

# Bipartite-play Dialogue Collection for Practical Automatic Evaluation of Dialogue Systems

Shiki Sato<sup>1,\*</sup> Yosuke Kishinami<sup>1,\*</sup> Hiroaki Sugiyama<sup>2</sup> Reina Akama<sup>1,3</sup>  
Ryoko Tokuhisa<sup>1</sup> Jun Suzuki<sup>1,3</sup>

<sup>1</sup>Tohoku University <sup>2</sup>NTT Communication Science Laboratories <sup>3</sup>RIKEN  
{shiki.sato.d1, akama, tokuhisa, jun.suzuki}@tohoku.ac.jp  
yosuke.kishinami.q8@dc.tohoku.ac.jp, h.sugi@ieee.org

## Abstract

Automation of dialogue system evaluation is a driving force for the efficient development of dialogue systems. This paper introduces the *bipartite-play* method, a dialogue collection method for automating dialogue system evaluation. It addresses the limitations of existing dialogue collection methods: (i) inability to compare with systems that are not publicly available, and (ii) vulnerability to cheating by intentionally selecting systems to be compared. Experimental results show that the automatic evaluation using the bipartite-play method mitigates these two drawbacks and correlates as strongly with human subjectivity as existing methods.

## 1 Introduction

The performance evaluation of dialogue systems is a crucial and challenging research topic for the dialogue research community. The community recommends human evaluation as the primary evaluation method, which is the gold standard but is time-consuming and costly. Moreover, reproducing the evaluation results is mostly impractical due to the unavailability of maintaining identical evaluators or identical evaluation conditions. Human evaluation is therefore unsuitable for evaluating daily updates of developing dialogue systems or comparing systems with non-public ones. Thus, constructing a better automatic evaluation method, which is both highly reproducible and low cost, is desirable. In particular, automating interactive evaluation, not static evaluation such as BLEU (Papineni et al., 2002), is attracting an increasing interest as static evaluation cannot capture diverse aspects of dialogue systems (Ghandeharioun et al., 2019).

An interactive evaluation framework consists of two phases: first, *collecting* the dialogues in which the systems to be evaluated (hereinafter called evaluation targets) talk to others (hereinafter called dia-

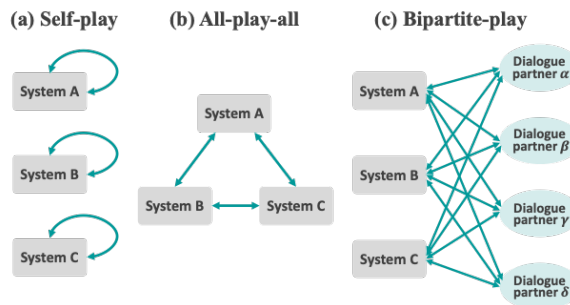


Figure 1: Dialogue collection methods. Here, the evaluation targets are System A, B, and C. (a) Self-play collects dialogues by talking to themselves (e.g., A-A and B-B). (b) All-play-all collects dialogues with other evaluation targets (e.g., A-B and A-C). (c) Our bipartite-play collects dialogues with fixed dialogue partners separated from the evaluation targets (e.g., A- $\alpha$  and A- $\beta$ ).

logue partners), then *rating* evaluation targets based on the quality of their utterances in the collected dialogues. Regarding the collecting (i.e., automating dialogue partners), **self-play** and **all-play-all** (Figure 1 (a) and (b)) are the current promising methods; All-play-all collects dialogues among multiple evaluation targets, while self-play collects dialogues with itself. Recently, Yang et al. (2022) have reported that all-play-all correlates with human evaluation strongly. However, all-play-all is not perfect and has at least two potential drawbacks: (i) the difficulty of comparison with publicly inaccessible systems and (ii) the vulnerability to cheating by choice of evaluation targets, i.e., with whom the evaluation target will talk (Section 3).

This paper addresses the above two drawbacks of the all-play-all method while maintaining the all-play-all method’s high correlation with human rating. Specifically, we propose the **bipartite-play** method, i.e., fixing and sharing a set of dialogue partners across studies as shown in Figure 1 (c) instead of assigning other evaluation targets as partners as shown in Figure 1 (b) (Section 4). The bipartite-play method offers (i) a fair comparison

\*Both authors contributed equally to this paper.

with publicly inaccessible systems as long as its developers use our method and (ii) prevention of cheating by an intentional choice of evaluation targets. Our experiments show that the bipartite-play method strongly correlates with humans as the all-play-all method while preventing the potential drawbacks in the all-play-all method.

## 2 Related Work

### 2.1 Automatic dialogue collection

**Self-play.** The self-play method collects dialogues where evaluation targets talk to themselves, i.e.,  $i \times 1 \times j$  dialogues in which collecting  $j$  dialogues for each of  $i$  evaluation targets. This method is cost-effective for interactive dialogue system evaluation since it does not require human interactions (Ghandeharioun et al., 2019; Deriu and Cieliebak, 2019). However, since there are few dialogue partners, it does not fully expose the characteristics of evaluation targets (Yang et al., 2022).

**All-play-all.** The all-play-all method collects dialogues between multiple evaluation targets, i.e.,  $i \times (i-1) \times j$  dialogues when collecting  $j$  dialogues for each of  $i$  evaluation targets (considering speaker order). This method also requires no human interactions. Compared to the self-play method, the all-play-all method’s dialogue partners are more diverse since it collects dialogues with other evaluation targets that result in various dialogues (Deriu et al., 2020; Yang et al., 2022). Additionally, direct interactions with evaluation targets make them easy to compare. Yang et al. (2022) experimentally showed that the evaluation using the all-play-all method correlates with human evaluation stronger than the self-play method.

### 2.2 Automatic dialogue rating

Compared with methods relying on reference responses (e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Lavie and Agarwal, 2007), Greedy Matching (Rus and Lintean, 2012), Vector Extrema (Forgues et al., 2014), and ADEM (Lowe et al., 2017)), reference-free methods, such as USR (Mehri and Eskenazi, 2020b), GPT-2 based evaluation (Pang et al., 2020), FED (Mehri and Eskenazi, 2020a), and DynaEval (Zhang et al., 2021), have attracted greater interest from the research community. For example, FED allows fine-grained practical evaluation of the system utterances without high-cost preparation,

such as training an evaluation model; it assesses system utterances for given dimensions, such as Fluency and Specificity, by guessing whether positive or negative responses are valid to the system utterance in terms of language model score (see Section 5.1). We also focus on reference-free evaluation, especially the FED metric, to automate the rating part of the interactive evaluation, as preparing references for automatically collected dialogues is impractical.

## 3 Limitations of All-play-all Method

The all-play-all method enables effective dialogue collection for system comparison, as described in Section 2. However, we point out that the current all-play-all method cannot handle the following two cases: First, when the group of evaluation targets includes unavailable systems. Since all-play-all requires the collection of dialogues with all evaluation targets, it is impossible to compare systems that are not released or that cannot be run by many researchers due to such computational resources. Second, when one attempts to boost their system’s performance by deploying an unfair evaluation setting. Our experiments (Section 6) reveal that one can intentionally improve the automatic evaluation results of desired systems by choosing evaluation targets to be compared when using the all-play-all method. If these potential drawbacks can be overcome, existing automated evaluation methods could be enhanced to be more versatile and practical.

## 4 Proposed Method: Bipartite-play

We introduce a new automatic dialogue collection method, called bipartite-play method, which updates the two aforementioned limitations.

### Evaluation targets vs fixed dialogue partners.

Considering the two drawbacks of the all-play-all method (Section 3), we propose fixing and sharing a set of publicly accessible systems as dialogue partners rather than assigning other evaluation targets as partners. Our idea is that even if evaluation targets do not talk to each other directly, dialogues in which evaluation targets talk to the same (shared) partners should be effective for system comparison. In this setting, the diversity of dialogue partners can be maintained by ensuring the diversity of pre-determined dialogue partners set. Sharing a set of dialogue partners allows a fair comparison with

publicly inaccessible systems as long as its developers use our method. Furthermore, predetermining a set of dialogue partners prevents cheating by an intentional choice of evaluation targets.

**Bipartite-play dialogue collection.** Given  $i$  evaluation targets, the bipartite-play method collects  $i \times k \times j$  dialogues by having evaluation targets interact  $j$  times with each of the  $k$  various dialogue partners predetermined for evaluation.

## 5 Bipartite-play based Evaluation Framework

Subsequently, we introduce bipartite-play to the interactive dialogue evaluation framework. We combine the dialogue collection based on the bipartite-play method with FED (Mehri and Eskenazi, 2020a), which is one of the modern and effective dialogue rating methods.

### 5.1 System evaluation procedure

Based on the  $m$  collected dialogues by the bipartite-play method, We assess an evaluation target for a dimension  $v$ . Specifically, we first evaluate the performance for  $v$  in a dialogue using the average score of the system’s utterances. We then determine the system’s whole performance for  $v$  using the average score of  $m$  dialogues. We compute the system utterances score using FED. This rating method evaluates the system’s utterances for  $v$  by guessing whether positive or negative responses for  $v$  are valid in terms of the language model as a response to the system’s utterance. The validity of each positive and negative response is automatically evaluated using a large-scale dialogue system. The evaluation value of  $v$  of the evaluation target’s utterance  $r$  for a context  $c$  is calculated as follows:

$$\sum_{p \in \mathcal{P}_v} D(c + r, p; \theta) - \sum_{n \in \mathcal{N}_v} D(c + r, n; \theta), \quad (1)$$

where  $\mathcal{P}_v$  and  $\mathcal{N}_v$  are the set of positive and negative responses for  $v$ , respectively.  $D(c, \cdot; \theta)$  is a function that calculates the probability of generating a response to  $c$  using a large-scale dialogue system with parameters  $\theta$ .

### 5.2 Preliminary experiment

We assess evaluation targets based on dimensions frequently deployed in recent research (Deriu et al., 2020; Adiwardana et al., 2020): Fluency, Specificity, and Sensibleness, additionally Overall. The applicability of FED to these dimensions is unclear

| Dimension    | FED    | w/o neg | w/o pos |
|--------------|--------|---------|---------|
| Fluency      | 0.121  | -0.145  | 0.171   |
| Specificity  | -0.022 | -0.364  | 0.340   |
| Sensibleness | 0.370  | -       | 0.370   |
| Overall      | 0.329  | -0.367  | 0.386   |

Table 1: Spearman’s rank correlation coefficients of the FED with human evaluation. “w/o pos” and “w/o neg” are the FED evaluations calculated without positive and negative responses respectively. “w/o neg” for Sensibleness is a missing value.

as Mehri and Eskenazi (2020a) cover only some of these dimensions. Therefore, as a preliminary experiment, we determine whether the FED evaluation for these dimensions correlates with humans.

**Dataset.** We created the dataset by collecting dialogues between the dialogue system and humans, then annotating the collected dialogues with a human evaluation score. Crowdsourcing<sup>1</sup> was employed in two processes. First, we collected dialogues between the 11 systems deployed as evaluation targets for the experiments in Section 6 and humans. We obtained 50 dialogues for each system, for 550 dialogues in total.<sup>2</sup> We then asked five workers to evaluate each collected dialogue with a five-point Likert scale for the question about each of the four dimensions.<sup>3</sup>

**FED evaluation settings.** We used the positive and negative responses manually created by Mehri and Eskenazi (2020a), and our additional responses for the FED evaluation. Also, we used Blender 9B from ParlAI (Miller et al., 2017) as a large-scale dialogue system to calculate FED scores. We used the four dimensions for which human evaluation scores were annotated in the constructed dataset.

**Results of FED evaluation.** Table 1 shows Spearman’s rank correlation coefficients between the FED and human evaluation results. We found that the FED evaluation using only the negative response correlates to some extent with human evaluation. Although Mehri and Eskenazi (2020a) proposed a method using positive and negative responses, we use only negative responses in subsequent experiments based on these results. Also, we

<sup>1</sup><https://www.mturk.com/>

<sup>2</sup>Starts with the human’s *Hi!* and continues for six turns.

<sup>3</sup>We asked workers *Are Bot’s responses fluent and grammatically correct?* (Fluency), *Are Bot’s responses specific and explicit in the given context?* (Specificity), *Are Bot’s responses sensible?* (Sensibleness), and *Is the overall impression of the chatbot good?* (Overall), and they answered from *Strongly disagree* (score 1) to *Strongly agree* (score 5).

---

**Evaluation targets:** Tfm-3B-Rdt-Bsm, Tfm-3B-Rdt-Msc, Tfm-3B-R2c-Bsm, Tfm-3B-Rdt-Lgu, GPT-345M-Wtx-Rdt, Tfm-89M-Ddc-Nft, Tfm-89M-Ddc-Crm, Tfm-89M-Ddc-Ddg, Tfm-89M-Ddc-Rdt, Tfm-89M-Ddc-Twt, PEn-256M-Rdt-Bst

---

**Partner systems:** Tfm-3B-Rdt-Slf, Tfm-3B-Rdt-Lgt, Tfm-3B-Rdt-Img, Tfm-3B-Rdt-Sfr, Tfm-1B-Rdt-Bsm, GPT-117M-Wtx-Rdt, GPT-762M-Wtx-Rdt, Tfm-406M-Rdt-Bsm, Tfm-406M-R2c-Bsm, Brt-406M-Rbt-Woi, Trm-89M-Ddc-Wow, Trm-89M-Ddc-Lgt, Trm-89M-Ddc-Emp, Trm-89M-Ddc-Cv2, Trm-89M-Rdt-Wow, Trm-89M-Rdt-Cv2, Trm-88M-Rdt-Bst, Trm-88M-Rdt-Cv2, PEn-256M-Rdt-Cv2, PEn-256M-Rdt-Emp, PEn-256M-Rdt-Wow, PEn-256M-Rdt-All, PEn-256M-Rdt-Bsm, B+F-256M-Rbt-Wow

---

\*Tfm: Transformer (Vaswani et al., 2017). GPT: DialoGPT (Zhang et al., 2020). PEn: PolyEncoder (Humeau et al., 2020). Brt: Bart (Lewis et al., 2020). B+F: FiD (Izcard and Grave, 2021) with Brt. Rdt: Pushshift Reddit Dataset (Baumgartner et al., 2020). R2c: R2C2 dataset (Shuster et al., 2022). Wtx: WebText dataset (Radford et al., 2019). Ddc: DodecaDialogue dataset (Shuster et al., 2020b). Rbt: Training dataset of RoBERTa (Liu et al., 2019). Bsm: Smith et al. (2020)’s multi-task dataset. Msc: Multi-Session Chat dataset (Xu et al., 2022). Lgu: LIGHT dataset (Urbanek et al., 2019) for unlikelihood training. Nft: No finetune. Crm: Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011). Ddg: DailyDialog dataset (Li et al., 2017). Twt: Tweets collected by Shuster et al. (2020b). Bst: BlendedSkillTalk dataset (Smith et al., 2020). Slf: Dialogues collected using the self-play method by Smith and Williams (2021). Lgt: LIGHT dataset. Img: Image-Chat dataset (Shuster et al., 2020a). Sfr: SaFeRDialogues dataset (Ung et al., 2022). Woi: Wizard of the Internet dataset (Komeili et al., 2022). Wow: Wizard of Wikipedia dataset (Dinan et al., 2019b). Emp: EmpatheticDialogues. All: Cv2+Emp+Wow. Cv2: ConvAI2 dataset (Dinan et al., 2019a).

Table 2: Dialogue systems for our experiments: 11 evaluation targets and 24 partner systems. Each system name represents [architecture]-[number of model parameters]-[pretrain data]-[finetune data].

found that the FED evaluation of Fluency correlates poorly with human evaluation, while the other dimensions correlate relatively well with human evaluation. However, the agreement rate for human evaluation is extremely low, and we consider Fluency evaluation with consistent results difficult even for humans.<sup>4</sup> One possible reason is that all systems have a high Fluency in neural response generation, so the difference in the Fluency of dialogues for each sample is small. Therefore, in the evaluation experiment of Section 6, we do not evaluate the Fluency dimension.

## 6 Experiments: System Evaluation

We show that the interactive automatic evaluation using the bipartite-play method correlates with humans as strongly as the all-play-all method, which has been reported to be an effective dialogue collection method but requires access to all evaluation targets. We first rank prepared evaluation targets by interactive human evaluation and then measure the correlation with the rankings by interactive automatic evaluations in the three dialogue collection methods: self-play, all-play-all, and bipartite-play.

### 6.1 Experimental settings

**Dialogue systems.** Table 2 shows the set of 11 evaluation targets and the set of 24 partner systems for the bipartite-play method with diverse architectures and training data from ParlAI.

**Dialogue collection settings.** For each of the three dialogue collection methods, We set the

<sup>4</sup>To compute inter-annotator agreement, we randomly divided the five annotators into two groups and calculated Spearman’s rank correlation coefficients between those groups. The results were 0.603 (Fluency), 0.835 (Specificity), 0.857 (Sensibleness), and 0.831 (Overall).

target-partner pairs for the self-play method, the all-play-all method, and the bipartite-play method. The resulting pairs are  $11 \times 1 = 11$ ,  $11 \times (11 - 1) = 110$ , and  $11 \times 24 = 264$ , respectively. A pair’s systems exchange utterances five times to form one dialogue following two given initial utterances, which we extracted from the initial parts of dialogues in the test set of the EmpatheticDialogues dataset (Rashkin et al., 2019). The evaluation target of each pair talks first. We found that ranking the 11 systems with the self-play method required 1,000 dialogues of each pair to converge in our settings, while the all-play-all method and the bipartite-play method required each pair’s 600 dialogues; we used these numbers of dialogues for the experiments.

**Interactive human evaluation.** We compute each evaluation target’s score for each of the three dimensions (i.e., Specificity, Sensibleness, and Overall) by averaging the manually annotated scores of 50 dialogues in Section 5.2. We then rank evaluation targets based on their averaged scores.

### 6.2 System evaluation results

Table 3 shows Spearman’s rank correlation coefficients of the automatic evaluations with the human evaluation. First, the automatic evaluation using the all-play-all method had a stronger correlation with humans than the self-play method; this is consistent with Yang et al. (2022)’s results. Second, the automatic evaluation with the bipartite-play method achieved the exact high correlation as the all-play-all method. This shows that the bipartite-play method enables reliable interactive automatic evaluation without direct interaction between evaluation targets.

Not requiring direct interaction makes system comparison across studies much easier. For in-



| Method         | Specificity | Sensibleness | Overall     |
|----------------|-------------|--------------|-------------|
| Self-play      | 0.83        | 0.70         | 0.77        |
| All-play-all   | <b>0.90</b> | <b>0.75</b>  | <b>0.85</b> |
| Bipartite-play | <b>0.90</b> | <b>0.75</b>  | <b>0.85</b> |

Table 3: Spearman’s rank correlation coefficients of the automatic evaluations using the three dialogue collection method with the human evaluation.

stance, with the same settings as our experiment, one can indirectly compare their systems with our evaluation targets by comparing systems’ FED scores. As one of the reference values, we present the FED scores of Tfm-3B-Rdt-Bsm, referred to as Blender 3B (Roller et al., 2021): 11.99 (Specificity), 14.48 (Sensibleness), and 3.99 (Overall).

### 6.3 Qualitative analysis of bipartite-play

Tables 4 and 5 show dialogue examples of Tfm-89M-Ddc-Ddg (an evaluation target) collected using the bipartite-play method. Tfm-89M-Ddc-Ddg talked with Tfm-1B-Rdt-Bsm, a high-performance system (Table 4), and GPT-117M-Wtx-Rdt, which is guessed to have relatively low performance in the set of dialogue partners (Table 5).

**Collecting dialogues.** Tfm-89M-Ddc-Ddg returned valid responses to high-quality utterances of Tfm-1B-Rdt-Bsm (Table 4). However, Tfm-89M-Ddc-Ddg generated semantically inappropriate responses (bolded parts) to low-quality utterances of GPT-117M-Wtx-Rdt (Table 5). These examples suggest that Tfm-89M-Ddc-Ddg can handle high-quality inputs but may be vulnerable to low-quality ones. Therefore, the bipartite-play method can reveal the diverse features of evaluation targets.

**Rating dialogues.** Regarding Tfm-89M-Ddc-Ddg, the scores of the dialogues in Tables 4 and 5 for the Overall dimension were 4.00 and 3.81, respectively; FED assigned higher scores to better dialogues than to mediocre ones. In these examples, FED succeeded in assigning appropriate ratings.

## 7 Discussion

The interactive automatic evaluation using the bipartite-play method correlates with humans as strongly as the all-play-all method, as described in Section 6, without direct interaction with evaluation targets. Additionally, we present another reason to use the bipartite-play method: one can sometimes cheat by choice of evaluation targets when using the all-play-all method.

|   |
|---|
| A: I love being in the "flow" – whether it’s in speaking to a group of people or working to help facilitate meetings. |
| B: What kind of work do you do?   |
| A: i’m a sales manager at a trading company.  |
| B: That sounds like a lot of fun. I bet you meet lots of interesting people. Do you like it?                          |
| A: i love it. i get to meet new people all the time.  |
| B: That’s great! I’m glad you enjoy it. It sounds like you have a great job.  |
| A: thanks. i’ve got a lot to learn.   |
| B: It’s always good to learn new things. What else do you like to do in your spare time?                              |
| A: i like to play chess. how about you?   |
| B: I love chess as well. My favorite game to play is checkers. Have you ever played that?                             |
| A: no, i haven’t. but i’d like to try.  |
| B: It’s a great game. You should give it a try. If you like chess, you’ll love it.                                    |

Table 4: Example of dialogue between an evaluation target (A; Tfm-89M-Ddc-Ddg) and one of the dialogue partners (B; Tfm-1B-Rdt-Bsm).

|   |
|---|
| A: I love being in the "flow" – whether it’s in speaking to a group of people or working to help facilitate meetings.     |
| B: What kind of work do you do?   |
| A: i’m a sales manager at a trading company.  |
| B: What type of job do you have? I’m looking to get a job as a software engineer in the next year or two.                 |
| A: i work in <b>the marketing department</b> . we’ve been in business for a long time.                                    |
| B: Where do you work? I’d love to be a part of the marketing team for a company like that. I’d be happy to work with you. |
| A: i work at a small company in the midwest.  |
| B: Where are you guys located? I’m in the Midwest and would love to work for a small business like that. Thanks!          |
| A: we’re in <b>los angeles</b> . we have a lot of offices here.   |
| B: Are you located in the US? If so, where are you located? If not, what state are you in? Thanks!                        |
| A: yes, we are in the u.s. and we are located in california.  |
| B: Are you in the U.S.? If so, what state do you live in? Thanks for your help!!  |

Table 5: Example of dialogue between an evaluation target (A; Tfm-89M-Ddc-Ddg) and one of the dialogue partners (B; GPT-117M-Wtx-Rdt). The evaluation target generated semantically inappropriate responses (bolded parts) to low-quality utterances of the partner.

### 7.1 How can we cheat on all-play-all?

We found that dialogues where systems frequently speak about the same things tend to receive low

ratings from the FED evaluation described in Section 5.2. This can be a desirable evaluation property since human evaluation is known to have the same tendency (Li et al., 2020). Therefore, we hypothesize that one could worsen a particular system’s ranking by forming an unfair set of evaluation targets where the system is likely to talk about repeated things.

## 7.2 Cheating examples

We show cheating using the all-play-all method following the above hypothesis. Specifically, based on Yang et al. (2022)’s observation that systems tend to speak repeated things in self-play (i.e., when talking with extremely similar systems), we attempt to worsen the ranking of a particular system by having the system talk with different but similar systems.

**Settings.** We form an unfair set of evaluation targets by collecting four systems, i.e., one whose rank we attempt to improve (favored system), another whose rank we attempt to worsen (unfavored system), and two systems similar to the unfavored system. We then check whether the ranking relationship between favored and unfavored ones changes from that of the original all-play-all evaluation (fair evaluation) in Section 6. In this unfair evaluation, unfavored systems have to construct dialogues with similar systems three out of four times, where repeated utterances are likely to occur as in self-play. We prepared two combinations of the unfavored system and its similar system: a series of DialoGPT (GPT-345M-Wtx-Rdt is the unfavored system, whose similar systems are GPT-124M-Wtx-Rdt and GPT-774M-Wtx-Rdt) and a series of Blender (Tfm-3B-Rdt-Bsm is the unfavored system, whose similar systems are Tfm-406M-Rdt-Bsm and Tfm-1B-Rdt-Bsm). We assigned each of all ten evaluation targets for the experiments in Section 6 except the unfavored one (GPT-345M-Wtx-Rdt or Tfm-3B-Rdt-Bsm) as a favored system. We focused on evaluation for Specificity, where the self-play property especially affects the results of automatic evaluation using the self-play method.

**Results.** Table 6 shows the change in the ranking relationship between favored and unfavored systems. The results show that we succeeded in intentionally improving the favored systems’ ranking in some cases. In this way, when using the all-play-all method, one can improve the automatic evaluation results of their systems by choice of evaluation targets. The bipartite-play method, fixing

| Fair \ Unfair | Favored wins | Favored loses |
|---------------|--------------|---------------|
| Favored wins  | 6            | <b>0</b>      |
| Favored loses | <b>2</b>     | 2             |

(a) Evaluation of 10 systems with DialoGPT series.

| Fair \ Unfair | Favored wins | Favored loses |
|---------------|--------------|---------------|
| Favored wins  | 1            | <b>0</b>      |
| Favored loses | <b>2</b>     | 7             |

(b) Evaluation of 10 systems with Blender series.

Table 6: Changes in the ranking relationship between favored versus unfavored systems by deploying unfair evaluation target sets instead of the original fair set. “Favored wins” means that a favored system was rated higher than the unfavored system. In both situations with the two unfair sets, the ranking was overturned in favor of the two favored systems out of ten.

and sharing a set of diverse partner systems, is one of the practical methods to prevent this cheating.

## 8 Conclusion

In this paper, we proposed the bipartite-play method as a dialogue collection method. The bipartite-play method can address the impossibility of comparison with publicly inaccessible systems and the vulnerability to cheating by intentional choice evaluation targets to improve the all-play-all method. For the proposed method, no dialogue with evaluation targets is required, thereby facilitating system comparison across studies and possibly enabling comparison with inaccessible systems. Our experiments showed that, compared with the evaluation using the all-play-all method, the automatic evaluation using the bipartite-play method correlates just as strongly with humans.

Although we formed a set of the bipartite-play method’s partner systems for the experiments considering its diversity of architectures and training data, it may still have some vulnerabilities. In future work, we will explore the property of the bot-bot dialogue further and refine the set of partner systems for the bipartite-play method.

## Acknowledgements

This work was mainly done under the NTT-Tohoku University collaborative research agreement. This work was also partly supported by JSPS KAKENHI Grant Numbers JP19H05693, JP21J22383, and JP22K17943 (training large-scale dialogue models, English proofreading, and registration fees).

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a Human-like Open-Domain Chatbot](#). In *aiXiv preprint arXiv:2001.09977*.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, Jeremy Blackburn, and Pushshift Io. 2020. [The Pushshift Reddit Dataset](#). In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM2020)*, volume 14, pages 830–839.
- Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. [Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs](#). In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (CMCL2011)*, pages 76–87.
- Jan 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 (INLG)*, pages 432–437.
- 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.
- 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. 2019a. [The Second Conversational Intelligence Challenge \(ConvAI2\)](#). In *aiXiv preprint arXiv:1902.00098*.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. [Wizard of Wikipedia: Knowledge-Powered Conversational agents](#). In *Proceedings of the 7th International Conference on Learning Representations (ICLR2019)*, pages 1–16.
- Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. [Bootstrapping dialog systems with word embeddings](#). In *Nips modern machine learning and natural language processing workshop*.
- Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. [Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems](#). In *Proceedings of the 2019 Conference on Neural Information Processing Systems (NeurIPS)*.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. [Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring](#).
- Gautier Izacard and Edouard Grave. 2021. [Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Main Volume*, pages 874–880.
- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. [Internet-Augmented Dialogue Generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL2022)*, volume 1, pages 8460–8478.
- Alon Lavie and Abhaya Agarwal. 2007. [METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, and Facebook Ai. 2020. [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pages 7871–7880.
- 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 (ACL2020)*, pages 4715–4728.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, Shuzi Niu, and Hong Kong. 2017. [Daily-Dialog: A Manually Labelled Multi-turn Dialogue Dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP2017)*, pages 986–995.
- 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, Veselin Stoyanov, and Paul G Allen. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#). In *aiXiv preprint arXiv:1907.11692*.
- Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. [Towards an Automatic Turing Test](#).



- Learning to Evaluate Dialogue Responses.** In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.
- Shikib Mehri and Maxine Eskenazi. 2020a. **Unsupervised Evaluation of Interactive Dialog with DialoGPT.** In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 225–235.
- Shikib Mehri and Maxine Eskenazi. 2020b. **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.
- Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. **ParIAI: A Dialog Research Software Platform.** In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 79–84.
- 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.
- 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 (ACL2002)*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. **Language Models are Unsupervised Multitask Learners.** *OpenAI Blog*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y. Lan Boureau. 2019. **Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset.** In *Proceedings of 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (ACL2019)*, pages 5370–5381.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. 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 (EACL)*, pages 300–325.
- Vasile Rus and Mihai Lintean. 2012. **A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics.** In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.
- Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020a. **Image-Chat: Engaging Grounded Conversations.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pages 2414–2429.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020b. **The Dialogue Dodecaathlon: Open-Domain Knowledge and Image Grounded Conversational Agents.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pages 2453–2470.
- Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Eth Zürich, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. **Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion.** In *aiXiv preprint arXiv:2203.13224 (2022)*.
- Eric Michael Smith and Adina Williams. 2021. **Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models.** In *aiXiv preprint arXiv:2109.03300 (2021)*.
- 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 (ACL2020)*, pages 2021–2030.
- Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. **SaFeR-Dialogues: Taking Feedback Gracefully after Conversational Safety Failures.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL2022)*, volume 1, pages 6462–6481.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. **Learning to Speak and Act in a Fantasy Text Adventure Game.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is All you Need.** In *Advances in Neural Information Processing Systems 30 (NIPS2017)*, pages 5998–6008.
- Jing Xu, Arthur Szlam, and Jason Weston. 2022. **Beyond Goldfish Memory: Long-Term Open-Domain Conversation.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL2022)*, volume 1, pages 5180–5197.



- Ruolan Yang, Zitong Li, Haifeng Tang, and Kenny Q Zhu. 2022. [ChatMatch: Evaluating Chatbots by Autonomous Chat Tournaments](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL2022)*, volume 1, pages 7579–7590.
- Chen Zhang, Yiming Chen, Luis Fernando D’Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. [DynaEval: Unifying Turn and Dialogue Level Evaluation](#). 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 5676–5689.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. 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 (ACL): System Demonstrations*, pages 270–278.