A Multi-Task Based Neural Model to Simulate Users in Goal-Oriented Dialogue Systems

To Eun Kim
University College London
London, United Kingdom
to.kim.17@ucl.ac.uk

Aldo Lipani
University College London
London, United Kingdom
aldo.lipani@ucl.ac.uk

ABSTRACT

A human-like user simulator that anticipates users' satisfaction scores, actions, and utterances can help goal-oriented dialogue systems in evaluating the conversation and refining their dialogue strategies. However, little work has experimented with user simulators which can generate users' utterances. In this paper, we propose a deep learning-based user simulator that predicts users' satisfaction scores and actions while also jointly generating users' utterances in a multi-task manner. In particular, we show that 1) the proposed deep text-to-text multi-task neural model achieves state-of-the-art performance in the users' satisfaction scores and actions prediction tasks, and 2) in an ablation analysis, user satisfaction score prediction, action prediction, and utterance generation tasks can boost the performance with each other via positive transfers across the tasks. The source code and model checkpoints used for the experiments run in this paper are available at the following weblink: https://github.com/kimdanny/user-simulation-t5.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Simulation types and techniques**; • **Information systems** → *Users and interactive retrieval.*

KEYWORDS

user simulation, multi-task learning, user satisfaction prediction, user action prediction, user utterance generation, task-oriented dialogue systems

ACM Reference Format:
To Eun Kim and Aldo Lipani. 2022. A Multi-Task Based Neural Model to Simulate Users in Goal-Oriented Dialogue Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531814

1 INTRODUCTION

A goal-oriented dialogue system is a conversational agent that interacts with users to solve specific tasks. One of the main challenges in building such conversational agents is to gather enough dialogues in order to train a model that is able to generalize and reach a satisfactory performance in terms of user satisfaction [5, 22]. Furthermore, the interactive nature of a dialogue system makes its evaluation challenging [1].

To tackle these limitations, user simulators, mimicking users' behavior, are needed in order to train and evaluate dialogue systems [1, 15]. For training, a user simulator can be used to generate a vast amount of synthetic dialogues, which then can be used for teaching dialogue strategies to the agent. For example, in reinforcement learning, a simulator can act as the environment that affects the agent with its own rewards [5, 11]. For evaluation, a user simulator can be used to interact with dialogue systems while tracking the generated dialogues with predicted satisfaction scores [15].

Another aspect of a user simulator is the modeling of the users' satisfaction. If a simulator can track the turn-level satisfaction score of users, the automatic evaluation of the agent's response based on the satisfaction score becomes possible [3, 5]. Along with the user satisfaction modeling, there have been some efforts to model the generation of users' utterances. One popular approach is an Agenda-Based User Simulation (ABUS) [23]. This probabilistic approach randomly sets a user goal and keeps it unchanged during the whole dialogue. Then, the agent's task is to let users achieve their goals by gradually figuring the users' needs out. However, as neural networks have shown promising results in various research domains, data-driven neural user simulators have demonstrated superior performance to the previous approaches [11].

Two recent papers have influenced our work. The first work includes user satisfaction analysis, and the second work includes user utterance generation. In the former, Sun et al. [25] investigated how users' satisfaction influences users' actions during the dialogue with an agent. They noticed that the system's unprofessional responses or failure to catch users' requirements can result in users' dissatisfaction. In the latter, Ben-David et al. [2] recommended the use of the T5 model [19] and used auxiliary tasks like utterance reordering and utterance generation, as an approach to improve the performance of the users' intents prediction task which are proven to be effective. Following this study, we hypothesize that the users' utterance generation task can give a positive transfer to the users' satisfaction scores and actions prediction tasks when trained together.

In this paper, we make the following contributions:

- We develop a neural architecture and train it to predict both satisfaction scores and actions at the same time in a multi-task learning (MTL) setting. We abbreviate this model as ***SatAct***. By doing this, we see an increase in performance for both prediction tasks through positive transfers across the tasks.

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531814

- We develop a neural architecture and train it to predict satisfaction scores and actions, and generate users' utterances at the same time in an MTL setting. We abbreviate this model as ***SatActUtt***.
- We perform an ablation study to further validate our hypothesis.

2 RELATED WORK

2.1 Satisfaction Modeling

Sun et al. [25] developed a model to predict users' satisfaction scores and actions by cascading one model specialized in the prediction of users' satisfaction scores to another model specialized in the prediction of users' actions. The latter uses the output of the former as an input. The results show that the Hierarchical Gated Recurrent Units (HiGRU) [9] and BERT [6] based models perform well in these two tasks. This analysis was performed using the conversational dataset named User Satisfaction Simulation (USS). This dataset contains turn-level annotations of users' satisfaction scores (on a 5-point scale) and actions [25], and it is built as an extension of 5 conversational datasets (MultiWOZ 2.1 [7], SGD [20], CCPE [18], JDDC [4], and ReDial [12]).

Another approach to model user satisfaction is to train a conversational agent with users' feedback in an online setting [8]. While conversing with users, the agent attempts to predict their users' satisfaction scores after every user utterance. However, the satisfaction score prediction is performed after each user utterance, so-called, an after-utterance (AU) prediction. This is in contrast with the before-utterance (BU) prediction. The AU prediction of user satisfaction can sometimes be easier than the BU prediction since there may be patterns in user utterances that indicate a clear dissatisfaction, e.g., "What are you talking about?" [8]. However, the BU prediction, although more difficult, can be used by the agent to prevent the user from having a bad experience beforehand by changing the course of the dialogue towards a less unsatisfactory one.

In this paper, we use the USS as our main dataset, because the turn-level satisfaction annotations are made before the users' last utterance (BU). We take HiGRU and BERT as our baseline models for the user satisfaction and action prediction tasks. Among the 5 different sources of datasets in USS, we do not use ReDial and JDDC. We do not use the former because it does not contain the annotation of users' actions. We do not use the latter because this source is in Chinese, whereas this paper we focuses on English. There are a few more differences we make in the preparation of the USS dataset, which will be discussed in Section 4.1.

2.2 Neural Language Models

After the introduction of the Transformer architecture [26], many encoder-based language models, such as BERT [6] were used in various tasks in Natural Language Processing. These models are usually pretrained with large corpora in a self-supervised way before being fine-tuned to specific downstream tasks. As these encoder-only architectures are more suitable for discriminative tasks, decoder-based models, such as GPT-2 [17] are used for generative tasks. Raffel et al. [19] experimented with an encoder-decoder model, named T5, where the target function is mapped into text. For example, for a regression task, the model is trained to generate a text like "v" where $v \in \mathbb{R}$, from a input sequence prepended with a task specific prefix. In other words, it converts every task into a text generation task, i.e., a regression task becomes a generation of a text-formatted target. This characteristic of T5 makes its use easy in a multi-task setting without the need to define task-specific heads.

The T5 model has been used in dialogue response generation tasks [2, 10, 14]. Kale and Rastogi [10] transformed every encoded system action into natural language, then concatenated them into a sequence to pass it into the T5 model. The T5 model is then fine-tuned to fuse the input sentences into one. Lin et al. [14] fine-tuned T5 in an end-to-end fashion for a response generation task, which resulted in state-of-the-art performance with fewer human annotations. Ben-David et al. [2] used the T5 model in a multi-task setting for users' intent prediction, training on the utterance reordering and generation tasks. They showed that these auxiliary tasks improved the performance of the users' intent prediction task.

2.3 Multi-Task Learning

There have been attempts to use a Multi-Task Learning (MTL) strategy for dialogue systems. Xu et al. [28] have shown that a BERT-based self-supervised MTL approach can significantly improve the performance on the multi-turn response selection task. This improvement was achieved by adding four auxiliary tasks. All four auxiliary tasks were proven to be useful as removing any of them decreased the performance of the model on the main task. Zhang et al. [29] developed DialogBERT. They pretrained BERT with five self-supervised tasks, which were found to be helpful in several other dialogue systems-related tasks.

In this paper, we fine-tune a T5 model in an MTL setting on the USS dataset to make a user simulator that predicts users' satisfaction scores and actions, and generates users' utterances. We hypothesize that the use of the pretrained T5 model fine-tuned in an MTL setting will allow the transfer of prior and task-specific knowledge across the tasks. Moreover, the simplistic nature of the T5 model allows the training of a regression task for satisfaction scores, classification task for actions, and generation task for users' utterances simultaneously.

3 MULTI-TASK LEARNING

We design three models based on T5 as illustrated in Figure 1: the *SatAct*, *SatActUtt* and *Utt* models. These models differ based on the considered learning tasks. In what follows we will indicate the dialogue history with $\mathcal{H}$, the satisfaction score with s, the action with a and the user-utterance with u.

3.1 SatAct

The goal of *SatAct* is to learn the function $\mathbf{M}(s, a|\mathcal{H})$, where an MTL based generative model $\mathbf{M}$ generates satisfaction scores and actions given a dialogue context $\mathcal{H}$. This model is used to compare the performance of an MTL based T5 model against the models developed by Sun et al. [25].

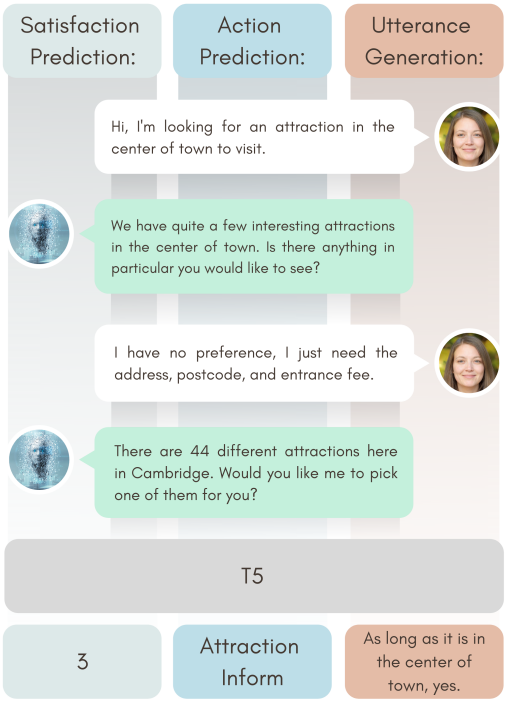

Figure 1: Multi-task learning with T5 and task-specific prefixes: "Satisfaction Prediction:", "Action Prediction:", and "Utterance Generation:". These are prepended to the dialogue context. Here we show a MultiWOZ 2.1 sample and the result generated by the fine-tuned T5 model.

3.2 SatActUtt

The goal of *SatActUtt* is to learn the function $\mathbf{M}(s, a, u|\mathcal{H})$, where an MTL based generative model $\mathbf{M}$ generates users' utterances along with satisfaction scores and actions given the dialogue context $\mathcal{H}$. Should this model be successful it would validate our hypothesis: adding utterance generation as an additional task to the *SatAct* model may provide a positive transfer across the tasks.

3.3 Utt

The goal of *Utt* is to learn the function $\mathbf{M}(u|\mathcal{H})$, where a generative model $\mathbf{M}$ generates only a user utterance given the dialogue context $\mathcal{H}$. The purpose of this model is to investigate the impact of satisfaction score and action prediction tasks on the utterance generation task. If the performance of *SatActUtt* is better than *Utt* then the transfer across the tasks is deemed positive.

4 EXPERIMENTS

4.1 Dataset Preparation

We are following the general preparation strategy of Sun et al. [25] with the exception of the use of different upsampling factors and the introduction of a new augmentation strategy. The former is used to mitigate a dataset imbalance and the latter to improve learning.

Table 1: Statistics of the dataset with upsampling factors.

	MultiWOZ 2.1	SGD	CCPE
#utterances	12 553	13 833	6 860
#(non-3) / #(3)	0.13	0.20	0.29
upsampling factor	7.5	5.0	3.5

First, we transform the USS dataset into:

$$\mathcal{D} = \{(\mathcal{H}_i, s_i, a_i, u_i)\}_{i=1}^{N} \quad (1)$$

where i is the index of a sample, $\mathcal{H}_i$ is the dialogue history, $s_i \in \{1, \ldots, 5\}$ is the user's satisfaction score, a_i is the user's action, and u_i is user's utterance. When training, we provide the models a dialogue history with up to 10 previous turns ({i-9, ..., i}) and let the models learn to predict the next turn ($i+1$). 10 turns are suitable considering the maximum token length of the T5-base model.

Second, we mitigate a dataset imbalance. As shown in Table 1, there is a large discrepancy between the ratios of non-3-rated and 3-rated satisfaction scores. To overcome this issue, Sun et al. [25] over-sampled the cases of non-3-ratings by a factor of 10. This factor was constant across the datasets. In our case, we calculate a specific upsampling factor for each dataset. In this way, we prevent a model from seeing too many non-3-ratings, which can harm the performance of the user simulator.

Third, while upsampling, rather than simply copying the same utterances, we randomly select one augmentation strategy among random deletion, random swap, random insertion, WordNet-based synonym replacement, and back-translation (to and from a random language). We perform this augmentation because it has been proven to be effective in several NLP tasks [27].

4.2 Training Details

We split the dataset into train, validation and test sets with a ratio of 8:1:1. When splitting the dataset, we ensured that each satisfaction scores were evenly spread across the splits. When training the model, we used T5-base (220M parameters, vocabulary size: 32 128) [19], and train up to 7 epochs with the following hyperparameters: *batch size* = 4, *Optimizer* = AdamW, *learning rate* = 1e-3. *max_length* is set to 10 when training a *SatAct*, while set to 100 when training *SatActUtt*. The loss function is defined as the negative log-likelihood.

Throughout the training phase, we evaluated an updated model with the validation set and saved the best model based on the validation loss. We used early stopping and linear learning rate scheduling. During the token generation phase, we set the *beam_size* to 5, the *top_p* to 0.95, and the *repetition_penalty* to 2.0. The training took around 6 hours for each model with an NVIDIA Tesla V100 GPU.

4.3 Evaluation Measures

For the evaluation of the models on the satisfaction score and action prediction tasks, we follow the previous work [25]. For the satisfaction score prediction, we use: Unweighted Average Recall (UAR), Cohen's Kappa, Spearman's Rho, and binary-F1-score (positive when satisfaction score > 2). For the action prediction, we use: Accuracy, Precision, Recall, and F1-score. We do not use the

Table 2: Performance for the User Satisfaction Score Prediction.

	MultiWOZ 2.1				SGD				CCPE			
	UAR	Kappa	Rho	F1	UAR	Kappa	Rho	F1	UAR	Kappa	Rho	F1
HiGRU (Sun *et. al.*)	0.225	0.143	**0.886**	0.238	0.293	0.118	0.451	0.086	0.237	0.167	0.881	0.274
BERT (Sun *et. al.*)	0.256	0.133	0.823	0.224	0.261	0.094	0.477	0.048	0.232	0.147	**0.891**	0.245
SatAct	0.535	**0.824**	0.873	**0.901**	0.449	0.619	0.681	0.713	0.222	0.094	0.347	0.165
SatActUtt	**0.572**	0.767	0.815	0.838	**0.608**	**0.763**	**0.822**	**0.847**	**0.437**	**0.612**	0.690	**0.734**

Table 3: Performance for the User Action Prediction.

	MultiWOZ 2.1				SGD				CCPE			
	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
HiGRU (Sun *et.al.*)	0.518	0.216	0.162	0.167	0.643	0.534	0.505	0.507	0.672	0.503	0.472	0.482
BERT (Sun *et.al.*)	0.519	0.255	0.183	0.191	0.661	0.570	0.572	**0.560**	**0.674**	**0.696**	**0.495**	**0.496**
SatAct	0.616	0.295	0.309	0.299	0.630	0.433	0.471	0.447	0.544	0.267	0.235	0.207
SatActUtt	**0.621**	**0.439**	**0.407**	**0.402**	**0.678**	**0.538**	**0.574**	**0.550**	0.623	0.458	0.429	0.430

Table 4: Performance for the User Utterance Generation.

		BLEU-1	BLEU-4	ROUGE-1-F	ROUGE-2-F	ROUGE-L-F	STS
MultiWOZ 2.1	*SatActUtt*	0.206	**0.019**	**0.193**	**0.065**	**0.182**	**0.327**
	Utt	**0.210**	0.018	0.192	0.063	0.180	0.318
SGD	*SatActUtt*	**0.281**	**0.050**	**0.269**	**0.125**	**0.263**	**0.403**
	Utt	0.267	0.048	0.251	0.117	0.247	0.394
CCPE	*SatActUtt*	**0.195**	**0.019**	**0.240**	**0.072**	**0.230**	**0.410**
	Utt	0.170	0.012	0.222	0.064	0.208	0.394

Table 5: Cross-domain UAR for the User Satisfaction Score Prediction.

Trained	Generate	MultiWOZ 2.1	SGD	CCPE
MultiWOZ 2.1	BERT	–	0.233	0.226
	SatActUtt	–	0.247	**0.275**
SGD	BERT	0.249	–	0.223
	SatActUtt	**0.280**	–	0.233
CCPE	BERT	0.213	0.216	–
	SatActUtt	0.266	**0.264**	–

accuracy measure for the satisfaction score prediction task due to the imbalance of the labels. For the utterance generation task, we use BLEU [16] and ROUGE [13] scores. For BLEU, we use the cumulative 1-gram (BLEU-1) and 4-gram (BLEU-4). For ROUGE, we use the ROUGE-1-F, ROUGE-2-F, and ROUGE-L-F. Also, we use the Semantic Textual Similarity (STS) score to evaluate the contextual similarity between the ground truth sentence and the generated user utterance. STS is defined as the cosine similarity between the sentence level embedding vector $\mathbf{e_1}$ from the ground truth utterance and embedding vector $\mathbf{e_2}$ from the model-generated utterance. The average of this similarity across the whole dataset is used for evaluating the performance of models trained on the user utterance generation task. The embeddings are taken from the pretrained *SRoBERTa-STSb-large* model [21].

5 RESULTS AND DISCUSSION

In Table 2, we show the performance of the models in predicting user satisfaction scores. In Table 3, we show the performance of the models in predicting user actions. In Table 4, we show the performance of the models in generating the user's utterances. Finally, in Table 5 we show the generalization ability of the models on satisfaction score prediction by testing them on different datasets.

First, as it is shown in Tables 2 and 3, MTL models achieve state-of-the-art performance in most of the metrics in both satisfaction score and action predictions by a large margin. Second, *SatActUtt* model is better than *SatAct* model in most cases. This means that utterance generation task has given a positive transfer to satisfaction and action prediction tasks. Third, in Table 4, *SatActUtt* model always beats the *Utt* model with the exception of one case. This means that the training on the satisfaction score and action prediction tasks also gives a positive effect on the training on the user utterance generation task. These results show that all three tasks can help each other to better simulate users. Moreover, in Table 5, the T5 model always shows better generalization ability than BERT in a cross-domain user satisfaction prediction task. This is most likely due to the larger number of parameters of the T5 model and the use of a bigger corpus used to pretrain it. Another noticeable result is that the T5 model does not work very well in CCPE. We believe that it is due to the small number of training samples in the dataset that hindered the model from being well fine-tuned on the CCPE domain.

One critical limitation of the proposed simulator is that it lacks the ability of modeling users' knowledge and mental status. For instance, for the ground-truth utterance: *"Yes, book the tickets, also I want places to go in town."*, our simulator generates: *"Yes, please book it for me."* However, the second part of the ground-truth sentence (*"also I want places to go in town."*) is very challenging to predict as it is a new topic that the user brought up. This is where a personal knowledge graph [1] or a memory augmented neural model [24] can come into play to incorporate the users' mental status or preferences into the simulator. Additionally, this reinforces the need for the design of better evaluation metrics for user utterance generation which are able to capture these subtleties.

6 CONCLUSION AND FUTURE WORK

In this paper, we have shown that the T5 model achieves state-of-the-art performance in predicting user satisfaction scores and actions in the USS dataset with cross-domain generalization ability. This is the first work that combines the user-utterance generation task with the user satisfaction score and action prediction tasks. Moreover, in our analysis, we proved that satisfaction score and action prediction, and utterance generation tasks give a positive transfer to each other when trained in an MTL setting. As future work, we plan to design new dialogue strategies that are able to anticipate potential users' dissatisfactions by using the predicted satisfaction scores.

REFERENCES

[1] Krisztian Balog. 2021. Conversational AI from an Information Retrieval Perspective: Remaining Challenges and a Case for User Simulation. In *Proc. of the Second International Conference on Design of Experimental Search & Information REtrieval Systems, Padova, Italy, September 15-18, 2021 (CEUR Workshop Proceedings, Vol. 2950)*. CEUR-WS.org, 80–90.

[2] Eyal Ben-David, Boaz Carmeli, and Ateret Anaby-Tavor. 2021. Improved Goal Oriented Dialogue via Utterance Generation and Look Ahead. arXiv:2110.12412 [cs.CL]

[3] Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint Turn and Dialogue level User Satisfaction Estimation on Multi-Domain Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 3897–3909.

[4] Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service. In *Proc. of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 459–466.

[5] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54, 1 (2020), 755–810.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, 4171–4186.

[7] 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 *Proc. of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 422–428.

[8] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazaré, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot! arXiv:1901.05415 [cs.CL]

[9] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. HiGRU: Hierarchical Gated Recurrent Units for Utterance-Level Emotion Recognition. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, 397–406.

[10] Mihir Kale and Abhinav Rastogi. 2020. Template Guided Text Generation for Task-Oriented Dialogue. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 6505–6520.

[11] Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 2018. Neural User Simulation for Corpus-based Policy Optimisation of Spoken Dialogue Systems. In *Proc. of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 60–69.

[12] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Proc. of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., 9748–9758.

[13] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81.

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

[15] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How Am I Doing?: Evaluating Conversational Search Systems Offline. *ACM Trans. Inf. Syst.* 39, 4, Article 51 (aug 2021), 22 pages.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, 311–318.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[18] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proc. of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 353–360.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[20] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8689–8696.

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3982–3992.

[22] Joel Ross, Andrew Zaldivar, Lilly C. Irani, and Bill Tomlinson. 2009. Who are the Turkers? Worker Demographics in Amazon Mechanical Turk.

[23] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Association for Computational Linguistics, 149–152.

[24] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. In *Association for the Advancement of Artificial Intelligence*. AAAI.

[25] Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. *Simulating User Satisfaction for the Evaluation of Task-Oriented Dialogue Systems*. Association for Computing Machinery, 2499–2506.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.

[27] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 6382–6388.

[28] Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues. *Proc. of the AAAI Conference on Artificial Intelligence* 35, 16 (May 2021), 14158–14166.

[29] Zhenyu Zhang, Tao Guo, and Meng Chen. 2021. *DialogueBERT: A Self-Supervised Learning Based Dialogue Pre-Training Encoder*. Association for Computing Machinery, 3647–3651.