Are Neural Open-Domain Dialog Systems Robust to Speech Recognition Errors in the Dialog History? An Empirical Study

Karthik Gopalakrishnan, Behnam Hedayatnia, Longshaokan Wang, Yang Liu, Dilek Hakkani-Tür

Amazon Alexa AI

{karthgop, behnam, longsha, yangliud, hakkanit}@amazon.com

Abstract

Large end-to-end neural open-domain chatbots are becoming increasingly popular. However, research on building such chatbots has typically assumed that the user input is written in nature and it is not clear whether these chatbots would seamlessly integrate with automatic speech recognition (ASR) models to serve the speech modality. We aim to bring attention to this important question by empirically studying the effects of various types of synthetic and actual ASR hypotheses in the dialog history on TransferTransfo, a state-of-the-art Generative Pre-trained Transformer (GPT) based neural open-domain dialog system from the NeurIPS ConvAI2 challenge. We observe that TransferTransfo trained on written data is very sensitive to such hypotheses introduced to the dialog history during inference time. As a baseline mitigation strategy, we introduce synthetic ASR hypotheses to the dialog history during training and observe marginal improvements, demonstrating the need for further research into techniques to make end-to-end open-domain chatbots fully speech-robust. To the best of our knowledge, this is the first study to evaluate the effects of synthetic and actual ASR hypotheses on a state-of-the-art neural open-domain dialog system and we hope it promotes speech-robustness as an evaluation criterion in open-domain dialog.

Index Terms: neural open-domain dialog systems, speech recognition, response generation, response selection

1. Introduction

Neural modeling approaches are prominent in research on both task-oriented and open-domain dialog. Traditional sequence-to-sequence models have been used for encoding the dialog history and predicting domains, intents, slot types, spans and more generally decoding full-fledged system responses. In recent years, large pre-trained Transformer-based models for natural language understanding (NLU) and natural language generation (NLG) have become ubiquitous [1, 2], leading to tremendous advances by fine-tuning towards these dialog tasks [3, 4].

In task-oriented speech-based dialog systems, the effect of ASR hypotheses has been widely studied and techniques have been devised to minimize the resulting downstream NLU errors [5, 6, 7, 8, 9, 10, 11]. More recently, end-to-end spoken language understanding approaches have been attempted to sidestep this problem [12, 13]. On the other hand, research in open-domain dialog is increasingly focusing on large, monolithic end-to-end neural models like Google's Meena [14] that are built using written data and evaluated on written interactions. Several written textual datasets have been created recently [15, 16, 17] to tackle various problems in open-domain dialog, including persona-grounding, knowledge-grounding and reasoning, and state-of-the-art chatbots have been built using them. However, it is not clear whether these written text-based open-domain chatbots would seamlessly integrate with ASR models to serve the speech modality, which is popular due to the ubiquity of voice assistants like Alexa, Google Assistant and Siri.

	Message
...	...
User	*That is insane,* ***but I heard that*** *is* ***cost*** *like* ***$110 million****, that is a* ***ton of*** *money*
ASR Output	*that is insane* ***by her Ladic*** *is* ***coast*** *like,* ***hungry and 10 medium****. That is a* ***tongue off*** *money.*
Written Text Model	that is crazy, i wonder if she can eat the food?
ASR Robust Model	yeah, it is. that is a lot of money. i think that the american people need that much money.

Figure 1: *A snippet from our augmented Topical-Chat audio test set, showing the ASR output for the user utterance in the dialog history and responses by two models: i) Written Text Model and ii) ASR Robust Model. Written Text Model is thrown off by the ASR output and responds about food, whereas ASR Robust Model responds correctly about money.*

Collecting large-scale written text-based dialog datasets is cheaper and more practical than collecting audio-based dialog datasets in many ways. But speech-robustness should be a factor of consideration when designing any (task-oriented or open-domain) dialog system intended to be deployed to the speech modality, even in the absence of audio-based training data. To bring attention to this important aspect in the open-domain dialog community, we empirically study the effects of various types of synthetic and actual ASR hypotheses in the dialog history on TransferTransfo (TF2) [18], a state-of-the-art neural open-domain dialog system based on the Generative Pre-trained Transformer (GPT) [19] from the NeurIPS ConvAI2 Conversational Intelligence Challenge [20]. We build off the Topical-Chat dataset [17] and perform two augmentations in our study: one creating *simulated* ASR hypotheses for the entire dataset, and another creating *actual* ASR hypotheses with a smaller audio-based analogue of the Topical-Chat test sets.

We observe that TF2 trained on written textual data is very sensitive to synthetic and actual ASR hypotheses introduced to the dialog history during inference time, with the sensitivity being particularly prominent for the task of response selection. As a baseline mitigation strategy, we introduce synthetic ASR hypotheses to the dialog history during training and observe marginal improvements, demonstrating the need for further research into techniques to make end-to-end open-domain chatbots fully speech-robust. Figure 1 shows a sample snippet with responses from TF2 models trained on written text and synthetic ASR hypotheses when fed speech-distorted dialog history.

A close work to ours in spirit is [21], which shows that

Transformer-based generative dialog models are insensitive to *unrealistic* perturbations like token-shuffling of the dialog history. Our work is more focused on evaluating the effects of introducing *realistic* perturbations to the dialog history in the form of synthetic and actual ASR hypotheses. Our augmentation of Topical-Chat, dubbed the Topical-Chat ASR dataset, is open-sourced[1] to enable open-domain dialog researchers to perform speech-robustness evaluation and fuel research into novel techniques to make monolithic neural open-domain dialog models more speech-robust.

2. Preliminaries

2.1. Notation

Let $D_t = [x_1, \ldots, x_t]$ denote a dialog history containing a sequence of t turns. Let $H_t = x_1 \oplus \cdots \oplus x_t$ denote a flattened sequence of all tokens in D_t. Let W be a truncate parameter for a flattened dialog history H, which retains at most W tokens from the end in H. When applied to H_t, we denote it as H_t^W. Finally, we denote a dialog example pair by (H_t^W, c_{t+1}), where c_{t+1} is a candidate response that is either the ground-truth response x_{t+1} at turn $t+1$ or a distractor response.

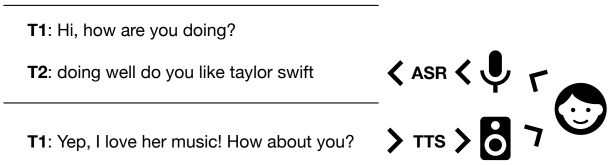
2.2. Data

For our experiments, we use the dialogs from Topical-Chat [17]. This is one of the largest and most diverse knowledge-grounded text-based open-domain dialog datasets publicly available today. Each dialog in Topical-Chat contains 20+ turns alternating between two Turkers and $\sim$19 tokens per turn on average. Topical-Chat has two types of test sets: *test freq* and *test rare*.

We performed two augmentations of Topical-Chat. First, a *simulated* augmentation wherein simulated errors are introduced at a corpus-level target word error rate (WER). For each data split (train/valid/test), we simulated ASR hypotheses for each turn in a dialog with four WER settings (0.1, 0.15, 0.2 and 0.3), each with a single seed for train and five random seeds for valid and test. We used the ASR error simulator method based on n-gram confusion matrix [22, 23] and trained the simulator on transcribed ASR output from an internal user study.

Second, an *actual* augmentation wherein two new test sets were created: *test freq audio* and *test rare audio*, which are smaller speech-based analogues of the original test sets. 40 dialogs corresponding to corpus-level distinct entity triplets were picked from each of the original test sets, and English-speaking human subjects of various ethnicities were asked to verbally read the dialogs with their own audio setup and record their audio, resulting in phonetically rich test sets. Two automated transcription systems (A and B) by Amazon were independently used to transcribe the collected audio, and each dialog transcription was aligned with the text of the original dialog based on edit distance followed by manual re-alignment to obtain the turn-level transcriptions. The Word Error Rate (WER) for the dialogs in *test freq audio* and *test rare audio* were in the range 0.1-0.39 by system A and in the range 0.08-0.35 by system B. For all experiments, we use the transcripts by system A.

Since our focus is on studying the effect of synthetic and actual ASR hypotheses for user utterances, we need to label the partners in a dialog example as *user* and *system*. For a dialog history D_t where turn $t+1$ is the target turn, we treat the Turker corresponding to turn $t+1$ as the *system* and the Turker corresponding to turn t as the *user*. This leads to system/user label assignments for all turns in D_t. Figure 2 shows this process.



T1: Hi, how are you doing?

T2: doing well do you like taylor swift

T1: Yep, I love her music! How about you?

Figure 2: *A dialog example where the target turn is by Turker T1. We treat the Turker corresponding to the target turn (T1) as the system, whose response is spoken to a user via a text-to-speech (TTS) engine. We treat the Turker corresponding to the previous turn (T2) as the user, whose spoken utterance is converted into text via ASR. We use synthetic and actual ASR hypotheses for the user turns: in this example, we remove punctuation from T2's only turn in the dialog history.*

2.3. Models

TransferTransfo [18] (TF2) is a state-of-the-art neural open-domain dialog system by Hugging Face based on the Generative Pre-trained Transformer (GPT) model [19] that won 1st place in automated evaluation and 2nd place in human evaluation at the NeurIPS ConvAI2 Conversational Intelligence Challenge [20].

In this system, GPT is fine-tuned on a dialog dataset in a multi-task learning fashion with the *language modeling* and *next utterance classification* tasks. Starter code for TransferTransfo along with the pre-trained model and BPE-based vocabulary is provided on GitHub by Hugging Face.

Language modeling: Trained with (H_t^W, x_{t+1}) pairs, this task is for learning to *generate* a response given a dialog history.

Next utterance classification: This task is for learning to *select* the ground-truth response x_{t+1} from a set of N candidate responses given a dialog history and is trained with $(H_t^W, \{c_{t+1}^i\})$ pairs, $i \in [1, \ldots, N]$. The set of candidate responses contains distractors alongside the ground-truth. In our experiments, we used $N = 5$ and randomly sampled turns from other dialogs in the same corpus (train/valid/test) as the one containing D_t to serve as distractors.

To remove confounding factors that may affect our study of the effect of synthetic and actual ASR hypotheses in the dialog history D_t, we avoid conditioning TF2 on knowledge from the reading sets provided to Turkers in Topical-Chat.

2.4. Training/Inference

Like [18], we use special tokens to identify speakers with their associated segments and initialize them with random embeddings to be learned during fine-tuning. We fine-tuned for a fixed number of 3 epochs with equal weight of 1.0 to the losses for both tasks. We wanted to fine-tune with a large dialog history and set $W = 256$. We used a train batch size of 2, performed gradient accumulation for 8 steps and gradient clipping with a max norm of 1.0, used the Adam optimizer and linearly decayed the learning rate from 6.25e-5 to 0 during the course of training. For simplicity, we set $H_t^W = x_t$ during inference time, i.e., we evaluate TF2's ability to generate and select given the last *user* turn. We used top-k, top-p nucleus sampling [24] with temperature T for decoding, where $k = 0$, $p = 0.9$ and $T = 0.7$. We also set a maximum decode length of 40 tokens.

3. Experimental Setup

Unlike written text (which we denote by **GOLD**), raw ASR hypotheses typically do not contain punctuation and casing. Since TF2 performs lowercasing during tokenization, we exclude cas-

[1] https://github.com/alexa/alexa-prize-topical-chat-dataset

ing from consideration, and evaluate our models on the following three variations of the dialog history D_t by using synthetic and actual ASR hypotheses:

- **NO-PUNC**: No punctuation in *user* turns in D_t.
- **WER-SIM**: Simulated errors are introduced to *user* turns in D_t at a corpus-level target WER.
- **REAL**: Actual ASR errors are introduced to *user* turns in D_t. This case is applicable for evaluation on *test freq audio* and *test rare audio* only.

We train these TF2 models as described in Section 2.4:

- **TF2-GOLD**: Trained on written text dialog examples.
- **TF2-NO-PUNC**: Trained on a NO-PUNC version of written text dialog examples.
- **TF2-WER-SIM**: For each WER, TF2 is trained on a WER-SIM version of written text dialog examples.

For the *language modeling* task, we use the standard automated metrics of perplexity (PPL) and unigram F1 between the generated and ground-truth response. For the *next utterance classification* task, we use recall $R_N@k$ [25] with $N = 5$ and $k = 1$, which measures the accuracy of selecting the ground-truth response from a set of candidate responses.

4. How Robust is TF2-GOLD?

To answer this question, we evaluate TF2-GOLD in various settings and empirically compare them against the GOLD setting.

4.1. NO-PUNC

Table 1 shows the results when evaluating TF2-GOLD in the NO-PUNC and GOLD settings on *test freq* and *test rare*. In order to better understand the effect of sentence segmentation, we report results separately for three versions of each test set:

- *single*: a subset of dialog examples where $H_t^W = x_t$ contains just a single sentence
- *multi*: a subset of dialog examples where $H_t^W = x_t$ contains more than one sentence
- *all*: *single* $\cup$ *multi*, i.e., all dialog examples

We observe a clear degradation in all metrics in the NO-PUNC setting relative to the GOLD setting, showing that TF2-GOLD relies on punctuation in the dialog history during both generation and selection. The degradation in PPL and $R_5@1$ when evaluating with the *multi* version of the test sets is greater than when evaluating with the *single* version, showing that the presence of sentence segmentation in dialog history comprised of multi-sentence utterances has a large impact on performance.

4.2. WER-SIM

Table 2 shows the results when evaluating TF2-GOLD in the WER-SIM and GOLD settings on *all* dialog examples of *test freq* and *test rare*. For each WER, we compute metrics using ASR hypotheses corresponding to all five seeds separately and report mean and standard deviation. We observe a significant degradation in all metrics in the WER-SIM setting relative to the GOLD setting, showing that TF2-GOLD is very sensitive to simulated ASR hypotheses in the dialog history. We also observe a significant degradation in metrics as the WER increases. The magnitude of degradation in $R_5@1$ shows that response selection relies heavily on the written form.

Table 1: *TF2-GOLD evaluated in NO-PUNC and GOLD settings on the Topical-Chat test set (freq / rare).*

Metric	**Version**	**NO-PUNC**	**GOLD**
PPL	*single*	18.0 / 23.9	17.9 / 23.6
	multi	19.2 / 25.3	18.8 / 24.7
	all	18.7 / 24.9	18.4 / 24.4
F1 (%)	*single*	16.6 / 16.2	16.4 / 16.4
	multi	15.9 / 15.5	16.0 / 15.6
	all	16.0 / 15.7	16.2 / 15.9
$R_5@1$ (%)	*single*	68.2 / 73.9	70.4 / 75.7
	multi	69.8 / 75.6	74.3 / 78.6
	all	69.7 / 74.8	73.7 / 77.7

Table 2: *TF2-GOLD evaluated in WER-SIM and GOLD settings on the Topical-Chat test set (freq / rare). For each WER, the top and bottom rows contain the mean and standard deviation of the computed metrics using simulated ASR hypotheses corresponding to all five seeds.*

WER	PPL	F1 (%)	$R_5@1$ (%)
0.1	19.26 / 25.49 (0.007 / 0.005)	15.59 / 15.38 (0.08 / 0.04)	66.59 / 71.53 (0.18 / 0.01)
0.15	19.55 / 25.81 (0.018 / 0.019)	15.42 / 15.32 (0.06 / 0.03)	65.08 / 70.13 (0.4 / 0.19)
0.2	19.83 / 26.11 (0.01 / 0.02)	15.31 / 15.1 (0.09 / 0.06)	63.75 / 68.42 (0.2 / 0.35)
0.3	20.46 / 26.83 (0.018 / 0.028)	14.89 / 14.77 (0.06 / 0.05)	59.96 / 64.47 (0.37 / 0.26)
GOLD	18.39 / 24.42	16.17 / 15.95	73.67 / 77.74

4.3. REAL

Table 3 shows the results when evaluating TF2-GOLD in the REAL and GOLD settings on *all* dialog examples of *test freq audio* and *test rare audio*. We observe a significant degradation in all metrics in the REAL setting relative to the GOLD setting, demonstrating that TF2-GOLD is very sensitive to actual ASR hypotheses in the dialog history.

Table 3: *TF2-GOLD evaluated in REAL and GOLD settings on our Topical-Chat audio test set (freq / rare).*

	PPL	F1 (%)	$R_5@1$ (%)
REAL	19.0 / 25.8	14.5 / 15.4	65.6 / 69.4
GOLD	17.4 / 24.1	15.5 / 16.1	76.3 / 78.6

5. Does Synthetic Training Help?

We now study the efficacy of synthetic training in making TF2 robust to *both* synthetic and actual ASR hypotheses.

5.1. NO-PUNC

We evaluate TF2-NO-PUNC in the NO-PUNC setting. Results for the NO-PUNC setting are in Table 4 and can be interpreted jointly with Table 1. We observe that TF2-NO-PUNC is more effective than TF2-GOLD at handling the NO-PUNC setting during inference time. There is a larger improvement in PPL and $R_5@1$ when evaluating with the *multi* version of the test sets than the *single* version, showing that NO-PUNC training is especially useful when the dialog history is comprised of multi-sentence utterances.

Table 4: *TF2-NO-PUNC evaluated in the NO-PUNC setting on the Topical-Chat test set (freq / rare).*

	PPL	F1 (%)	$R_5@1$ (%)
single	17.7 / 23.6	16.6 / 16.3	69.1 / 74.4
multi	18.7 / 24.7	16.0 / 15.6	72.8 / 76.9
all	18.3 / 24.4	16.2 / 15.9	70.9 / 76.4

5.2. WER-SIM

We evaluate TF2-WER-SIM in the WER-SIM setting. Results for the WER-SIM setting are in Table 5 and can be interpreted jointly with Table 2. We observe that TF2-WER-SIM is more effective than TF2-GOLD at handling the WER-SIM setting during inference time.

Table 5: *TF2-WER-SIM evaluated in the WER-SIM setting on the Topical-Chat test set (freq / rare). Each row refers to a model trained with the corresponding WER: the top and bottom rows contain the mean and standard deviation of the computed metrics using ASR hypotheses corresponding to all five seeds.*

WER	PPL	F1 (%)	$R_5@1$ (%)
0.1	18.69 / 24.92 (0.006 / 0.006)	16.19 / 15.94 (0.06 / 0.05)	69.77 / 74.32 (0.17 / 0.1)
0.15	19.52 / 25.89 (0.018 / 0.013)	15.48 / 15.17 (0.07 / 0.07)	67.7 / 71.99 (0.04 / 0.2)
0.2	19.36 / 25.8 (0.005 / 0.018)	15.75 / 15.48 (0.02 / 0.09)	66.74 / 71.45 (0.31 / 0.3)
0.3	19.59 / 26.21 (0.009 / 0.015)	15.56 / 15.23 (0.11 / 0.08)	64.18 / 68.17 (0.28 / 0.15)

5.3. REAL: Automated Evaluation

We evaluate TF2-NO-PUNC and TF2-WER-SIM in the REAL setting. Results are in Table 6 and can be interpreted jointly with Table 3. We observe that TF2-NO-PUNC and TF2-WER-SIM are generally more effective than TF2-GOLD at handling the REAL setting during inference time.

Table 6: *TF2-NO-PUNC and TF2-WER-SIM evaluated in the REAL setting on our Topical-Chat audio test set (freq / rare).*

Model	PPL	F1 (%)	$R_5@1$ (%)
TF2-NO-PUNC	18.5 / 25.2	15.6 / 15.1	66.4 / 71.3
0.1	18.31 / 25.27	15.64 / 15.65	66.93 / 71.09
0.15	19.11 / 26.02	15.48 / 15.14	67.16 / 70.76
0.2	18.77 / 25.73	15.11 / 15.49	65.68 / 71.09
0.3	18.6 / 25.65	15.64 / 15.35	66.14 / 71.21

Our evaluation shows that synthetic training provides reasonable improvements when the nature of errors matches during training and inference and very marginal improvements otherwise. An analysis of our *simulated* and *actual* ASR augmented data showed there are more insertion and deletion errors in the simulated data, and only about 10% of substitution errors in *actual* appear in *simulated* ASR augmented data. But regardless of the existence of training-inference match/mismatch, the perfomance gap with TF2-GOLD in the GOLD setting (particularly prominent in $R_5@1$) demonstrates that there is still a need for creative error-specific and/or error-agnostic techniques to make monolithic neural open-domain dialog models *demonstrably* robust to non-trivial target errors.

5.4. REAL: Human Evaluation

We also performed a human evaluation of the *language modeling* task from the two tasks in TF2, specifically generating responses from TF2-GOLD, TF2-NO-PUNC and TF2-WER-SIM in the REAL setting on the Topical-Chat audio test sets. 100 unique snippets were prepared from each test set, where each snippet contained a *written-text* dialog history and responses from all four models. But, the responses were obtained by feeding the models the *error-distorted* version of the dialog history. For each test set, a pair of annotators were asked to annotate snippets 1-60 and 40-100 respectively, thus providing 20 overlapping annotations to compute inter-annotator agreement. The annotators were asked to annotate on a 5-point nominal scale (1: Strongly Disagree, 2: Disagree, 3: Neither Agree Nor Disagree, 4: Agree, 5: Strongly Agree) whether each response was appropriate (AP) for the provided *written-text* dialog history.

The goal of this human evaluation was to see if there's any perceptible difference in responses generated from TF2-GOLD and the top 3 synthetic models with the least PPL in the REAL setting from Section 5.3 when fed error-distorted dialog history.

Table 7: *Human evaluation of responses by TF2-GOLD, TF2-NO-PUNC and TF2-WER-SIM (0.1 and 0.3) in the REAL setting on our Topical-Chat audio test set (freq above, rare below). AP = appropriateness. We report mean and margin of error interval at 95% confidence.*

Model	**GOLD**	**NO-PUNC**	**0.1**	**0.3**
AP	3.2 ± 0.25 3.6 ± 0.25	3.4 ± 0.26 3.7 ± 0.23	3.2 ± 0.24 3.7 ± 0.23	3.1 ± 0.25 3.7 ± 0.23

We bucketed ratings 1-2 and 4-5 when computing Fleiss' kappa annotator agreement and got scores of 0.54 and 0.38 on the two test sets. We observe in Table 7 that TF2-NO-PUNC is marginally better than TF2-GOLD for both test sets and TF2-WER-SIM is marginally better than TF2-GOLD on one test set.

6. Conclusion

We empirically studied the effects of synthetic and actual ASR hypotheses in the dialog history on TF2, a large state-of-the-art text-based neural open-domain dialog system from the NeurIPS ConvAI2 challenge. We observed that TF2 trained on written data is very sensitive to such hypotheses introduced to the dialog history during inference time, demonstrating that text-based neural open-domain chatbots may not be very effective at serving the speech modality as-is. We observed that training TF2 with synthetic ASR hypotheses makes it more robust to both synthetic and actual ASR hypotheses during inference time, with considerable room for improvement left for future work. Our augmentation of Topical-Chat, dubbed the Topical-Chat ASR dataset, is open-sourced and we hope our work sparks discussion and further research into modality-centric and modality-agnostic open-domain dialog systems.

7. References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019.

[3] J.-G. Zhang, K. Hashimoto, C.-S. Wu, Y. Wan, P. S. Yu, R. Socher, and C. Xiong, "Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking," *arXiv preprint arXiv:1910.03544*, 2019.

[4] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," *arXiv preprint arXiv:1911.00536*, 2019.

[5] X. Yang and J. Liu, "Using word confusion networks for slot filling in spoken language understanding," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] R. Masumura, Y. Ijima, T. Asami, H. Masataki, and R. Higashinaka, "Neural confnet classification: Fully neural network based spoken utterance classification using word confusion networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6039–6043.

[7] O. Z. Khan, J.-P. Robichaud, P. A. Crook, and R. Sarikaya, "Hypotheses ranking and state tracking for a multi-domain dialog system using multiple asr alternates," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] P. G. Shivakumar, M. Yang, and P. Georgiou, "Spoken language intent detection using confusion2vec," *arXiv preprint arXiv:1904.03576*, 2019.

[9] R. Schumann and P. Angkititrakul, "Incorporating asr errors with attention-based, jointly trained rnn for intent detection and slot filling," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6059–6063.

[10] D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur, "Beyond asr 1-best: Using word confusion networks in spoken language understanding," *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.

[11] B. Thomson, K. Yu, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, and S. Young, "Evaluating semantic-level confidence scores with multiple hypotheses," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[12] P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. Moreno, R. Prabhavalkar, Z. Qu, and A. Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 720–726.

[13] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5754–5758.

[14] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu *et al.*, "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.

[15] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" *arXiv preprint arXiv:1801.07243*, 2018.

[16] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," *arXiv preprint arXiv:1811.01241*, 2018.

[17] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, and D. Hakkani-Tür, "Topical-chat: Towards knowledge-grounded open-domain conversations," *Proc. Interspeech 2019*, pp. 1891–1895, 2019.

[18] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *arXiv preprint arXiv:1901.08149*, 2019.

[19] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[20] E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe *et al.*, "The second conversational intelligence challenge (convai2)," in *The NeurIPS'18 Competition*. Springer, 2020, pp. 187–208.

[21] C. Sankar, S. Subramanian, C. Pal, S. Chandar, and Y. Bengio, "Do neural dialog systems use the conversation history effectively? an empirical study," *arXiv preprint arXiv:1906.01603*, 2019.

[22] M. Fazel-Zarandi, L. Wang, A. Tiwari, and S. Matsoukas, "Investigation of error simulation techniques for learning dialog policies for conversational error recovery," *arXiv preprint arXiv:1911.03378*, 2019.

[23] J. Schatzmann, B. Thomson, and S. Young, "Error simulation for training statistical dialogue systems," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 526–531.

[24] A. Holtzman, J. Buys, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

[25] M. Henderson, I. Vulić, D. Gerz, I. Casanueva, P. Budzianowski, S. Coope, G. Spithourakis, T.-H. Wen, N. Mrkšić, and P.-H. Su, "Training neural response selection for task-oriented dialogue systems," *arXiv preprint arXiv:1906.01543*, 2019.