UNIVERSITÀ DI TRENTO

SiS LAB Signals & Interactive Systems

Are LLMs Robust for Spoken Dialogues?

Seyed Mahed Mousavi[1], Gabriel Roccabruna[1], **Simone Alghisi[1]**,
Massimo Rizzoli[1], Mirco Ravanelli[2], Giuseppe Riccardi[1]

[2] Concordia University, Mila-Quebec AI Institute, Canada | [1] Signals and Interactive Systems Lab, University of Trento, Italy

Presented at The 14th International Workshop on Spoken Dialogue Systems Technology

Introduction

- LLMs have demonstrated SOTA performance in dialogue modeling.
- LLMs are **trained** on written textual data from different sources.
- But, **spoken dialogues are different** from written dialogues: **1) conversational style and vocabulary**; **2) disfluencies**; **3) ASR errors.**
- The **robustness** of LLMs for **spoken dialogue** is ***understudied.***

Approach

1 Classified realistic **ASR errors** from **transcriptions of spoken TODs** and modelled their **distribution.**

2 **Generated a training set of noisy dialogues** by simulating ASR error in a large dataset of written dialogues.

3 **Fine-tuned two LLMs** (GPT-2 and T5) for the tasks of response generation and DST **using original and noisy dialogues.**

4 Performed **a comparative analysis** using both **automatic** (perplexity & Joint Goal Accuracy) and **human evaluation.**

Datasets

1 **MultiWOZ 2.1**, a dataset of **10k multi-domain written TODs.**

2 2 spoken versions of MultiWOZ development sets (valid & test sets) provided by DSTC 11.

- **Human Verbatim (HV)**: speakers read out loud the written turns.
  "I need you to find a hotel so I have a place to stay. It doesn't need to include internet, but it should include free parking."
- **Human-Paraphrased (HP)**: speakers paraphrased the written turns.
  "Can you find a place to stay? Internet not needed, but parking needed."

1. ASR Error Simulation

1 Automatically transcribed spoken Human-Verbatim dialogues.

2 Computed the statistics of ASR errors in three categories: **a) Insertions, b) Deletions,** and **c) Substitutions.**

3 Automatically injected the observed errors in written TODs by corrupting each token based on the calculated statistics.

Error Injection

Data	Insertions	Deletions	Substitutions
HV Transcriptions	2.3	1.9	8.1
MultiWOZ 2.1			
Error-Injected Train	2.1	5.9	8.0
Error-Injected Dev	2.1	6.0	8.3

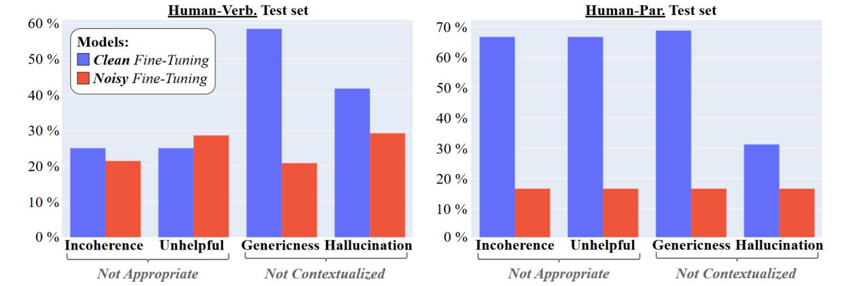
2. Evaluation - Dialogue State Tracking

- Introducing noise in the dialogue **slightly worsens** the model's performance.
- Introducing noise in the slot values **increases** the model robustness to noisy data slightly.

Models	H-Verb.	H-Par.
T5 Small Fine-Tuned on		
Clean Dialogues	**21.19**	**19.93**
Noisy Dialogues	19.07	18.08
Noisy Dialogues + 20% of Slot Values	19.72	18.38
Noisy Dialogues + 50% of Slot Values	20.09	18.73

3. Human Evaluation - Response Generation

The model **fine-tuned on noisy dialogues** generates **less generic**, **less hallucinated**, and **more coherent and helpful** responses.

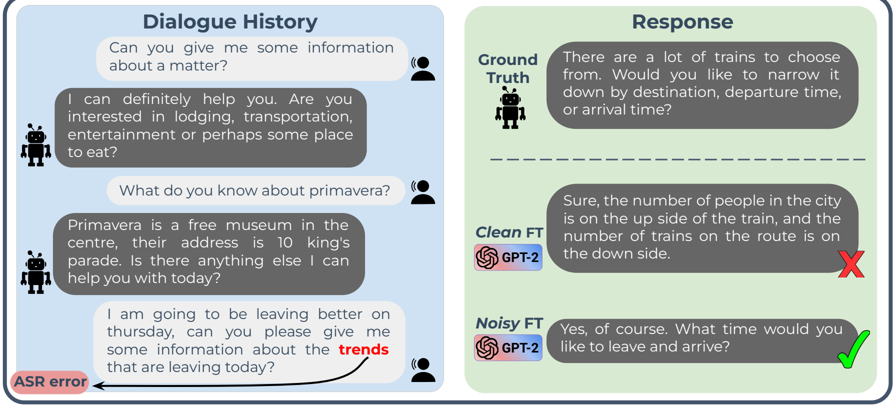


4. Human Eval. - Response Generation

- The model fine-tuned on noisy data produces **more appropriate** responses by 15%.
- Helps the model to **produce more contextualized** responses for both verbatim and paraphrased spoken settings.

Models	Correctness		Appropriatenes s		Contextualizatio n	
	H-Ver.	H-Par.	H-Ver.	H-Par.	H-Ver.	H-Par.
Ground Truth	0.85	0.70	1.00	1.00	1.00	0.95
GPT-2 Medium						
Clean Fine-Tuning	**1.00**	0.85	0.95	0.80	0.90	0.85
Noisy Fine-Tuning	0.95	0.85	0.95	**0.95**	**0.95**	**0.95**

5. Example - Clean VS Noisy Fine-Tuning



Dialogue History

Can you give me some information about a matter?

I can definitely help you. Are you interested in lodging, transportation, entertainment or perhaps some place to eat?

What do you know about primavera?

Primavera is a free museum in the centre, their address is 10 king's parade. Is there anything else I can help you with today?

I am going to be leaving better on thursday, can you please give me some information about the **trends** that are leaving today?

ASR error

Response

Ground Truth: There are a lot of trains to choose from. Would you like to narrow it down by destination, departure time, or arrival time?

***Clean* FT** GPT-2: Sure, the number of people in the city is on the up side of the train, and the number of trains on the route is on the down side. ✗

***Noisy* FT** GPT-2: Yes, of course. What time would you like to leave and arrive? ✓

ICT DAYS