EmoUS: Simulating User Emotions in Task-Oriented Dialogues

Hsien-Chin Lin
Shutong Feng
Christian Geishauser
linh@hhu.de
Heinrich Heine University Düsseldorf
Düsseldorf, Germany

Nurul Lubis
Carel van Niekerk
Michael Heck
Heinrich Heine University Düsseldorf
Düsseldorf, Germany

Benjamin Ruppik
Renato Vukovic
Milica Gašić
gasic@hhu.de
Heinrich Heine University Düsseldorf
Düsseldorf, Germany

ABSTRACT

Existing user simulators (USs) for task-oriented dialogue systems only model user behaviour on semantic and natural language levels without considering the user persona and emotions. Optimising dialogue systems with generic user policies, which cannot model diverse user behaviour driven by different emotional states, may result in a high drop-off rate when deployed in the real world. Thus, we present EmoUS, a user simulator that learns to simulate user emotions alongside user behaviour. EmoUS generates user emotions, semantic actions, and natural language responses based on the user goal, the dialogue history, and the user persona. By analysing what kind of system behaviour elicits what kind of user emotions, we show that EmoUS can be used as a probe to evaluate a variety of dialogue systems and in particular their effect on the user's emotional state. Developing such methods is important in the age of large language model chat-bots and rising ethical concerns.

CCS CONCEPTS

• **Human-centered computing** → **User models**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

dialogue system, user simulation, emotion simulation

ACM Reference Format:
Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2023. EmoUS: Simulating User Emotions in Task-Oriented Dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3539618.3592092

1 INTRODUCTION

Task-oriented dialogue systems (DSs) help users accomplish their goals, such as searching for nearby restaurants or booking a hotel. Proficient DSs are often trained via reinforcement learning (RL), which demands a large number of interactions between the system and users, making training with real users expensive and time-consuming. Therefore, utilizing user simulators (USs) to build a controlled interactive environment becomes attractive [6].

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592092

Despite recent USs in task-oriented dialogues properly modelling user extrinsic behaviour in terms of semantic actions and natural language [17, 36], a crucial aspect is still lacking: the user intrinsic state such as user persona and the emotional state. A generic user policy may lead to limited linguistic diversity and fails to capture diverse actions driven by varying user emotions. Adjusting the probability distribution of user actions in rule-based USs is a popular method to address diversity [13], but real users differ in more ways than just action preferences. Training USs by supervised learning with different initialisation [35] or by RL with varying reward functions can also form various user policies [17], but that can only provide diverse extrinsic behaviour, e.g. the action length in each turn or the semantic content.

In this work, we propose a user simulator that models the user emotional state conditioned on the dialogue context and the user persona. More specifically, our contributions are as follows:

- We propose an **emo**tional **u**ser **s**imulator that we call *EmoUS*[1]. The *EmoUS* response includes the user emotion, semantic actions, and natural language utterances. To the best of our knowledge, this is the first user simulator with user emotion for task-oriented dialogue systems.
- EmoUS exhibits an increased linguistic diversity for the same context by modelling the user policy and emotion jointly,
- The user emotion of EmoUS provides valuable insights for evaluating DSs, offering a more subtle and detailed understanding beyond a simple measure of task success.

2 RELATED WORK

The effectiveness of a task-oriented dialogue policy trained by RL with a US is greatly affected by the quality of the US [27]. Rule-based USs are commonly used to train DSs, such as the agenda-based US (ABUS) [28]. ABUS models the user goal as a stack-like agenda, ordered by the priority of the user actions updated by hand-crafted stacking and popping rules. While its action probability distribution can be manipulated to simulate different user behaviour [13], it only generates semantic actions without natural language generation or emotion prediction. Moreover, designing rules for complex scenarios is labour-intensive and transferring these rules to new domains can be challenging. To address these limitations, data-driven USs have been developed, which learn user policy directly from data. The sequence-to-sequence (Seq2Seq) model structure is the most common framework. The input sequence may include the dialogue history and user goal as a list of features or plain text. The output sequence can be semantic actions or natural language utterances [7, 10, 15, 17, 18, 37, 38]. Tang et al. [35] train USs by supervised

[1] https://gitlab.cs.uni-duesseldorf.de/general/dsml/emous-public

learning with different initialisation to create different user policies. Lin et al. [17] proposed GenTUS, an ontology-independent US which generates natural language utterances as well as the underlying semantic actions for interpretability. Its behaviour can be configured by RL with different reward functions. These USs can simulate extrinsic user behaviour, e.g. actions and utterances, but intrinsic user states are neglected, e.g. satisfaction level and emotional status [24].

In comparison to generating responses with given emotions [3, 21, 33] or recognising user satisfaction classification after receiving user utterances [1, 8, 11, 12, 30, 32], the user satisfaction modelling should predict intrinsic user states first then generates actions or utterances. Sun et al. [34] and Deng et al. [4] investigate how user satisfaction impacts user behaviour on the semantic level. Pan et al. [23] transfer the emotion from chit-chat to task-oriented dialogues utilising data augmentation. Kim and Lipani [14] proposed SatActUtt, which generates users' satisfaction, action (only with intent and domain), and utterance based on dialogue history as multi-task learning. We consider SatActUtt as our baseline as it is the first US modelling both intrinsic and extrinsic user behaviour. While SatActUtt can predict user satisfaction scores adequately based on dialogue history, it does not include the user goal. This makes it difficult to train a dialogue system. In addition, it only considers satisfaction and dissatisfaction, disregarding aspects such as different emotion elicitors or user personas [19, 22]. Feng et al. [9] annotated a task-oriented dialogue dataset with 7 user emotions inspired by Ortony, Clore and Collins (OCC) emotion model [22]. It considers user conduct and emotion elicitors for human-human and human-machine task-oriented dialogues, representing a more fine-grained user intrinsic state for task-oriented dialogues.

3 SIMULATING USER EMOTION IN TASK-ORIENTED DIALOGUES

Task-oriented DSs are underpinned by an *ontology* which is typically composed of *intents*, *domains*, *slots*, and *values*. *Intents* define user or system global intentions of their respective actions in each turn. Users and systems may have different intents, e.g., systems can *confirm* user's request and users can *negate* system's proposal. *Domains* are the topics that can be discussed in the conversation. They can be further specified by *slots* and each can take a number of *values*. We assume that the users of task-oriented dialogues have a *goal* they want to achieve, which is defined as $G = \{d_1 : [(s_1, v_1), (s_2, v_2), \dots], d_2 : [(s_3, v_3), \dots], \dots\}$, where domain d_i, slot s_i and value v_i are selected from the ontology.

Semantic *user actions* and *system actions* are composed of tuples, $(intent, domain, slot, value)$. Semantic actions can be transformed into natural language utterances. User *emotion* in task-oriented dialogues may be triggered by different elicitors, or related to different user personas. For example, the system not responding adequately may lead to the user being dissatisfied [9]. A user *persona* represents users' attitudes and feelings towards certain events, such as feeling fearful after a robbery [20] or includes users' conduct, i.e. how users behave when communicating, e.g. are they polite or impolite. For example, the persona of a polite user who is feeling excited to visit a museum is *persona* = {user: polite, attraction: excited}. The user persona can be derived from dialogue history during training and sampled from a distribution for inference.

User simulation with emotion can be viewed as a Seq2Seq problem. For each turn, EmoUS predicts the user emotion based on the context information, e.g. the dialogue history, the user goal, and the user persona, and generates semantic actions and natural language responses based as follows.

3.1 Model structure

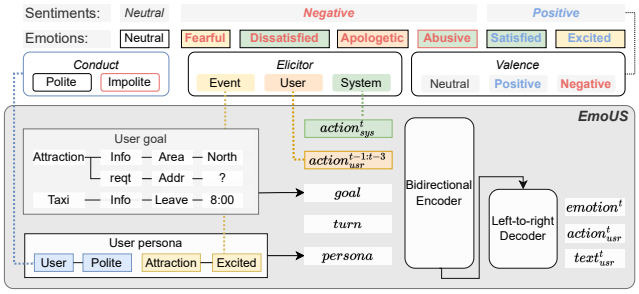

Figure 1: The model structure of EmoUS

EmoUS builds upon GenTUS and additionally incorporates user persona and emotion prediction. More specifically, EmoUS takes the system action $action^t_{sys}$, user history $action^{t-1:t-3}_{usr}$, user goal *goal*, turn information *turn* and the user persona *persona* as input and generates user emotion *emotion*, semantic actions $action^t_{usr}$, and an utterance $text^t_{usr}$ as output at turn t (as shown in Fig. 1). By introducing different user personas and emotions, more diverse user behaviours on both semantic and linguistic aspects can be simulated even in the same context.

EmoUS considers the three aspects of user emotions in task-oriented dialogues according to EmoWOZ [9], namely *elicitor*, *conduct*, and *valence*. The emotion elicitor can be an event, the system, or the user. Their respective information can be captured from the event *persona* attribute, system $action^t_{sys}$, and user $action^{t-1:t-3}_{usr}$. The user conduct, whether polite or impolite, is recorded as a user persona. The valence aspect, or the sentiment polarity of each emotion, is informed intrinsically in the emotion prediction.

Following the setting in Lin et al. [17], the input and output sequences are represented as JSON-formatted strings, composed of natural language tokens. In this way, EmoUS achieves ontology independence and can transfer to unseen domains.[2] Then we train EmoUS as a Seq2Seq model and leverage BART [16], a transformer-based natural language generator with a bidirectional encoder and a left-to-right decoder. BART demonstrates impressive performance in a range of language-related tasks.

4 EXPERIMENTAL SETUP

The aim of our experiments is to demonstrate that EmoUS is able to generate user emotion, semantic actions, and utterances based on the context of the conversation and the user persona. Furthermore, we show that the emotion prediction of EmoUS is a valuable tool

[2]As this property is directly inherited from GenTUS, we do not examine it in our experiments.

for evaluating DSs. We conduct our experiments on EmoWOZ [9]. It contains user emotion annotations for human-human dialogues from MultiWOZ [2] and $1k$ human-machine dialogues between volunteers and an RNN-based dialogue policy trained on MultiWOZ.

4.1 Supervised learning for emotion simulation

Our model is inherited from Huggingface's transformers [39] and trained on EmoWOZ. To measure the emotion prediction performance, we calculate the macro-F1 score of sentiments and emotions. We compare sentiment prediction against SatActUtt [14], a user model which predicts sentiment, user action (composed with intent and domain only), and utterances based on the dialogue history.

Following the setting of Lin et al. [17], we evaluate the performance of modelling user semantic actions by F1 score and turn accuracy and the natural language generation (NLG) performance by slot error rate (SER), sacre-BLEU score [25] and self-BLEU score [41]. SER measures the agreement between the semantic actions and the corresponding utterance. $SER = (m + h)/N$, where N is the total number of slots in semantic actions, m and h stand for the number of missing and hallucinated slots. The self-BLEU evaluates the diversity of generated utterances in the following way. After generating a sentence for every data point, we calculate a BLEU score by treating all other generated sentences as references. By averaging these scores, we get the self-BLEU score where the lower score implies a higher diversity.

4.2 Interacting with DS

We estimate the generalisation ability of a US by cross-model evaluation, where a DS trained with a particular US is evaluated by different USs [29]. Policies of different DSs are trained with various USs, including the agenda-based US (ABUS) with T5 [26] natural language generator (ABUS-T5), GenTUS, and EmoUS, by proximal policy optimisation (PPO) [31], a simple and stable RL algorithm, for 200 epochs, each of which consists of 1000 dialogue turns. Each policy is trained on 5 random seeds and the performance is averaged. The DSs also include a natural language understanding module composed with BERT [5] for understanding users' utterances and a rule-based dialogue state tracker for tracking the users' states under the ConvLab-3 framework [40].

We also analyse how different system behaviour elicit user emotions. For this purpose, we used $1k$ dialogues between EmoUS and a DS trained by EmoUS. We categorised various system behaviour in the following groups: *confirm* - the system repeats the slots and values informed by the user, *no_confirm* - the system does not repeat this information, *miss_info* - the system requests the information just mentioned by the user, *neglect* - the system does not respond to the user request, *reply* - the system responds to the user request, and *loop* - the system takes identical actions for two turns in a row.

5 EXPERIMENTAL RESULTS

5.1 User emotion modelling

As shown in Table 1, EmoUS outperforms SatActUtt on sentiment prediction by 0.314 on macro-F1 score. This is not unexpected as EmoUS includes the user goal in inputs and the user sentiment in task-oriented dialogues is centred around the user goal [9]. In addition, the performance of sentiment prediction between EmoUS and EmoUS$_{noPersona}$ is similar, but the emotion prediction improves by 0.202 on the macro-F1 score when including the user persona. This indicates that considering the user persona improves the performance of user emotions triggered by different elicitors.

Table 1: Performance for emotion and sentiment prediction of different models by measuring macro-F1 score.

model	sentiment	emotion
SatActUtt	0.379	-
EmoUS$_{noPersona}$	0.673	0.299
EmoUS	**0.693**	**0.501**

We demonstrate that user emotion simulation can be further configured by multiplying different weights w on the probability of *neutral*, i.e. *neutral* is more likely to be selected with a higher weight. As shown in Fig. 2, EmoUS is purely neutral without any emotion as $w = 1.5$. As the weight decreases, EmoUS achieves the best performance on *fearful* as $w = 0.95$, best on *dissatisfied* as $w = 0.9$, and best on *apologetic* as $w = 0.85$. Thus, we can change the probability distribution of emotions in the user response, inducing different user behaviour, by modifying the weight of emotions.

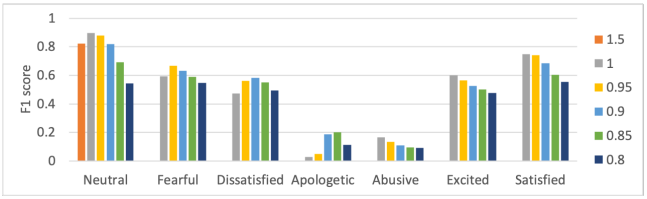

Figure 2: Different weights of the neutral emotion will have different F1-score on each user emotion.

5.2 User action prediction

The results of user action prediction are shown in Table 2, where EmoUS$_{emo}$ generates semantic actions based on golden emotions. EmoUS is superior to SatActUtt because EmoUS can generate semantic actions following the user goal, whereas SatActUtt does not consider the user goal. Additionally, EmoUS is still comparable to GenTUS despite it models a more complex task, simulating user emotions and semantic actions jointly. Moreover, EmoUS$_{emo}$ surpasses GenTUS since EmoUS$_{emo}$ generates semantic actions utilising more information then GenTUS, such as the user persona and golden emotions.

5.3 Natural language evaluation

The NLG results are shown in Table 3, where GenTUS$_{act}$ generates utterances based on golden semantic actions and EmoUS$_{emo+act}$ is based on golden emotion and semantic actions. On the other hand, GenTUS and EmoUS are generated based on their prediction. The Sacre-BLEU is calculated with golden utterances.

Although SatActUtt generates the most diverse utterances with the lowest Self-BLEU score, it also has the lowest Sacre-BLEU score,

Table 2: Performance for user action prediction.

model	Intents+domains F1	Intents+domains ACC	Full action F1	Full action ACC
GenTUS	0.890	0.854	0.762	0.600
SatActUtt	0.317	0.221	-	-
EmoUS	0.892	0.857	0.764	0.603
EmoUS$_{emo}$	**0.904**	**0.867**	**0.775**	**0.611**

which means it by and large generates random responses irrelevant to the user goal. On the other hand, EmoUS$_{emo+act}$ has a comparable Sacre-BLEU and SER with GenTUS$_{act}$ and a much lower Self-BLEU score, which means EmoUS is able to generate more diverse responses than GenTUS but still follows the user goal and maintains the agreement between the semantics and the language.

Table 3: The NLG performance on EmoWOZ of different models. The arrow directions represent which trend is better.

model	SER↓	Sacre-BLEU↑	Self-BLEU↓
Human	0.054	-	0.770
GenTUS	0.116	-	0.950
GenTUS$_{act}$	**0.092**	**19.61**	0.930
SatActUtt	-	2.90	**0.433**
EmoUS	0.118	-	0.715
EmoUS$_{emo+act}$	0.096	16.91	0.708

5.4 Cross-model evaluation

As shown in Table 4, the DS trained with EmoUS performs comparably to the DS trained with ABUS-T5 when evaluating with ABUS-T5 (0.62 vs 0.63 success rate), and similarly to the DS trained with GenTUS when evaluating with GenTUS (both at 0.53 success rate). However, the DS trained with EmoUS outperforms the DS trained with ABUS-T5 by 7% absolute and the DS trained with GenTUS 5% absolute on success rate when evaluating with EmoUS (success rates of 0.52 vs 0.45 and 0.47 respectively). This indicates that EmoUS provides a better learning environment and makes DSs trained with it perform well when evaluated on diverse USs.

Table 4: The success rates of policies trained on EmoUS, GenTUS, and ABUS with T5 NLG (ABUS-T5) when tested on various USs. Each pair is evaluated by 400 dialogues on 5 seeds, which is 2K dialogues in total.

US for training	US for evaluation ABUS-T5	GenTUS	EmoUS
ABUS-T5	0.63	0.48	0.45
GenTUS	0.60	0.53	0.47
EmoUS	0.62	0.53	**0.52**

5.5 System behaviour with the user emotions

In 1k dialogues between EmoUS and a DS trained by it, the system behaviour *confirm*, *no_confirm*, and *miss_info* elicit neutral emotion. As systems respond properly, e.g. *reply*, users are likely to feel satisfied, but when systems behave unprofessionally, e.g. *neglect* and *loop*, users may feel dissatisfied (see Table 5). This observation is in line with the user study conducted by Sun et al. [34].

Furthermore, we plot the average user sentiment per turn where positive = +1, neutral = 0, and negative = −1, for each dialogue outcome. As expected, users are more positive in successful dialogues and more negative in failed dialogues on average (see Fig. 3). In addition, we also notice a drop in sentiment around turn 6, which suggests the user may feel impatience after that.

Table 5: Proportion of neutral and system-eliciting emotions triggered by various system behaviour.

System behaviour	User emotion Neutral	Dissatisfied	Satisfied
confirm	86.00%	2.20%	11.80%
no_confirm	71.80%	16.60%	11.60%
miss_info	79.20%	11.10%	9.70%
neglect	27.10%	65.00%	7.90%
reply	51.50%	4.10%	44.40%
loop	28.60%	65.90%	5.50%

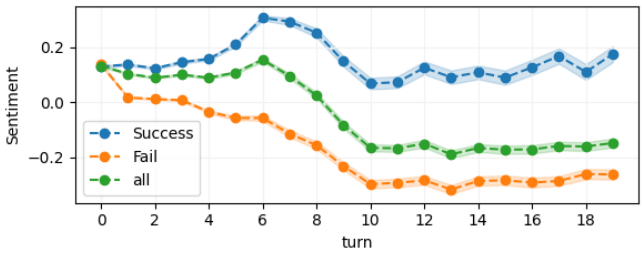

Figure 3: The average user sentiment in different turns.

6 CONCLUSION

We present EmoUS, a simulated user that generates emotional and thus more diverse output in task-oriented dialogues. It can be further configured by manipulating different weights for each emotion or different user personas. Our results show that EmoUS is useful to examine the impact of dialogue systems on the user's emotional state. Developing such probes is of particular importance with the increasing usage of dialogue systems and the rising ethical concerns of large language model chat-bots.

In future, the correlations between personas and emotions should be investigated, e.g. whether polite users show more satisfaction even though system responses are inadequate. Human evaluation should also be conducted to address the validity of our simulation. Furthermore, we plan to utilise EmoUS for the development of emotion-sensitive DSs.

REFERENCES

[1] Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019. Multi-domain Conversation Quality Evaluation via User Satisfaction Estimation. arXiv:1911.08567 [cs.LG]

[2] 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*. Association for Computational Linguistics, Brussels, Belgium, 5016–5026. https://doi.org/10.18653/v1/D18-1547

[3] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3734–3743. https://doi.org/10.18653/v1/N19-1374

[4] Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022. User Satisfaction Estimation with Sequential Dialogue Act Modeling in Goal-Oriented Conversational Systems. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2998–3008. https://doi.org/10.1145/3485447.3512020

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] W. Eckert, E. Levin, and R. Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. 80–87. https://doi.org/10.1109/ASRU.1997.658991

[7] Layla El Asri, Jing He, and Kaheer Suleman. 2016. A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems. *Interspeech 2016* (2016), 1151–1155.

[8] Klaus-Peter Engelbrecht, Florian Gödde, Felix Hartard, Hamed Ketabdar, and Sebastian Möller. 2009. Modeling User Satisfaction with Hidden Markov Models. In *Proceedings of the SIGDIAL 2009 Conference*. Association for Computational Linguistics, London, UK, 170–177. https://aclanthology.org/W09-3926

[9] Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4096–4113. https://aclanthology.org/2022.lrec-1.436

[10] Izzeddin Gür, Dilek Hakkani-Tür, Gokhan Tür, and Pararth Shah. 2018. User modeling for task oriented dialogues. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 900–906.

[11] Sunao Hara, Norihide Kitaoka, and Kazuya Takeda. 2010. Estimation Method of User Satisfaction Using N-gram-based Dialog History Model for Spoken Dialog System. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/579_Paper.pdf

[12] Ryuichiro Higashinaka, Yasuhiro Minami, Kohji Dohsaka, and Toyomi Meguro. 2010. Modeling User Satisfaction Transitions in Dialogues from Overall Ratings. In *Proceedings of the SIGDIAL 2010 Conference*. Association for Computational Linguistics, Tokyo, Japan, 18–27. https://aclanthology.org/W10-4304

[13] Simon Keizer, Milica Gašić, Filip Jurčíček, François Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Parameter estimation for agenda-based user simulation. In *Proceedings of the SIGDIAL 2010 Conference*. Association for Computational Linguistics, Tokyo, Japan, 116–123. https://aclanthology.org/W10-4323

[14] 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*. 2115–2119.

[15] Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 2018. Neural User Simulation for Corpus-based Policy Optimisation of Spoken Dialogue Systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Melbourne, Australia, 60–69. https://doi.org/10.18653/v1/W18-5007

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[17] Hsien-chin Lin, Christian Geishauser, Shutong Feng, Nurul Lubis, Carel van Niekerk, Michael Heck, and Milica Gasic. 2022. GenTUS: Simulating User Behaviour and Language in Task-oriented Dialogues with Generative Transformers. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, 270–282. https://aclanthology.org/2022.sigdial-1.28

[18] Hsien-chin Lin, Nurul Lubis, Songbo Hu, Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, and Milica Gasic. 2021. Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Singapore and Online, 445–456. https://aclanthology.org/2021.sigdial-1.47

[19] Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, Ayu Purwarianti, and Satoshi Nakamura. 2016. Emotion and its triggers in human spoken dialogue: Recognition and analysis. *Situated Dialog in Speech-Based Human-Computer Interaction* (2016), 103–110.

[20] François Mairesse and Marilyn Walker. 2006. Automatic Recognition of Personality in Conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, New York City, USA, 85–88. https://aclanthology.org/N06-2022

[21] Yanying Mao, Fei Cai, Yupu Guo, and Honghui Chen. 2022. Incorporating Emotion for Response Generation in Multi-Turn Dialogues. *Applied Intelligence* 52, 7 (may 2022), 7218–7229. https://doi.org/10.1007/s10489-021-02819-z

[22] Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571299

[23] Yan Pan, Mingyang Ma, Bernhard Pflugfelder, and Georg Groh. 2022. User Satisfaction Modeling with Domain Adaptation in Task-oriented Dialogue Systems. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, 630–636. https://aclanthology.org/2022.sigdial-1.59

[24] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* 7 (2019), 100943–100953.

[25] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 186–191. https://www.aclweb.org/anthology/W18-6319

[26] 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. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[27] Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*. 45–54.

[28] 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, Rochester, New York, 149–152. https://www.aclweb.org/anthology/N07-2038

[29] Jost Schatztnann, Matthew N Stuttle, Karl Weilhammer, and Steve Young. 2005. Effects of the user model on simulation-based learning of dialogue strategies. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 220–225.

[30] Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication* 74 (2015), 12–36.

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[32] Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. Emotionflow: Capture the Dialogue Level Emotion Transitions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8542–8546. https://doi.org/10.1109/ICASSP43922.2022.9746464

[33] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating Responses with a Specific Emotion in Dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3685–3695. https://doi.org/10.18653/v1/P19-1359

[34] 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. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2499–2506.

[35] Zhiwen Tang, Hrishikesh Kulkarni, and Grace Hui Yang. 2021. High-Quality Dialogue Diversification by Intermittent Short Extension Ensembles. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 1861–1872. https://doi.org/10.18653/v1/2021.findings-acl.163

[36] Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. Transferable Dialogue Systems and User Simulators. In *Proceedings of the 59th Annual Meeting*

of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 152–166.

[37] Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. Transferable Dialogue Systems and User Simulators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 152–166. https://doi.org/10.18653/v1/2021.acl-long.13

[38] Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. 2022. A Unified Dialogue User Simulator for Few-shot Data Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3788–3799. https://aclanthology.org/2022.findings-emnlp.277

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[40] Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, et al. 2022. ConvLab-3: A Flexible Dialogue System Toolkit Based on a Unified Data Format. *arXiv preprint arXiv:2211.17148* (2022).

[41] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1097–1100.