Utilizing Prior Knowledge to Improve Automatic Speech Recognition in Human-Robot Interactive Scenarios

Pradip Pramanick
TCS Research
India

Chayan Sarkar
TCS Research
India

ABSTRACT

The prolificacy of human-robot interaction not only depends on a robot's ability to understand the intent and content of the human utterance but also gets impacted by the automatic speech recognition (ASR) system. Modern ASR can provide highly accurate (grammatically and syntactically) translation. Yet, the general purpose ASR often misses out on the semantics of the translation by incorrect word prediction due to open-vocabulary modeling. ASR inaccuracy can have significant repercussions as this can lead to a completely different action by the robot in the real world. Can any prior knowledge be helpful in such a scenario? In this work, we explore how prior knowledge can be utilized in ASR decoding. Using our experiments, we demonstrate how our system can significantly improve ASR translation for robotic task instruction.

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; *Knowledge representation and reasoning*.

KEYWORDS

ASR, HRI, embodied agent, robotics knowledge, cognitive robot

ACM Reference Format:
Pradip Pramanick and Chayan Sarkar. 2023. Utilizing Prior Knowledge to Improve Automatic Speech Recognition in Human-Robot Interactive Scenarios. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580129

1 INTRODUCTION

Spoken interaction is an important part of a system that aims to enable a seamless way of instructing a robot [17, 19]. The recent success of automatic speech recognition (ASR) systems [9] has paved the way for realistic applications of ASR in human robot interaction [11, 24]. However, the efficacy of an ASR system deployed in a robot depends upon various factors [11] and its accuracy is predominantly affected by noise, speaker's accent and distance [15]. Research on ASR primarily focus on improving transcription accuracy for general-purpose applications and existing ASR systems that are publicly available (commercial or otherwise), follow the same general approach of modeling and training. Previous attempts at improving speech recognition in robots primarily focus on either improving the quality of the received speech signal or exploit controllable acoustic characteristics [15].

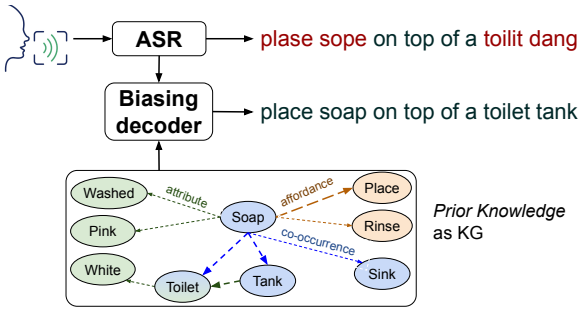

Figure 1: A scenario that shows the typical errors made by an ASR model. Our approach improves the transcription using relevant prior knowledge.

This work aims to improve ASR accuracy, assuming the ASR is utilized to transcribe natural language instructions given to a robot [20, 21]. Specifically, we introduce the problem of incorporating domain-specific prior knowledge about objects in the environment while performing inference with a pre-trained ASR model. We consider three types of relational knowledge about objects - affordance, physical attributes, and co-occurrence relations (spatial attributes). Even though a particular instance of such knowledge would be domain-specific, the knowledge types are common to human-robot interaction that involves natural language [2, 10, 12, 22].

Figure 1 illustrates a scenario where a robot takes a spoken instruction[1], and transcribe it incorrectly by an ASR model. The figure also shows a part of a KG, previously obtained by the robot, and illustrates how a subset of the words in the instruction are present as nodes in the KG, with the edges denoting the particular type of the relationship. The existing works that include ASR in robots, do not use such factual knowledge. We propose a method to include the KG during ASR inference, which is based on an approach to bias the *beam search decoding* process of ASR [8, 18, 25, 26]. However, we propose significant modifications to the biasing method to make it suitable for biasing using a KG. We summarize our main contributions as the following.

- To the best of our knowledge, this is the first work to utilize semantic knowledge to improve ASR accuracy for robotics application.

[1]The example is from a benchmark dataset, introduced in [23].

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580129

- We propose a new biasing of ASR decoder using KG and show significant improvements in speech recognition accuracy, when applied to transcribing spoken natural language instructions.

2 RELATED WORKS

Along with improving the general purpose ASR's accuracy, the existing works on improving speech recognition in robots mostly focus on improving the acoustic characteristics of the speech signal, either by filtering noisy signals or training with noisy audio data [15]. Some work also focus on minimizing the effects of noise in speech directed towards a robot in motion. Kennedy *et al.* [7] experiment with different microphone types, distance and speaker angles. Some works have also explored controlling the robot's behavior (pausing, turn-taking, etc.) while interacting to improve ASR accuracy [13, 24]. While these prior works introduce approaches and guidelines for effectively using ASR models in robots, there is little progress in the area of using of semantic knowledge to improve ASR. Oneata *et al.* [16] propose a method to perform speech recognition using visual context in unmanned aerial vehicles. Their approach combines features extracted from images using a recurrent neural network during ASR inference. This prior visual information can be considered as knowledge, but such knowledge is highly contextual and may not always be relevant. Thus, their approach is unsuitable to apply on a static, long-term semantic knowledge representation such as a KG.

Recently, Pramanick *et al.* [18] proposed a shallow-fusion biasing approach to improve ASR inference in robots, using visual context captured from the robot's camera. Their system also uses contextual knowledge from vision and uses a shallow-fusion biasing method to include the visual context during ASR inference. Shallow fusion biasing has been effectively applied for improving speech recognition accuracy for rare words and proper nouns, e.g., person names [5, 6, 8, 25, 26]. However, all these modifications of the shallow-fusion biasing model use a static biasing vocabulary. Although, [18] dynamically change the biasing vocabulary with the robot's movement, the dynamism is only per-inference, and the bias vocabulary remains the same during the entire decoding process. Such an approach cannot be applied optimally to this problem as we are interested in dynamic biasing using partial facts (nodes) from a KG, whose validity (i.e., being an edge in the KG) can only be checked in the future time steps during decoding. We show that the standard shallow-fusion biasing model with a static vocabulary is sub-optimal in Section 5.3.

3 OVERVIEW

We follow a well-known approach of ASR modeling for open-vocabulary speech recognition, named the CTC model [3]. Given a speech input $\mathbf{x}$, discretized into l' time-steps, the CTC model produces a shorter sequence of the probability distribution of output labels $\mathbf{y}$ that can be decoded to find the optimal output sequence, i.e., the transcription,

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{l} P(y_i|\mathbf{x}), l < l'. \tag{1}$$

The output labels are usually chosen to be sub-word units, such as *Grapheme* (characters) or *WordPiece* to perform open-vocabulary speech recognition [1]. The equation above is usually approximated using a beam search algorithm [3]. We propose a method to bias the output of the beam search using prior knowledge. We propose novel refinements to the existing method of introducing bias in ASR, namely the *shallow-fusion biasing* model [25, 26]. In the following, we describe our approach to obtain and represent the prior knowledge and utilize the same to dynamically bias the beam search. Figure 2 shows an outline of our approach with an example.

4 KNOWLEDGE-AIDED INFERENCE

Utilizing the prior knowledge in ASR decoding requires the knowledge data to be represented pertinently. In this section, we describe the knowledge representation before detailing our proposed method of dynamic biasing using this prior knowledge.

4.1 Knowledge Representation

We represent the prior knowledge as a directed graph, subsequently referred as a knowledge graph (KG) throughout the paper. Each node in the KG is a unique natural language token and an edge denotes a particular relationship between the tokens (as observed by a data collector or an automated annotator). Thus a pair of connected nodes represent a rational fact. As we do not assume an unobserved relationship to be false, we do not consider negative facts and ignore unobserved relationships altogether. In this work, we demonstrate our approach by considering three binary relationships corresponding to three distinct types of edges in the KG. We summarize the edge types in Table 1. In our experiments, we extract the KG from a pre-collected corpus of natural language instructions given to a robot, which involves executing various household tasks [23]. However, such a KG can also be constructed from visual observations during exploration [2, 22] or even defined manually. In the following, we briefly describe the process of extracting the KG from the corpus.

We primarily use a dependency parser[2] to extract the relationships from a given natural language instruction. In particular, we convert the syntactic dependency graph of a given text into a set of nodes and edges using a set of rules. Assuming an instruction is made up of n tokens, $I = t_{1:n}$, an edge type between a pair of tokens (t_i, t_j) is defined as following.

$$\forall (t_i, t_j) \in I, \begin{cases} t_i \xRightarrow{affordance} t_j, & \text{if } t_i \xrightarrow{dobj} t_j \\ t_i \xRightarrow{attribute} t_j, & \text{if } t_i \xrightarrow{amod,\, compound} t_j \\ t_i \xRightarrow{co-occurrence} t_j, & \text{if } t_i \xRightarrow{affordance} t_x \\ & \text{and } t_j \xrightarrow{pobj} t_x, t_x \in I \end{cases}$$

where $\rightarrow$ denotes a dependency relation in the given dependency graph. The interpretation of the dependency relations used to construct the rules can be found in [14]. We use a rule-based approach for fast inference during beam search. However, it is also possible to consider the edge extraction as a supervised learning problem.

[2] https://spacy.io/api/dependencyparser

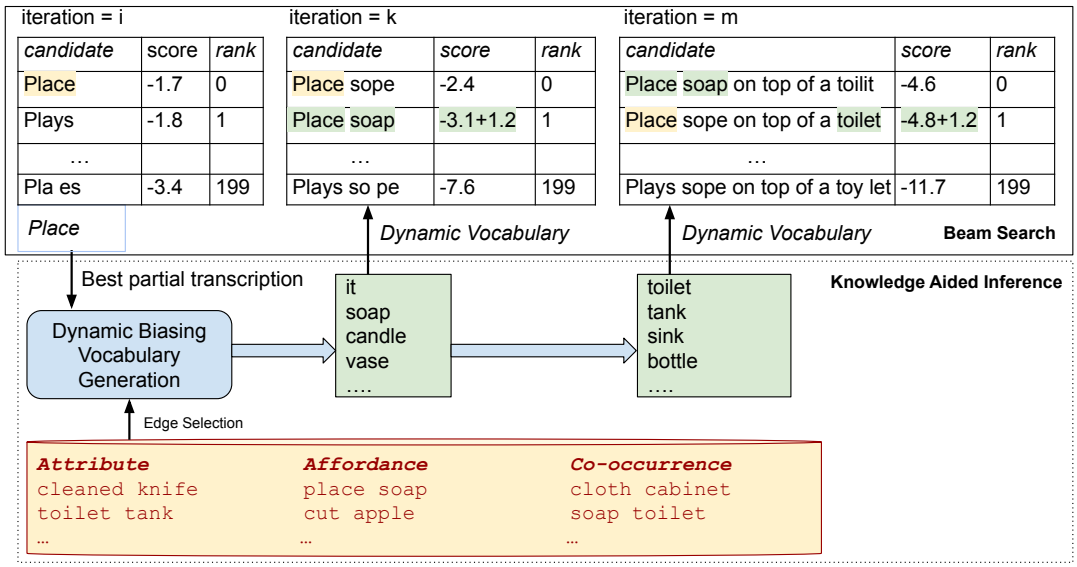

Figure 2: An overview of our approach that shows a hypothetical scenario of a beam search.

Table 1: Types of edges in the KG.

Edge-type	Description	Examples
Affordance	Denotes that an action is applicable on a object.	rinse → sponge, cut → apple.
Attribute	A physical or an abstract property of an object.	small → lamp, dining → table.
Co-occurrence	A spatial relationship between two objects.	book → desk, knife → table.

4.2 Dynamic Biasing

The beam search decoding algorithm attempts to find an approximately optimal solution for Eq. 1, i.e., finds the output label sequence with the maximum likelihood under the approximation constraint of keeping at most N candidates (partial transcriptions) in each iteration of the search. As shown in Figure 2, each iteration extends the current candidates with $|L|$ sub-word labels from the ASR network's output, to generate $N \times |L|$ candidates and scores them according to the CTC objective [3],

$$S_c = log(P_{CTC}(c|\mathbf{x})). \tag{2}$$

It is common to use an n-gram language model by a log-linear interpolation with this score to improve the transcription accuracy [1, 3]. Shallow-fusion biasing further improves the chances of correctly recognizing proper nouns, such as person names, which is usually under-represented in the standard ASR training datasets [6]. Shallow-fusion biasing methods generally boosts the scores of certain candidates. These candidates contain words or phrases from a pre-set biasing vocabulary, assuming such a biasing vocabulary is obtained before performing the beam search.

We observe that the KG can also be utilized to bias the beam search and thus propose a method to do so. However, the standard shallow fusion biasing model, if applied directly, i.e., by simply treating the nodes of the entire KB as a large biasing vocabulary, would result in sub-optimal decoding. Instead, we propose a dynamic biasing vocabulary generation method that predicts which edges in the KG are relevant during a certain point in the beam search, and restricts the biasing vocabulary to a subset of the nodes in the KG. We use the dependency parsing graph of the top ranked partial transcription to extract nodes that are part of the KG. Given the current node (the last in the sequence), we generate a dynamic biasing vocabulary by selecting the nodes in the KG that has an edge with the current node. As illustrated using the example of decoding the spoken instruction - *'place soap on top of a toilet tank'* in Figure 2, given the partial transcription *'place'*, the dynamic biasing vocabulary contains nodes having an edge with the *place* node. Similarly, given the partial transcription *'place soap'*, the biasing vocabulary changes to only the nodes having an edge with the *soap* node, and so on. Our biasing model re-scores a candidate if it finds an edge from the KG to be present the candidate. In other words, an observed relationship between a pair of nodes in a partial transcription, improves its score. We formally define the rescoring function as,

$$\mathcal{R}(S_c) = \begin{cases} S_c + \lambda & \exists e \in T', e \in KG \\ S_c & otherwise \end{cases} \tag{3}$$

where e represents an edge, T' represents a partially decoded transcription, and λ is a hyper-parameter.

5 EVALUATION

In this section, we evaluate our dynamic biasing approach and compare it with a couple of baseline systems.

5.1 Data

We perform experiments on a well-known dataset for evaluating natural language instruction following the capabilities of a robot,

Table 2: ASR performance compared to baseline models.

Model	WER	WERR
Wav2Vec2	13.05	—
Wav2Vec2 + LM	6.01	53.95
KG-Bias (dynamic)	**5.47**	**58.08**
KG-Bias (affordance)	5.54	57.55
KG-Bias (attribute)	5.68	56.48
KG-Bias (co-occurrence)	5.79	55.63
KG-Bias (static)	5.55	57.47

namely the Alfred dataset [23]. In particular, we use the textual instructions in the *train* subset to build the KG. To evaluate speech recognition, we first convert the *valid-seen* and *valid-unseen* subsets into spoken instructions, using text-to-speech models. Specifically, we use the *gTTS*[3] model to generate spoken instructions from valid-seen and use the *Tacotron2*[4] model for valid-unseen. We remove any text that is present in the intersection of these three subsets. Finally, we perform all hyper-parameter tuning on valid-seen and evaluate on valid-unseen. This ensures that the evaluation is done in a *zero-shot* manner, both in terms of unseen text and audio of an unknown speaker. The valid-seen and the valid-unseen sets contain 789 and 749 spoken instructions, respectively.

5.2 Baselines & Ablation

We compare our approach with two baseline models. We also perform ablation experiments by restricting to a single edge type and biasing using the entire KG. We perform modifications to a standard CTC beam search implementation[5] as described below.

- Wav2Vec2 - We use a pre-trained wav2vec2-base model [1].
- Wav2Vec2 + LM - We train a 3-gram language model (LM) using KenLM [4] from the *train* set of Alfred [23]. Then we combine this LM with Wav2Vec2 using a log-linear interpolation.
- KG-Bias (affordance) - We perform shallow-fusion biasing using only the affordance relationships in the KG.
- KG-Bias (attribute) - We perform biasing using only the attribute relationships in the KG.
- KG-Bias (co-occurrence) - We perform biasing using only the attribute relationships in the KG.
- KG-Bias (static) - We use a fixed biasing vocabulary for decoding all the instructions, obtained by including all the nodes in the KG.

5.3 Results

We use a beam width of 100 for all the models. We find all other optimal hyper-parameters using bayesian optimization[6], performed on the valid-seen set. We use the standard ASR evaluation metrics – word error rate (WER) and relative reduction in WER, i.e., WERR. The results are shown in Table 2.

[3]https://pypi.org/project/gTTS
[4]https://github.com/mozilla/TTS
[5]https://github.com/PaddlePaddle/PaddleSpeech
[6]https://ax.dev/docs/bayesopt.html

The basic ASR model without any modifications during inference, i.e., the standard Wav2Vec2 ASR system achieves around 13% WER. This shows that even in noise-less audio and consistent pronunciation (due to TTS models), general-purpose ASR models are unable to accurately transcribe spoken instructions for a robotics domain. We find a large improvement of approximately 54% relative improvement by using a language model, which is trained from in-domain textual examples.

Further, we find a substantial improvement by using our approach, i.e., shallow-fusion biasing using a pre-collected knowledge graph, named as *KG-Bias (dynamic)* in Table 2. We obtain the lowest WER of 5.47 and the highest relative reduction in WER compared to the Wav2Vec2 model (around 58%). Compared to the LM interpolated baseline model, the relative reduction in WER is 9%.

In our first ablation experiment, we consider using only affordance relationships in the KG to bias the inference, which results in a slight increase in WER, 1.3% compared to using the entire KG. However, we still find that the WER is lesser (better) than both the standard and the LM-interpolated model. Similarly, we find that by separately biasing using attribute and co-occurrence relationships, the WER slightly increases by 3.7% and 5.5% relative to the full KG model. Again, both of these relationships obtain better speech recognition accuracy than the baselines. Interestingly, we find that biasing using affordance relationships yields better results than both attribute and co-occurrence relationships. The co-occurrence relationships seem to have the least effect on biasing. However, these observations could be specific to this dataset and due to the vocabulary coverage of the KG. Finally, we also show the results of static biasing using the full KG, which achieves a WER of 5.55. However, this approach clearly results in sub-optimal decoding, as we get almost similar and slightly better result from the *KG-Bias (affordance)* model that doesn't use the entire KG.

5.4 Limitations

Our experiments show promising results on the viability of using prior knowledge represented as a KG. However, our approach is naturally limited by the vocabulary coverage of the KG. In our experiments, we extract the KG from the *train* split of *Alfred*, which results in a high coverage. In the test set, 21% of the words are OOVs, i.e., absent in the KG. Future works can explore other ways of extracting the KG and inference methods to use a sparser KG. Also, future experiments can include human speech instead of TTS with variations in noise and accent.

6 CONCLUSION

In this paper, we present a method to improve an automatic speech recognition system that is to be deployed in a robot. In particular, we propose a novel way of utilizing a knowledge graph, containing information about relationships among objects, attributes, and actions, during the inference using the speech recognition system. We find promising results from experiments on a dataset of spoken natural language instructions given to a robot. There are several possibilities for extending this work, by improving the knowledge representation and reasoning. Moreover, the general idea of biasing inference using prior knowledge could be explored in other robotic applications.

REFERENCES

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 12449–12460.

[2] Angel Andres Daruna, Weiyu Liu, Zsolt Kira, and S. Chernova. 2019. RoboCSE: Robot Common Sense Embedding. *2019 International Conference on Robotics and Automation (ICRA)*, 9777–9783.

[3] Awni Y Hannun, Andrew L Maas, Daniel Jurafsky, and Andrew Y Ng. 2014. First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. *arXiv preprint arXiv:1408.2873* (2014).

[4] Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 187–197.

[5] Young Mo Kang and Yingbo Zhou. 2020. Fast and robust unsupervised contextual biasing for speech recognition. *arXiv preprint arXiv:2005.01677* (2020). https://doi.org/10.48550/ARXIV.2005.01677

[6] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5828.

[7] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 82–90.

[8] Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, et al. 2021. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. *arXiv preprint arXiv:2104.02194* (2021).

[9] Jinyu Li. 2022. Recent Advances in End-to-End Automatic Speech Recognition. *APSIPA Transactions on Signal and Information Processing* 11, 1 (2022).

[10] Weiyu Liu, Dhruva Bansal, Angel Andres Daruna, and Sonia Chernova. 2021. Learning Instance-Level N-Ary Semantic Knowledge At Scale For Robots Operating in Everyday Environments.. In *Robotics: Science and Systems*.

[11] Matthew Marge, Carol Espy-Wilson, Nigel G Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé III, Debadeepta Dey, et al. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language* 71 (2022), 101255.

[12] So Yeon Min, Devendra Singh Chaplot, Pradeep Kumar Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. 2021. FILM: Following Instructions in Language with Modular Methods. In *International Conference on Learning Representations*.

[13] Omar Mubin, Joshua Henderson, and Christoph Bartneck. 2014. You just do not understand me! Speech Recognition in Human Robot Interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 637–642.

[14] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 1659–1666.

[15] José Novoa, Rodrigo Mahu, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, and Néstor Becerra Yoma. 2021. Automatic speech recognition for indoor hri scenarios. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–30.

[16] Dan Oneaţă and Horia Cucu. 2021. Multimodal speech recognition for unmanned aerial vehicles. *Computers & Electrical Engineering* 90 (2021), 106943. https://doi.org/10.1016/j.compeleceng.2020.106943

[17] Pradip Pramanick, Hrishav Bakul Barua, and Chayan Sarkar. 2020. DeComplex: Task planning from complex natural instructions by a collocating robot. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6894–6901.

[18] Pradip Pramanick and Chayan Sarkar. 2022. Can Visual Context Improve Automatic Speech Recognition for an Embodied Agent?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing(EMNLP)*. Association for Computational Linguistics.

[19] Pradip Pramanick, Chayan Sarkar, P Balamuralidhar, Ajay Kattepur, Indrajit Bhattacharya, and Arpan Pal. 2019. Enabling human-like task identification from natural conversation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6196–6203.

[20] Pradip Pramanick, Chayan Sarkar, Snehasis Banerjee, and Brojeshwar Bhowmick. 2022. Talk-to-Resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot. *Robotics and Autonomous Systems* 155 (2022), 104183.

[21] Pradip Pramanick, Chayan Sarkar, and Indrajit Bhattacharya. 2019. Your instruction may be crisp, but not clear to me!. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–8.

[22] Pradip Pramanick, Chayan Sarkar, Sayan Paul, Ruddra dev Roychoudhury, and Brojeshwar Bhowmick. 2022. DoRO: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters* 7, 4 (2022), 10826–10833.

[23] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10740–10749.

[24] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.

[25] Ian Williams, Anjuli Kannan, Petar S Aleksic, David Rybach, and Tara N Sainath. 2018. Contextual Speech Recognition in End-to-end Neural Network Systems Using Beam Search.. In *Interspeech*. 2227–2231.

[26] Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-Fusion End-to-End Contextual Biasing. In *Proc. Interspeech 2019*. 1418–1422.