‘*What are you referring to?*’ Evaluating the ability of multi-modal dialogue models to process clarificational exchanges

Anonymous ACL submission

Abstract

Referential ambiguities arise in dialogue when a referring expression does not uniquely identify the intended referent for the addressee. Addressees usually detect such ambiguities immediately and work with the speaker to *repair* it using meta-communicative, clarificational exchanges (CE): a *Clarification Request* (CR) and a response (see Fig. 1). Here, we argue that the ability to generate and respond to CRs imposes specific constraints on the architecture and objective functions of multi-modal, visually grounded dialogue models. We use the SIMMC 2.0 dataset to evaluate the ability of different state of the art model architectures to process CEs, with a metric that probes the contextual updates that arise from them in the model. We find that language-based models are able to encode simple multi-modal semantic information and process some CEs, excelling with those related to the dialogue history, whilst multi-modal models can use additional learning objectives to obtain disentangled object representations, which become crucial to handle complex referential ambiguities across modalities overall[1].

1 Introduction

In dialogue, people work together on a moment by moment basis to achieve shared understanding and coordination (Clark, 1996; Clark and Brennan, 1991; Goodwin, 1981; Healey et al., 2018; Mills, 2007). A key mechanism people use to repair misunderstandings when they occur is via meta-communicative, clarificational exchanges (CE): a clarification request (CR) followed by a response (see Fig. 1). CRs are a highly complex phenomenon: they are multi-modal (Benotti and Blackburn, 2021), highly context-dependent with different forms and interpretations (Purver, 2004; Purver and Ginzburg, 2004), and can occur at different levels of communication on Clark’s (1996) joint action ladder (Schlangen, 2004; Benotti and Blackburn, 2021). But while the crucial role of generating and responding to CRs in dialogue systems has long been recognised (San-Segundo et al., 2001; Rieser and Moore, 2005; Rodríguez and Schlangen, 2004; Rieser and Lemon, 2006), CRs still remain an understudied phenomenon (Benotti and Blackburn, 2021), especially in the context of recent successes in multi-modal dialogue modelling (Suglia et al., 2021; Wang et al., 2020; Chen et al., 2020; Guo et al., 2022; Das et al., 2017; Chen et al., 2021; Agarwal et al., 2020). There is recent work related to identifying when to pose a CR (Madureira and Schlangen, 2023; Zhu et al., 2021; Shi et al., 2022), but few evaluate the ability of models to process their responses (Gervits et al., 2021; Aliannejadi et al., 2021).

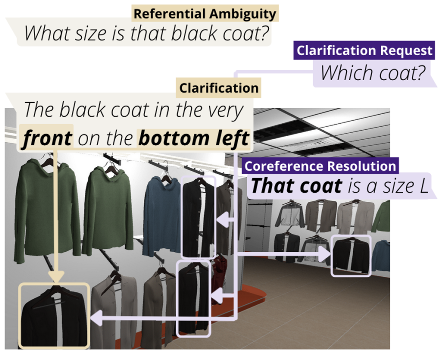

Figure 1: Example referential ambiguity and clarification in SIMMC 2.0 dialogues.

In this paper, we use CRs as a testbed for studying and evaluating different neural dialogue model architectures (see also Madureira and Schlangen (2023)). We focus on *referential CRs* occurring at level three of Clark’s (1996) action ladder: that of *understanding*. We provide a framework for evaluating how well multi-modal dialogue models are able to exploit referential CEs to resolve ambiguous referential descriptions. We use this framework to probe several state-of-the-art models proposed for the SIMMC 2.0 Challenge (Kottur et al., 2021) trained to resolve situated multi-modal

[1]We release our code and evaluation experiments AnonymisedURL

coreferences with CEs found in the SIMMC 2.0 dataset itself.

The results indicate that the ability of a model to exploit CRs to resolve referential ambiguities depends on the level of granularity of the model's cross-modal representations, i.e. how well information about different object attributes is represented. In particular, we find that the model that includes a training objective designed for predicting object attributes in a multi-task setup performs significantly better than the rest which was not optimised with this objective. This is in line with findings in Suglia et al. (2020) who show that having disentangled object representations (Bengio et al., 2013) allows models to better partition the search space of potential referents; and thereby better exploit effective object attributes in disambiguation.

2 Dataset

We used the SIMMC 2.0 dataset (Kottur et al., 2021), which is a collection of multi-modal task-oriented dialogues, where both the system and the agent are situated in the same virtual environment. The dataset dialogues have a high degree of ambiguity and use rich referring expressions due to the overlap of many similar-looking objects (e.g., 5 red t-shirts in view); dialogues with references to multiple and previously discussed objects (mean 4.5 unique objects referenced per dialogue, SD: 2.4); and changing points of view throughout dialogues with partially observed objects. Thus, referential ambiguities in both the visual and conversational contexts are common. Refer to Appendices A and C for more details. Furthermore, other common datasets do not contain coordination phenomena exhibited in SIMMC 2.0 (i.e. GuessWhat?! (de Vries et al., 2017)) or have a mixture of CRs which focuses solely on multi-modal referential ambiguities (e.g., Photobook (Haber et al., 2019)).

2.1 CRs in SIMMC 2.0

We focus on the clarificational sub-dialogues from the SIMMC 2.0 dataset. During the challenge, the dataset authors proposed several tasks, two of which are relevant here: Multi-modal Disambiguation (detecting whether the system has enough information to identify a unique object or is ambiguous) and Multi-modal Coreference Resolution (find the objects mentioned by the user). The dataset provides annotations that mark whether a turn is ambiguous or not, and which objects are referred to. Models were implicitly required to handle them as part of longer conversations, although the challenge did not explore clarifications in-depth. We choose this dataset for studying CRs for two main reasons: 1) it contains complex multi-modal dialogues with gold labels for referential ambiguity; 2) it focuses on tasks such as disambiguation and coreference resolution in multi-modal settings that are directly related with the problem of CR resolution.

2.2 Clarification Taxonomy

To evaluate how models handle CEs, we need to understand their ability to exploit fine-grained contextual information across modalities beyond level three of Clark's (1996) action ladder. Therefore, we derive a taxonomy of different types of clarifications depending on the information or *Disambiguating Property* exploited to resolve them: 1) **Individual Property**, such as object colour or state (i.e., *"The red jacket hanging"*); 2) **Dialogue History**, such as referring to previously mentioned objects (i.e., *"the one you recommended"*); and 3) **Relational**, such as position or their relation to other objects in the scene (i.e., *"the left shirt, next to the central rack"*).

These types are not mutually exclusive, and thus we often find that CRs are resolved with complementary information (i.e., *"The green dress on the right"*). Refer to Appendix C for discourse and taxonomy samples.

3 Experimental Setup

3.1 Clarification Extraction and Tagging

This section gives a summary of how we extracted the clarifications from the SIMMC 2.0 dataset using the gold annotations and tagged them using our taxonomy from Section 2.2.

When a turn is annotated as ambiguous, the system generates a CR (e.g., *"which one do you mean?"*). We label as **Before-CR** the user utterances preceding a CR (the user gave ambiguous information); whereas we label as **After-CR** the following user utterances that resolve the ambiguity. We obtain a subset of CEs (10% of all system turns are CRs) which we use for the analysis. Finally, we use a keyword-based method to tag the disambiguating properties exploited for clarifications (cf. Appendix B).

3.2 Metrics

We follow the SIMMC 2.0 evaluation protocol and measure coreference resolution performance using **Object F1**, derived as the mean of recall and precision for the predicted objects at each turn, as defined in (Kottur et al., 2021).

Along with object F1, we look at the difference in F1 between the turns before and after a clarification. Intuitively, a model that can process clarifications will improve after one, reflecting a higher F1 in the set of turns after a CR. Similarly, the turns before a CR may perform poorly, signalling confusion or uncertainty in general. We take this as the **Relative Delta** Δ to compare it across models.

3.3 Models

For our evaluation, we selected publicly available state-of-the-art models that took part in the SIMMC 2.0 challenge[2]. We give the relevant model details below, but please refer to original papers for additional architectural information.

Language-based We use two GPT-2-based (Radford et al., 2019) models: the Baseline ($Baseline_{GPT-2}$) from Kottur et al. (2021) (36.6% Object F1 ↑); and an improved version from one of the challenge participant teams (Hemanthage and Lemon, 2022), $GroundedLan_{GPT-2}$ (67.8% F1 ↑). Both models are similar and treat the task as a generation task, and are jointly trained with other goals in the challenge (coreference resolution, dialogue state tracking and response generation).

Vision-and-Language We take LXMERT-based (Tan and Bansal, 2019) model (ANON) ($VisLan_{LXMERT}$, 68.6% F1 ↑) that combines the images from the visual scenes and the dialogue to predict the coreferenced objects at each turn. It extracts object attributes from a Detectron2 model (Wu et al., 2019) to use as textual descriptions along with the visual features. For each object in the scene, it outputs a probability for the object being referenced in that turn and selects those above a threshold. This model is only trained on coreference resolution.

Language-Vision-and-Relational We use the model of the coreference challenge winner team (Lee et al., 2022) ($MultiTask_{BART}$, 74% F1 ↑), a BART-based model (Lewis et al., 2020) trained to handle all challenge tasks. A pretrained ResNet model (He et al., 2016) encodes each object along with its non-visual attributes, a learnable embedding that is later mapped to match the dimension of BART. The model is jointly optimised on multiple tasks, including several secondary tasks that enable learning disentangled object representations (Bengio et al., 2013) through object attribute slot prediction for each coreferenced object. The object location is also encoded through the bounding box information and a location embedding layer. Finally, the canonical object IDs are used to ground relations between the object locations, the visual and non-visual attributes.

4 Experiments

Referential Ambiguities Firstly, we explore whether referential ambiguities are an issue for models and if clarifications are thus needed. From the initial two rows of Table 1, we observe that, aside from the $Baseline_{GPT-2}$ model, all other models perform worse in turns **Before-CR** than when evaluating **All Turns**. This implies that indeed those utterances lack information to uniquely identify the referent objects, causing referential ambiguities for models and a lower object F1.

We also find that the F1 is higher in turns **After-CR** compared to turns **Before-CR** in all models but $Baseline_{GPT-2}$. This suggests that models can at least process clarifications in some cases. The $VisLan_{LXMERT}$ and $MultiTask_{BART}$ models even benefit with increased performance in **After-CR** turns compared to **All Turns**.

Regarding the surprisingly high scores for the $Baseline_{GPT-2}$ in turns **Before-CR** and low for **After-CR** , we suspect that it is due to the model exploiting linguistic phenomena along with smart use of previously mentioned objects and their canonical IDs, as explained in (ANON). The model's performance drops dramatically when is crucial to carry over cross-turn information and ground it in dialogue which is required **After-CR** .

Disambiguating Properties Using the CR taxonomy (cf. Section 2.2), we probe how models perform at exploiting different information with subsets of clarifications (bottom of Table 1).

All models but the baseline show a similar performance in **Before-CR** turns that exploit an Individual Property. $GroundedLan_{GPT-2}$ and $VisLan_{LXMERT}$ show a moderate F1 increase in the following **After-CR** turns, whereas $MultiTask_{BART}$ obtains a more substantial improvement (+11.3% Δ). Individual object properties in this dataset relate to concepts in the visual context which may be difficult to see or complex to understand beyond colour or shape (e.g., long sleeve or folded).

The $GroundedLan_{GPT-2}$ model implicitly encodes object attributes using a global object ID, which allows the model to learn latent information during training that carries over to evaluation

[2]Not all models were public and some had missing code or weights.

Model	$Baseline_{GPT-2}$			$GroundedLan_{GPT-2}$			$VisLan_{LXMERT}$			$MultiTask_{BART}$		
Split	**Before-CR**	**After-CR**	Δ	**Before-CR**	**After-CR**	Δ	**Before-CR**	**After-CR**	Δ	**Before-CR**	**After-CR**	Δ
All Turns	34.3 (.01)			67.8 (.01)			68.6 (.01)			74.0 (.01)		
CR Turns	36.4 (.01)	29.1 (.01)	-20.1%	64.8 (.01)	67.7 (.01)	+4.4%	65.7 (.01)	69.2 (.01)	+5.4%	66.9 (.01)	74.3 (.01)	+11.1%
Disambiguating Property												
Individual Property	35.4 (.02)	27.4 (.01)	-22.7%	65.0 (.02)	68.0 (.02)	+4.6%	65.1 (.02)	69.3 (.01)	+6.4%	68.0 (.02)	**75.7** (.01)	+11.3%
Dialogue History	47.6 (.04)	43.7 (.04)	-8.2%	81.7 (.03)	82.1 (.03)	+0.4%	81.7 (.03)	**84.6** (.03)	+3.5%	67.2 (.04)	75.7 (.04)	+12.6%
Relational Context	32.9 (.02)	25.0 (.02)	-24.1%	62.4 (.02)	63.7 (.02)	+2.1%	62.7 (.02)	65.0 (.02)	+3.7%	66.5 (.02)	**72.6** (.02)	+9.1%

Table 1: Evaluation results for models at handling CEs with different disambiguating properties. Measured in **Object F1** $\uparrow$ (SD) and **Relative Delta** Δ.

sets (i.e. <OBJ_256>). On the other hand, the $VisLan_{LXMERT}$ model encodes colours and shapes explicitly using textual descriptions (i.e. `blue hoodie`) and implicitly in the visual region of interest features, which explains the slightly higher performance in these particular clarifications. However, the vision module of $VisLan_{LXMERT}$ is not explicitly trained to detect complex properties, only attributes such as colours or shapes (i.e. blue hoodie), and is instead left to the visual features to represent this information.

The multi-task learning objectives of $MultiTask_{BART}$ help the model obtain more fine-grained disentangled representations than using vision alone which helps in resolving ambiguities related to individual properties. Suglia et al. (2020) suggests that exploiting explicit object attributes reduces the potential referents and thus may also lead to improvements in solving CRs.

$GroundedLan_{GPT-2}$ and $VisLan_{LXMERT}$ models perform well when the clarifications are related to the dialogue context. Their initial F1 (+81%) suggests that they are able to carry information across turns particularly well and may not even need a CR in these cases. Both models also improve in **After-CR** turns, with $VisLan_{LXMERT}$ reaching the highest score for this category. On the other hand, $MultiTask_{BART}$ improves its performance to 75.7% F1 (+12.6% Δ), but it does not display the same ability to exploit the linguistic context as the other models. This is likely due to the multi-task formulation involving specific loss functions which focus on visual and relational information only. Thus, the model obtains strong visual and relational object representations, whilst affecting the quality of BART's pre-trained language representations.

Relational clarifications seem to be the most difficult type to process for models, with the lowest F1 scores overall. The $MultiTask_{BART}$ model is able to exploit this information considerably better than the other models and improves by a +9.1% to 72.6%. This is an important strength of the model which extends its ability to encode visual attributes of the objects with information about the relationships between the objects in the scene. For instance, this model is able to capture the positions of the objects in the scene and how they relate to each other. The $VisLan_{LXMERT}$ model encodes positional information such as bounding box coordinates too, but it is not able to learn from them (ANON). This is justified by previous research by (Salin et al., 2022) that shows how multi-modal models struggle with concepts such as position, and that they rely on language bias instead.

5 Conclusion

Referential ambiguities are common in situated human conversations. We sometimes cannot fully understand or identify a referred object or event, and thus we engage in clarification exchanges to resolve the ambiguity. In this paper, we analyse how several state-of-the-art models treat clarifications in situated multi-modal dialogues using the SIMMC 2.0 dataset. We classify the types of clarifications by the disambiguating property exploited and then evaluate the models with subsets of the data.

We find that language-based models perform well, yet struggle to benefit from clarifications. On the other hand, vision seems to be an important (but not essential) addition for models, which helps processing multi-modal CEs. Paired with a strong dialogue context, these types of models can perform reasonably well and carry information across turns to better handle clarifications. Finally, encoding relations between objects and their locations, and using additional learning objectives to predict attribute slots seems the strongest architecture for models to handle CEs.

Based on these results, to create improved models that can resolve referential ambiguities in situated dialogues, we need *holistic object-centric representations* that contain information about attributes and properties (Seitzer et al., 2022), and that can *dynamically* change to reflect the information exchanges available in the dialogue context.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Luciana Benotti and Patrick Blackburn. 2021. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

Feilong Chen, Fandong Meng, Xiuyi Chen, Peng Li, and Jie Zhou. 2021. Multimodal incremental transformer with visual grounding for visual dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 436–446, Online. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

H. H. Clark and S. A. Brennan. 1991. *Grounding in communication*, pages 127–149. Washington: APA Books.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Felix Gervits, Gordon Briggs, Antonio Roque, Genki A. Kadomatsu, Dean Thurston, Matthias Scheutz, and Matthew Marge. 2021. Decision-theoretic question generation for situated reference resolution: An empirical study and computational model. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 150–158, New York, NY, USA. Association for Computing Machinery.

Charles Goodwin. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York.

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. 2022. Gravlbert: Graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 285–297.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science (topiCS)*, 10(2).

Bhathiya Hemanthage and Oliver Lemon. 2022. Global-local information-aware multimodal grounding with GPT for co-reference resolution. In *AAAI 2022 DSTC10 Workshop*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, and Kee-Eung Kim. 2022. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the CoDraw dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2303–2319, Dubrovnik, Croatia. Association for Computational Linguistics.

Gregory J. Mills. 2007. *Semantic co-ordination in dialogue: the role of direct interaction.* Ph.D. thesis, Queen Mary University of London.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, University of London.

Matthew Purver and Jonathan Ginzburg. 2004. Clarifying noun phrase semantics. *Journal of Semantics*, 21(3):283–339.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Verena Rieser and Oliver Lemon. 2006. Using machine learning to explore human multimodal clarification strategies. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 659–666, Sydney, Australia. Association for Computational Linguistics.

Verena Rieser and Johanna Moore. 2005. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 239–246, Ann Arbor. Association for Computational Linguistics.

Kepa Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, Barcelona, Spain.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11248–11257.

Ruben San-Segundo, Juan M. Montero, J. Ferreiros, R. Córdoba, and José M. Pardo. 2001. Designing confirmation mechanisms and error recover techniques in a railway information system for Spanish. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 136–139, Aalborg, Denmark. Association for Computational Linguistics.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 136–143, Boston. Association for Computational Linguistics.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. 2022. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute actions or ask clarification questions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2060–2070, Seattle, United States. Association for Computational Linguistics.

Alessandro Suglia, Yonatan Bisk, Ioannis Konstas, Antonio Vergari, Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon. 2021. An empirical study on the generalization power of neural representations learned via visual guessing games. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2135–2144.

Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon. 2020. CompGuessWhat?!: A multi-task evaluation framework for grounded language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7625–7641, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yue Wang, Shafiq Joty, Michael Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. 2020. VD-BERT: A Unified Vision and Dialog Transformer with BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3325–3338, Online. Association for Computational Linguistics.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. `https://github.com/facebookresearch/detectron2`.

Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. 2021. Self-motivated communication agent for real-world vision-dialog navigation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1574–1583.

A Additional Dataset Details

In the SIMMC 2.0 dataset (Kottur et al., 2021), the agents acts as the shopping assistant to a user in a virtual shop. It encompasses the domains of fashion and furniture over 11,244 dialogues and it was collected using a mix of dialogue self-play and crowd-sourcing. The dataset is originally split into `train/dev/devtest/test-std` with 65% / 5% / 15% / 15% of the dialogues respectively.

Each dialogue is complemented by images of the environment scene and rich metadata annotations. Some dialogue have multiple scene images with partially overlapping sets of objects, requiring models to carry over information from previous viewpoints. On average, dialogues have 5.2 utterance pairs (*user-assistant*) and associated scenes have a mean of 27.6 objects, with some of them reaching up to a maximum of 141 items.

Since the gold data from the `test-std` split is not available, we used the `devtest` data for our evaluation. Thus, some of the model object F1 scores may differ from their respective papers by a few decimals.

B Additional CR Details

B.1 Clarification Tagging Method

The algorithm for CR tagging is based on manual annotations using the dev set, and then creating a set of keywords and regexes that would automatically find the disambiguating property used. **Individual Properties** include mentions of: colour (*blue*), object types (*jacket*), style (*floral*), brand names (*Yogi Fit*), states (*folded*) and other (*long-sleeve*). The metadata provides all of this information that we use as keywords. Other information such as **Relational Context** was based on positional keywords (*left, top*) and relational with objects (*next to*) or the scene (*farthest*). **Dialogue History** was based on linguistic cues and the use of common structures (i.e. *"...in my cart"*, *"you mentioned"*). We left some CRs as unclassified ($<0.06\%$ of the data) because they do not provide any meaningful additional information or are out of scope (i.e., *"What is that lamp made of? - Sorry, which one? - I'm not sure, I think it's a lamp"*).

During tagging, we applied the algorithm to each clarification, including both the system CR and the user response, as the information is sometimes scattered across turns (i.e., see SYS CR in Dialogue Sample 1). We manually verified this method and designed unit tests to ensure correct classification as the tagging algorithm increased in complexity.

C Corpus Examples

Dialogue Sample 1

USR *Hello, do you have any jackets for me to look at?*
SYS *Sure, what do you think of the light grey jacket hanging up high at the back left?*
USR *Do you have anything with a similar size range to the black sweater beside the light grey jacket?*
SYS *Sorry, I don't have anything similar to that*
Before-CR USR *What size is that sweater anyways?* [Referential Ambiguity]
CR SYS *The black one?* [Clarification Request]
After-CR USR *Yes exactly* [Clarification]
`- Tags in CE: Individual Property`

Dialogue Sample 2

USR *I need some help picking out a jacket.*
SYS *Do you like the look of that solid black one hanging in the bottom row up front on the right?*
USR *I do like the look of that one. Will you please put that in my cart?*
SYS *Sure! I'll add that black jacket to your cart right now.*
USR *Now could we look for a grey coat?*
SYS *There's a grey coat hanging directly above the black jacket you like.*
Before-CR USR *Actually first I'd like to know which brand makes that jacket. If you have any data on what other customers think of it, that'd be good to hear too.* [Referential Ambiguity]
CR SYS *Sorry, which jacket specifically are you curious about?* [Clarification Request]
After-CR USR *The black one up front next to the skirt* [Clarification]
`- Tags in CE: Individual Property, Relational`
SYS *That jacket is a Yogi Fit original, and other buyers have rated it a 4.0 on average*
USR *Are there any super ritzy options with stellar reviews?*
SYS *I don't see any, sorry. Is there something else I can help you with?*
USR *I'd like to find a dress in extra large.*
SYS *I have a feeling you'll like the dresses on this first rack up front and center. Would you wear the purple one on the left, the brown one on the right, or the grey one in front?*
Before-CR USR *Woops, I didn't even ask you what sizes that jacket came in.* [Referential Ambiguity]
CR SYS *Which jacket do you want to know the sizes of?* [Clarification Request]
After-CR USR *That black one that I had you put in my cart* [Clarification]
`- Tags in CE: Individual Property, Dialogue History`

Sample dialogues with CEs from the SIMMC 2.0 dataset.

Split	Mean Candidate Objects (SD)
All Turns	2.98 (5.01)
CR Turns	5.18 (5.48)
Disambiguating Property	
Individual Property	5.52 (5.69)
Dialogue History	4.57 (4.82)
Relational Context	5.78 (5.91)

Table 2: Statistics about the level of ambiguity in SIMMC based on candidate objects of the same type.