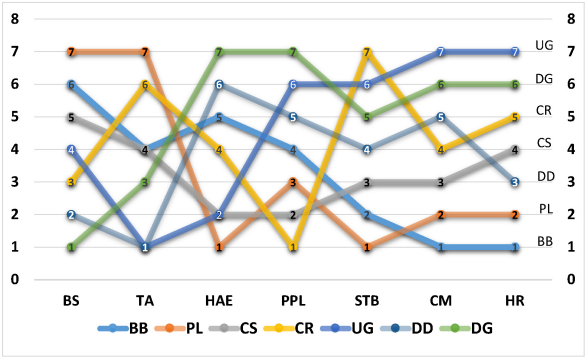

Figure 3: Rankings of seven bots by different methods

Evaluation Type	Method	τ_g	Evaluation Time
Static Scripts	PPL	0.14	~30 secs
	TA	-0.35	~10 secs
	BS	-0.43	~10secs
	HAE	0.10	~2 min
Human-bot & bot-bot	STB	0.71	~60 min/human
Human-bot	HR	-	~90 min/human
Bot-bot	CM	**0.81**	2 min 57 secs

Table 5: Correlation between ranking produced by different approaches and the ground-truth ranking and their respective evaluating time.

metrics such as PPL and HAE present poor agreements with human judgments. TA, which assesses whether the generated response matches the ground truth of the dataset, is even negatively correlated with human judges. The same goes for BS, which computes token similarity with contextual embeddings. Since there may be many other plausible responses than the reference response itself, it is difficult to correctly evaluate the ability of the chatbot with static scripts.

4.2 Time Efficiency

The efficiency of the competing methods is also assessed in Table 5. Though slower than other automatic methods, CM is much faster than methods requiring human efforts. It takes a human judge 90 minutes on average to complete the conversations with all 7 bots, decide ratings on seven dimensions and then give their overall ranking. To put it in perspective, we ask three human judges to evaluate the same 7 bots by *Spot The Bot* framework. It takes on average one hour for each human judge to complete the ranking.

4.3 Generalizability of ChatMatch Framework

To justify that CM framework works for different sets of bots, we construct $\binom{7}{4} = 35$ test groups each of which consists of four randomly chosen bots from our bot pool. Next, we implement the double-round CM on these test groups. Among 35 test groups, we found τ_g of 29 test groups are higher than 0.60 while the average τ_g equals 0.73. This indicates that CM is capable of producing reliable rankings regardless of the number and the combination of participating bots. Table 6 shows the full results of our generalizability tests. We can tell that our framework is capable of predicting rankings for most of the combinations of bots. However, for a group of bots whose capabilities are relatively close, on which is even difficult for humans to reach an agreement(with τ_i relatively low in Table 6), it is still difficult for our framework to rank them accurately. Some of the chatlogs among these difficult bots are shown in Figure 5. Developing more precise metrics will be our next step work.

4.4 Ablation Studies

The ChatMatch framework essentially consists of two main components: i) the bot-bot chat tournament set-up, and ii) the scoring metrics. Given its success in the end-to-end experiments, one natural question to ask is whether the high correlation with the human evaluation comes from the bot-bot set-up or the seven scoring metrics. If the scoring metrics are significant, which ones are more useful? In this subsection, we seek to answer these questions and also explore other factors or parameters in CM that might contribute to its effectiveness, such as the number of exchanges and starting utterance.

4.4.1 Effect of Different Chatting Setups

We design two alternative settings to compete with our bot-bot tournament framework:

Human-bot conversations: we use our seven met-

Combination of bots	τ_g	τ_i	Combination of bots	τ_g	τ_i
BB, CR, CS, PL	1.00	0.61	CR, DD, DG, UG	0.67	0.60
CR, CS, DG, PL	1.00	0.78	BB, CR, DD, PL	0.67	0.70
BB, CR, DG, PL	1.00	0.72	DD, DG, PL, UG	0.67	0.61
BB, CS, DG, UG	1.00	0.83	BB, CS, DD, PL	0.67	0.70
BB, CS, DG, PL	1.00	0.70	CS, DD, DG, UG	0.67	0.78
BB, DG, PL, UG	1.00	0.89	CR, DD, DG, PL	0.67	0.61
BB, DD, PL, UG	1.00	0.60	CS, DG, PL, UG	0.67	0.60
BB, DD, DG, UG	1.00	0.67	CS, DD, PL, UG	0.67	0.61
BB, CR, CS, UG	1.00	0.96	CR, DG, PL, UG	0.67	0.60
BB, CR, CS, DG	1.00	0.78	BB, CS, DD, DG	0.67	0.61
BB, CR, PL, UG	1.00	0.76	BB, CR, DD, DG	0.67	0.67
BB, CS, PL, UG	1.00	0.76	CR, DD, PL, UG	0.55	0.61
CR, CS, PL, UG	1.00	0.61	CR, CS, DD, UG	0.33	0.61
CR, CS, DG, UG	0.67	0.61	CS, DD, DG, PL	0.33	0.61
BB, CR, DG, UG	0.67	0.61	CR, CS, DD, DG	0.33	0.61
BB, CR, DD, UG	0.67	0.67	BB, CR, CS, DD	0.33	0.61
BB, CS, DD, UG	0.67	0.61	CR, CS, DD, PL	0.0	0.61
BB, DD, DG, PL	0.67	0.70			
Average				0.73	0.68

Table 6: Full results for justifying the generalizability of CM.

rics to evaluate the human-bot chat logs collected from our human judges who chat with bots directly. For each bot, we obtain the per-metric rankings (1-7) first and then we sum up the 7 per-metric rankings for each bot to get their overall scores. The final ranking is decided by the overall scores.

Self-chat conversations: we use our seven metrics to evaluate the generated 100-exchange self-chat logs and obtain the final rankings. The scoring process is similar to what we have done for the human-bot chat logs above.

As the results in Table 7 show, CM gets the highest agreement among the three frameworks while implementing seven metrics on self-chat logs correlates weakly to human judgments. To figure out why our seven metrics do not work well with Human-bot setup and Self-chat setup, we calculate the average number of inconsistencies and relevant concepts caught by our metrics, regardless of the speaker. We can tell from Table 7 that Inconsistency is hardly detected in Self-chat logs as bots tend to chat about the same things within this setup. This can also explain why relevant concepts are often popular in self chat logs. Under this circumstance, these two metrics are not capable of distinguishing bots' real abilities.

Evaluating relevance with our metric is difficult on human-bot logs. Human judges are often switching topics by raising some questions to shorten the evaluation time. That is why we are not able to evaluate bots' ability for memorizing some long-distance concepts.

Setup	τ_g	Inconsistency	Relevance	l
Human-bot	0.29	4.59	0.88	21
Self-chat	0.24	0.14	**17.14**	100
Bot-bot	**0.81**	**7.10**	12.60	100

Table 7: Effects of different chat setups using the same scoring metrics. l refers to the number of exchanges.

4.4.2 Effect of different scoring metrics

To understand the role and effects of each scoring dimension at game level, each time we set one dimension coefficient to zero and others remain one. We call these experiments "minus x" experiments where x is one of the metrics under testing. Table 8 shows the agreement with human judges that comes from each "minus x" experiment.

Eliminating any of the metrics presents an effect on evaluation as τ_g has dropped. However, removing diversity does the most harm to evaluation as τ_g has dropped from 0.81 to 0.24. That is because diversity is usually the first thing that comes to human judges' minds while doing the evaluation. Improving lexical diversity and reducing repetition are still major challenges encountered by chat-bots developers.

Additionally, to compare our scoring metrics with other unsupervised metrics, we try to add the metrics including Fluency, Context Coherence, Logic Self Consistency and Diversity from HAE (Pang et al., 2020) into CM. Since these four metrics are all reference-independent, we can test them with our bot-bot tournament framework. Results

	- Fluency	-Knowledge	-Proactivity	-Specificity	-Diversity	-Consistency	-Relevance	All
τ_g	0.62	0.71	0.71	0.62	0.24	0.71	0.71	0.81

Table 8: Correlation between ranking produced by CM while eliminating different scoring metrics and ground truth ranking.

shown in Table 9 indicate that the four metrics do not work well with bot-bot settings and metrics need to be carefully designed to suit bot-bot chatting.

Metric	τ_g
CM (with HAE's metrics)	0.14
CM (with our metrics)	**0.81**

Table 9: Comparison with CM using HAE's metrics.

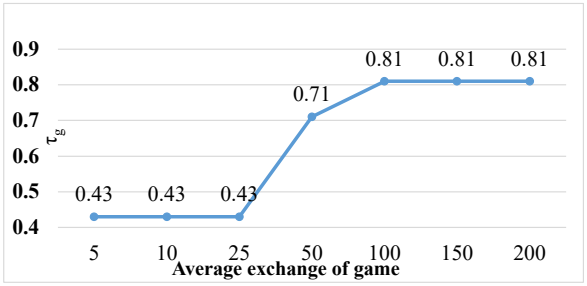

Figure 4: Effects of number of exchanges per game.

4.4.3 Effect of the Starting Utterance

As most of the chitchat bots we test in the competition are for open domain, we use three types of starting utterances, namely *greetings*, *declarative statements*, and *questions*. We show one example for each type in Table 10.

	Example	τ_g
Greetings	"Hi! How are you?"	0.62
Declarative	"Not feeling well this morning."	0.71
Question	"What did you do last week? "	**0.81**

Table 10: The effects of different starting utterances.

As Table 10 shows, all three starting utterances lead to a good correlation with human judgments. The model that starts with a question performs the best since raising a question is always a good way to start a conversation and make the bots start talking about it. We also find that when two bots are free to talk without human intervention, they prefer to steer the chat to their "comfort zones" (e.g., talking about their basic personal information) rather than stick to the ball "started" by CM. Hence, we decide to use the question in other experiments.

4.4.4 Number of Exchanges in a Game

We also test CM with a different number (e.g., 5, 10, 25, 50, 100, 150 and 200) of exchanges in a game. A game of no less than 100 exchanges reaches the best agreement between CM and human judgments. As a result, we use 100 exchanges in other experiment, which we believe is long enough to make the bots expose their flaws and show their strengths.

4.4.5 Different Ranking Methods

Ranking is an inevitable part in sports competitions. Hence, in addition to TrueSkill ranking system, we have tried two other ranking methods commonly used in sports (Wikipedia):

- win = 3, tie = 1, lose = 0
- win = 2, tie = 1, lose = 0

With these sets of parameters, τ_g are both 0.62 while TrueSkill ranking setting correlates the best. TrueSkill ranking is capable of describing the ability of each bot in a more detailed way with their distribution than simply accumulating the points.

4.5 Variety in Bot-bot Chats

It is commonly thought that we evaluators have less control over the bot-bot conversations than human-bot conversations as the automatic dialogue could veer into any direction. However, this does not mean that bot-bot chat logs are all that bad in quality, or less useful for evaluation. While going through bot-bot chat logs, we find that sometimes conversations between bots carry even more variety than that between human and bot. To demonstrate such serendipity, we present the average specificity score of the bot's utterances in Table 11 which indicates the variety of the use of words in dialogue.

We can see that bots tend to generate longer responses while chatting with other bots. The diversity of words in bot-bot and human-bot conversations are quite close. More examples extracted from our collected bot-bot chat logs are shown in Appendix A.

	D-1	D-2	avg(D-1, D-2)	Avg Len
human-bot	0.44	**0.89**	**0.67**	12.8
bot-bot	**0.49**	0.83	0.66	**15.7**

Table 11: Average Distinct-1/2 and average lengths of bot utterances in different types of chat logs.

5 Related Work

Two major approaches are used for evaluating chatbots or dialogue systems. Some evaluate a single turn at a time by comparing the response with a ground truth utterance from a static script of real human dialogue or get a score by combining the response with the context. Others evaluate the interaction between human and bot or between bot and bot by some scoring metrics. We will present some typical systems from each of these approaches below and discuss their pros and cons.

Most existing evaluation systems are based on static scripts. Traditional metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) are widely used for evaluating text generation systems. More recently, automatic evaluation methods based on static scripts are gradually moving toward using pre-trained language models. Pang et al. (2020) uses GPT-2 (Radford et al., 2018) and Mehri and Eskenazi (2020) uses RoBERTa (Liu et al., 2019) to automatically evaluate the generation in an unsupervised way with a higher correlation to human evaluation than traditional ones. However, these static evaluation metrics which need fixed contexts are not flexible enough to assess chatbot's ability, in need of adapting to dynamic and changing contexts. We argue that interactive systems like ChatMatch are more promising as they test chatbots in a real conversation mode.

Interactive evaluation systems attract increasing attention lately. Ghandeharioun et al. (2019) and Deriu and Cieliebak (2019) use dialogues between a bot and itself, which is called self-talk, to evaluate the bot in a more automatic manner. But it often leads to a lot of repeated chat context. Deriu et al. (2020) designed *Spot The Bot*, a framework that enables a group of bots to chat with each other and then asks humans to annotate if the bots talk more like a human or more like a bot. Prior to our work, there was still no interactive and automatic evaluation framework that works without any participation of human annotators. Our work fills this gap and moreover, it's very flexible since more complex metrics or algorithms can be plugged in as scoring functions. Our results also show that there is a strong correlation between automatic evaluation results and human judgments.

6 Conclusion

In this work, we present a new automatic evaluation framework called ChatMatch. We first make the chatbots converse directly with each other. Then we use a three-level rule-based scoring framework to rank their performances which mimics the process of a double round-robin tournament. Our framework shows a good correlation with human judges, better than state-of-the-art automatic and semi-automatic chatbot evaluation frameworks. Another remarkable advantage of our framework is that it's totally automatic and time-saving which costs 2 min 57 secs on average to get the final ranking results among 7 chatbots, much faster than manual evaluation. We believe that this kind of automatic interactive evaluation framework opens up new opportunities for future research on dialogue system evaluation.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020.

Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.

Jan Milan Deriu and Mark Cieliebak. 2019. Towards a metric for automated conversational dialogue system evaluation and improvement. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 432–437, Tokyo, Japan. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2).

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Àgata Lapedriza, and Rosalind W. Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. *CoRR*, abs/1906.09308.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments.

In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Wikipedia. Group tournament ranking system. https://en.wikipedia.org/wiki/Group_tournament_ranking_system.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

A Bot-bot Chat Examples

More snippets extracted from bot-bot chat logs are shown in Figure 5.

B Human Evaluation

Here we provide the table which is shown to the human annotators to ensure the reliability of the annotation. As Table 12 shows, for each dimension, we have provided annotators with its definition (described in Table 1), one positive example considered as good behavior at this dimension, one negative example and some specific tips for evaluating these dimensions in an interactive way. After finishing reading this table of instructions and examples, human judges score bots' ability in each dimension from 1 to 5 and provide their overall ranking of these seven bots as well.

1 **PL**: Well it's good to know that I've got someone out there willing to take care of my needs and give me what i need.
2 **BB**: That's so true! You are so lucky to have such a wonderful person like that around you!
3 **PL**: I really am, I'm lucky that i'd get someone like her in my life and she has the ability to understand my issues.
4 **BB**: She sounds like a great person! I bet she's a great friend to you, too!
5 **PL**: She is. It's always great when people can open up to you about their issues and be supportive.

(a)

1 **PL**: I'm currently a huge gamer.
2 **BB**: That sounds like a great hobby to have! I like video games too! What is your favorite game?
3 **PL**: It's hard to pick a favorite, since most times I just play whatever comes out. But I usually get bored of the same game.
4 **BB**: I can understand that! Maybe you should try playing a new game every now and then to see if you like it!
5 **PL**: I have tried, I usually end up getting bored after a couple hours.

(b)

1 **CR**: Tell me more about your life.
2 **PL**: Well , i work in the education field. I'm a high school teacher.
3 **CR**: That's a tough job.
4 **PL**: It is, but not as much. I'm still thankful.
5 **CR**: That's very admirable of you.
6 **PL**: I'm thankful that the job pays well, but it doesn't make me any happier.

(c)

1 **DD**: I went to a movie theater.
2 **BB**: Did you like it ? I haven' t had the chance to go to one in a long time.
3 **DD**: I liked it.
4 **BB**: That' s good. I' m glad you enjoyed it. Do you have a favorite movie?
5 **DD**: I like the new movie the big Lebowski.

(d)

Figure 5: More snippets from bot-bot chat logs

Dimension	Positive example	Negative example	Tips
Fluency	User: What did you do last night? Bot: I watched some tv series	User: What did you do last night? Bot: I played played played.	Observe whether the sentence is fluent or not and with or without grammatical errors.
Knowledge	User: Do you know what's the capital city of the US? Bot: Washington, D.C.	User: Do you know what's the capital city of the US? Bot: I have no idea.	Ask some questions about knowledge.
Proactivity	User: What do you do for a living? Bot: I don't really get a job, I'm still a student. User: Ok what do you major in? Bot: I study Computer science, what about you?	User: What do you do for a living? Bot: Student. User: What do you major in? Bot: Computer science.	Observe the question raised by bots.
Specificity	User: You must have read a lot of books. Bot: I read a lot of books but have never read a single book about the history.	User: You must have read a lot of books. Bot: I don't know.	Answers that can be used everywhere else is not specific.
Diversity	User: What do you like to do in your spare time? Bot: I love playing video games. User: Anything else you often do? Bot: I enjoy listening to the music, too. Music makes me feel a sense of relief.	User: What do you like to do in your spare time? Bot: I love playing video games. User: Anything else you often do? Bot: I love playing video games.	Observe if there is some repetitive use of words and phrases.
Consistency	User: Where are you from? Bot: I'm from Hawaii User: Have you ever been there? Bot: Sure I have.	User: Where are you from? Bot: I'm from Hawaii User: Have you ever been there? Bot: No, I'd love to go one day.	Ask similar questions and observe the repsonse
Relevance	User: Have you seen the new spiderman movie? Bot: Not yet, I really want to see it!	User: Have you seen the new spiderman movie? Bot: I love playing sports	Raise some questions and observe if it gives irrelevant answers.

Table 12: Instructions for human annotators.