LEEETs-Dial: Linguistic Entrainment in End-to-End Task-oriented Dialogue systems

Nalin Kumar and **Ondřej Dušek**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
kumarnal@student.cuni.cz, odusek@ufal.mff.cuni.cz

Abstract

Linguistic entrainment, or alignment, represents a phenomenon where linguistic patterns employed by conversational participants converge to one another. While entrainment has been shown to produce a more natural user experience, most dialogue systems do not have any provisions for it. In this work, we introduce methods for achieving dialogue entrainment in a GPT-2-based end-to-end task-oriented dialogue system through the utilization of shared vocabulary. We experiment with training instance weighting, entrainment-specific loss, and additional conditioning to generate responses that align with the user. We demonstrate that all three approaches produce substantially better entrainment than the base, non-entrainment-optimized model, as confirmed by both automated and manual evaluation metrics.

"reservation" vs. "booking" (SNG076)

Input: Yes. I'll need a reservation for 3 nights starting Sunday. There are 3 people in my party.

Dataset: I have made reservations for 3 starting on Sunday for 3 nights.
Baseline: I'm sorry, but the booking was unsuccessful. Would you like to try another hotel?
Ours: Yes, your reservation is all set for 3 nights starting Sunday. Your reference number is 00000032. Do you need anything else?

"depart" vs. "leave" (SNG0256)

Input: Hi, I'm looking to catch a train to Norwich. Are there any that depart at or around 17:00?

Dataset: There are departures most days from Cambridge at 17:36. Would you like any more information of train times or help booking a train?
Baseline: Yes, there are 49 trains that leave Norwich on that day. What day would you like to travel?
Ours: I have a train departing Cambridge at 17:36 and arriving Norwich at 18:55. Would you like to book this train?

Figure 1: Examples of linguistic entrainment in task-oriented dialogues from the MultiWOZ 2.1 dataset (dialogue IDs shown in brackets). While the responses in the dataset reuse the same words, a base model (Base-CE) produces a non-matching expression, hampering dialogue fluency. Our approach (LK-CE(0)) keeps the appropriate reuse. See Section 4 for model details.

1 Introduction

During a natural dialogue, speakers adapt (entrain, align) to the way of speaking of their conversational partners, thereby establishing a shared understanding. This was shown to correlate with dialogue success (Nenkova et al., 2008) and it occurs at multiple linguistic levels: speakers synchronize their speech rate and phonetic patterns (Ostrand and Chodroff, 2021), adopt shared lexical terms (Brennan, 1996; Friedberg et al., 2012) and employ similar syntactic constructions (Reitter et al., 2006). Consequently, to facilitate successful and natural conversations, achieving entrainment is desirable in task-oriented dialogue systems (DSs), where the aim is to assist users in accomplishing tasks such as reserving tickets or venues. In DSs, entrainment can be useful for aligning on common lexical terms, increasing dialogue naturalness and possibly leading to higher efficiency. For instance, consider the two examples in Figure 1 – using "reservation" vs. "booking" or "depart" vs. "leave". However, few prior works attempted at entrainment in DSs, mostly with rule-based or modular DSs only (Lopes et al., 2013, 2015; Hu et al., 2014; Dušek and Jurčíček, 2016).

Recent years have seen significant advancements in task-oriented DSs through end-to-end neural models, fully trainable from data (Wen et al., 2017; Bordes et al., 2016; Lei et al., 2018). Use of pre-trained language models (LMs) yielded more fluent responses while simultaneously ensuring the comprehension of user intents and achieving successful dialogues (Lee, 2021; Yang et al., 2021; He et al., 2022). However, the generated responses often suffer from low diversity compared to human-human dialogues (Nekvinda and Dušek, 2021), and the DSs lack any dedicated support or mechanisms for entrainment, as their training relies on cross-entropy or other objectives that focus on dialogue content rather than phrasing.

Using the GPT-2-based two-stage system

Findings of the Association for Computational Linguistics: NAACL 2024, pages 727–735

AuGPT (Kulhánek et al., 2021) as our task-oriented end-to-end baseline DS, we propose the following three approaches to improve entrainment:

- a data-centric approach assigning higher weight to high-entrainment training instances via two straightforward weighting functions,
- an additional loss function to boost the probability of user tokens in generated responses,
- additional keyword-based generation conditioning to increase lexical entrainment.

We show that all our proposed approaches increase entrainment while minimally affecting other dialogue metrics; instance weighting and keyword conditioning also show improved human rankings. Our experimental code is released on GitHub.[1]

2 Related Works

Linguistic entrainment has been studied for decades (Garrod and Anderson, 1987; Brennan and Clark, 1996). In DSs, Reitter et al. (2006) modeled syntactic entrainment, while Nenkova et al. (2008) showed the correlation of high-frequency word entrainment with dialogue naturalness and success. Lopes et al. (2013) and (Hu et al., 2014) used rules to entrain lexical or syntactic choices of a spoken DS to the user; Lopes et al. (2015) used a statistical model based on handcrafted features. Work in statistical entrainment methods is limited; the only work known to us by Dušek and Jurčíček (2016) modified an LSTM-based response generator to adapt to the user's lexical choices.

State-of-the-art in task-oriented DSs is dominated by end-to-end systems based on pretrained neural LMs (Peng et al., 2021), which generate the belief state and the final response in sequence (Lei et al., 2018, cf. Section 3). Extensions involve using belief state differences (Lin et al., 2020), explicit system actions (Hosseini-Asl et al., 2020; Yang et al., 2021), contrastive classifiers (Peng et al., 2021) or data augmentation (Kulhánek et al., 2021). While a few techniques improve output diversity (Nekvinda and Dušek, 2021), none of them targets entrainment. Despite their recent popularity, prompted large LMs still underperform compared to finetuned LMs (Hudeček and Dusek, 2023).

[1] https://github.com/knalin55/LEEETs-Dial

3 Proposed Approaches

As our baseline model, we choose AuGPT (Kulhánek et al., 2021), a GPT-2 (Radford et al., 2019) based task-oriented end-to-end DS, which models dialogue as a sequence-to-sequence task. Same as other contemporary end-to-end systems, AuGPT works in two steps: (1) *generating belief state* (user-preferred slot values) from dialogue history and user input, and (2) *generating response* based on dialogue history, user input, generated belief state and database results (which are based on the belief state). We modify the response generation step.

Our modifications address primarily lexical entrainment and involve instance weighting (Section 3.1), an additional loss based on user input tokens (Section 3.2), and further conditioning on user keyword tokens on model input (Section 3.3).

3.1 Instance Weighting (IW)

We prioritize ground truth responses with greater overlap between the system and the user (i.e. higher entrainment) during training, by assigning them a higher weight. We use a simple 1-gram precision to quantify the lexical user-system overlap.

We explore two weight functions: (1) A discrete one with a simple threshold τ to distinguish high-entrainment training instances:

$$W_1(p) = 1 \text{ if } p \leq \tau, 10 \text{ otherwise}$$

(2) A continuous function modifying sigmoid:

$$W_2(p) = \frac{10}{1 + \exp(w \cdot (\beta - p))} + \epsilon$$

Here, w denotes a scaling factor (spread) and β is the average entrainment for the training data, centering the distribution. We add a small ϵ to avoid zero weight in instances with no entrainment.

3.2 User Likelihood Loss (ULL)

To increase lexical entrainment, we introduce a user-likelihood loss to increase the probability of reusing user tokens in the system output.

For a set of user tokens $U = \{u_1, u_2, \ldots, u_n\}$, we increase their likelihood by minimizing the loss:

$$L_t(p(.|x_{<t}), U) = -\alpha \cdot \log\left(\sum_{u \in U} p(u|x_t)\right)$$

Decreasing L_t means an increase in the probability $p(u|x_t)$. We add L_t to the base loss (Section 4.3) and use α to control the weight of user tokens.

3.3 Conditioning on Lexical Keywords (LK)

To enforce the reuse of user tokens, we introduce an additional section at the end of the AuGPT input sequence (i.e., after database results), called "keywords". During training, we include all overlapping tokens as keywords, so the model learns to incorporate them in its outputs.

During inference, we determine the keywords to be reused from the input user tokens using self-attention scores from the last encoder layer. We first calculate the mean across all attention heads. For each $u_i \in U = \{u_1, u_2, \ldots, u_n\}$, we compute the score $S(u_i) = \sum_{j,j \neq i} M_{ji}$, where M is the mean of last layer's attention heads. We then include as keywords all tokens u_i with scores $S(u_i) \geq t \cdot S_{max}$, where $S_{max} = max(S(u)|u \in U)$, with the threshold $t < 1$.

To smoothly expose the keywords to the model, we use a blending parameter σ (Roller et al., 2021; Nekvinda and Dušek, 2022), i.e., with the probability σ, we pass attention-scores-based keywords (as discussed in the previous paragraph) instead of overlapping tokens from the training instance.

4 Experiments

4.1 Data & Training Setup

For our experiments, we choose the MultiWOZ 2.1 dataset (Budzianowski et al., 2018; Eric et al., 2020), one of the most prominent task-oriented benchmarks with 10k dialogues spanning over 7 domains. As the dataset was created by online human-human dialogues, it does include naturally occurring entrainment and is thus suitable for the experiments, which we confirmed by an initial manual inspection and by computing entrainment metrics (cf. Section 4.4 and Table 1).[2]

We train all models for 10 epochs and keep the best checkpoint using the average of two token-level accuracies: accuracy against the ground-truth response (response contents) and against the user input (entrainment). We report test set scores averaged over 5 runs with different random seeds.

4.2 Baselines

Base We use Kulhánek et al. (2021)'s AuGPT as our base model. We start from the publicly available checkpoint pretrained on Taskmaster (Byrne et al., 2019) and Schema-guided Dialogue (Rastogi et al., 2020).[3] We then experiment with the choice of loss functions: In addition to the base cross-entropy loss (CE), we also consider the unlikelihood loss (Welleck et al., 2020) (CE+Unl).

D&J16 As an additional baseline, we use AuGPT with our own reimplementation of the decoding approach originally used by Dušek and Jurčíček (2016) in an LSTM-based context, which generates multiple outputs via beam search and then reranks them based on 1- and 2-gram match with the context. We use beam size 15.

GPT-4 For comparison with an LLM-based approach, we also include results for prompting GPT-4 (details are given in Appendix A). To limit experiment cost, we only use a sample of 200 instances from the test set.

4.3 Our Model Variants

$\mathbf{IW_i}$-loss We experiment with both functions defined in Section 3.1. Given that the dataset exhibits an 18.1% lexical overlap with user inputs (1-gram precision, lex-p_1; cf. Section 4.4), we set 25% as a desirable value.[4] Thus, we keep $\tau = 25.0$ for W_1. To spread W_2 almost to 0 and keep its mid-point around the dataset's 1-gram precision, we assign $\beta = 18.1$ and $w = 0.8$. We use $\epsilon = 0.1$. Thus, we have, $W_2(14.3) \approx 1.1$, $W_2(18.1) \approx 5.1$, and $W_2(25) \approx 10.06$.

ULL(α) For the choice of α in ULL, we start with $\alpha = 0.1$, and we gradually increase it to 0.5.We need a balanced combination of ULL and CE losses, as high α could lead to responses that are repetitive or identical to the user inputs. Additionally, as using ULL with CE only resulted in nonsensical repeats of user tokens, we only report scores for ULL with CE+Unl.

LK-loss (σ) For generation conditioned on keywords, we keep the threshold t as 0.1. We experiment with $\sigma \in \{0, 0.05, 0.5\}$.

4.4 Automatic Evaluation Metrics

We report the standard MultiWOZ metrics from Nekvinda and Dušek (2021) (*inform*, *success*,

[2]Note that other contemporary task-oriented sets, e.g., Schema-guided Dialogue (Rastogi et al., 2020), are not suitable as their dialogue structures were set by rules and crowd workers only paraphrased isolated utterances.

[3]https://huggingface.co/jkulhanek/augpt-bigdata

[4]We could not find any earlier work that discusses an ideal extent of lexical entrainment in such a context. We thus aimed at a slightly higher value than what is found in the data. Since our experiments showed promising results from the start, we did not optimize this parameter any further.

BLEU, and *delexicalized BLEU*) to evaluate state tracking and response generation. For lexical entrainment, we use 1-gram precision (lex-p_1) and recall (lex-r_1) against user input. For syntactic entrainment, we report the 2-gram (syn-p_2) and 3-gram precision (syn-p_3) scores on the POS tags of the user tokens and generated responses (i.e., matching part-of-speech patterns). We also use 50MFC, a variant of the metric introduced by Nenkova et al. (2008), measuring entrainment on the 50 most frequent words in the corpus:

$$50\text{MFC} = -\sum_{w \in 50\text{MF}} \left| \frac{\text{count}_S(w)}{|S|} - \frac{\text{count}_U(w)}{|U|} \right|$$

50MFC sums the differences in relative frequencies of 50 most frequent words in user and system utterances. It ranges from -2 to 0, with 0 being the perfect alignment. The idea is to measure entrainment on frequent, domain-independent words. We report average metrics from five runs with different random initializations, along with standard deviations.

4.5 Human Evaluation Setup

We run a small-scale in-house evaluation to complement the automatic evaluation scores. We use relative ranking by naturalness on a sample of 100 outputs. We select models from each group with better trade-offs between success rates and entrainment. We use the best-entraining model among the five runs. We report mean ranking (R_m) and proportions of instances with ranks 1,2,6,7 ($R_{1/2/6/7}$).

5 Results

5.1 Automatic Evaluation

Table 1 shows that all our approaches outperform the Base experiments on entrainment metrics. Although the models are primarily trained to increase the lexical entrainment, this also results in improved syntactic entrainment. As our methods do not differentiate between domain-specific terms and common words, the alignment on common words is also slightly improved in most setups, as shown by 50MFC scores. While the D&J16 reranking gets even better entrainment scores, its BLEU performance is low, as optimizing for 1/2-gram precision produces very terse outputs.

Models using IW do not only improve entrainment, but also maintain similar MultiWOZ scores to the baseline. In particular, IW_1-CE has substantially better lexical (lex-p_1 and lex-r_1) and syntactic (syn-p_2 and syn-p_3) entrainment while even maintaining a slightly better inform and success rates. Using IW_2 and/or Unl yields slightly lower success rates, with similar entrainment scores.

For ULL, entrainment scores show a positive correlation with the choice of α's while MultiWOZ scores decrease with an increase in α, but the drop is very slight for 0.1 and 0.2. This is not surprising, as with increasing α, the model gets more focused on aligning to the user and less on dialogue success. ULL(0.2) seems to have the best tradeoff.

The LK approach generally has high entrainment; the blending approach helps keep the keywords consistent during training and inference and is necessary to maintain good MultiWOZ scores.

The full results of the comparison with GPT-4 on the smaller data sample are shown in Table 3 in Appendix A. While GPT-4's responses look fluent and accurate and get high coverage of the user input tokens (lex-r_1) and even good syntactic entrainment (syn-p_2, syn-p_3), they are substantially longer, leading to lower precision-based lex-p_1 and BLEU scores. In addition, GPT-4 occasionally fails to follow the instructions given in the prompt.

5.2 Human Evaluation

Table 2 shows manual evaluation scores for selected setups. Here, IW_1-CE performs best on mean ranking and is most frequently ranked first, along with LK-CE. Despite similar numbers in Table 1, we see a noticeable difference between the scores of IW_1-CE and IW_2-CE. This can be attributed to the higher variance in lex-r_1, resulting in the outputs from the best run of IW_1-CE surpassing the quality of IW_2-CE. The generated responses from ULL experiments were often not fluent enough, hence their lower ranking. While their entrainment metrics are high, they only capture token-level alignment and are not directly related to fluency. In some of the examples, the outputs achieved high scores by simply repeated phrases from the user input. The human ranking here corresponds with the lowered MultiWOZ success rates, showing that entrainment cannot override the main dialogue objective. The outputs of the D&J16 reranking method were shorter, less polite, and less interactive, which resulted in the worst overall ranking. Appendix B illustrates this on a few sample outputs.

Model	MultiWOZ				Linguistic entrainment				
	inform	success	bleu	delex bleu	lex-p$_1$	lex-r$_1$	syn-p$_2$	syn-p$_3$	50MFC
Ground truth	-	-	-	-	18.1	21.4	13.0	3.8	-0.69
Base-CE	83.5$_{\pm0.7}$	65.8$_{\pm1.9}$	**15.7**$_{\pm0.5}$	17.4$_{\pm0.5}$	20.7$_{\pm0.4}$	24.5$_{\pm0.5}$	14.8$_{\pm0.2}$	5.0$_{\pm0.2}$	-0.31$_{\pm0.01}$
Base-CE+Unl	80.5$_{\pm2.7}$	65.1$_{\pm1.0}$	15.1$_{\pm0.8}$	16.8$_{\pm1.0}$	21.1$_{\pm1.1}$	23.8$_{\pm1.0}$	15.1$_{\pm0.5}$	5.0$_{\pm0.4}$	-0.31$_{\pm0.01}$
D&J16	**85.7**	63.6	10.6	11.5	**31.9**	26.1	**23.1**	**10.4**	-0.32
IW$_1$-CE	84.5$_{\pm1.9}$	**68.6**$_{\pm3.3}$	14.9$_{\pm1.0}$	16.3$_{\pm1.3}$	22.9$_{\pm0.7}$	30.9$_{\pm1.5}$	16.4$_{\pm0.1}$	5.9$_{\pm0.1}$	-0.31$_{\pm0.00}$
IW$_1$-CE+Unl	79.1$_{\pm3.0}$	64.4$_{\pm2.7}$	15.5$_{\pm0.7}$	**17.5**$_{\pm1.0}$	22.0$_{\pm0.7}$	26.7$_{\pm0.8}$	15.7$_{\pm0.3}$	5.4$_{\pm0.3}$	-0.31$_{\pm0.01}$
IW$_2$-CE	82.6$_{\pm3.7}$	67.7$_{\pm2.5}$	15.3$_{\pm0.9}$	16.9$_{\pm1.1}$	22.9$_{\pm0.9}$	29.8$_{\pm0.8}$	16.4$_{\pm0.5}$	5.8$_{\pm0.3}$	-0.31$_{\pm0.01}$
IW$_2$-CE+Unl	79.2$_{\pm2.0}$	64.1$_{\pm2.4}$	15.4$_{\pm0.9}$	17.3$_{\pm1.1}$	22.7$_{\pm0.9}$	28.0$_{\pm1.0}$	16.2$_{\pm0.5}$	5.6$_{\pm0.3}$	-0.31$_{\pm0.01}$
ULL (0.10)	80.6$_{\pm2.6}$	65.4$_{\pm2.2}$	15.5$_{\pm0.5}$	17.3$_{\pm0.6}$	22.8$_{\pm0.7}$	26.9$_{\pm0.8}$	16.0$_{\pm0.5}$	5.4$_{\pm0.3}$	-0.30$_{\pm0.01}$
ULL (0.20)	81.6$_{\pm2.0}$	65.3$_{\pm1.3}$	15.3$_{\pm0.7}$	17.0$_{\pm0.7}$	23.7$_{\pm0.2}$	29.4$_{\pm1.0}$	16.2$_{\pm0.1}$	5.7$_{\pm0.1}$	-0.29$_{\pm0.01}$
ULL (0.25)	81.6$_{\pm1.9}$	63.6$_{\pm2.4}$	14.6$_{\pm0.6}$	16.1$_{\pm0.6}$	24.7$_{\pm0.2}$	31.6$_{\pm1.5}$	16.9$_{\pm0.1}$	6.1$_{\pm0.1}$	-0.27$_{\pm0.01}$
ULL (0.30)	81.7$_{\pm2.9}$	61.5$_{\pm4.2}$	13.3$_{\pm0.5}$	14.8$_{\pm0.5}$	26.5$_{\pm0.8}$	34.6$_{\pm1.9}$	18.3$_{\pm1.0}$	7.2$_{\pm0.8}$	-0.25$_{\pm0.00}$
ULL (0.40)	80.2$_{\pm2.3}$	53.6$_{\pm3.3}$	11.8$_{\pm0.4}$	12.9$_{\pm0.4}$	27.9$_{\pm0.6}$	40.0$_{\pm0.7}$	19.0$_{\pm0.5}$	7.9$_{\pm0.3}$	-0.21$_{\pm0.01}$
ULL (0.50)	78.6$_{\pm2.7}$	45.7$_{\pm6.0}$	9.2$_{\pm1.1}$	9.9$_{\pm1.1}$	29.6$_{\pm1.7}$	**45.8**$_{\pm0.7}$	20.8$_{\pm0.5}$	9.5$_{\pm0.3}$	**-0.19**$_{\pm0.01}$
LK-CE (0)	77.4$_{\pm3.4}$	57.2$_{\pm5.6}$	11.3$_{\pm0.5}$	11.8$_{\pm0.6}$	26.3$_{\pm0.6}$	37.4$_{\pm2.1}$	17.2$_{\pm0.2}$	6.6$_{\pm0.2}$	-0.27$_{\pm0.01}$
LK-CE (0.05)	83.3$_{\pm0.9}$	66.3$_{\pm1.7}$	12.8$_{\pm0.1}$	13.9$_{\pm0.2}$	25.8$_{\pm0.4}$	33.6$_{\pm1.0}$	17.0$_{\pm0.3}$	6.5$_{\pm0.2}$	-0.29$_{\pm0.01}$
LK-CE (0.5)	83.3$_{\pm2.8}$	65.2$_{\pm1.6}$	14.6$_{\pm0.3}$	16.1$_{\pm0.4}$	22.6$_{\pm0.7}$	27.6$_{\pm0.4}$	15.5$_{\pm0.8}$	5.4$_{\pm0.5}$	-0.30$_{\pm0.01}$
LK-CE+Unl (0)	76.8$_{\pm2.5}$	59.4$_{\pm4.0}$	11.1$_{\pm0.4}$	11.7$_{\pm0.5}$	27.6$_{\pm0.6}$	39.3$_{\pm0.7}$	17.9$_{\pm0.4}$	7.1$_{\pm0.3}$	-0.27$_{\pm0.01}$
LK-CE+Unl (0.05)	82.4$_{\pm0.8}$	64.3$_{\pm2.9}$	12.1$_{\pm0.4}$	13.0$_{\pm0.4}$	25.1$_{\pm0.1}$	33.3$_{\pm0.2}$	16.6$_{\pm0.1}$	6.3$_{\pm0.1}$	-0.28$_{\pm0.01}$
LK-CE+Unl (0.5)	82.0$_{\pm0.8}$	65.2$_{\pm1.0}$	14.0$_{\pm0.1}$	15.6$_{\pm0.2}$	23.0$_{\pm0.3}$	27.9$_{\pm0.8}$	15.3$_{\pm0.3}$	5.3$_{\pm0.2}$	-0.29$_{\pm0.01}$

Table 1: Automatic metric results for state tracking, response generation and entrainment on the full MultiWOZ 2.1 test set (cf. Section 4.4 for metrics and Sections 4.2 and 4.3 for system variants). Except for D&J16, figures shown are averages of five runs with different random initializations, with standard deviations shown in subscript.

Model	R_m	R_1	R_2	R_6	R_7
base-CE	4.18	5	12	15	12
D&J16	5.35	1	7	26	30
IW$_1$-CE	**3.16**	26	18	12	3
IW$_2$-CE	3.77	20	15	13	15
LK-CE (0.05)	3.25	26	21	7	10
ULL (0.20)	4.17	15	10	16	11
ULL (0.25)	4.13	7	17	11	19

Table 2: Manual evaluation for generated responses on a sample of 100 outputs – mean rank R_m, and number of cases out of 100 where each system is ranked first (R_1), second (R_2), second to last (R_6) and last (R_7).

6 Conclusion

Although previous research showed that linguistic entrainment helps dialogue success, its application in end-to-end task-oriented dialogue systems has been largely overlooked. To address this gap, we introduced three techniques aimed at improving lexical entrainment of system responses to user inputs: (1) We show that prioritizing training instances with higher system-user word overlap improves entrainment, with comparable success rates. (2) We explore using user tokens' likelihood loss to control entrainment. While entrainment increases, both naturalness and correctness of outputs suffer with higher loss weight. (3) We additionally condition generation on user tokens likely to be reused (based on self-attention weights). We blend self-attention-selected tokens with true response tokens at training time to prime the model to use them. This yields responses with high fluency and better entrainment. The blending is necessary to maintain high dialogue success rate. In general, all methods seem to work successfully in aligning both domain-dependent and independent words.

In the future, we plan to incorporate longer context and focus more on syntactical entrainment. We also plan to use retrieval-augmented generation (Nekvinda and Dušek, 2022).

Acknowledgements

This research was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) and by the HumanE-AI-Net project (EC Horizon 2020, Grant Agreement H2020-FETFLAG-2018-2020 no. 952026, micro-project "Use of dialog context to boost ASR/NLG/TTS and improve the overall quality of voice dialog systems"). It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

Limitations

The proposed methods focus exclusively on addressing lexical entrainment in dialogues, overlooking entrainment at different linguistic levels. Additionally, the study is conducted and evaluated only at the response level despite the possibility of entrainment occurring across the entire dialogue.

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Susan E Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of the 1996 International Symposium on Spoken Dialogue*, 96:41–44.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. ArXiv: 1810.00278.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525.

Ondřej Dušek and Filip Jurčíček. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 185–190, Los Angeles. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Heather Friedberg, Diane Litman, and Susannah B. F. Paletz. 2012. Lexical entrainment and success in student engineering groups. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 404–409.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. GALAXY: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Zhichao Hu, Gabrielle Halberg, C. Jimenez, and M. Walker. 2014. Entrainment in pedestrian direction giving: How many kinds of entrainment. In *Proceedings of the IWSDS'2014 Workshop on Spoken Dialog Systems*, pages 90–101.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online. Association for Computational Linguistics.

Yohan Lee. 2021. Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1296–1303.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Melbourne, Australia. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2013. Automated two-way entrainment to improve spoken dialog system performance. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8372–8376. IEEE.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112.

Tomáš Nekvinda and Ondřej Dušek. 2022. AARGH! end-to-end retrieval-generation for task-oriented dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 283–297, Edinburgh, UK. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, Flavours of Success: The Case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT, Short Papers*, pages 169–172, Columbus, Ohio. Association for Computational Linguistics.

Rachel Ostrand and Eleanor Chodroff. 2021. It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue. *Journal of Phonetics*, 88:101074.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

David Reitter, Frank Keller, and Johanna D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124, New York City, USA. 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.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards fully end-to-end task-oriented dialog system with GPT-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.

A Comparison with GPT-4

Table 3 shows evaluation metric scores on randomly selected 200 examples from the test set and compares our results to GPT-4-based responses, given gold-standard dialogue states. The responses from GPT-4 look fluent and factually accurate and have the best coverage over the user inputs, as reported by lex-r_1 scores. Also, they seem to preserve the syntactic structure of the user inputs better than the other models, as evident by syn-p_2 & syn-p_3 scores. We observed this behavior even though GPT-4 was not specifically prompted to align syntactically. However, the generated outputs are substantially longer, leading to lower lex-p_1 and BLEU scores. Furthermore, GPT-4 struggles in several cases to generate appropriately delexicalized responses, further lowering the BLEU scores. Although we evaluated multiple variants of the prompt with instructions, GPT-4 still was not guaranteed to give an appropriate response. Figure 2 shows our prompt for GPT-4, along with a few example outputs, documenting both well-aligned and inappropriate responses.

B Sample Dialogues

Sample dialogues, with outputs of the different variants of our systems, are shown in Figure 3.

Model	bleu	delex bleu	lex-p$_1$	lex-r$_1$	syn-p$_2$	syn-p$_3$	50MFC	length ratio
base-CE	16.25	18.82	20.74	0.24	17.90	5.78	-0.13	1.10
D&J16	10.78	11.87	**33.06**	0.29	20.25	8.72	-0.17	0.76
GPT-4	3.21	2.67	17.29	**0.54**	**26.83**	**11.51**	-0.22	3.03
IW$_1$-CE	16.47	18.62	23.96	0.32	24.94	9.02	-0.15	1.26
IW$_2$-CE	14.42	16.39	22.34	0.28	21.49	5.93	**-0.10**	1.25
LK-CE (0.05)	12.12	13.77	26.21	0.32	21.31	8.06	-0.11	1.12
ULL (0.20)	14.02	16.18	23.65	0.29	18.59	6.49	-0.15	1.07
ULL (0.25)	**17.37**	**19.46**	26.62	0.33	23.33	9.33	-0.16	1.10

Table 3: Automatic evaluation scores for 200 randomly selected responses with GPT-4 as a baseline. Note that the scores for other systems differ from Table 1 as the evaluated sample is different. We only show scores for selected system variants with the best-entraining of the five runs. We evaluate the responses on MultiWOZ BLEU and delexicalized BLEU, entrainment metrics (as in Table 1), and length ratio of the generated responses to the gold ones.

	Prompt:	Generate a lexically aligned response to the user given the context, dialogue state and database results. Make sure to replace the named entities in the generated response with the given slot names, if any. <context> CONTEXT <dialogue state> DS <database results> DB_RESULTS <slot names> SLOTNAMES
1.	**User:**	Hi, I am trying to plan my trip and could use some help with a particular restaurant.
	GPT-4:	Hello! I'd be happy to help you with finding the perfect restaurant for your trip. Please provide some details about your preferences, such as the location, cuisine, and price range. Once I have this information, I will search through our database to find the most suitable option for you.
2.	**User:**	Please book for 4 people starting Friday for 2 nights.
	GPT-4:	I have successfully booked a room at the Home From Home hotel for 4 people, starting on friday for a duration of 2 nights. The hotel is in the moderate price range. Please note down the booking reference number: 00000021. If you need any further assistance, feel free to ask.
3.	**User:**	Great and what will be the cost of each ticket?
	GPT-4 (delexicalized):	The cost of each ticket is [price]14 Pounds[/price].
	GPT-4:	The cost of each ticket is 8.08 pounds14 Pounds[/price].
4.	**User:**	Yeas, what to recommend if I want to see good architecture in the west part of town?
	GPT-4:	I recommend you visit the west part of Cambridge to explore its stunning architecture. However, I do not have specific attractions in that area at the moment. You can still discover many beautiful buildings and landmarks during your visit. Enjoy your stay at the A and B Guest House with your group of 6 people for 4 nights starting from Tuesday. If you need any further assistance, feel free to ask.

Figure 2: Our GPT-4 prompt template (top) and example outputs (bottom). The CONTEXT, DS, DB_RESULTS and SLOTNAMES variables are filled in according to the current dialogue context, the (gold-standard) dialogue state and database results, and the slot names for the current domain. Examples 1 and 2 are quite fluent and syntactically aligned to the user. In Example 3, the model struggles with using slot placeholders and their correct values. This issue was quite frequent in our limited observation, even after trying several different prompts. Example 4 shows a self-contradicting response from the model.

Context: [...]
User: Yes please, Could you book 8 tickets and provide the reference number.
Ground Truth: Sure, the train is actually number TR0378 and your reference number is K3KI84VN. The price will be 188.8 GBP payable at the station. Is there anything else today?

Base-CE: Booking was successful, the total fee is 188.80 pounds payable at the station. Reference number is: 00000071.
D&J16: Booking was successful, the total fee is 188.80 pounds payable at the station. Your reference number is 00000071
IW1-CE: I have booked your 8 tickets. Your reference number is 00000071. Is there anything else I can help you with?
ULL (0.25): Yes I have booked 8 tickets and the reference number is 00000071
LK-CE (0.05): I was able to book 8 tickets for you. Your reference number is 00000071. Is there anything else I can help you with?

Context: [...]
User: Thanks. I'm also looking for a train for Saturday. Can you assist me with that?
Ground Truth: Yes, I can. Could you give me more information, such as, where you will departing from and the approximate time?

Base-CE: Sure, where will you be departing from and going to?
D&J16: Where are you departing from?
IW1-CE: I can help you with that. Where will you be departing from and arriving to?
ULL (0.25): Sure, I can assist you with that. Where are you departing from?
LK-CE (0.05): I can help with that. Where will the train depart from and where would you like to go?

Context: [...]
User: I would like one in the centre please.
Ground Truth: Christ's College, Corpus Christi, King's College, and Pembroke College are some of the colleges in that area

Base-CE: Christ's College is located in the centre and has free admission
D&J16: Christ's College is in the centre of town.
IW1-CE: There are 13 colleges in the centre. I recommend Christ's College. Would you like more information on it?
ULL (0.25): I would recommend Christ's College. Would you like the address?
LK-CE (0.05): Christ's College is one of the best college in the centre. Would you like more information?

Figure 3: In the first example, entrainment methods effectively produce more natural and less automated-like outputs, even when the ground truth response itself looks less natural. In the second example, the model outputs employing entrainment methods adeptly incorporate the phrases *Can you*, *assist me*, and *with that*, whereas the reranking method, D&J16, yields a lesser natural output. In the third example, the phrase *in the centre* is consistently present in almost every output, but D&J16 and Base-CE struggle to sustain the conversation. Conversely, the other methods successfully continue the conversation with improved entrainment.