Multimodal Coreference Resolution

Alejandro Santorum Varela
Svetlana Stoyanchev, Kate Knill

UNIVERSITY OF CAMBRIDGE

TOSHIBA

Introduction

Hi, do you have short sleeve t-shirts?

Yes, we do. What color do you prefer?

White. Similar to the one next to the black hoodie.
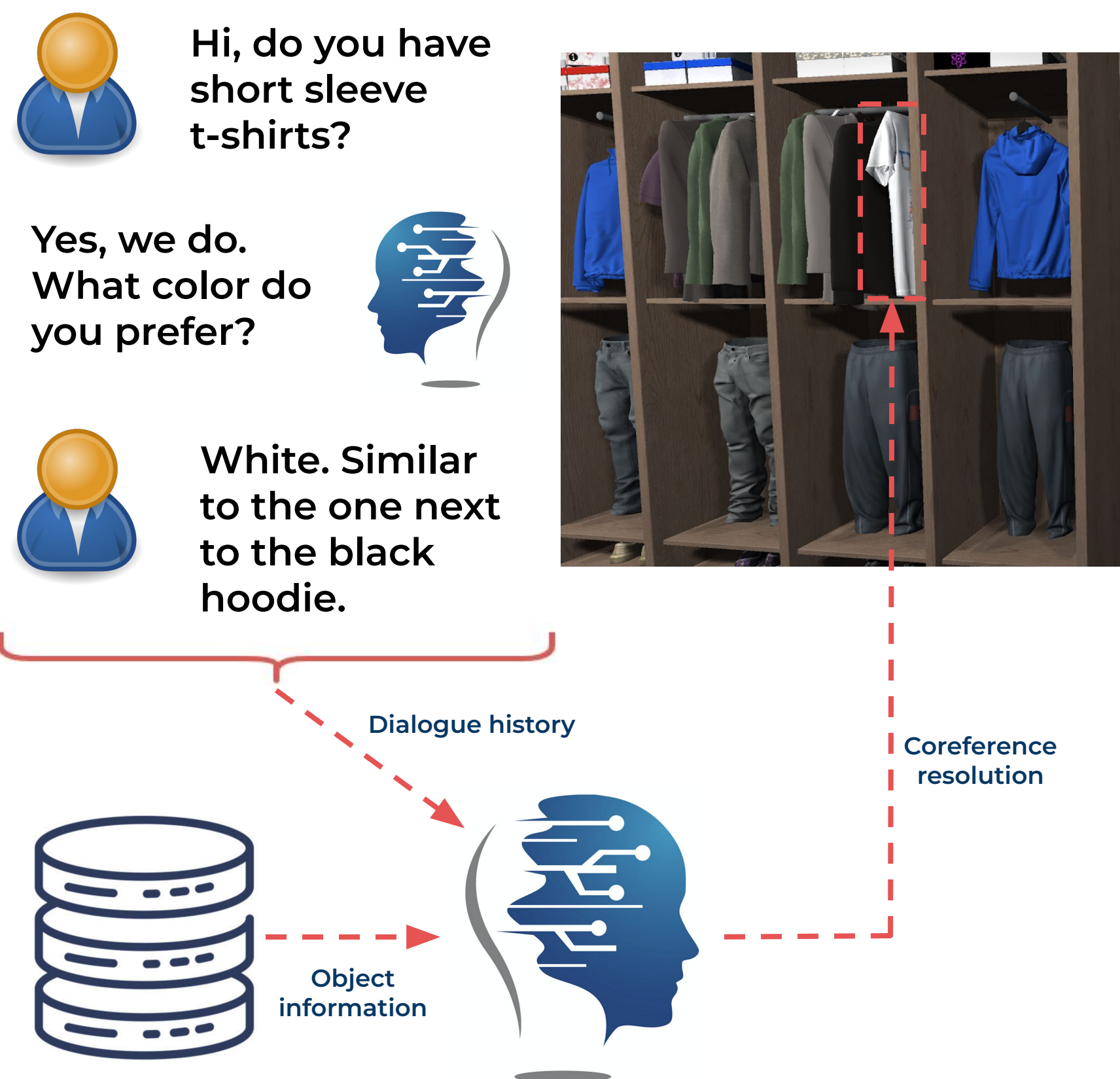
Dialogue history

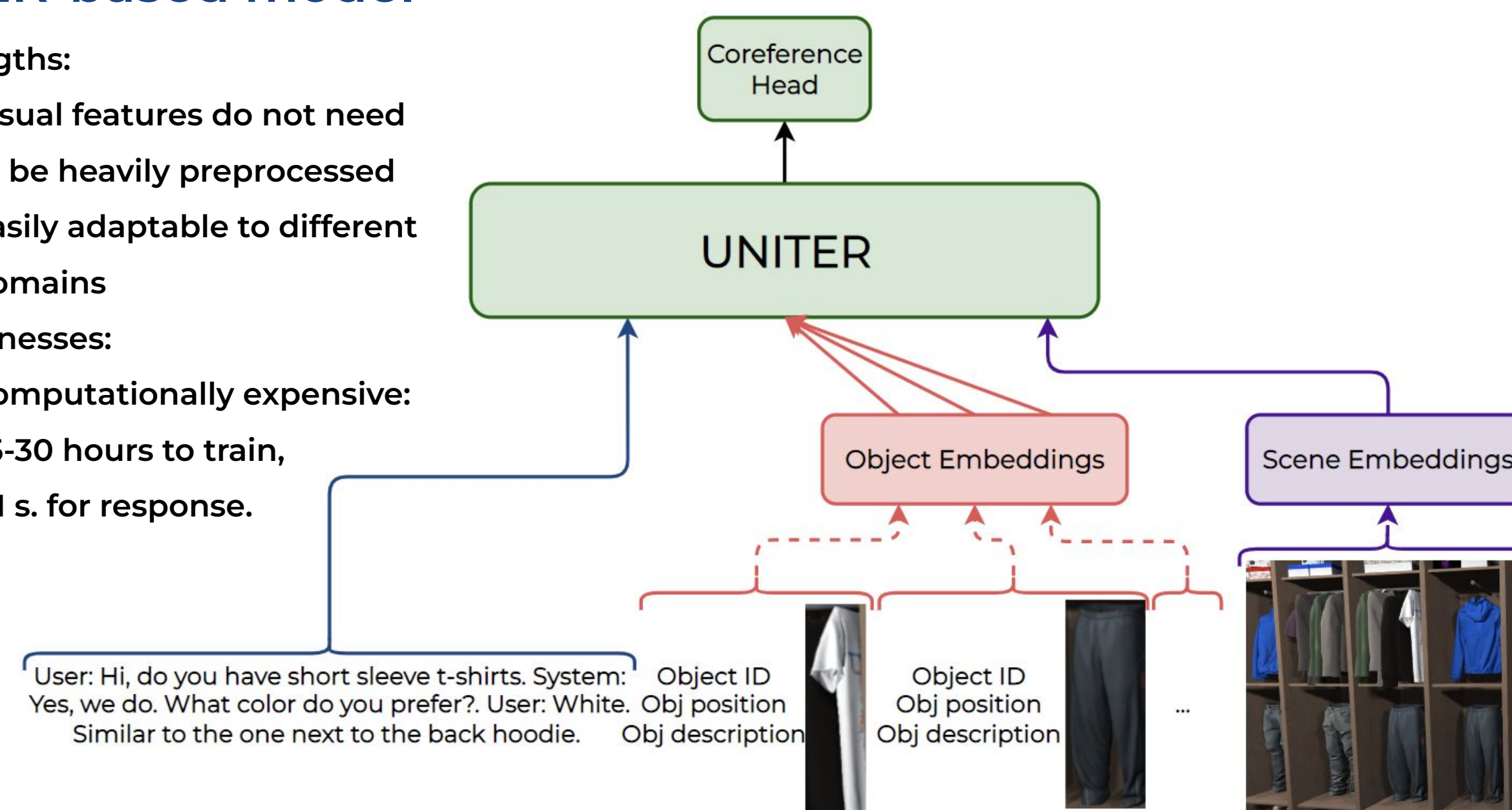Coreference resolution

Object information

- Applications:
  - E-commerce virtual assistant: answer customer inquiries about objects.
  - In-site interpreter: improved translations using both scene and textual context.
  - Boost other natural language tasks, like question answering or generation.

- SIMMC2 dataset published by Facebook Research is used for investigation.
  - It contains dialogues, object descriptions and scene images.

- The 10th Dialog System Technology Challenge (DSTC10) partially focused on the multimodal coreference resolution task.
  - SIMMC2 dataset was used for the competition.
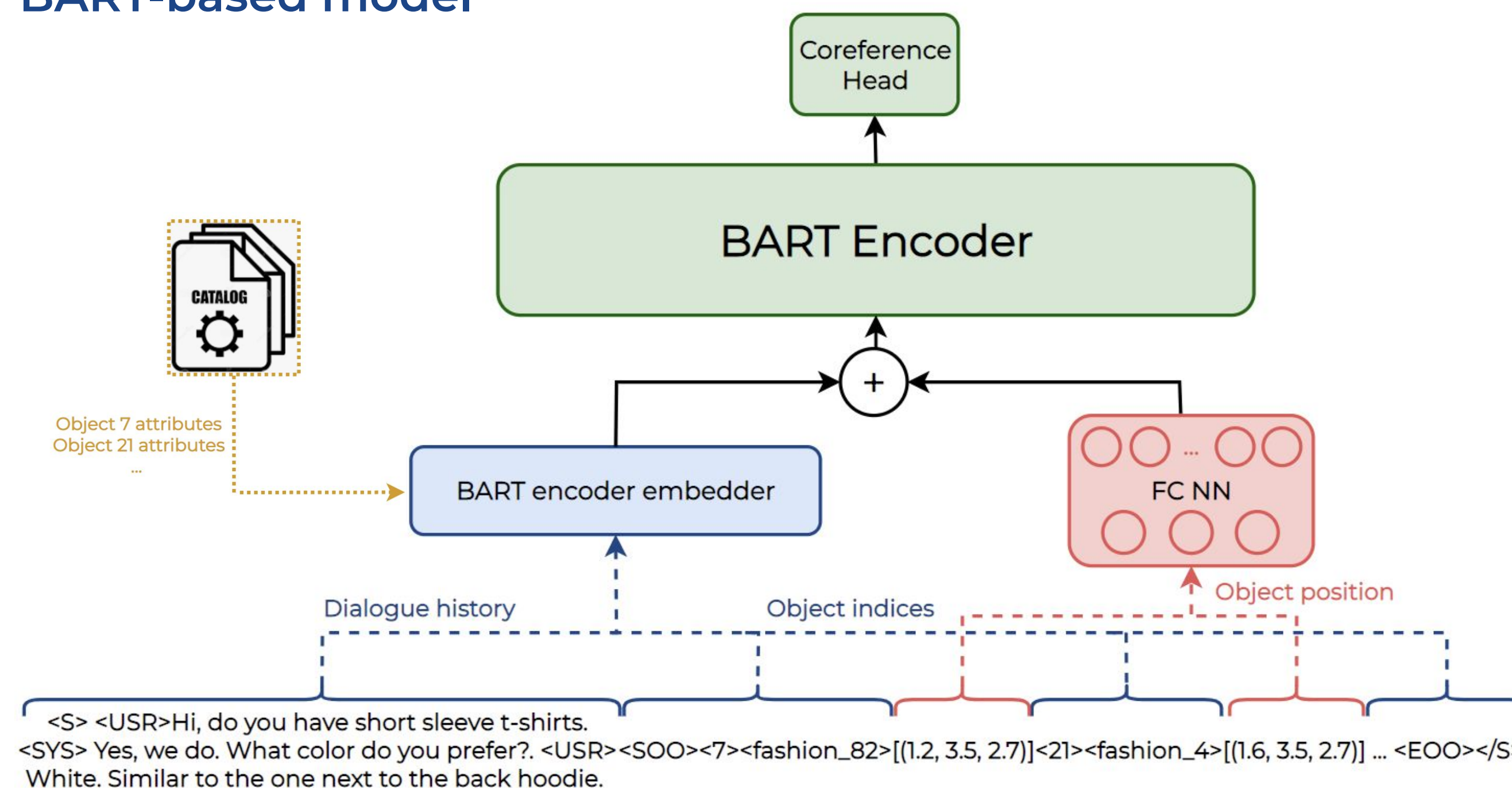 - Best performing systems are studied as a enhanced baseline.

SOTA MMCR Systems

UNITER-based model

- Strengths:
  - Visual features do not need to be heavily preprocessed
  - Easily adaptable to different domains
- Weaknesses:
 - Computationally expensive: 25-30 hours to train, ~ 1 s. for response.

Coreference Head

UNITER

Object Embeddings

Scene Embeddings

User: Hi, do you have short sleeve t-shirts. System: Yes, we do. What color do you prefer?. User: White. Similar to the one next to the back hoodie.

Object ID
Obj position
Obj description

Object ID
Obj position
Obj description

...

BART-based model

Coreference Head

BART Encoder

CATALOG

Object 7 attributes
Object 21 attributes
...

+

BART encoder embedder

FC NN

Object position

Dialogue history

Object indices

<S> <USR>Hi, do you have short sleeve t-shirts.
<SYS> Yes, we do. What color do you prefer?. <USR><SOO><7><fashion_82>[(1.2, 3.5, 2.7)]<21><fashion_4>[(1.6, 3.5, 2.7)] ... <EOO></S>
White. Similar to the one next to the back hoodie.

- Strengths:
  - Winner of DSTC10 on this task.
  - Computationally cheaper: around 5 hours to train, <0.5 seconds for response.

- Weaknesses:
  - Bad at handling objects not seen in training.
  - Scene images need to be described in natural language to be used.

Proposed improvements

- Include object descriptions in the input of the BART-based model. ✓

- Provide image embeddings to improve the coreference head of the BART-based model. ⧗

Bart Encoder Output

Scene Image Embedder

Coref Head
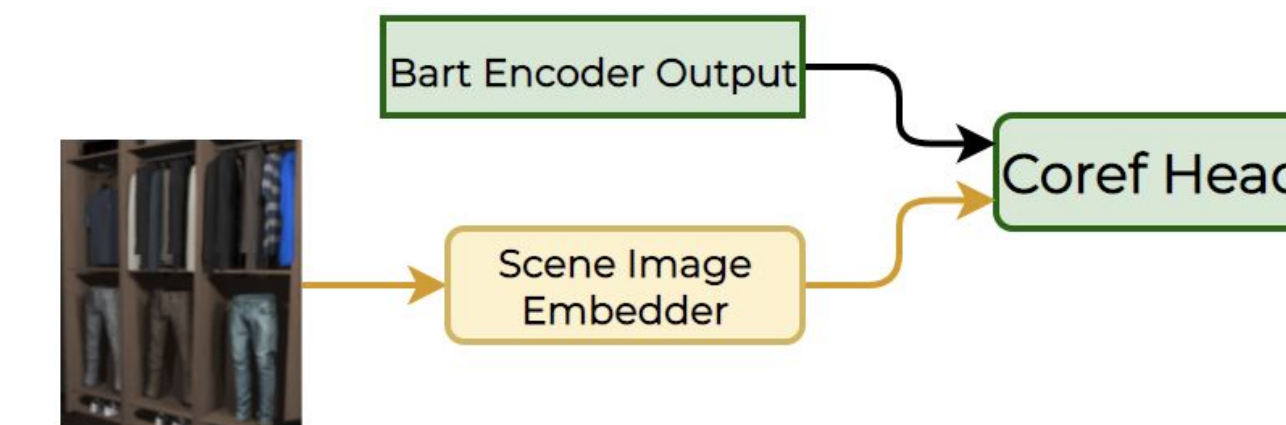
- Suppress object IDs in UNITER-based model to make it scene-independent. ⧗

Results

Model	Object F1-Score
GPT-2 Baseline (Facebook Research)	36.6%
UNITER-based DSTC10 submission (New York Uni. Shanghai)	67.4%
UNITER-based + previously mentioned objects (NYU Shanghai)	72.8%
BART-based DSTC10 submission (KAIST & Samsung Research)	74.3%
UNITER-based + prev. objects + removing obj. IDs (Ours)	**75.8%**
BART-based using object descriptions (Ours)	**76.1%**

Multimodal Coreference Resolution performance on devtest split

References

[1] Satwik Kottur et.al. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. *Association for Computational Linguistics*. 2021.

[2] Yichen Huang et. al. UNITER-Based Situated Coreference Resolution with Rich Multimodal Input. *Computing Research Repository*. 2021.

[3] Haeju Lee et. al. Tackling Situated Multi-Modal Task-Oriented Dialogs with a Single Transformer Model. *Association for Computational Linguistics*. 2021.

Acknowledgments