GENERAL INSTRUCTIONS

- **Authors: Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.**
- **Authors: We cannot accept new source files as corrections for your article. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.**
- **Authors: If you need an invoice or have any other billing questions, please contact apcinquiries@ieee.org as they handle these requests**

QUERIES

Q1. Author: Please provide the correct ORCID for the authors Alejandro Santorum Varela and Svetlana Stoyanchev
Q2. Author: Please confirm or add details for any funding or financial support for the research of this article.
Q3. Author: Please provide the name of the corresponding author.
Q4. Author: Please check and confirm whether the authors affiliations in the first footnote are correct as set.
Q5. Author: Please update Refs. [3], [4], and [30].
Q6. Author: Please provide page range for Refs. [5], [19], and [29].
Q7. Author: Please provide the volume number/page range for Ref. [18].
Q8. Author: Please provide complete bibliographic details for Ref. [23].

Alejandro Santorum Varela - alejandro.santorum@gmail.com (Corresponding author)
Svetlana Stoyanchev - svetlana.stoyanchev@toshiba.eu
Simon Keizer - simon.keizer@toshiba.eu
Rama Doddipatla - rama.doddipatla@toshiba.eu
Kate Knill - kate.knill@eng.cam.ac.uk

In this document we address the **queries prior to the final publication of the manuscript** "Entity Resolution in Situated Dialog With Unimodal and Multimodal Transformers" with DOI 10.1109/TASLP.2023.3304468.

Q1. - The ORCID of the requested authors are:

- Alejandro Santorum Varela with ORCID 0000-0003-3555-9501 (this is actually set correctly in the proof).
- Svetlana Stoyanchev has no ORCID.

Q2. - We would like to explicitly add that the Toshiba Cambridge Research Laboratory provided the computing resources for the investigation.

Q3. - The corresponding author is **Alejandro Santorum Varela** with email alejandro.santorum@gmail.com

Q4. - The names and affiliations are ok, however, some emails got recently updated and it would be great to update that in the manuscript too. In particular, emails at @crl.toshiba.co.uk are now at **@toshiba.eu**

Q5. - The updated latex code for references [3], [4] and [30] are (respectively):

```
@article{uniter_solution,
        author = {Yichen Huang and Yuchen Wang and Yik{-}Cheung Tam},
        title = {UNITER-Based Situated Coreference Resolution with Rich Multimodal
    Input},
        journal = {CoRR},
        volume = {abs/2112.03521},
        year = {2021},
        url = {https://arxiv.org/abs/2112.03521},
        eprinttype = {arXiv},
        eprint = {2112.03521},
        timestamp = {Mon, 13 Dec 2021 17:51:48 +0100}
}

@article{heriot_watt_uni,
        title={Exploring Multi-Modal Representations for Ambiguity Detection \&
    Coreference Resolution in the SIMMC 2.0 Challenge},
        author={Javier Chiyah-Garcia and Alessandro Suglia and José Lopes and Arash
    Eshghi and Helen Hastie},
        year={2023},
```

```
      journal={CoRR},
      eprint={2202.12645},
      archivePrefix={arXiv},
      primaryClass={cs.CL}
}

@article{kakao_paper,
      author    = {Joosung Lee and
                   Kijong Han},
      title     = {Multimodal Interactions Using Pretrained Unimodal Models for {SIMMC}
                   2.0},
      journal   = {CoRR},
      volume    = {abs/2112.05328},
      year      = {2021},
      url       = {https://arxiv.org/abs/2112.05328},
      eprinttype    = {arXiv},
      eprint    = {2112.05328},
      timestamp     = {Tue, 14 Dec 2021 14:21:31 +0100}
}
```

Q6. - The page ranges for refs [5], [19] and [29] are:

- [5]: DK/NA
- [19]: pages={41--51},
- [29]: pages={59--66},

Q7. - Updated latex entry for ref [18], including volume, is :

```
@article{mmcr_correlation_between_spaces,
      author    = {Qibin Zheng and
                   Xingchun Diao and
                   Jianjun Cao and
                   Xiaolei Zhou and
                   Yi Liu and
                   Hongmei Li},
      title     = {Multi-modal space structure: a new kind of latent correlation for
                   multi-modal entity resolution},
      journal   = {CoRR},
      volume    = {abs/1804.08010},
      year      = {2018},
      url       = {http://arxiv.org/abs/1804.08010},
      eprinttype    = {arXiv},
      eprint    = {1804.08010},
      timestamp     = {Wed, 04 Dec 2019 13:45:46 +0100}
}
```

Q8. - The complete bibliographic information for ref [23] is:

```
@article{unity_tech_environment,
        title={A history of the {U}nity game engine},
        author={Haas, John K},
        year={2014},
        publisher={Worcester Polytechnic Institute}
}
```

Entity Resolution in Situated Dialog With Unimodal and Multimodal Transformers

Alejandro Santorum Varela, Svetlana Stoyanchev, Simon Keizer, Rama Doddipatla, and Kate Knill, *Senior Member, IEEE*

***Abstract*—In this work we address the entity resolution task for situated multimodal dialog investigating how a unimodal approach, which uses only textual information as input (representing visual attributes as text), compares to a multimodal system, which processes both text and visual information. We analyze two of the top performing models presented in the Tenth Dialog Systems Technology Challenge and propose modifications that enhance their performance on the multimodal coreference resolution task. We evaluate these approaches on in- and out-of-domain settings by training the models on the fashion domain and testing on the furniture domain, and vice-versa, to assess the generalizability of the models. Through systematic analysis, we show that while both systems achieve similar performance on in-domain scenarios, the multimodal system generalizes better to out-of-domain settings. A combination strategy of enhanced unimodal and multimodal systems achieves F1 = 0.80 (5% absolute gain compared to the best performing system). Finally, human performance on the same task is evaluated on a small subset, suggesting that the performance of the current automatic models is on par with people on this task.**

***Index Terms*—BART, DSTC10, multimodal coreference resolution, SIMMC2.0, transformers, UNITER.**

I. Introduction

In CONVERSATIONS speakers often refer to real world entities in order to convey information to their communication partner. The same entity may be referred to using different referring expressions throughout the conversation. A new entity may be introduced with a specific noun phrase mentioning the entity by name or properties, e.g *'The Rice Boat is a nice Indian restaurant'*. Subsequent utterances may then 'co-refer' to that same entity, e.g., using an under-specified noun phrase *'Is **that restaurant** in the city centre?'*, or just a pronoun '*Is **it** in the city centre*?'. Coreference or, more generally, entity resolution is the Natural Language Processing task of identifying the referent (a previously introduced entity, e.g. a particular restaurant) given a referring expression (e.g., *'it'*).

AI agents have to interpret references to objects to communicate successfully with the user. In previous research, entity resolution has been addressed for spoken or chat-based dialog agents, such as information search dialog systems [1]. References in spoken or chat-based dialog systems are resolved based on language features. In spoken dialog systems prosodic information may also be helpful for resolving references [2].

In contrast to the language-only setting of a typical single-modality dialog system, situated AI agents, such as robots and virtual humans, are co-present with the user in the same physical or virtual environment. In addition to the language modality, situated agents also have access to a visual scene through the video camera of the interface. In this co-present setting, it is natural for both the user and the system to refer to visible object properties such as color, shape, or relative position to identify an object, e.g. *'the red blouse on the left'*. Entity resolution in situated systems involves determining which physical object the user is referring to in an utterance, if any. In such systems successful entity resolution requires the use of both language and visual channels.

The Situated and Interactive Multimodal Conversation Track at *The Tenth Dialog System Technology Challenge* (DSTC10) involved four subtasks: 1) detection of ambiguous utterances, 2) coreference resolution, 3) state tracking, and 4) response generation, using the SIMMC2.0 dataset.[1] The dataset consists of semi-automatically collected customer-agent dialogs situated in two virtual environments: a clothing shop and a furniture shop. In this work, we use this dataset to analyze the approaches to multimodal coreference resolution or identifying the objects (furniture or clothing items) that the user is referring to. We chose to focus on the coreference sub-task which is essential for any situated AI agent that has to interpret natural language commands.

One of the approaches to multimodal entity resolution in the challenge involved applying multimodal transformers pretrained on both visual and text information, such as captioned images [3], [4]. In this approach, a single multimodal transformer model processes both visual and text information to determine which objects in the scene were referred to by the user. Another approach involved a unimodal text-only transformer that processed features representing visual attributes of the objects in a scene in the form of text. Interestingly, the top performing method in the competition followed the latter approach [5].

Manuscript received 14 November 2022; revised 19 April 2023; accepted 29 July 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Seungwhan Moon.

Alejandro Santorum Varela and Kate Knill are with the Department of Engineering, Cambridge University, CB2 1TN Cambridge, U.K. (e-mail: alejandro.santorum@gmail.com; kate.knill@eng.cam.ac.uk).

Svetlana Stoyanchev, Simon Keizer, and Rama Doddipatla are with the Speech Technology Group, Toshiba Europe Limited Cambridge Research Laboratory, CB4 0GZ Cambridge, U.K. (e-mail: svetlana.stoyanchev@crl.toshiba.co.uk; simon.keizer@crl.toshiba.co.uk; rama.doddipatla@crl.toshiba.co.uk).

Digital Object Identifier 10.1109/TASLP.2023.3304468

[1] https://github.com/facebookresearch/simmc2

In this work, we analyze and extend two of the top performing systems from the DSTC10 competition: 1) the multimodal UNITER-based [3] and 2) the text-only BART-based [5] models. In addition to using standard data splits of the dataset, which combines the two domains, we evaluate the models in out-of domain settings to determine their ability to generalize. We show that the multimodal transformer approach is better at generalizing to unseen scenes than the unimodal transformer approach. Additionally, we evaluate the effect of encoding non-visual and visual object attributes within the transformer-based systems and analyze the models' performance on newly mentioned objects (which requires scene understanding) and on previously mentioned objects (which may be resolved using the dialogue context). We improve the models by adding an auxiliary task head that identifies the number of referred items in the last user turn. Finally, we show that combining the two approaches further improves the performance on this task to F1 $= 0.80$ on the SIMMC2.0 dataset.

The contributions of this work include:

- Evaluation of the unimodal and multimodal transformer approaches in an out-of-domain setting to determine the generalization ability of the models.
- Extending each of the models and combining them to achieve an improved performance over the proposed models on the coreference resolution task.
- Human performance assessment on the SIMMC2 dataset showing the models already perform on par with people.

In Section II, we recap the prior related work, Section III describes the SIMMC2.0 dataset used in the experiments. In Section IV we present the baseline methods and the proposed modifications. All the experiments and evaluation results are shown in Section V, and we perform error analysis of the models in Section VI. We conclude in Section VII summarizing and highlighting the main contributions of this work.

II. Related Work

Coreference/reference/entity resolution and grounding are the terminology used in the literature to describe the task of identifying which entity is being referred to. An entity may be a physical object, a grouping of objects (a table, chairs, cars, etc.) or an abstraction (thought, conclusion, etc.). Reference resolution is an essential part of language understanding. While *coreference resolution* assumes multiple references to the same entity, *reference resolution* more generally describes resolving each occurrence of a reference.

Reference resolution in text and dialog has been extensively addressed in past research [6], [7], [8], [9], [10], as well as annotation techniques for coreference [11]. [12] tackles coreference resolution of noun phrases in unrestricted text, and [13] presents an approach to pronoun resolution based on syntactic paths. In text or language-only dialog, entities referred to are present in the mind of a reader or a dialog participant but not in a shared visual space. In a situated dialog, such as human-robot interaction, discussed entities include actual items present in the physical space shared by the dialog participants. Resolving these object mentions requires the use of both text and visual features [14], [15], similarly to the other multimodal tasks, such as Visual Question Answering [16].

TABLE I
SIMMC2 Dataset Statistics and Metadatas

	Fashion	**Furniture**	**All**
Dataset statistics			
No. dialogs	$\sim$ 7.2K	$\sim$ 4K	$\sim$ 11.2K
No. objs. per scene	$\sim$ 32	$\sim$ 11	$\sim$ 20
Metadata			
Visual attributes	assetType, pattern and sleeveLength	-	type and color
Non-visual attrs.	size	materials	brand, price and customerReview

In recent years, multimodal entity resolution has been extensively addressed due to the advances in robotics. As in other Natural Language tasks, the use of Large Language Models based on *Transformers* architectures [17] has been shown effective for multimodal coreference. [18] proposes a semi-supervised learning model to correlate embedding space structures of each modality in coreference resolution. In situated settings, gestures play an important role for entity resolution. In [19], the authors analyze which of the gestures provides extra information for the dialog and in [20], human perception of automatically generated referring expressions is examined.

The Ninth and the Tenth Dialog Systems Technology Challenges included a track for situated multimodal conversational AI [21], [22] set in a simulated virtual shop environment with coreference resolution as one of the tasks. In this work we investigate two of the approaches to entity resolution participating in the DSTC10 [3], [5], analyzing them in a cross-domain setting and on mentioned and new object references.

III. Data

The SIMMC2 dataset presents task-oriented dialogs between a virtual shop assistant and a user. Each dialog is situated in a virtual fashion or furniture shop and is associated with one or more virtual shopping scenes simulated using a VR environment [23]. The dialogs were generated semi-automatically by first simulating the dialog structure and then manually editing the content of the utterances [22].

The dataset contains 7.2 K and 4 K dialogs in the fashion and furniture domains respectively. Fashion scenes have on average around 32 objects per scene while the furniture scenes have 11, making the fashion domain more challenging for the entity resolution task. The objects in each scene are annotated with their bounding boxes and IDs, which are linked to the metadata containing information about the objects in the form of visual and non-visual attributes (see Table I).

Each user and system utterance is annotated with the IDs of the objects that are mentioned in this utterance. For example, the referring expression in the system's utterance *'I have these two pink ones, one on top and one on the bottom, do you like them?'* identifies the objects #1 and #45 in the corresponding scene (see Fig. 1). The goal of the entity resolution system is to identify which objects the current user utterance is referring to, given the dialog context and the object IDs referenced in it, the visual

Fig. 1. Example of a SIMMC2 scene situated in a fashion shop.

scene (including the scene image and the object locations), and the metadata of the objects.

SIMMC2.0 is split into three sets: train (64%), dev (5%) and devtest (15%). Following the competition rules, we train the models on the training set, tune the parameters on the dev set, and report results on the devtest set.

IV. Method

The approaches to coreference resolution used by the participants of the competition differed in the methods of processing visual information. One approach was to use pretrained multimodal transformers, such as LXMERT [24], or UNITER [25], to jointly encode the image information and the dialog history [3], [4]. Alternatively, the input was processed by a text-only transformer model, such as BART [26], BERT [27], or GPT2 [28]. Visual features were represented as text extracted from an image with an off-the-shelf image processing model [29] or from the metadata [5], [30].

The scores achieved by the challenge participants on the coreference resolution task varied from $F1 = 0.52$ to $F1 = 0.75$. The multimodal methods that processed images generally achieved higher scores, indicating that the use of visual input cannot be fully replaced by the metadata information. However, one of the text-only transformer-based approaches that did not process images was very effective, with a reported score of $F1 = 0.75$ [5]. In this method, the authors pre-train an object encoder with a contrastive learning technique similar to [31], but using object IDs instead of images. In this work, we compare this unimodal BART-based approach with the top performing multimodal UNITER-based approach [3]. In the rest of this article we refer to them as *BART-based* and *UNITER-based* methods respectively or simply BART/UNITER.

A. UNITER-Based Model [3]

UNiversal Image-TExt Representation (UNITER) [25] is a multimodal encoder based on a transformer architecture pre-trained over four image-text datasets (COCO [32], Visual Genome [33], Conceptual Captions [34], and SBU Captions [35]). It is aimed at enabling downstream tasks with joint multimodal embeddings. The UNITER-based model for the multimodal coreference resolution task (see Fig. 2) takes as input the dialog history including the last user utterance (U), the encoding of each of the objects in the scene (O), and the image of the overall scene (S). The dialog history is encoded using the pre-trained BERT [27] encoder. The cropped images of each object in the scene as well as the full scene image are encoded using the visual pre-trained CLIP model [31]. Next, each object in the scene is encoded with the combined information consisting of its CLIP-generated image embedding, 3D object coordinates, non-visual metadata attributes, binary flags indicating whether the object is present in the scene and whether it was previously mentioned, and object ID. The output of UNITER (H) provides a hidden multimodal representation for each object within the dialogue context (H_obj), which is passed into a dense layer producing a logit Z for each object in the scene. The logits are transformed into probabilities (Y) using the Sigmoid function $\sigma(Z)$, which are then used to classify each object with a binary label indicating whether the object is present or absent in the last user's reference. The input encoding, dense layer, and sigmoid layer steps are shown in (1).

$$\begin{aligned} H = \langle H\_dial, H\_obj, H\_s \rangle &= \text{UNITER\_Enc}(U, O, S) \\ Z &= \text{Dense}(H\_obj) \\ Y &= \sigma(Z) \end{aligned} \tag{1}$$

UNITER parameters are fine-tuned while the object embedder, scene embedder and dense layer parameters are trained on the binary coreference annotations of the SIMMC2.0 dataset. For full details of the model please see the original article [25].

B. BART-Based Model [5]

In the proposed solution for the DSTC10 competition, an encoder-decoder BART model addresses all four tasks of the SIMMC2.0 challenge at the same time: disambiguation, coreference resolution, state tracking, and response generation. While it was shown that multi-task learning can benefit each individual task, we observe that for the SIMMC2.0 dataset learning multiple tasks does not benefit the performance on coreference resolution. Hence, in our experiments we use a simplified model with a single coreference prediction task and omit the BART decoder used for the response generation task.

Unlike the UNITER-based approach, the BART-based approach does not directly process images of objects or scenes. Instead, the dialog history, object IDs, and object locations are encoded using a pre-trained BART model (see Fig. 3). In parallel with training the model for the coreference tasks, a contrastive encoder is trained with a contrastive learning objective [31] by maximizing the cosine similarity between the object ID embedding and text attributes embeddings of the corresponding object and by minimizing the cosine similarity between the ID embedding and the attribute embeddings of other objects. This approach allows to associate objects in a scene with the corresponding visual metadata descriptions without having to use images or visual attributes at inference time.[2] The embedded objects IDs, that are expected to encode visual and non-visual attributes of each object, are combined with the embeddings of the corresponding object

[2]The use of visual attributes was disallowed by the competition rules.

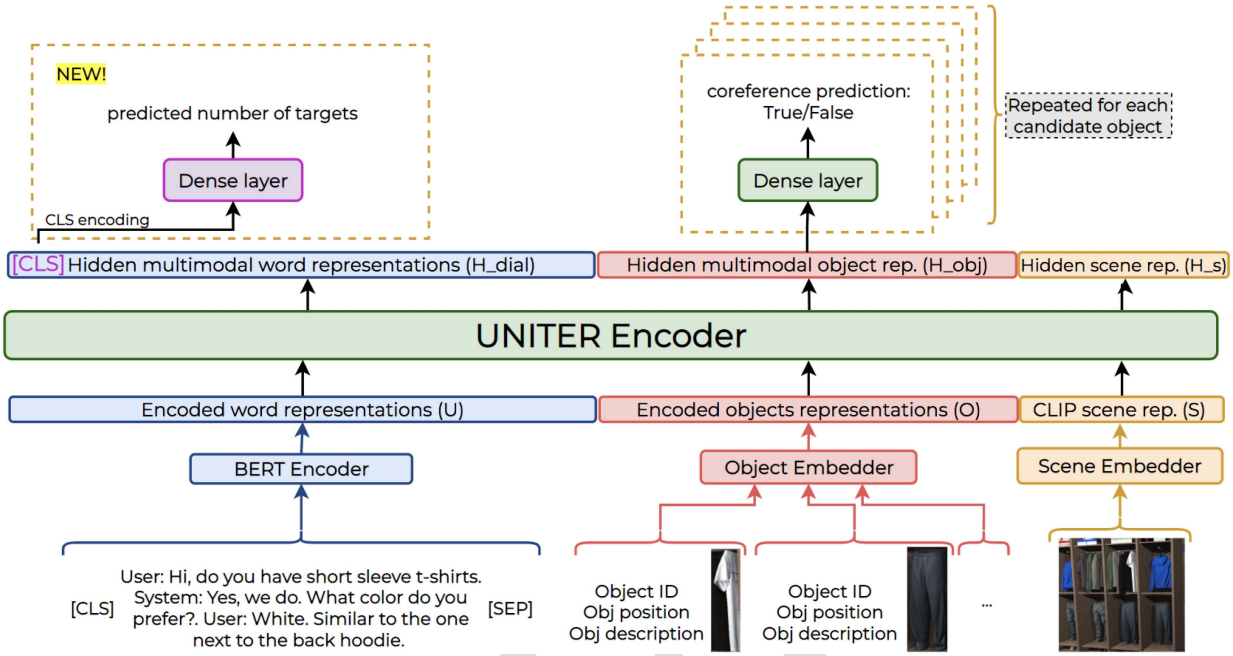

Fig. 2. UNITER-based model for coreference resolution [3].

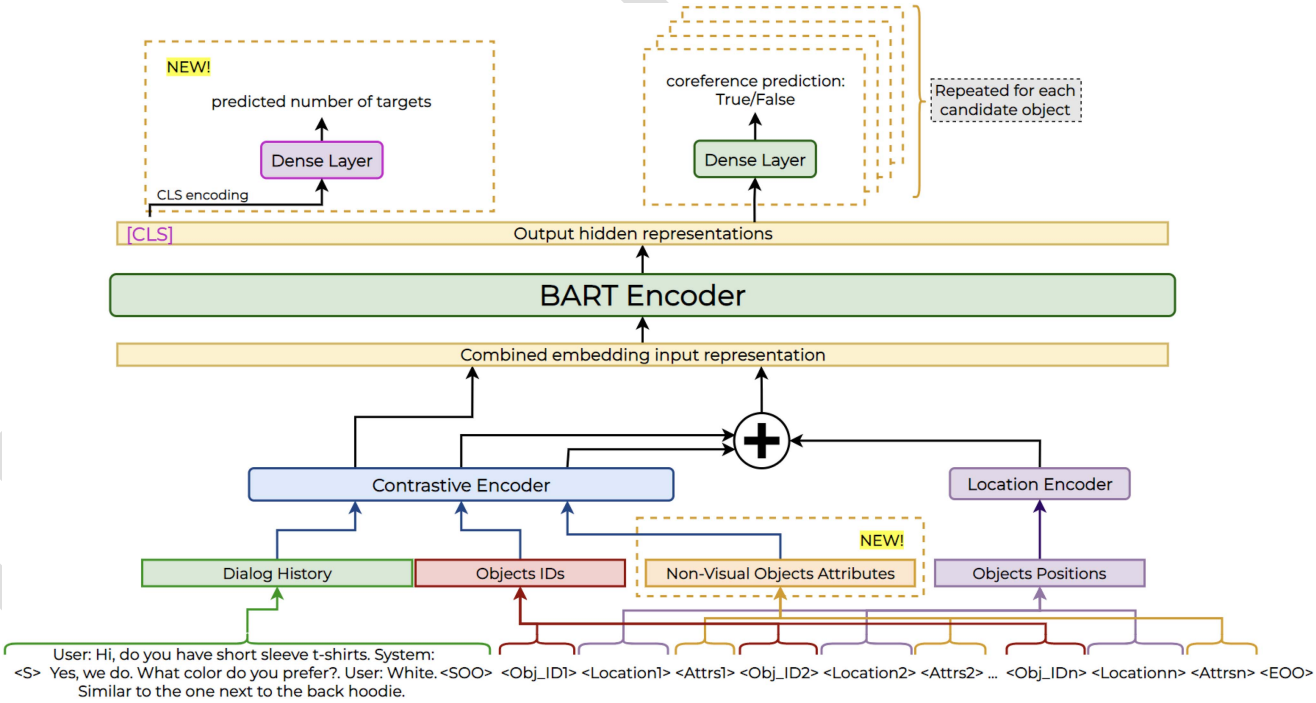

Fig. 3. BART-based model for coreference resolution [5].

positions. Then, the result is joined with the embedded word sequence of the dialog history and passed to the BART encoder.

Similarly to the UNITER-based model, the output of the BART encoder is passed into a dense layer followed by a sigmoid that outputs a binary prediction for each object (see (1)). The BART encoder parameters are fine-tuned while the location encoder and the coreference dense layer parameters are trained on the binary coreference resolution task using cross entropy loss. As an auxiliary task, the original model also predicted *empty coref*, a binary value for the whole utterance (0 if the utterance contains at least one entity reference and 1 otherwise). Our ablation study showed that this did not improve the performance, but rather than removing it, *empty coref* prediction was replaced with prediction of the number of referred objects (see Section IV-C).

TABLE II
EXAMPLE SIMMC2.0 DIALOG BASED ON A FASHION SHOP SCENE IN FIG. 1

Speaker	Utterance	Objects Mentioned
User	What sweaters do you have with good ratings?	
System	I have these two pink ones, one on top and one on the bottom, do you like them?	1, 45
User	How much is the bottom one and who makes it?	PREDICT: 45

TABLE III
EXAMPLES OF UTTERANCES AND THE NUMBER OF ITEMS THEY REFER TO

"How about giving me a look at some hats I'd like?"	0
"The one next to the black hoodie."	1
"I want the blue coat on the right and the red shirt on top."	2
"How about these two jeans and the red blouse on the rack?"	3

The contrastive learning procedure for encoding object IDs differentiates the BART-based approach from the other approaches in the DSTC10 challenge. While this method is effective for detecting objects whose IDs were used in pre-training, it cannot generalize to unseen objects, as confirmed by our experiments.

C. Predicting the Number of Referred Objects

In both UNITER-based and BART-based models, the presence of a reference to an object is predicted for each object in a scene individually, without taking into account the other objects. We observe that the model can *underpredict* (predicting too few objects) or *overpredict* (predict too many objects). To address these errors we propose a postprocessing method which relies on knowing the number of objects that the user's utterance refers to.

By analyzing user utterances independently of the dialog and scene context, we can detect the number of objects the user refers to. For example, the first utterance in Table III is a general question and does not refer to any specific objects. The second one refers to one object as it mentions *'the one'*. As the third utterance contains a conjunction *'and'*, we can infer that it refers to two objects. We use an auxiliary learning task to predict the number of objects referred to by the user, $N \in \{0, 1, 2, > 2\}$, using multi-class classification. This approach is similar to the one used by the original BART-based model, which, among other tasks, predicts whether the user's turn referred to an item using an auxiliary learning task. Instead of making a binary prediction, we use a multi-class prediction.

The predicted number of objects N is used to heuristically post-process the model's outputs: if N is larger than the number of initially predicted target objects, we force the model to make more predictions, using the estimated probabilities from the *softmax* function applied to the *dense* layer's output.[3]

[3]Users rarely referred to more than 3 objects in the data, hence we limited the classification to four classes.

TABLE IV
SIMMC2.0 DATA SPLITS USED FOR THE EXPERIMENTS

Dataset Name	Domain	∼ objs/ scene	Selection method
ALL-train	Both	24	SIMMC2.0 *train* split
ALL-test	Both	25	SIMMC2.0 *devtest* split
FASH-14K	Fashion	33	Subset of ALL-train&ALL-test excluding `held-out` scene and the data from **FASH-9K-ID**.
FURN-12K	Furniture	10	Subset of ALL-train&ALL-test
FASH-6K	Fashion	31	Subset of ALL-test
FURN-2K	Furniture	9	Subset of ALL-test
FASH-9K-ID	Fashion	33	Random select from ALL-train&ALL-test excluding `held-out` scene
FASH-9K-IDHO	Fashion	24	Subset of ALL-train&ALL-test set in `held-out` shop.
FURN-9K-OOD	Furniture	9	ALL-train & ALL-test
ALL-100	Both	26	100 randomly selected examples from ALL-test

`held-Out` is a randomly selected scene from fashion domain (`cloth_store_1498649_woman`).

V. EXPERIMENTS

We present experimental results on the SIMMC2.0 dataset for the multimodal UNITER-based and unimodal BART-based approaches in the following settings: the *in-domain* setting, where train and test data contain the same set of scenes, the *in-domain-held-out* setting, where no one scene appears in both train and test data, and the *cross-domain* setting with different domains in train and test data. We also describe how the proposed modifications to the models improve reference resolution performance. In addition, we explore the use of visual attributes in the BART-based model. Finally, the models' performance is compared with human performance on a randomly selected subset of examples.

The official *train*, *dev*, and *devtest* splits of the SIMMC2.0 dataset contain both fashion and furniture domains. Because the same scenes from each domain occur across all data splits, we refer to the evaluation of the models trained and tested on the official data split as the *'in-domain'* experimental condition. To evaluate the models on unseen scenes in the same (*'in-domain-held-out'*) or different (*'cross-domain'*) domain, we create the data splits described in Table IV.

A. In-Domain Evaluation

Table V shows the reported and replicated (*Ours* column) performance for the baseline and the original versions of UNITER-based and BART-based models trained and tested on the official *train/devtest* split. Our replicated results for each of the models closely match the results reported by the corresponding papers [3], [5], [22].

Both of the original models use scene-specific object IDs and the BART-based model also uses global IDs as input during training and inference. The reliance on the IDs may prevent the models from generalizing. A model that learns references to

TABLE V
EVALUATION OF MULTIMODAL COREFERENCE RESOLUTION ON IN-DOMAIN CONDITION

Model	Description	Reported (F1)	Ours (F1)
	Original		
GPT-2	[22] (baseline)	0.366	0.381
UNITER	[3]	0.728	0.726
BART	[5]	0.743	0.742
	Modified		
UNITER	Remove object IDs	-	**0.758**
BART	Loss: coref task + *empty coref*	-	0.748
BART	Loss: coref task	-	0.748
BART	Loss: coref task + #obj. heur.	-	0.752
BART	Loss: coref task + non-visual	-	0.760
BART	Loss: coref task + #obj. heur. + non-visual	-	**0.763**
BART	Loss: coref task + #obj. heur. + non-visual + visual	-	**0.775**†

All models are trained and tested on the official train and devtest data splits of SIMMC2.0. The model marked with † is not legal for DSTC10 as it uses visual metadata attributes.

objects based on their IDs rather than visual properties would not be able to detect a reference to a new object that was not seen during training. The BART-based model's use of object IDs is inherent to the method, as it pre-trains a contrastive encoder that maps global object IDs to the visual and non-visual features (see Section IV-B).

To test whether the UNITER-based model relies on the (scene-level) object IDs, we removed them from the input by excluding them from the object embedding. Interestingly, this modification results in an increase of the F1 score from 0.726 to 0.758, a 3% absolute increase. This result indicates that the UNITER-based model does not rely on the object IDs and therefore has the potential to generalize across domains.

As multitask learning has been shown to improve the performance of an individual task [36], the original BART-based model uses multi-task learning to jointly train a model for disambiguation, coreference resolution, state tracking, and response generation. In this work, we focus on coreference and evaluate a single-task variant of the BART-based model by modifying the loss function to only include the coreference loss. Contrary to the expectation, using a single-task loss function does not decrease the performance: the F1 score slightly increases from 0.742 to 0.748. Removing the *empty coref.* auxiliary task does not affect the result, but using the new proposed auxiliary task of predicting the number of objects with the corresponding heuristics (see Section IV-C) increases the performance to 0.752.

Next, we investigate using `metadata` features as input to the models (see Table I). The metadata includes a list of non-visual and visual attributes for each object in the scenes that can be referenced during the dialog (see Table I). Only non-visual attributes are allowed to be used at inference time in the DSTC10 competition, as the visual attributes are expected to be recognized from the scene image. The original UNITER-based model includes non-visual attributes as input to the object encoder. While the original BART-based model learns the object IDs embedding using contrastive learning, it does not take non-visual features as input during inference. We hypothesize that in addition to pretraining with visual and non-visual attributes, adding visual and non-visual descriptions of objects in the scene directly as input may help the model at inference time. We show including non-visual object features improves the BART model performance to $F1 = 0.763$ and including visual features further improves it to $F1 = 0.775$.

B. Evaluation Across Scenes and Domains

An entity resolution model for a situated multimodal interface should, ideally, generalize to new in-domain settings (*in-domain-held-out* experimental condition) as well as new domains (*cross-domain* experimental condition) by relying on generic visual attributes, such as color, relative position, and object type. While objects of different type may appear in scenes of different domains (furniture vs. clothing items), pre-training a multimodal transformer on a large generic dataset may give the UNITER-based model generalization ability. However, BART-based model relies on object ID embeddings that may prevent it from generalizing across domains. To assess the models' ability to generalize across scenes and domains we train and evaluate both models on different data subsets.

We first assess cross-domain generalization by comparing *in-domain* and *cross-domain* evaluation conditions (see top part of Table VI). The *in-domain* models are trained on the ALL-train split, which contains both fashion and furniture examples, and the *cross-domain* models are trained on a subset with only one domain (FASH-14 K or FURN-12 K). We observe that the performances of both UNITER and BART-based models is higher when tested on the furniture than on the fashion domain. The two models in the *in-domain* setting achieve an $F1$ of $0.843/0.861$ on furniture and $0.736/0.731$ on fashion respectively. This difference can be explained by the higher complexity of the fashion domain scenes with an average of 31 objects in fashion scenes and only 9 in the furniture scenes (see Table IV). For both models the performance is lower in the *cross-domain* setting than in the *in-domain* setting. However, the UNITER-based model achieves higher performance than the BART-based model in the cross-domain condition on each of the domains. In particular, in the *cross-domain* condition tested on the fashion domain, the UNITER model performance is $F1 = 0.425$ while the BART-based model scores $F1 = 0.194$. This result shows that the UNITER-based model, which extracts object information from images, has a better potential of generalizing to a new domain than the BART-based model, which fails to generalize due to its reliance on object IDs.

Next, we assess generalization to novel scenes in the same domain. A reorganization of a fashion store results in a new arrangement (or 'scene') composed of existing items. A robust model should still be able to resolve the user's references to the items in the new scene. To test this generalization ability, we construct a dataset FASH-14 K that excludes all examples associated with one `held-out` scene and use it for training. [4] We compare the models by testing them on 1) the fashion subset that contains

[4]Note that the objects that occur in `held-out` scene also occur in the training set.

TABLE VI
EVALUATION ACROSS SCENES AND DOMAINS

Experimental Condition	Test domain	Train set	Test set	UNITER-based (F1 ± std. error)	BART-based (F1 ± std. error)
In-domain vs. Cross-domain					
In-domain	Furniture	ALL-train	FURN-2K	0.843 ± 0.009	**0.861** ± 0.008
Cross-domain	Furniture	FASH-14K	FURN-2K	**0.525** ± 0.011	0.457 ± 0.010
In-domain	Fashion	ALL-train	FASH-6K	**0.736** ± 0.006	0.731 ± 0.006
Cross-domain	Fashion	FURN-12K	FASH-6K	**0.425** ± 0.006	0.194 ± 0.005
In-domain vs. Cross-domain vs. In-domain-held-out					
In-domain	Fashion	FASH-14K	FASH-9K-ID	**0.694** ± 0.005	0.675 ± 0.005
Cross-domain	Furniture	FASH-14K	FURN-9K-OOD	**0.549** ± 0.006	0.373 ± 0.005
In-doman-held-out	Fashion	FASH-14K	FASH-9K-IDHO	0.621 ± 0.005	**0.740** ± 0.005

In-domain: train and test data contain the same set of scenes; cross-domain: no domain overlap between train and test data; in-domain-held-out: no scene overlap between train and test data.

TABLE VII
EVALUATION OF THE AUXILIARY TRAINING TASK WITH THE ANALYSIS ON MENTIONED AND NEW OBJECTS

	UNITER	BART	UNITER w/ Aux. task	BART w/ Aux. task
Mentioned objects	0.837 ± 0.005	0.796 ± 0.005	**0.844** ± 0.005	0.827 ± 0.005
New objects	0.644 ± 0.008	**0.722** ± 0.006	0.644 ± 0.008	0.700 ± 0.006
Overall	0 758 ± 0 005	0 760 ± 0 005	0 761 ± 0 005	0 763 ± 0 004

The models are trained on all-train and tested on the standard all-test data splits. The metrics include the F1 score and standard error.

the same scenes as the training set (*in-domain* condition), 2) the furniture subset (*cross-domain* condition), and 3) the dataset with `held-out` scenes only (*in-domain-held-out* condition). The results are shown in the bottom part of Table IV. For the *in-domain* condition the UNITER and BART models achieve similar performance with $F1$ of $0.694/0.675$. In the *cross-domain* condition we again observe that the UNITER-based model performance is higher than that of the BART-based model, with an $F1$ of 0.549 vs. 0.373. In the *in-domain-held-out* condition, the UNITER-based model performance drops by 7% points to 0.621 while the BART-based model performance increases to 0.740. We observe that in-domain test set FASH-9K-IDHO has lower complexity with 24 objects per scene in comparison to the out-of domain dataset FASH-9K-ID, which has 33, making it potentially an easier task for both models. Nevertheless, the result shows that the BART's contrastive encoder for object IDs is effective in reference resolution on new scenes with seen objects. The drop in performance of the UNITER-based model on the `held-out` dataset indicates that the model may not effectively resolve references to relative positions.

C. Mentioned vs. New References

A reference to an entity in a dialog can be referring to a *mentioned* object or to a *new* object, not yet mentioned in previous discourse. In the example dialog shown in Table II, the last user utterance *'How much is the bottom one and who makes it?'* refers to an object previously mentioned by the system. However, a user may have said *'I would like to see the black sweater at the top next to the pink one'* referring to a new item not mentioned previously.[5] In many of the examples, resolving a reference to a previously mentioned object does not require visual information from the corresponding scene. However, a reference to a new object in a multimodal context always requires processing of the visual scene including the recognition of the object attributes and their relative position. Table VII shows the performance breakdown on the references to the *mentioned* and *new* objects for both models. We observe that both UNITER and BART have higher $F1$ scores on mentioned objects $(0.837/0.796)$ than on new objects $(0.644/0.722)$.

Next, we evaluate the effect of using the auxiliary task of predicting the number of objects referenced in an utterance (see Section IV-C) on the *mentioned* and *new* references. The auxiliary task of predicting the number of target objects has an accuracy of 99.1% and 98.1% for UNITER and BART models respectively. Note that in 50.4% of the examples the user does not refer to any objects, in 23.1% they refer to one single object and in 26.5% to two objects. Without the auxiliary task, both models get similar performance with $F1 = 0.76$. We observe that the auxiliary task benefits the the performance on *mentioned* objects. UNITER's F1 increases from 0.837 to 0.844 and BART's increases from 0.796 to 0.827, a relative increase of 1% and 4% respectively. The performance on the *new* objects, however, is unchanged for UNITER and decreased for BART. This can be caused by the reference detection probabilities on new objects being less accurate, highlighting the challenge of processing visual information required for new objects.

The SIMMC2.0 dataset includes metadata information with visual and non-visual attributes (see Table I). Non-visual attributes are acceptable by the DSTC10 competition rules, but visual attributes are not as the models are encouraged to extract them from the visual scenes. Table VIII shows how the use of these attributes affects the performance of the BART-based model on *mentioned* and *new* objects. We observe that adding

[5] 58.5% of the target objects were previously *mentioned* in the dialog and 41.5% of them were *new* in the SIMMC2.0 dataset.

TABLE VIII
EFFECT OF THE NON-VISUAL (NON-VIS.) AND VISUAL ATTRIBUTES FOR BART-BASED MODEL

	BART-based (F1 ± std. error)		
w/ Aux. task	✓	✓	✓
w/ non-vis. attrs.		✓	✓
w/ visual attrs.			✓
Mentioned	0.812 ± 0.005	0.827 ± 0.005	**0.835** ± 0.005
New	0.693 ± 0.006	0.700 ± 0.006	**0.715** ± 0.006
Overall	0.752 ± 0.005	0.763 ± 0.004	**0.775** ± 0.004

TABLE IX
RESULTS AFTER COMBINING UNITER AND BART-BASED MODELS ON THE STANDARD SIMMC2.0 *DEVTEST* SET AND COMPARABLE TO THE RESULTS IN TABLE V

UNITER for Mentioned w/ obj. heur.	BART for New w/ obj. heur	F1 Score overall
✓	✓	0.789 ± 0.005
✓		**0.800** ± 0.005

TABLE X
COMPARING UNITER-BASED AND BART-BASED MODEL WITH HUMAN PERFORMANCE ON A SUBSET OF 100 RANDOMLY SELECTED EXAMPLES

	UNITER w/ Aux. task	BART	Human
Mentioned	0.873	0.844	**0.886**
New	0.632	**0.776**	0.610
Overall	0.811	0.820	**0.822**

DIALOGUE HISTORY of example No. 1731
User : Could you recommend a coat for me? System : There is a black coat in the front on the right, a grey option in the section just behind it, and a light grey option on the left in the second to last section. <SOM> <6>, <12>, <7> <EOM> User : I'll take that grey option, as well as the black option you showed me.

Fig. 4. Two referred coats are not identified, marked in red rectangles.

non-visual attributes benefit the performance on mentioned objects the most increasing it by 1.5% absolute points. In contrast, addition of the visual attributes leads to the larger increase (also 1.5% absolute points) in the performance on the *new* objects. This supports the hypothesis that when referring to new objects, processing of the visual information is especially critical.

We observe that the UNITER-based system has a higher score than the BART-based system on *mentioned* objects while BART-based system has a higher score on *new* objects. If we combine the strengths of the two approaches by using UNITER's prediction for the *mentioned* and BART's prediction for the *new* objects, we achieve $F1 = 0.80$ for the best model combination (see Table IX), an absolute increase of 3/4% points in comparison with the best performing BART/UNITER-based models reported in Table V.

When deploying such a combination in a system, we would have to keep track of the objects mentioned by the system and the user. While the objects mentioned by the system are known, the objects referred to by the user are automatically identified in previous turns and may contain errors. Hence, the performance of this combined system may degrade due to the propagation of the errors, but not below the performance of the inferior system.

D. Human Evaluation

The model combination achieves $F1 = 0.80$ on the coreference resolution task on *in-domain* condition leaving ample room for improvement. To assess the upper bound for the model performance on SIMMC2.0 dataset, we asked people to perform the annotations using the same information as the models. We sampled 100 random examples from the SIMMC2.0 `devtest` set and had three members of the team annotate them. Each example consists of the dialog segment between the user and the assistant, the multimodal context (IDs of the mentioned items), and the current scene image with the objects' bounding boxes with their corresponding IDs.[6] Three members of the team completed the task by annotating the object IDs referred to in the last user utterance in each dialog segment. The inter-annotator agreement [37] on this task was high with an average pairwise $\kappa = 0.86$.

Table X shows the human and model performance on the set of 100 examples. We observe that the overall human performance was $F1 = 0.822$, which is very close to the UNITER and BART's $F1$ of 0.811 and 0.820 on the same test set. Similarly to the models, the annotators did better than the models on *mentioned* objects and worse on *new* objects, achieving only $F1 = 0.610$.

Although the human evaluation was performed on a small dataset, the results suggest that the improved models are already achieving near human-level performance on this dataset.

VI. ERROR ANALYSIS

We manually examined errors made by the models and by the human annotators. The following examples illustrate the common errors. In the following images, red rectangles indicate the ground truth and green rectangles indicate the model's prediction.

Missing references: Fig. 4 illustrates an example where a BART model is able to recognize the references in a user's utterance. The model assigns a higher probability to the objects #6 and #12, but these probabilities are below the threshold and therefore not considered a positive prediction. This type of error is aimed to be addressed with the auxiliary task of predicting the number of references alongside the heuristics to post-process the models predictions.

[6]The users did not have access to the non-visual metadata features which the models had and which lead to a 1.5% absolute increase for the BART model (see Table VIII).

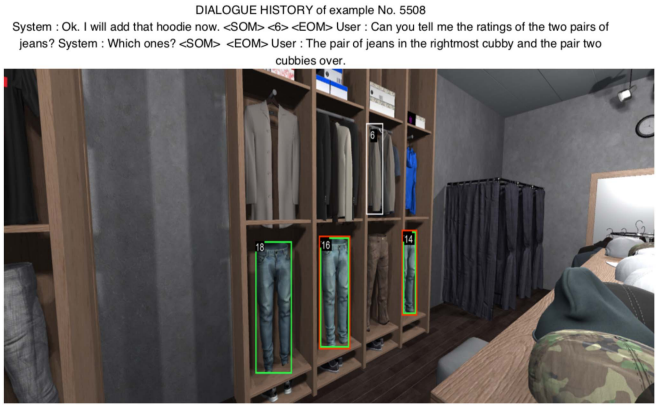

Fig. 5. All jeans are predicted because they have equal descriptions.

Fig. 6. Example is wrongly annotated: it has no ground-truth targets.

False positives: Fig. 5 illustrates a common mistake where referring expressions describing relative locations are leading to false positive prediction. In the image, several jeans are shown, of which #14 and #16 are the true targets. The model, however, also predicts #18, which is therefore a false positive.

While pre-trained large language models might be able to know spacial expressions like *'on the left'*, *'on top of the table'*, etc., they still struggle to identify descriptions such as *'the rightmost cubby'* and *'two cubbies over'*. Not only are the expressions more complex, but they are also dependent on potentially unknown items, such as a cubby, a rack or a shelf. Large language models may not be pre-trained on such vocabulary, and SIMMC2 does not include metadata descriptions of the background objects that are not part of the domain merchandise.

Dataset annotation errors: We have observed that some examples in the SIMMC2 dataset are not correctly annotated, lowering the performance of automatic metrics where the prediction may have been correct. Fig. 6 shows an example where the ground-truth targets are missing when they should be present. The user is clearly referring to the area rug and the model correctly identifies it.

Human annotator errors: The human annotators in our experiment have an average turn-level error rate of 12%. In fact, 7 examples out of 100 were wrongly classified by all three the human annotators, and 4 examples out of 100 contained errors by two of the annotators. This indicates that the annotators are

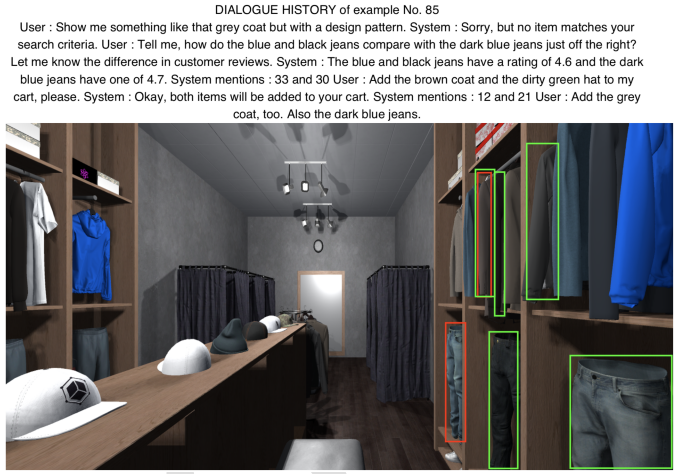

Fig. 7. Annotators fail to identify correctly all referred objects since there are multiple options matching the description.

struggling mainly with the same subset of examples. The first source of errors comes from the fact that SIMMC2 contains annotation noise, as illustrated in Fig. 6. On the other hand, some of the correctly annotated examples are difficult to resolve because either an object is too small, occluded, or cropped, or the color is not clearly displayed, or the reference is ambiguous.

Fig. 7 shows an ambiguous example where all three human annotators failed to identify the ground truth targets (the pair of jeans at the far back of the first cubby). Most annotators usually targeted the trousers in the middle or the rightmost part since they are darker and the dialog mentions some "dark blue jeans". Moreover, the other referred object is a grey coat, but there are plenty of other items matching that description in the upper shelf.

To sum up, we have seen that some references are challenging even for human annotators since there are many objects matching the provided description in the conversation. Additionally, the dataset contains incorrectly annotated examples and ambiguities that can be resolved in the future versions. However, the models are still struggling with some of cases that a human could label correctly.

VII. Conclusion

In this work we address multimodal entity (or coreference) resolution, an essential component for a situated system with access to visual and speech input. We analyze, compare, and improve two of the publicly released models that participated in the DSTC10 competition [22], a multimodal UNITER-based model [3] and a unimodal BART-based model [5]. We enhance these systems and evaluate on in-domain and cross-domain conditions where training and test datasets have disjoint sets of objects.

Our experiments show that the unimodal approach outperforms the multimodal one on *in-domain* experiments, indicating that contrastive learning of attributes from metadata employed by the authors is an effective way to learn visual features. However, this method is only effective when objects are seen in training. As expected, the multimodal approach outperformed

the unimodal one in a *cross-domain* setting, showing a potential for generalization.

Analyzing the performance on objects *mentioned* in the previous discourse and *new* objects, we reveal that the unimodal approach performs better on *new* references while the multimodal one performs better on previously *mentioned* objects. We introduced a new auxiliary task predicting the number of objects mentioned in an utterance and heuristic adjustment of the models' output. The experiments showed that our proposal leads to improvement for both models on *mentioned* objects. A combination of the two methods where we use the unimodal prediction on *mentioned* and multimodal prediction on *new* objects achieves $F1 = 0.80$, a new SOTA on coreference prediction on the SIMMC2.0 dataset. Finally, our analysis of human performance on this task suggests that the current models perform already on par with people on a subset of dialogues sampled from this dataset.

ACKNOWLEDGMENT

This article was produced by Toshiba Europe Ltd. (Cambridge, U.K.) in collaboration with the Cambridge University Engineering Department.

REFERENCES

[1] S. Stoyanchev, S. Keizer, and R. Doddipatla, "Action state update approach to dialogue management," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7398–7402.

[2] I. Roesiger, S. Stehwien, A. Riester, and N. T. Vu, "Improving coreference resolution with automatically predicted prosodic information," in *Proc. Workshop Speech-Centric Natural Lang. Process.*, 2017, pp. 78–83. [Online]. Available: https://aclanthology.org/W17-4610

[3] Y. Huang, Y. Wang, and Y. Tam, "UNITER-based situated coreference resolution with rich multimodal input," 2021, *arXiv:2112.03521*.

[4] F. J. Chiyah-Garcia, A. Suglia, J. Lopes, A. Eshghi, and H. Hastie, "Exploring multi-modal representations for ambiguity detection & coreference resolution in the SIMMC 2.0 challenge," 2022, *arXiv:2202.12645*.

[5] H. Lee et al., "Tackling situated multi-modal task-oriented dialogs with a single transformer model," in *Proc. AAAI Conf. Artif. Intell. Workshop*, 2022.

[6] M. Poesio, R. Stuckardt, and Y. Versley, *Anaphora Resolution - Algorithms, Resources, and Applications, Theory and Applications of Natural Language Processing*. Berlin, Germany: Springer, 2016.

[7] A. H. Anderson et al., "The HCRC map task corpus," *Lang. Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[8] W. Monroe et al., "Colors in context: A pragmatic neural model for grounded language understanding," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 325–338, 2017.

[9] A. J. Stent and S. Bangalore, "Interaction between dialog structure and coreference resolution," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2010, pp. 342–347.

[10] L. E. Asri et al., "Frames: A corpus for adding memory to goal-oriented dialogue systems," in *Proc. 18th Annu. SIGDIAL Meeting Discourse Dialogue*, 2017, pp. 207–219.

[11] K. v. Deemter and R. Kibble, "On coreferring: Coreference in MUC and related annotation schemes," *Comput. Linguistics*, vol. 26, no. 4, pp. 629–637, 2000. [Online]. Available: https://aclanthology.org/J00-4005

[12] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases," *Comput. Linguistics*, vol. 27, no. 4, pp. 521–544, 2001.

[13] S. Bergsma and D. Lin, "Bootstrapping path-based pronoun resolution," in *Proc. 21st Int. Conf. Comput. Linguistics 44th Annu. Meeting Assoc. Comput. Linguistics*, 2006, pp. 33–40.

[14] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 153–169.

[15] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei, "Linking people in videos with "their" names using coreference resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 95–110. [Online]. Available: http://vision.stanford.edu/pdf/vignesh14.pdf

[16] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal, "Visual reference resolution using attention memory for visual dialog," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3722–3732.

[17] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[18] Q. Zheng, X. Diao, J. Cao, X. Zhou, Y. Liu, and H. Li, "Multi-modal coreference resolution with the correlation between space structures," *Artif. Intell.*, 2018.

[19] A. Kumar et al., "Multimodal coreference resolution for exploratory data visualization dialogue: Context-based annotation and gesture identification," in *Proc. 21st Workshop Semantics Pragmatics Dialogue*, 2017.

[20] I. V. d. Sluis, S. Luz, W. Breitfuß, M. Ishizuka, and H. Prendinger, "Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world," *Int. J. Hum.- Comput. Stud.*, vol. 70, pp. 611–629, 2012.

[21] S. Moon et al., "Situated and interactive multimodal conversations," in *Proc. 28th Int. Conf. Comput. Linguistics. Int. Committee Comput. Linguistics*, 2020, pp. 1103–1121.

[22] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi, "SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4903–4912.

[23] Unity, "Unity technologies," 2019.

[24] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 5100–5111.

[25] Y.-C. Chen et al., "Uniter: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 104–120.

[26] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter, Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Documentation*, vol. 1, no. 8, 2019, Art. no. 9.

[29] S. Agarwal, O. Dusek, I. Konstas, and V. Rieser, "A knowledge-grounded multimodal search-based conversational agent," in *Proc. Workshop 36th AAAI Conf. Artif. Intell.*, 2018.

[30] J. Lee and K. Han, "Multimodal interactions using pretrained unimodal models for SIMMC 2.0," 2021, *arXiv:2112.05328*.

[31] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[32] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[33] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, 2017.

[34] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

[35] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. 24th Int. Conf. Neural Image Process. Syst.*, 2011, pp. 1143–1151.

[36] A. Maurer, M. Pontil, and B. Romera-Paredes, "The benefit of multitask representation learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2853–2884, Jan. 2016.

[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.