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

Anonymous ACL submission

Abstract

Linguistic entrainment, or alignment, represents a phenomenon where linguistic patterns employed by conversational participants converge to one another. While alignment 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 alignment in a GPT-2-based end-to-end dialogue system through the utilization of shared vocabulary. We experiment with training instance weighting, alignment-specific loss, and additional conditioning to generate responses that align with the user. By comparing different entrainment techniques on the MultiWOZ dataset, we demonstrate that all three approaches produce significantly better-aligned results than the baseline, as confirmed by both automated and manual evaluation metrics.

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. However, few prior works attempted this, 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., 2016; Bordes et al., 2016; Lei et al., 2018). Use of pretrained language models 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 aligning responses, as their training relies on cross-entropy or other objectives that focus on dialogue content rather than phrasing.

User: I would like a taxi from Saint John's college to Pizza Hut Fen Ditton.
BS: taxi {departure = saint john's college, destination = pizza hut fenditton}

Generic Response: What time do you want to leave?
Preferred Response: What time would you like to leave?

Figure 1: An example of linguistic entrainment. The preferred response has the same syntactic construction as the user input, along with overlapping function words.

We propose multiple approaches to ensure alignment for end-to-end dialogue models. We employ the GPT-2-based two-stage system AuGPT (Kulhánek et al., 2021) as our task-oriented end-to-end baseline DS. We investigate and utilize various techniques, encompassing both data-centric approaches and the incorporation of additional objectives, to improve dialogue alignment while maintaining the success rate of generated dialogues. Our proposed methods outperform the baseline on automated and manual evaluation metrics, showing improved entrainment to user inputs. Our contributions can be summarized as follows:

- We propose a data-centric approach to promote better-aligned training instances, assigning them a higher weight during training via two straightforward weighting functions.
- We introduce an additional loss function that

maximizes the probability of user tokens within the generated responses.

- We present additional keyword-based generation conditioning, designed to allow dialogue alignment at the lexical level.

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

2 Related Works

Linguistic entrainment has been studied for quite some time (Brennan and Clark, 1996; Garrod and Anderson, 1987). In dialogue systems, Reitter et al. (2006) modeled entrainment of syntactic rules, while Nenkova et al. (2008) showed the correlation of entraining high-frequency words with dialogue naturalness and success. Lopes et al. (2013) and (Hu et al., 2014) proposed rules for adapting the lexical or syntactic choices of the system to that of the user in spoken DS; 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 natural language generator for adapting to the user's lexical choices.

Current state-of-the-art in task-oriented DSs is dominated by end-to-end systems based on pretrained neural language models (Peng et al., 2021) which generate in two stages (Lei et al., 2018), generating the belief state and the final response in sequence (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 of the techniques improve output diversity (Nekvinda and Dušek, 2021), none of them target entrainment.

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 dialogues as a sequence-to-sequence task. Similar to other contemporary end-to-end systems, AuGPT processes dialogues in two steps: (1) *generating belief state* (user-preferred slot values) from dialogue history and user input, and (2) *generating response* using a sequence of dialogue history, user input, generated belief state and database results (which are based on the belief state). We make modifications to the response generation step.

Our modifications address lexical and syntactic entrainment and can be categorized into three groups: 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 that exhibit a greater degree of overlap between the system and the user (i.e. better alignment) 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: discrete and continuous. The discrete one uses a simple threshold τ to distinguish well-aligned training instances:

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

For a continuous weight function, we modify the sigmoid function as follows:

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

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

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 $U = \{u_1, u_2, \dots, u_n\}$, we increase their likelihood by minimizing the loss:

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

Decreasing L_t means an increase in the probability $p(u|x_t)$. We add L_t to the base loss (Section 4.2) and use α to control the weight of user tokens.

3.3 Conditioning Generation on Lexical Keywords (LK)

To enforce reusing of user tokens, we introduce an additional section at the end of the AuGPT input sequence (i.e., after database results), called

[1] URL will be provided in the final version.

"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 $t < 1$.

4 Experiments

4.1 Data & Training Setup

We experiment on the publicly available MultiWOZ 2.1 dataset (Eric et al., 2020). MultiWOZ is a collection of 10k task-oriented human-human written dialogues spanning over 7 domains.

We train all models for 10 epochs and consider the best checkpoint using the average of two token-level accuracies: accuracy against the ground-truth response (response contents) and against the user input (user alignment). We report scores on the MultiWOZ 2.1 test set, averaged over 5 runs with different random seeds.

4.2 Model Variants

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).[2] We then experiment with the choice of loss functions: In addition to the base cross-entropy loss (CE), we also consider the unlikelihood loss introduced by Welleck et al. (2019) (CE+Unl).

$\mathbf{IW_i}$-loss We experiment with both weight functions defined in Section 3.1. The dataset exhibits a 1-gram precision of 18.1, and we posit that a desirable threshold would be 25.0. 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 user-likelihood loss, we experiment with $\alpha \in \{0.1, 0.2, 0.25, 0.3, 0.4, 0.5\}$. We only report scores with CE+Unl since using CE only resulted in nonsensical repeats of user tokens.

LK For generation conditioned on keywords, we keep the threshold t as 0.1.

4.3 Automatic Evaluation Metrics

For overall dialogue quality, we use the standard MultiWOZ metrics from Nekvinda and Dušek (2021). We report *inform*, *success*, *BLEU*, and *delexicalized BLEU* to evaluate state tracking and response generation. To capture 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 entrainment metric introduced by Nenkova et al. (2008), measuring alignment on the 50 most frequent words in the corpus:

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

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 alignment on frequent, domain-independent words.

4.4 Human Evaluation Setup

We perform a small-scale in-house evaluation. 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 alignment (six in total). Based on the alignment scores, we use the best model among the five runs for manual evaluation. We report mean ranking (R_m), as well as proportions of instances where the generated outputs are ranked 1st (R_1), 2nd (R_2), 5th (R_5), and 6th (R_6).

5 Results

5.1 Automatic Evaluation

Table 1 shows that all three approaches outperform the baseline on linguistic entrainment metrics. The models trained using IW also have statistically similar MultiWOZ scores to the baseline models. In particular, IW_1-CE has significantly better lexical (lex-p_1 and lex-r_1) and syntactic (syn-p_2 and syn-p_3) alignment while even maintaining a slightly

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

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	$83.5_{\pm 0.7}$	$65.8_{\pm 1.9}$	$\mathbf{15.7}_{\pm 0.5}$	$17.4_{\pm 0.5}$	$20.7_{\pm 0.4}$	$24.5_{\pm 0.5}$	$14.8_{\pm 0.2}$	$5.0_{\pm 0.2}$	$-0.71_{\pm 0.01}$
base-(CE+Unl)	$80.5_{\pm 2.7}$	$65.1_{\pm 1.0}$	$15.1_{\pm 0.8}$	$16.8_{\pm 1.0}$	$21.1_{\pm 1.1}$	$23.8_{\pm 1.0}$	$15.1_{\pm 0.5}$	$5.0_{\pm 0.4}$	$-0.71_{\pm 0.01}$
IW$_1$-CE	$\mathbf{84.5}_{\pm 1.9}$	$\mathbf{68.6}_{\pm 3.3}$	$14.9_{\pm 1.0}$	$16.3_{\pm 1.3}$	$22.9_{\pm 0.7}$	$30.9_{\pm 1.5}$	$16.4_{\pm 0.1}$	$5.9_{\pm 0.1}$	$-0.69_{\pm 0.01}$
IW$_1$-(CE+Unl)	$79.1_{\pm 3.0}$	$64.4_{\pm 2.7}$	$15.5_{\pm 0.7}$	$\mathbf{17.5}_{\pm 1.0}$	$22.0_{\pm 0.7}$	$26.7_{\pm 0.8}$	$15.7_{\pm 0.3}$	$5.4_{\pm 0.3}$	$-0.70_{\pm 0.01}$
IW$_2$-CE	$82.6_{\pm 3.7}$	$67.7_{\pm 2.5}$	$15.3_{\pm 0.9}$	$16.9_{\pm 1.1}$	$22.9_{\pm 0.9}$	$29.8_{\pm 0.8}$	$16.4_{\pm 0.5}$	$5.8_{\pm 0.3}$	$-0.69_{\pm 0.01}$
IW$_2$-(CE+Unl)	$79.2_{\pm 2.0}$	$64.1_{\pm 2.4}$	$15.4_{\pm 0.9}$	$17.3_{\pm 1.1}$	$22.7_{\pm 0.9}$	$28.0_{\pm 1.0}$	$16.2_{\pm 0.5}$	$5.6_{\pm 0.3}$	$-0.69_{\pm 0.00}$
ULL (0.10)	$80.6_{\pm 2.6}$	$65.4_{\pm 2.2}$	$15.5_{\pm 0.5}$	$17.3_{\pm 0.6}$	$22.8_{\pm 0.7}$	$26.9_{\pm 0.8}$	$16.0_{\pm 0.5}$	$5.4_{\pm 0.3}$	$-0.69_{\pm 0.00}$
ULL (0.20)	$81.6_{\pm 2.0}$	$65.3_{\pm 1.3}$	$15.3_{\pm 0.7}$	$17.0_{\pm 0.7}$	$23.7_{\pm 0.2}$	$29.4_{\pm 1.0}$	$16.2_{\pm 0.1}$	$5.7_{\pm 0.1}$	$-0.67_{\pm 0.01}$
ULL (0.25)	$81.6_{\pm 1.9}$	$63.6_{\pm 2.4}$	$14.6_{\pm 0.6}$	$16.1_{\pm 0.6}$	$24.7_{\pm 0.2}$	$31.6_{\pm 1.5}$	$16.9_{\pm 0.1}$	$6.1_{\pm 0.1}$	$-0.65_{\pm 0.01}$
ULL (0.30)	$81.7_{\pm 2.9}$	$61.5_{\pm 4.2}$	$13.3_{\pm 0.5}$	$14.8_{\pm 0.5}$	$26.5_{\pm 0.8}$	$34.6_{\pm 1.9}$	$18.3_{\pm 1.0}$	$7.2_{\pm 0.8}$	$-0.62_{\pm 0.01}$
ULL (0.40)	$80.2_{\pm 2.3}$	$53.6_{\pm 3.3}$	$11.8_{\pm 0.4}$	$12.9_{\pm 0.4}$	$27.9_{\pm 0.6}$	$40.0_{\pm 0.7}$	$19.0_{\pm 0.5}$	$7.9_{\pm 0.3}$	$-0.57_{\pm 0.01}$
ULL (0.50)	$78.6_{\pm 2.7}$	$45.7_{\pm 6.0}$	$9.2_{\pm 1.1}$	$9.9_{\pm 1.1}$	$\mathbf{29.6}_{\pm 1.7}$	$\mathbf{45.8}_{\pm 0.7}$	$\mathbf{20.8}_{\pm 0.5}$	$\mathbf{9.5}_{\pm 0.3}$	$\mathbf{-0.52}_{\pm 0.01}$
LK-CE	$77.4_{\pm 3.4}$	$57.2_{\pm 5.6}$	$11.3_{\pm 0.5}$	$11.8_{\pm 0.6}$	$26.3_{\pm 0.6}$	$37.4_{\pm 2.1}$	$17.2_{\pm 0.2}$	$6.6_{\pm 0.2}$	$-0.65_{\pm 0.01}$
LK-(CE+Unl)	$76.8_{\pm 2.5}$	$59.4_{\pm 4.0}$	$11.1_{\pm 0.4}$	$11.7_{\pm 0.5}$	$27.6_{\pm 0.6}$	$39.3_{\pm 0.7}$	$17.9_{\pm 0.4}$	$7.1_{\pm 0.3}$	$-0.65_{\pm 0.01}$

Table 1: Automatic evaluation metric scores of state tracking and response generation on MultiWOZ. We use 1-gram precision and recall for evaluating lexical entrainment.

better inform and success rates. Using IW$_2$ and/or Unl results in slightly lowered success rates, with similar alignment scores.

For ULL, alignment scores show a positive correlation with the choice of α's while MultiWOZ scores decrease with increasing 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-generated outputs have high alignment but lower MultiWOZ scores than the baseline. This can be attributed to the inconsistent "keywords" values during training and inference.

Model	R_m	R_1	R_2	R_5	R_6
base-CE	3.93	7	12	33	10
IW$_1$-CE	2.97	19	25	11	8
IW$_2$-CE	3.67	16	14	19	21
ULL (0.20)	3.89	9	15	16	22
ULL (0.25)	3.80	9	18	17	22
LK-CE	**2.74**	40	16	4	17

Table 2: Manual evaluation scores for generated responses – mean rank R_m, and number of cases out of 100, where the given system is ranked first (R_1), second (R_2), second to last (R_5) and last (R_6).

5.2 Human Evaluation

Table 2 shows manual evaluation scores for selected setups. Here, LK-CE performs best on mean ranking and is most frequently ranked first. 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 low ranking.

6 Conclusion

Although previous research indicated that linguistic entrainment helps achieving successful dialogue, its application in end-to-end task-oriented dialogue systems has been largely neglected. To address this gap, we introduce three techniques aimed at aligning the system's responses with user inputs: (1) We show that prioritizing training instances with better alignment helps achieve better output alignments, with comparable success rates. (2) We explore the use of user tokens' likelihood loss to control alignment. While alignment is increased, with increased loss weight, both naturalness and correctness of outputs suffer. (3) We additionally condition generation on user tokens likely to be reused (based on self-attention weights). This yields responses with high fluency and better alignment, but with a slightly lower success rate.

In the future, we plan to incorporate longer context and add approaches focused solely on syntactical entrainment. We also plan to try approaches based on retrieval-augmented generation (Nekvinda and Dušek, 2022).

References

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Susan E Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 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.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358*.

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.

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.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

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 Appendix: Sample Dialogues

User: I would like to leave Cambridge to Bishops Stortford.
Ground Truth: I have one leaving at 17:29, arriving in Bishop's Stortford at 18:07. Does that work?

Base-CE: I have train TR0969 leaving at 17:29 and arriving at 18:07. Would that work for you?
IW1-CE: I have 4 trains that match your criteria. The first leaves at 17:29 and arrives at 18:07. Would you like me to book you a ticket?
ULL (0.25): I have a train that leaves at 17:29. Would you like me to book that for you?
LK-CE: TR0969 will leave Cambridge at 17:29 and arrive in Bishop's Stortford at 18:07. Would you like to book a ticket?

User: Can I get a taxi to take me from the restaurant to the hotel?
Ground Truth: What time would you like that taxi to pick you up?

Base-CE: Sure! What time would you like to leave or arrive by?
IW1-CE: Sure, when would you like the taxi to pick you up?
ULL (0.25): a red volvo will pick you up at cafe uno and take you to the hotel. Contact number is 53623942635. Can I help with anything else?
LK-CE: I can get that taxi for you. What time would you like to leave the restaurant?

Context: [...]
User: Yes. please book me a table for friday at 13:00 for 6 people.
Ground Truth: I have booked your table. The Reference number is: RQS2GVAQ

Base-CE: I have made your reservation. The reference number is 00000001. Is there anything else I can help you with?
IW1-CE: Your table is reserved for Friday at 13:00 for 6 people. The reference number is 00000001.
ULL (0.25): I have made those reservations. Your reference number is 00000001.
LK-CE: I was able to book your table for Friday at 13:00 for 6 people. Your reference number is 00000001.

Figure 2: In the first example, the alignment models modify and use the phrase *I would like* in their responses. For LK-CE, due to the presence of input keywords *Cambridge* and *Bishops Stortford*, it retains those terms as well. In the second example, IW1-CE fails to align to the phrase *Can I get*. In the third example, the ground truth and the User align on *book me a table*. However, the baseline fails to capture it because of the lack of a significant number of such training instances. The ULL model fails to align completely, while LK-CE aligns on both the significant phrases in the user input.