Evaluation Metrics for NLG and TTS in Task-Oriented Dialog

PhD. Thesis Proposal

Ondřej Plátek
Faculty of Mathematics and Physics, Charles University,
Malostranské Náměstí 25
118 00 Prague, Czech Republic

Abstract

Our thesis proposal explores the evaluation of Task-oriented Dialogue (ToD) systems and Text-to-Speech Synthesis (TTS) using automatic metrics. Our aim is to eventually integrate these metrics into the evaluation process of Spoken Dialogue Systems. We built the TTS and open-ended dialog metrics based on Self-Supervised Learning (SSL) Models. We further enhance the models by contrastive losses, yielding improved ranking and regression performance compared to human-annotated preferences. We established TTS and Chat trainable metrics closely approaching the State-of-the-Art (SOTA). We are also working on publishing our research for Data-to-Text (D2T) and ToD Natural Language Generation (NLG) factuality trainable metrics. Although all the metrics are close to SOTA, they exhibit similar limitations.

Our focus lies on addressing these limitations: (1) The ambiguous generalization of new data (domains) leads to occasional catastrophic failures, and (2) the unreliable performance at the utterance level despite satisfactory ranking system performance on the dataset level. Additionally, we address the broader concern of (3) the lack of interpretability in neural network metrics.

To address these issues, we propose a series of experiments aimed at resolving the identified limitations and enhancing the evaluation process for D2T and ToD NLG. We hope our research on automatic metrics for NLG and TTS will contribute to developing more reliable NLG methods and Spoken Dialogue Systems.

1 Introduction

In recent years, there has been a surge of interest in the practical applications of Natural Language Generation (NLG) and Text-to-Speech (TTS) systems, resulting in a growing demand for efficient and reliable evaluation methods. We will first provide a background, which will motivate our goals, and finally, we will introduce the structure of this proposal.

1.1 Background

The ability to rapidly assess NLG and TTS systems cost-effectively has become a pressing need for researchers and industry practitioners. At the same time, the users demand applicability across a diverse set of domains or at least an easy-to-use adaptation method to new data.

The traditional approach of human evaluation, involving the subjective assessment of dialogue response appropriateness, the factuality of data-to-text generation, and the naturalness of speech synthesis, remains indispensable. However, it is inherently challenging to conduct, time-consuming, and often difficult to reproduce human evaluation consistently across different contexts.

Automatic evaluation metrics are inherently reproducible, cheap to use, fast to compute, and typically more consistent than human annotators.

1.2 Goals

Therefore, we are interested in **designing and improving neural metrics**, which have the following properties: **(1) High correlation with human judgment**; In contrast to many established non-trainable metrics that do not correlate well with human judgments. Notable examples of still used metrics are BLEU (Papineni et al., 2002) for NLG (Lubis et al., 2022; Lowe et al., 2017a) and MCD (Kominek et al., 2008) for TTS evaluation. Neural Network(NN) based metrics are excellent in discovering the correlation between the evaluated data and its score pairs – especially if they have seen enough in-domain training pairs. The NN-based metrics are successfully

used for evaluation in machine translation (MT) (Rei et al., 2022a; Kocmi and Federmann, 2023), summarization (Krubiński and Pecina, 2022), chat (Zheng et al., 2023; Plátek et al., 2023) or TTS (Saeki et al., 2022; Huang et al., 2022a; Plátek and Dušek, 2023)

(2) Predicting interpretable structured outputs is an interesting and underexplored research area. There are several established evaluation methodologies based on strictly structured labels; dialogue state items annotation(Williams et al., 2013), or fine-grained machine translation labels MQM (Freitag et al., 2021), aligning MT source and target sentence (Dou and Neubig, 2021), and automatic speech recognition (ASR) used as a proxy for Intelligibility of TTS systems (Spille et al., 2018). However, obtaining such labeling is tedious and expensive on new datasets, but the foundational models can predict such labels, especially if transfer learning using few-shot examples is possible.(Peng et al., 2021; Radford et al.; Ouyang et al., 2022) Predicting tying a metric with structured labels allow better interpretability of the metrics because the well-structured labels are typically better than the global quality of interests. The better-defined labels are easier to annotate and can be easily inspected. E.g., in Task-oriented Dialogue (TOD) systems, we can either ask annotators to rate appropriateness or annotate the dialogue state with dialogue state items (Žilka et al., 2013) and compute based on the dialogue state if the user's goals were achieved in the conversation. Note that it is possible to build neural models for predicting dialogue state items using DST (Wu et al., 2020; Plátek et al., 2016) and at the same time predicting the appropriateness directly (Plátek et al., 2023). We hypothesize that DST annotations will have a higher inter-annotator agreement.

(3) Reproducible metrics. We suggest that automatic metrics (as released software) should support easy reproduction of any experiments for which the metrics were used previously. We stress that properly versioning the metrics and their models is essential for reproducibility. We aim to provide open-source implementations with publicly available models for every trainable metric we develop. We try to follow software best practices and semantic versioning for each of our metrics and its models.

We want to focus on solving the following known drawbacks of current automatic metrics:

(4) Unclear generalization to new data or domain; It is often the case that NN models, including automatic metrics, tend to fail catastrophically in new domains.

Current automatic metrics should improve their **(5) utterance level predictions**. Their performance in ranking systems on the dataset level is already close to or better than human-level performance on many tasks (Huang et al., 2022a; Rei et al., 2022a; Kocmi and Federmann, 2023), but utterance-level still lags behind.

We also plan to explore **(6) uncertainty estimates techniques** especially on the utterance level. Using uncertainty estimates should allow us to detect catastrophic failures for particular utterances.

1.3 Contents of this Proposal

We introduced our research's motivation and main goals in the previous section. The following sections introduce the tasks which we would like to evaluate:

- Data-to-text (D2T) Natural Language Generation (NLG) task in Section 2.1.
- NLG for Task-oriented-Dialogue(ToD) in Section 2.2.
- Text-to-Speech Synthesis (TTS) used for isolated prompts and for dialogue context. It is described in Section 3.1 and 3.1 respectively.

Next, we introduce our research conducted so far in Section 3:

- Section 3.1 present our work (Plátek and Dušek, 2023) for Speech Synthesis evaluation in the VoiceMos challenge dataset (Huang et al., 2022a). According to the official leaderboard, our submission ranked fourth on the Main Track with the value of 0.935 in the system Spearman Correlation Coefficient benchmark. Similarly, we finished third with a value of 0.937 for the OOD track.
- Section 3.2 describes our submission (Plátek et al., 2023) to DSTC11 Track 4 – ChatEval challenge (Rodríguez-Cantelar et al., 2023a). We finished second out of six teams.[1]
- In Section 3.3, we reflect our lessons learned from reproducing a human MT evaluation in work (Vamvas and Sennrich, 2022), but also from our own research.

[1] See our *team6_t2t_s2* results at dstc11.

Finally, in Section 4 and 5, we outline our immediate and long-term plans for our research.

2 Building Automatic Metrics for D2T, ToD, Chat and TTS systems

We aim to improve methods for evaluating the data-to-text, task-oriented dialogue, chat, and text-to-speech systems. We plan to use transformer-based (Vaswani et al., 2017) models pretrained using self-supervised learning (Radford et al.; Touvron et al., 2023; Ouyang et al., 2022) for our neural based metrics since they offer well-performing representation trained without any labels. For textual domains, encoder-decoder models similar to T5 (Raffel et al., 2020) or decoder-only models like GPT2 (Radford et al.) are mostly used. For speech input, the self-supervised transformer encoder models with clustering objectives are used Wav2vec 2.0 (Baevski et al., 2020) Wav2vec 2.0 and HuBERT (Hsu et al., 2021). However, we will not focus on the details of the models but rather on their evaluation for a given task and their typical errors and flaws, which are often valid across the tasks.

Although each task requires a dedicated pretrained model for its best performance, many possibilities exist for reusing developed methods and algorithms across the tasks. As we research evaluation and benchmarking methods, we realize the advantage of high-quality test datasets which discriminate the evaluated systems and show the problems of the low-performing systems. For each of our outlined goals in Section 1.2, we hope to select the task where the problem is benchmarked easiest and hopefully improved with the least effort.

In this section, we will briefly introduce the tasks which we aim to evaluate. From now on, we will consistently use the term *benchmarking* for the evaluation of the performance of the automatic metrics and *evaluation* for the evaluation of the individual tasks.

2.1 Data-to-Text (D2T) NLG

The D2T NLG systems convert structured data to text in natural language, capturing its intended meaning. The challenge in the D2T NLG is to be understandable, fluent, and factually correct, not only on the sentence level but also on larger segments of texts based on the intended applications. There are similar tasks like image description (Karpathy and Fei-Fei) and text summarization (El-Kassas et al., 2021). However, they do not require such a degree of consistency of style for a given domain, and their input modality is not so sparse.

We work with WebNLG dataset (Gardent et al., 2017) and its extended variant (Castro Ferreira et al., 2018), which contains RDF triples and their corresponding text descriptions from DBPedia (Auer et al., 2007). The triple in the dataset represents a *subject*, a *predicate*, and an *object*. See Figure 1.

In Section refFactuality we aim to evaluate the generated text description by finding corresponding RDF triples alignment per sentence

The Rotowire dataset (Wiseman et al., 2017) defines the D2T NLG task as generating relatively long summaries from a tabular data where only subset of data should be summarized. The model should be able to verbalize several summary statistics from multiple entries or highlight the most important facts which make the task very challenging but also hard to evaluate. See Figure 1.

In industry, NLG is still often realized with a set of handcrafted templates which are selected heuristically (Rudnicky et al., 1999). The variability of the generated utterances is limited, and the scalability is poor. Therefore corpus-based methods (Oh and Rudnicky, 2000; Mairesse and Young, 2014) modeled by neural networks are used in academia since 2015(Wen et al., 2015, 2016). The black-box-like models based on neural networks suffer from hallucinations and omissions – adding and removing facts not grounded in the context. The boom of LLMs trained for chat (Ouyang et al., 2022; Touvron et al., 2023) brought more fluent and robust models to NLG, which suffer less from hallucinations. However, the problem is only reduced and not solved. It remains an open question how much hallucinations are acceptable for different applications. We focus on evaluating factuality which we describe in Section 4.1.

2.2 NLG in Task-oriented-Dialogue

The Table 1 shows an example of a task-oriented dialogue with annotation of dialogue state in the form of dialogue acts.

DST is example of label InteNsive tasks useful for ToD Evaluation.

Dialogue State Tracking (DST) Dialogue state is used to keep track of the dialogue history, approximating the meaning of history with a predefined discrete set of dialogue acts. Dialogue State Trackers update the state with correct values after each

reference The Washington Wizards defeated the Charlotte Hornets, 107 – 93, at Capital One Arena on Saturday. The Wizards (42 – 34) and the Hornets (34 – 43) engaged in quite a battle in this one, as they went back and forth in the first half before Washington pulled away for good with a strong scoring output in the third quarter. Washington jumped ahead 30 – 27 in the first, but Charlotte won the second quarter 24 – 20 so they ended with a one – point lead at the break (51 – 50). The big difference in this game came in the third quarter, as the Wizards established a huge 37 – 23 supremacy over that span that made it 87 – 74 after three and they never looked back afterwards. Washington won the final quarter 20 – 19, and ended cruising their way to celebrate with a much – needed win in John Wall's return. Otto Porter Jr. was the best player for the Wizards in this game, as he finished with 26 points, 11 rebounds, two assists and two blocks. Bradley Beal chipped in with 22 points and three assists. John Wall looked a bit rusty at times, but for the most part he looked every bit of the All-Star he is and delivered a double – double of 15 points, 14 assists and three boards. Mike Scott chipped in with 15 points off the bench. Dwight Howard led the way for the Hornets with 22 points and 13 rebounds, but he couldn't do everything on his own. Michael Kidd-Gilchrist and Marvin Williams each added 10 points, while Guillermo Hernangomez added 11 points off the bench. Malik Monk had 17 points off the bench as well, and Kemba Walker struggled after finishing with seven points on 3 – 9 from the field.

data

team	entity	period	AST	BLK	DOUBLE	DREB	FG3A	FG3M	FG3_PCT	FGA	FGM	FG_PCT	FTA	FTM	FT_PCT	MIN	OREB	PF	PTS	STL	TOV	TREB	+/-
home	all	game	30	4		33	39	18	46	88	40	45	12	9	75	4	11	23	107	11	12	44	
		H1	95	10		66	810	44	5	2424	128	5	30	20	67	6060	53		3020	52	51	119	
		H2	115	03		1011	129	73	57	2416	146	6	36	25	69	6060	21		3720	13	06	1032	
		Q1	9	1		6	8	4	50	24	12	50	3	2	67	60	5		30	5	5	11	
		Q2	5	0		6	10	4	40	24	8	33	0	0	0	60	3		20	2	1	9	
		Q3	11	0		10	12	7	58	24	14	58	3	2	67	60	2		37	1	0	12	
		Q4	5	3		11	9	3	33	16	6	38	6	5	83	60	1		20	3	6	12	
		OT	0	0		0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	
	Otto Porter	game	30	4		33	39	18	46	88	40	45	12	9	75	4	11	23	107	11	12	44	
		H1	95	10		66	810	44	5	2424	128	5	30	20	67	6060	53		3020	52	51	119	
		H2	115	03		1011	129	73	57	2416	146	6	36	25	69	6060	21		3720	13	06	1032	
		Q1	9	1		6	8	4	50	24	12	50	3	2	67	60	5		30	5	5	11	
		Q2	5	0		6	10	4	40	24	8	33	0	0	0	60	3		20	2	1	9	
		Q3	11	0		10	12	7	58	24	14	58	3	2	67	60	2		37	1	0	12	
		Q4	5	3		11	9	3	33	16	6	38	6	5	83	60	1		20	3	6	12	
		OT	0	0		0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	
	Bradley Beal	game	30	4		33	39	18	46	88	40	45	12	9	75	4	11	23	107	11	12	44	
		H1	95	10		66	810	44	5	2424	128	5	30	20	67	6060	53		3020	52	51	119	
		H2	115	03		1011	129	73	57	2416	146	6	36	25	69	6060	21		3720	13	06	1032	
		Q1	9	1		6	8	4	50	24	12	50	3	2	67	60	5		30	5	5	11	
		Q2	5	0		6	10	4	40	24	8	33	0	0	0	60	3		20	2	1	9	
		Q3	11	0		10	12	7	58	24	14	58	3	2	67	60	2		37	1	0	12	
		Q4	5	3		11	9	3	33	16	6	38	6	5	83	60	1		20	3	6	12	
		OT	0	0		0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	
	John Wall	game	30	4		33	39	18	46	88	40	45	12	9	75	4	11	23	107	11	12	44	

reference Acta Mathematica Hungarica, or Acta Math. Hungar., has a LCCN number of 83646315 and a ISSN number of 1588-2632.

data

subject	predicate	object
Acta Mathematica Hungarica	LCCN number	83646315
Acta Mathematica Hungarica	abbreviation	Acta Math. Hungar.
Acta Mathematica Hungarica	issn number	1588-2632

Figure 1: The Tabgenie toolkit unifies the dataset format and displays the dataset in tabular view. (Kasner et al., 2023). The top image shows example from the Rotowire dataset (Wiseman et al., 2017), and the bottom image shows an example from the WebNLG dataset (Gardent et al., 2017).

USER:	*I would like a cheap restaurant.*	inform(price=cheap)
SYSTEM:	*Golden plate is cheap in city-center.*	inform(name=Golden plate)
USER:	*What is the cuisine?*	request(cuisine)
SYSTEM:	*They serve chinese food.*	inform(cuisine=chinese)
USER:	*Sounds good. Bye!*	goodbye()
SYSTEM:	*Have a great day.*	goodbye()

Table 1: Example of task-oriented dialogue in the restaurant reservation domain. Utterance representations as dialogue acts are depicted on the right. Intents are highlighted in orange, slot names in blue and respective values in green. Note the hallucination in the first system response in the red font, which is not grounded in dialogue act items for the first turn.

turn. Žilka et al. (2013) provides a comparison of different data-driven models for dialogue state tracking, including rule-based methods. Neural networks successfully predict the state values but also can estimate the state value distributions (Mrkšić et al., 2016; Zhong et al., 2018) The probabilistic distributions are useful for downstream tasks, e.g., dialogue policy learning. However, the dialogue state labels can also be useful for evaluating the system response or clustering the conversations.

Dialogue Policy (DP) The main interest of our research in ToD is the evaluation of the dialogue policy (DP). Its responsibility is to decide which action (which API to call or how what to say) the system should take at each turn. In other words, the DP does the high-level planning of the dialogue. The policy decision can thus be framed as a classification task (Gašić and Young, 2013). Learning the policy just from the offline data might not produce robust policy due to low variability in the data. Therefore, many works model the dialogue as a partially observable Markov decision process (Gašić et al., 2010; Thomson and Young, 2010). Reinforcement learning techniques are then applied to learn the policy and incorporate human feedback (Peng et al., 2017; Su et al., 2016).

Natural Language Generation (NLG) Natural Language Generation in ToD is very simple if the system does not perform DP and NLG at the same time. However, the trend is to use end-to-end models which (Kulhánek et al., 2021; Yang et al., 2021) are responsible for generating the reply based on the dialogue history and optional API query results for a given dialogue turn. Such methods model turn meaning implicitly since they generate the turn word by word.

Datasets with a large number of ToD conversations have been non-existent up to very recently. The release of large datasets for both text (Zhang et al., 2023) and spoken ToD (Si et al., 2023) favors using large models and encourages evaluation of multiple-domains systems. The de facto standard still is the MultiWOZ dataset (Budzianowski et al., 2018) which is a multi-domain dataset. See the example dialogue containing just one domain in Table 1. The reason for the lack of ToD datasets is that annotating user goals and dialogue states is expensive. In addition, one needs to design carefully the database and its API so the task is not trivial but also understandable and easy-to-use for researchers.

2.3 *Chat* – aka Open-ended Dialogue

We define *chat* as an open-ended dialogue (Rodríguez-Cantelar et al., 2023a) where the user and the system can talk about any topic. The role of the system is not only to inform the user but also to be empathetic (Rashkin et al., 2019). The instruction-tuned large language models (Wang et al., 2022; Chia et al., 2023) with human feedback (Ouyang et al., 2022; Touvron et al., 2023) improved on one side following dynamical (open-ended) users' goals. On the other side, equally important was training with human feedback to avoid toxic responses (Ouyang et al., 2022) and to be empathetic (Rashkin et al., 2019).

2.4 Synthesized Speech in Isolated Prompts and Dialogue

Speech synthesis research (Josef Psutka et al., 2006; Taylor, 2009) has recently focused mainly on so-called parametric models (Zen et al., 2009; Shen et al., 2018). Since 2016, neural models dominate parametric synthesis and have outperformed all other approaches to speech synthesis in terms of the naturalness of synthesized speech. Currently, the best quality read synthesis is represented by VITS (Kim et al., 2021), Tacotron 2 (Shen et al., 2021) models, which achieve top rankings on popular read datasets LJ Speech (Keith Ito and Linda Johnson, 2017) and VCTK (Veaux, Christophe et al.,

2017).

So far, publications dealing with conversational speech synthesis are rather scarce. The closest to conversational synthesis are the following models that focus on controlling prosody:

- The RyanSpeech dataset (Zandie et al., 2021) dataset and model were recorded using a single high-quality voice. It uses utterances from dialogues, but it does not contain whole dialogues.
- MixerTTS (Tatanov et al., 2021) that includes language model conditioning and ablation analysis on the LJ Speech dataset.
- VALL-E (Wang et al., 2023) introduces a language modeling approach to TTS, and the model can be prompted by audio and text.

3 Our Work So Far

In our recent works (Plátek and Dušek, 2023; Plátek et al., 2023), we have focused on the evaluation of synthesized speech and responses for chat – open domain dialogues. The evaluation task are similar in the sense that up to recently both task solely relied on human evaluation because of the complexity of the tasks and, thus hard to define a single objective. The neural metrics which we investigated are trained to approximate the human evaluation but have the following interesting properties: (1) are relatively robust to new data and systems, (2) are reproducible and consistent, and (3) fast and cheap.

3.1 Automatically Evaluating Short Speech Prompts

The work (Plátek and Dušek, 2023) presents the MooseNet metric, which predicts the Mean Opinion Score (MOS) from a single utterance of synthesized speech. We present the MooseNet metric, which predicts the Mean Opinion Score (MOS) from a single utterance of synthesized speech. Using MOS from recruited listeners is a well-established standard for evaluating text-to-speech (TTS) and voice conversion (VC) systems (, ITU-T), and MOS prediction metrics (Huang et al., 2022b) are a way to automate this process. The organizers of the 2022 VoiceMOS Challenge (Huang et al., 2022a) released a large dataset with MOS annotations for TTS and VC systems' outputs (called BVCC), so that MOS prediction metrics can be trained in a supervised manner. One of the aims of the VoiceMOS challenge was investigating the use of self-supervised learning (SSL) speech models (Baevski et al., 2020; Babu et al., 2022) finetuned for the MOS prediction task. Using SSL models requires fewer annotated utterances than training NN models from scratch, but fine-tuning on a limited number of examples may lead to overfitting to the audio channel and speech properties of the training data, hurting performance on non-matching examples. To investigate this, the VoiceMOS challenge included two tracks: The main track with 4,974 training utterances and the Out-of-Domain (OOD) training set with only 136 utterances, intended to evaluate the applicability of MOS predictors trained on the main track to a new domain.

While the VoiceMOS main track data proved to be large enough for building a robust SSL-based MOS predictor and SSL-based models are the current state of the art on the task (Cooper et al., 2022; Saeki et al., 2022; Huang et al., 2022b; Tseng et al., 2022), it is still unclear how many MOS annotated utterances are really needed for finetuning an SSL model. By experimenting on both VoiceMOS main and ODD track datasets, we investigate if pretraining on a larger dataset is crucial for fine-tuning the MOS predictor to a new small dataset and how much data is needed for it. We show in a simple ablation study that even with 5% training data, SSL fine-tuning outperforms the previous non-SSL state-of-the-art LDNet model (Huang et al., 2022b). In addition, we present an even more effective low-resource alternative approach to the traditional finetuning paradigm of SSL models by reframing the MOS-prediction regression as classification and introducing Probabilistic Linear Discriminative Analysis (PLDA). In contrast to previous systems, PLDA performs very well even for a few hundred annotated utterances. Furthermore, its projections are computed very fast on a single CPU and thus require minimal resources for training and inference. Importantly, PLDA can be easily combined with existing neural network models.

Our contributions are the following:

(1) We introduce a new SSL-based neural network MOS prediction model, dubbed MooseNet, which is based on models of Cooper et al. (Cooper et al., 2022) and Saeki et al. (Saeki et al., 2022) and further improves model training, optimizing hyperparameters and introducing multi-task learning. The MooseNet neural network reaches near state-of-the-art performance on the VoiceMOS data. See the architecture in Figure 3.

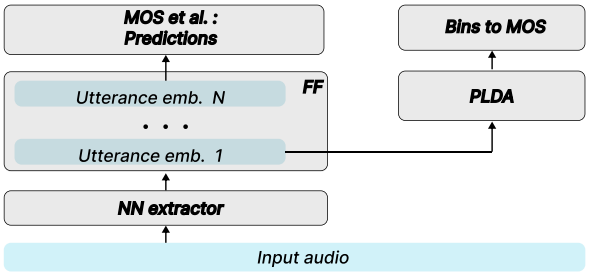

Figure 2: PLDA can use any layer after global pooling as utterance level embedding as its features.

(2) We introduce PLDA as a convenient method for adapting pre-trained models to downstream tasks. We demonstrate the use of PLDA on several variants of SSL models (Baevski et al., 2020; Babu et al., 2022). See Figure 2.

(3) In ablation studies on VoiceMOS data, we investigate the performance of PLDA and several strong neural baselines based on the amount of available data. We show that PLDA consistently improves SSL models, matching state-of-the-art on VoiceMOS. Models without finetuning to the MOS prediction task as well as specifically fine-tuned models, benefit from using PLDA.

(4) Our best results of 0.929+-0.005 Spearman Correlation Coefficient for the main track and 0.956+0.011 for the Out-of-Domain track are competitive with SOTA models.

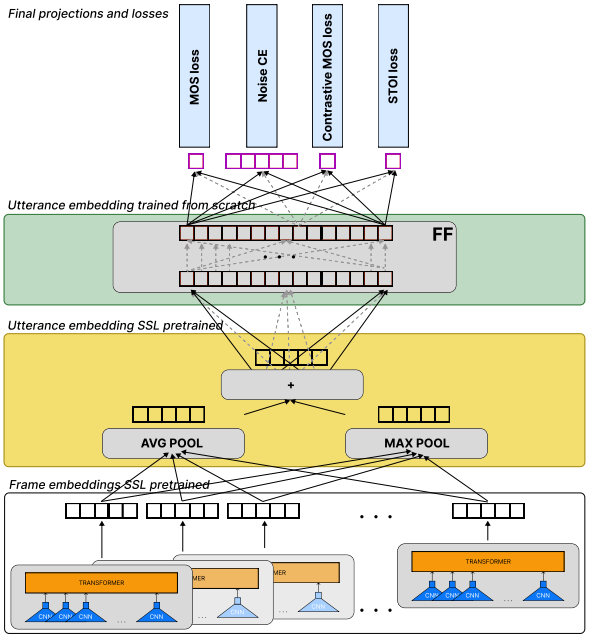

Figure 3: MooseNet architecture is based on pre-trained SSL models. Frame-level embeddings are transformed to utterance level by global pooling. FF layers and final projections are the only parameters trained from scratch.

Towards Evaluating Speech in Dialogue Despite our belief that models similar to works (Tatanov et al., 2021; Wang et al., 2023) are capable of modeling dialogue context well and thus produce matching prosody, we are unaware of any work with the notable exception of (Zandie et al., 2021), which evaluates the convenience of prosody and dialogue context.

3.2 Three Ways of Using Large Language Models to Evaluate Chat

Our work (Plátek et al., 2023) describes the systems submitted by *team6* for ChatEval, the DSTC 11 Track 4 competition aimed at evaluating open-domain chat.[2] We participated in Task 2, which focuses on evaluating multiple criteria on the level of individual dialogue turns. The task of evaluating responses in a chat is challenging because it requires an understanding of the interlocutor's roles (pragmatics), the conversation's context, and the response's meaning (semantics). See Table 2. At the same time, the conversations are often ungrammatical (Rodríguez-Cantelar et al., 2023b) and vary in style (Zhang et al., 2018). The commonly used metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), or BERTScore (Zhang et al., 2019), are based on comparison to human references and thus correlate poorly with human judgments on the turn-level, as they penalize many correct responses for a given chat context (Zhao et al., 2017). Previous reference-less metrics based on neural networks and language models still do not reach sufficient correlations with human judgements (Zhang et al., 2020; Lowe et al., 2017b).

In our work, we followed up on the recent development of pretrained Large Language Models (LLMs) with instruction finetuning (Brown et al., 2020), which have been found to be capable evaluators in machine translation, summarization as well as dialogue (Kocmi and Federmann, 2023). Therefore, we applied LLMs and specific prompting to elicit ratings for the multiple qualities evaluated in DSTC11 Track 4 Task 2: appropriateness, content richness, grammatical correctness, and relevance. We present three different systems used for our three submissions, all of which are based on LLMs and few-shot prompting: (1) We evaluate a straightforward approach with manually designed fixed

[2]Results & task description at chateval.org/dstc11. Our experimental code is available at github.com/oplatek/chateval-llm.

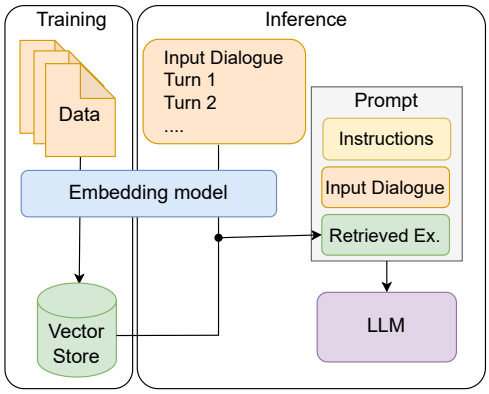

Figure 4: The architecture of the vector store approach with LLM. During training, we construct the vector store from embedded annotated dialogues. At inference time, the input dialogue is embedded, and most similar examples from the vector store are retrieved to be included in the prompt.

prompts for off-the-shelf open LLMs checkpoints. (2) We train a simple feed-forward regression neural network (FNN) on top of frozen LLM embeddings to predict the turn-level metrics scores. (3) We used the ChatGPT API and few-shot examples retrieved dynamically from the development set to improve the prompting performance. As no data annotated with the target metrics were available for the challenge, we heuristically mapped existing annotations from the development set to the target metrics, and we manually annotated a small rehearsal dataset for hyperparameter search.

Based on the human annotations released after the challenge finished, our *team6* achieved second place thanks to our third method. In an ablation study conducted after the challenge, our method 3 – dynamically prompted chatGPT with few-shot examples achieved a Spearman Correlation Coefficient (SCC) for ranking systems of 0.6136. The Llama 2 7B Chat Human-Feedback trained model (Touvron et al., 2023) – the best open model we experimented with – achieved system SCC 0.3914. This approach showed that LLM prompting is a viable option for prototyping chat evaluation.

3.3 Lessons Learned from ReproHum Project and our Experiments

We participated in ReproHum project (Belz et al., 2023), which aimed at reproducing human evaluation in NLP papers. We quickly realized that errors, misconceptions, and lack of best practices in human evaluation are prevalent in the field. We see it as a natural process when exploring new ideas and methods since the best practices are not yet established. However, we realize that to develop a widely used tool for evaluation, one of its core values should be reproducibility. We also realized that the more any work is relied upon, the more it is tested and studied, and the fewer errors and design flaws it has.

4 Immediate Plans

During our work on *chat* evaluation, we saw that hallucinations are a common problem even for chat data. The second example of *chat* conversation in Table 2 shows that the system contradicts even itself so itors last reply is not well grounded in the conversation history in this case its first response. In Section 4.1, we will describe our plan for benchmarking the factualness of D2T NLG and NLG for ToD on MultiWOZ, where we have alignment annotations for the mentioned facts.

4.1 Evaluating Factuality in D2T and in ToD Responses

Our proposed experiments are straightforward but stand on several critical observations. We propose (1) to build a hallucination detection system based on pretrained LLMs (Touvron et al., 2023; Ouyang et al., 2022) and (2) benchmark the hallucination detector on datasets using annotations - alignments for each factual entity.

Regarding (1), we plan to build the datasets from the Enriched WebNLG (Castro Ferreira et al., 2018) and MultiWOZ (Budzianowski et al., 2018) datasets. Both datasets contain alignments for the structured input data and the generated text. The datasets were designed to train factual systems so they contain minimal factual errors. We will introduce factual errors by perturbing the structured input data. Adding facts to input creates omissions assuming the gold natural language text is kept intact. Similarly, carefully deleting facts from input creates hallucinations.

Regarding (2), our first proposal for the hallucination detection system is to use pretrained LLMs, which are known to be effective in performing (Ouyang et al., 2022) string operations. We simply intend to prompt LLMs to do Natural Language Inference and ask which RDF triples in WebNLG or which triples representing dialogue state item are used in corresponding the natural language text. We would like to explore different adaptation strategies

Dialogue Turns	Appr	Rich.	Gram.	Rel
My boss gave me a 10 raise just last month And it was a nice surprise	5	5	5	-
It's great and he might think you're doing a great job	5	5	5	5
We have always been very nice He has always been very supportive of me	4	5	5	5
That's a good thing	4	3	5	4
do you have any pets?	5	4	3	-
I am retired so I love to travel so pets would slow me down	4	4	3	4
I understand that my idea of traveling is a hot hot bubble bath	3	4	2	2
Yes I have dogs and cats I like to take them with me on trips	2	4	2	2

Table 2: Two examples of complete conversations from the DSTC 11 ChatEval challenge set are annotated with turn-level metrics: appropriateness, content richness, grammatical correctness, and relevance. The context for each turn are the previous turns (lines) in the conversation. The second conversation at the bottom of the table shows an inappropriate response in the last turn because the last response contradicts previous responses of the system.

of LLMs for the task, including QLoRa (Dettmers et al., 2023), or Chain-of-Though prompting (Wei et al., 2023)

4.2 Comparison of Prosody in Dialogue Context

We are interested in evaluating synthesized speech for ToD. We hypothesize that the prosody of synthesized speech is important for every user, and the users notice the inconvenient use of prosody. However, non-existing datasets are available for benchmarking prosody convenience in ToD or read speech for D2T NLG.

We intend to build a dataset for benchmarking prosody in ToD. However, we need to verify that untrained users can notice the difference in prosody. We plan to use Prolific crowdsourcing service to perform a human study to verify our hypothesis for the English SpokenWoz dataset (Si et al., 2023).

Using crowdsource workers is important for several reasons:

- We can target native or non-native speakers of English.
- The knowledge of the task does not bias them.
- The annotation of the task has the potential to scale for creating a larger dataset beyond this initial experiment.

The high-level overview of the experiment is as follows:

1. We manually select 100 conversations from the SpokenWoz task-oriented dataset recorded in Wizard-of-Oz style between *agent* and an *user*. We will select them manually based on our perceived importance of prosody in the conversation.
2. We will synthesize the system replies from the conversations using a state-of-the-art TTS (Wang et al., 2023) system without access to the dialogue context enforcing uninformed reading style.
3. We will also convert the original human voice to the same speaker using (Wang et al., 2023), obtaining a pair of utterances different only in prosody.
4. We will randomly replace one or more system prompts in the original conversation with the prepared synthesized speech, randomly choosing the read or conversational prosody.
5. The annotators will be asked to rate inappropriate responses based on the audio.
6. Finally, we will evaluate the overall accuracy and inter-annotator agreement of the annotators.

5 Long-term Projects

In this section, we present research topics that require substantial effort.

5.1 Uncertainty Estimates for Trainable Metrics

During our work on MooseNet (Plátek and Dušek, 2023), we realized that for some utterances, the model fails catastrophically, and the researcher would not notice unless she started listening to recordings. Following the work of (Lakshminarayanan et al., 2017), we attempted to use an ensemble of models of the same architecture and use their variance for uncertainty predictions. We measure the correlation of the uncertainty estimates with the model's error rate on utterance level. In our

informal experiments, we benchmarked the performance using the Spearman correlation coefficient and obtained a disappointing performance of 0.2.

We plan to similarly benchmark all our baselines (plain & conversational TTS, factuality for ToD, and appropriateness for chat) against several proposed techniques for uncertainty estimates.

- Deep ensemble based method (Lakshminarayanan et al., 2017).
- Methods based on dropout (Dawalatabad et al., 2022; Rei et al., 2022b).
- Identifying weaknesses in the evaluation metric model by using the task model influence to follow the metric (Amrhein and Sennrich, 2022).

An obvious addition is implementing a toy univariate regression and classification example to study the techniques on well-understood and limited tasks. We welcome any suggestions for the uncertainty estimation techniques since we are unaware of any silver bullet.

5.2 Time to return to User Simulator Evaluation for ToD?

The widely used ToD dialogue policy corpus-based evaluation on existing conversations suffers from a mismatch between the actual model policy for the current system actions and the other system's actions from the corpus. Such evaluation is also able to cover a wide range of user behaviors. Following the work of (Lubis et al., 2022), we will call the problem of evaluating the success of a dialogue system using corpus-based methods a ***Context mismatch***.

The work of (Lubis et al., 2022) shows that standard corpus-based metrics (Nekvinda and Dušek, 2021; Budzianowski et al., 2018) correlate poorly with human evaluation, but the evaluation using the user simulator shows a strong correlation with human judgment on MultiWOZ.[3]

The work of (Cheng et al., 2022) shows that the user simulator (US) jointly optimized with dialogue policy (DP) allowed the model to achieve almost 98% success rate on MultiWOZ 2.0, effectively training the US to be a cooperative and adapted user for given dialogue system. The US and dialogue model were finetuned using reinforcement learning to receive a positive reward when the US and the dialogue policy model achieved the user's goal. The authors consequently concluded that MultiWOZ 2.0 is too easy a dataset for dialogue policy evaluation. We respectfully disagree with this conclusion; human users can hardly optimize their behavior to a newly deployed dialogue system, and frequently the user is not cooperative at all. Such US does not model large variability of human users, which in reality range from cooperative to adversarial in one dimension, from talkative to laser-focused to achieve their user goal, from native English speakers to hard-to-understand foreigners, etc. However, we agree that it is useful to show that the goals in the MultiWOZ dataset are solvable with current models with the US simulating a user that cooperates and knows the deployed system.

In sharp contrast, the thorough work of (Liu et al., 2022) shows that training and evaluating dialogue policy (and NLG) systems with multiple user simulators is beneficial, and there is room for improvement even on the simplistic MultiWOZ dataset. Their ablation study introduces "Out-of-Domain" (OOD) evaluation using US not used during training. The OOD US scored the DP trained with the single US, and scored the DP with the US used during training. The OOD evaluation suffers from 3% to 43% degradation, but if the dialogue policy is trained using multiple simulators, the degradation is only from 1% to 8%.[4] **More importantly, the authors show that multiple-user simulators correlate well with human evaluation.**[5]

We see the US evaluation as a principal way how to evaluate ToD system. However, we see the following problems with using the user simulator to evaluate dialogue policy and NLG models.

- Up until recently, the user simulators were time-consuming to develop.
- The currently available user simulators are costly to train and run at inference.
- The user simulators are not controllable and can not be used to evaluate the dialogue policy on a specific user type or specific problems.
- The evaluation is not directly comparable for any two dialogue policy models on individual conversations. It is unclear how to compare

[3]The absolute value of Fless' Kappa for the best corpus-based method Success Match is 0.623 versus 0.991 for the success measure obtained using ConvLab 2 user simulator.

[4]For details, see Table 2.

[5]See the *Avg.* column in Table 2 and the human evaluation of Table 3.

two dialogue histories turn by turn even for the same user goal, except for the final success rate metrics.

We propose to solve the following by

- Using pretrained decoder-only models simular to (Cheng et al., 2022) to build a user simulator.
- Leveraging Parameter Efficient Training using QLoRa (Dettmers et al., 2023) to train multiple diverse user simulators, which can run in inference mode with minimal overhead to a single model.
- Diversifying the user simulators by focusing on a particular style, dataset, etc. We plan to explore combining the properties using techniques such as Elastic Weight Removal (Daheim et al., 2023).

Comparing multiple dialogue policies models and always strictly compare the models on discovered errors. **On which conversations is the model strictly better? Is the model A better than the model B for the given goal? Which conversations are hard for the dialogue policy models?** The questions were partially inspired by the Chatbot Arena (Zheng et al., 2023) and (Liu et al., 2022), which used multiple user simulators. We propose to evaluate multiple dialogue policy models $DP_i; i \in 1, ..., M$ using multiple user simulators $US_j \in 1, ..., S$. We assume that we will evaluate the dialogue success rate SR on the MultiWOZ dataset for simplicity. We will have a set of goals compatible with all the user simulators $G_i \in 1, ..., G$. For each pair of (DP_i, US_j), we will jointly finetune the user simulator for the given dialogue policy model to optimize SR using reinforcement learning to have the *cooperative* user simulator that has the highest chance to achieve the perfect SR similarly too (Cheng et al., 2022). We categorize the goals into the following categories based on the SR scores:

- For the G_{easy} easy goals, every evaluated model and user simulator combination achieved a perfect success rate.
- For the G_{reach} reachable goals, the DP models were able to achieve perfect SR only with the *cooperatively-finetuned* user simulators. We plan to finetune each user simulator according to one interpretable property IP. We say the IP property helps to achieve the SR measure if *cooperatively-finetuned* user simulator was first finetuned to IP and then to SR.

We suggest collecting the *failure* conversations that do not achieve perfect SR as DP_i was conversing with US_j, and we introduce the task of **dialogue continuation**. Our motivation is two-fold. First, we want to focus on hard conversations; second, we want to compare the dialogue policy models on the same conversation histories. Assuming a finished conversation C of length len_C we will iteratively remove the last turns and run the same US_j simulator with different models. If the original model DP_i cannot finish any conversation history of length $len_C - k$ for $k \in 1, ..., Hard - Last$ over multiple randomized runs with the US_j, we say that the model DP_i cannot solve the last $Hard - Last$ turns for given conversation C. Alternative model DP_{better} is considered strictly better if it achieves perfect SR by continuing from a longer conversation history of length $len_C - eps; eps < Hard - Last$.[6]

The proposed approach combines the corpus-based approach and the user simulator approach. We plan to evaluate if the proposed approach can identify hard examples from existing validation and tests on MultiWOZ and if it is able to pinpoint problematic dialog histories or even user and system actions. Finally, we will evaluate if it correlates well with human judgment.

6 Conclusion

Our thesis proposal described current challenges in the evaluation of NLG and Synthesized Speech in task-oriented dialogue. We motivated and stated our goals, presented our first experiments and proposed future work. Our main focus lies in automatic neural metrics for NLG and TTS systems that tend to correlate much better with human judgment than previous metrics. However, neural metrics bring their own problems; e.g., the lack of interpretability and the need to maintain the trained model in addition to software may complicate reproducibility. We are especially interested in improving the confidence estimation of the neural metrics and improving the metrics performance on utterance/turn level.

[6]Epsilon is

7 Acknowledgements

We thank Ondřej Dušek for his guidance, valuable feedback, and suggestions, Vojtěch Hudeček for many discussions. We also co-authors Vojta, Zdeněk Kasner, Patricia Schmidtová, and Mateusz Lango for their work.

This research was supported by Charles University projects GAUK 40222 and SVV 260575, and by the European Research Council (Grant agreement No. 101039303 NG-NLG). It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth and Sports project No. LM2018101).

References

Chantal Amrhein and Rico Sennrich. 2022. Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *NeurIPS*, volume 33, pages 12449–12460. Curran Associates, Inc.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Huiyuan Lai, Chris van der Lee, Emiel van Miltenburg, Yiru Li, Saad Mahamood, Margot Mieskes, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Pablo Mosteiro Romero, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP. ArXiv:2305.01633 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Pawe l 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. *arXiv:1810.00278 [cs]*. ArXiv: 1810.00278.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. 2022. Is MultiWOZ a Solved Task? An Interactive TOD Evaluation Framework with User Simulator. ArXiv:2210.14529 [cs].

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models. ArXiv:2306.04757 [cs].

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization Ability of MOS Prediction Networks. In *ICASSP*, pages 8442–8446.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic Weight Removal for Faithful and Abstractive Dialogue Generation. ArXiv:2303.17574 [cs].

Nauman Dawalatabad, Sameer Khurana, Antoine Laurent, and James Glass. 2022. On Unsupervised Uncertainty-Driven Speech Pseudo-Label Filtering and Model Calibration. ArXiv:2211.07795 [cs, eess].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv:2305.14314 [cs].

Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. ArXiv:2101.08231 [cs].

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG Challenge: Generating Text from RDF Data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Milica Gašić, Filip Jurčíček, Simon Keizer, François Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2010. Gaussian processes for fast policy optimisation of pomdp-based dialogue managers. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 201–204. Association for Computational Linguistics.

Milica Gašić and Steve Young. 2013. Gaussian processes for pomdp-based dialogue manager optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):28–40.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM TASLP*, 29:3451–3460.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022a. The VoiceMOS Challenge 2022. Technical Report arXiv:2203.11389, arXiv. ArXiv:2203.11389 [cs, eess] type: article.

Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. 2022b. LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech. In *ICASSP*, pages 896–900. IEEE.

International Telecommunications Union (ITU-T). 2006. ITUT Recommendation: Vocabulary for Performance and Quality of Service.

Josef Psutka, Luděk Müller, Jindřich Matoušek, and Vlasta Radová. 2006. *Mluvíme s počítačem česky*.

Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions.

Zdeněk Kasner, Ekaterina Garanina, Ondrej Platek, and Ondrej Dusek. 2023. TabGenie: A Toolkit for Table-to-Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 444–455, Toronto, Canada. Association for Computational Linguistics.

Keith Ito and Linda Johnson. 2017. The lj speech dataset.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *arXiv:2106.06103 [cs, eess]*. ArXiv: 2106.06103.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. ArXiv:2302.14520 [cs].

John Kominek, Tanja Schultz, and Alan W Black. 2008. SYNTHESIZER VOICE QUALITY OF NEW LANGUAGES CALIBRATED WITH MEAN MEL CEPSTRAL DISTORTION. page 6.

Mateusz Krubiński and Pavel Pecina. 2022. From COMET to COMES – Can Summary Evaluation Benefit from Translation Evaluation? In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 21–31, Online. Association for Computational Linguistics.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Dialogue with Pre-trained Language Models and Data Augmentation. *arXiv:2102.05126 [cs]*. ArXiv: 2102.05126.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. ArXiv:1612.01474 [cs, stat].

Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. 2022. One cannot stand for everyone! Leveraging Multiple User Simulators to train Task-oriented Dialogue Systems.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017a. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017b. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Michael Heck, Shutong Feng, and Milica Gasic. 2022. Dialogue Evaluation with Offline Reinforcement Learning. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 478–489, Edinburgh, UK. Association for Computational Linguistics.

François Mairesse and Steve Young. 2014. Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, 40(4):763–799.

Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777*.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, Flavours of Success: The Case of MultiWOZ. *arXiv:2106.05555 [cs]*. ArXiv: 2106.05555.

Alice Oh and Alexander Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In *ANLP-NAACL 2000 Workshop: Conversational Systems*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. *arXiv preprint arXiv:1704.03084*.

Ondřej Plátek, Petr Bělohlávek, Vojtěch Hudeček, and Filip Jurčíček. 2016. Recurrent Neural Networks for Dialogue State Tracking. ArXiv:1606.08733 [cs].

Ondřej Plátek and Ondřej Dušek. 2023. MooseNet: A Trainable Metric for Synthesized Speech with a PLDA Module. ArXiv:2301.07087 [cs, eess].

Ondřej Plátek, Hudeček Vojtěch, Patricia Schmidtová, Lango Mateusz, and Ondřej Dušek. 2023. Three Ways of Using Large Language Models to Evaluate Chat. ArXiv:2301.07087 [cs, eess].

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

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. ArXiv:1910.10683 [cs, stat].

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Ricardo Rei, José G. C. De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. ArXiv:2209.06243 [cs].

Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander Rudnicky. 2023a. Overview of Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems at DSTC 11 Track 4. ArXiv:2306.12794 [cs].

Mario Rodríguez-Cantelar, Chen Zhang, Chengguang Tang, Ke Shi, Sarik Ghazarian, João Sedoc, Luis Fernando D'Haro, and Alexander Rudnicky. 2023b. Overview of Robust and Multilingual Automatic Evaluation Metrics for Open-Domain Dialogue Systems at DSTC 11 Track 4. ArXiv:2306.12794 [cs].

Alexander I. Rudnicky, Eric H. Thayer, Paul C. Constantinides, Chris Tchou, R. Shern, Kevin A. Lenzo, Wei Xu, and Alice Oh. 1999. Creating natural dialogs in the Carnegie Mellon Communicator system. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 1531–1534.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. Technical report. ArXiv:2204.02152 [cs, eess] type: article.

Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. 2021. Non-Attentive Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modeling. *arXiv:2010.04301 [cs]*. ArXiv: 2010.04301.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. ISSN: 2379-190X.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. ArXiv:2305.13040 [cs].

Constantin Spille, Stephan D. Ewert, Birger Kollmeier, and Bernd T. Meyer. 2018. Predicting speech intelligibility with deep neural networks. *Computer Speech & Language*, 48:51–66.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. *arXiv preprint arXiv:1605.07669*.

Oktai Tatanov, Stanislav Beliaev, and Boris Ginsburg. 2021. Mixer-TTS: non-autoregressive, fast and compact text-to-speech model conditioned on language model embeddings. *arXiv:2110.03584 [eess]*. ArXiv: 2110.03584.

Paul Taylor. 2009. *Text-to-speech synthesis*. Cambridge university press.

Blaise Thomson and Steve Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Wei-Cheng Tseng, Wei-Tsung Kao, and Hung yi Lee. 2022. DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores. In *Interspeech*, pages 4541–4545.

Jannis Vamvas and Rico Sennrich. 2022. As Little as Possible, as Much as Necessary: Detecting Over- and Undertranslations with Contrastive Conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention As All You Need. *NeurIPS*, 30.

Veaux, Christophe, Yamagishi, Junichi, and MacDonald, Kirsten. 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. ArXiv:2301.02111 [cs, eess].

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. ArXiv:2204.07705 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 120–129, San Diego, California. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in Data-to-Document Generation. ArXiv:1707.08052 [cs].

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. ArXiv:2012.03539 [cs].

Rohola Zandie, Mohammad H. Mahoor, Julia Madsen, and Eshrat S. Emamian. 2021. RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis. *arXiv:2106.08468 [cs, eess]*. ArXiv: 2106.08468.

Heiga Zen, Keiichi Tokuda, and Alan W Black. 2009. Statistical Parametric Speech Synthesis. page 24.

Chen Zhang, Luis D'Haro, Rafael Banchs, Thomas Friedrichs, and Haizhou Li. 2020. Deep am-fm: Toolkit for automatic dialogue evaluation. *Conversational Dialogue Systems for the Next Decade*, pages 53–69.

Jianguo Zhang, Kun Qian, Zhiwei Liu, Shelby Heinecke, Rui Meng, Ye Liu, Zhou Yu, Huan Wang, Silvio Savarese, and Caiming Xiong. 2023. DialogStudio: Towards Richest and Most Diverse Unified Dataset Collection for Conversational AI. ArXiv:2307.10172 [cs].

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? ArXiv:1801.07243 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. ArXiv:2306.05685 [cs].

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive dialogue state tracker. *arXiv preprint arXiv:1805.09655*.

Lukáš Žilka, David Marek, Matěj Korvas, and Filip Jurcicek. 2013. Comparison of bayesian discriminative and generative models for dialogue state tracking. In *Proceedings of the SIGDIAL 2013 Conference*, pages 452–456.