Leveraging Large Language Models for Exploiting ASR Uncertainty

Pranay Dighe*, Yi Su*, Shangshang Zheng, Yunshu Liu, Vineet Garg, Xiaochuan Niu, Ahmed Tewfik (* = equal contribution)
ICASSP 2024 · Apple

Abstract

Problem
Large language models (LLMs) are good at NLP tasks, but for spoken language understanding (SLU), need to handle uncertainty from speech recognition (ASR) errors.

Goal: Make minimal changes to underlying ASR and LLM models, so that they can be shared across multiple tasks.

Proposed Approach: Prompt LLMs with n-best list of ASR hypotheses instead of error-prone 1-best hypothesis.

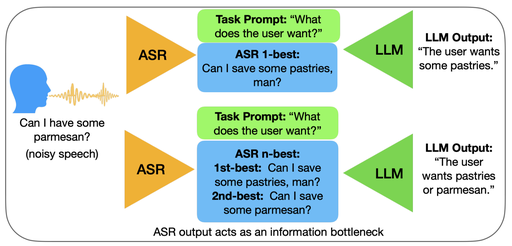


Figure: A toy NLP task which demonstrates that prompting the LLM with n-best ASR hypotheses allows it to exploit ASR uncertainty to better solve the downstream NLP task.

Tasks and Datasets

Task 1: Device-Directed Speech Detection (DDSD)

Definition: Binary classification task to identify if a spoken utterance is directed towards a voice assistant or a human.
Dataset: Internal dataset with train (weakly-labeled, ~107k utterances per class) and eval (human-graded, 12,771 device-directed, 2,274 human-directed) partitions.

Task 2: Keyword Spotting

Definition: Multi-class classification task to identify which command keyword (e.g. "yes", "no", "up", "down") if any, was spoken in the utterance.
Dataset: Google Speech Commands (GSC) 35 keywords, 10 considered in-domain commands. Train partition (~85k utterances) and test partition (~11k utterances).

Proposed ASR+LLM Approach

Approach

1. **Obtain n-Best ASR hypotheses** with costs
   - Condensed representation of ASR lattice.
   - Hypotheses separated by newlines.
   - Costs indicate ASR confidence.
2. **Design prompting templates**
   - Explain concept of n-best lists to LLM.
   - Task description and input formatting.

Table: Examples of n-Best ASR hypotheses

utterance-prompt	GSC	DDSD
1-best	*"nive"*	*"shuffle play U2"*
n-best	*"hive [-47.8]* *five [-46.8]* *bye [-44.0]* *hive [-31.5]"*	*"shuffle play U2 [-84.4]* *shuffle play Kito [-83.1]* *shuffle play Buku [-82.9]* *shuffle play Kitu [-82.8]"*
Ground-truth	*"five"*	*"shuffle play U2"*

Table: Prompting Template for DDSD Task

Prefix	Infix	Suffix
1-best: *"Determine whether the following spoken utterance is directed towards a voice assistant or a human being."*	*"Typical spoken utterances directed towards the voice assistant are commands to fulfill a task or queries to get some information."*	**binary-target:** *"Answer only from the following categories ['1', '0'] where '1' indicates that the utterance is directed towards the voice assistant and '0' indicates that the utterance is directed towards a human being."*
n-best: *"In this task, we provide an n-best list of ASR hypotheses for a spoken utterance. Each of the hypothesis is separated by a newline character. The cost of each hypothesis is at the end in the format '[cost]' where a low cost indicates that we are more confident about that ASR hypothesis. Determine whether the following spoken utterance is directed towards a voice assistant or a human being by taking into account all the n-best hypotheses."*		**0-100 scale:** *"Answer on a scale of 0 to 100 where a score of '100' indicates that the utterance is directed towards the voice assistant and '0' indicates that the utterance is directed towards a human being. Your answer should only contain an integer between 0 and 100."*

3.Finetuning Procedure
- Base Model: a Vicuna-7B-v1.3, a pre-trained instruction-tuned LLaMA large language model, as the base model.
- Finetuned Low-Rank Adaptation (LoRA) adapters on the LLM.
- LoRA adapters have only 4.1M parameters (0.06% of LLM's 7B parameters) which makes finetuning compute efficient.

4.DDSD Classification Tasks:
- Binary DDSD: Finetuned to output binary labels (0/1) for device/human directed
- Scale 0-100 DDSD: Finetuned to output scores on 0-100 scale (from a teacher LatticeRNN model).

5.Keyword Spotting Task: Finetuned to output 10 keywords or "OOV".

Analysis and Results

DDSD Analysis
- Prompting w/o finetuning works best for n=1, and only for binary classification task.
- With finetuning, the LLM can leverage n-best ASR hypothesis for better performance for both binary as well as a 0-100 scale task.

n	Binary Target				100-Scale Task			
	Base Model		Finetuned		Base Model		Finetuned	
	TPR	FPR	TPR	FPR	FPR95	EER	FPR95	EER
1	91.5	30.0	90.5	8.3	84.7	32.0	53.6	10.7
2	85.1	32.7	91.3	5.2	90.3	55.2	30.0	8.2
4	87.2	46.4	91.8	5.2	85.8	46.0	13.1	7.6
8	85.9	43.0	91.8	4.7	81.9	36.3	11.3	7.5
16	85.0	42.0	92.0	4.9	84.7	36.7	10.5	7.4

Table: DDSD Rresults

Table: Results on Keyword Spotting Task

Examples of some corrections made by the LLM are "app"→"up", "Lyft"→"left", and "call"→"go"

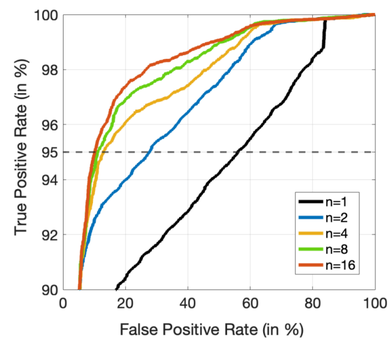
System	yes	no	up	down	left	right	on	off	stop	go	OOV	Total Acc.
Baseline	98.6/98.1	83.4/98.3	99.6/55.1	100.0/87.2	100.0/88.3	100.0/87.1	100.0/78.5	99.5/90.3	99.0/98.8	99.7/81.3	93.0/99.3	94.5
FT n=1	98.3/98.6	94.4/95.6	97.8/84.5	99.5/89.4	99.8/95.2	99.7/97.5	98.2/82.3	96.4/93.3	99.0/99.0	97.9/91.3	96.4/99.5	97.0
FT n=8	99.3/99.0	94.1/98.0	97.8/85.7	99.5/92.1	99.0/95.2	99.7/98.2	98.5/83.8	97.7/93.5	99.3/98.8	99.0/94.8	96.9/99.5	97.5

Results

LLM Response with and without Finetuning

DDSD Prompt: *Task-prompt (n-best Prefix + Infix + binary-target Suffix) + "score of the Hawaii game [-144.2] \n what's the score of the Hawaii game [-135.1]"*
LLM without Finetuning: *"Based on the provided n-best list, the most likely hypothesis for the spoken utterance is: 'score of the Hawaii game [-144.2]'. This hypothesis is the first in the list, indicating that it is the most likely interpretation of the spoken utterance. The cost associated with this hypothesis is '-144.2', which suggests that the ASR system is not very confident about this hypothesis. Given the content of the utterance, it is more likely to be directed towards a voice assistant rather than a human being. The utterance contains a question about the Hawaii game, which is a sports-related topic that is commonly addressed to voice assistants. Therefore, the answer is '1'."*
LLM with Finetuning: *"1"*

Figure: DDSD ROC Curve using Scale 0-100 Outputs



Conclusion

- LLMs can effectively leverage n-best ASR hypotheses via prompting and LoRA finetuning.
- Significant improvement over just using 1-best ASR for intent classification and keyword spotting tasks.
- Efficient method to exploit ASR uncertainty without redesigning ASR and LLM models.

References

- Hu et al. "LoRA: Low-rank Adaptation of Large Language Models." ICLR 2022.
- Warden. Speech Commands Dataset. arXiv 2018.