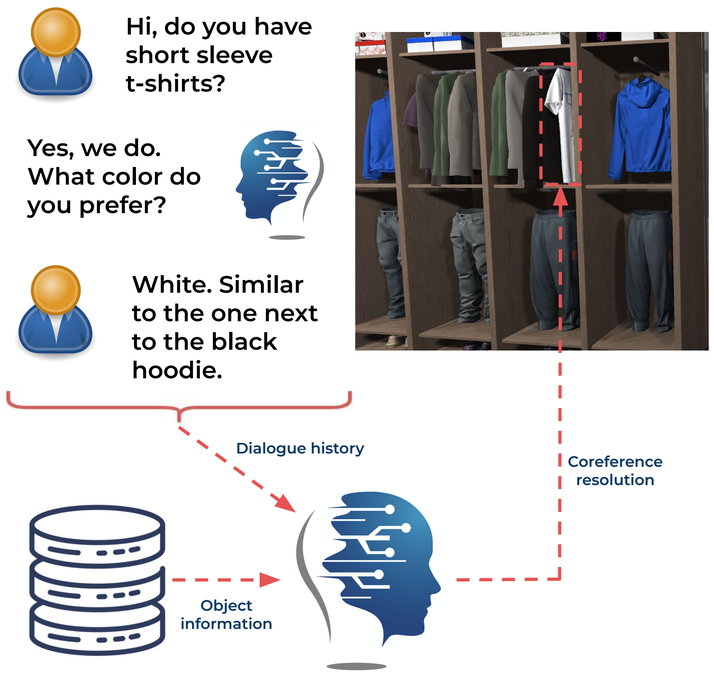
Multimodal Coreference Resolution

Alejandro Santorum Varela
Svetlana Stoyanchev, Kate Knill

UNIVERSITY OF CAMBRIDGE

TOSHIBA

Introduction

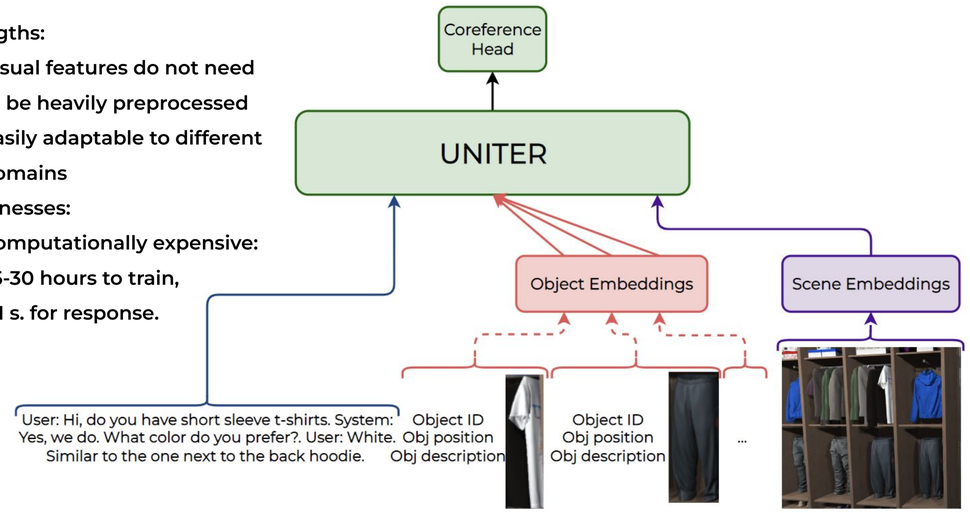


- Applications:
  - E-commerce virtual assistant: answer customer inquiries about objects.
  - In-site interpreter: improved translations using both scene and textual context.
  - Boost other natural language tasks, like question answering or generation.
- SIMMC2 dataset published by Facebook Research 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.

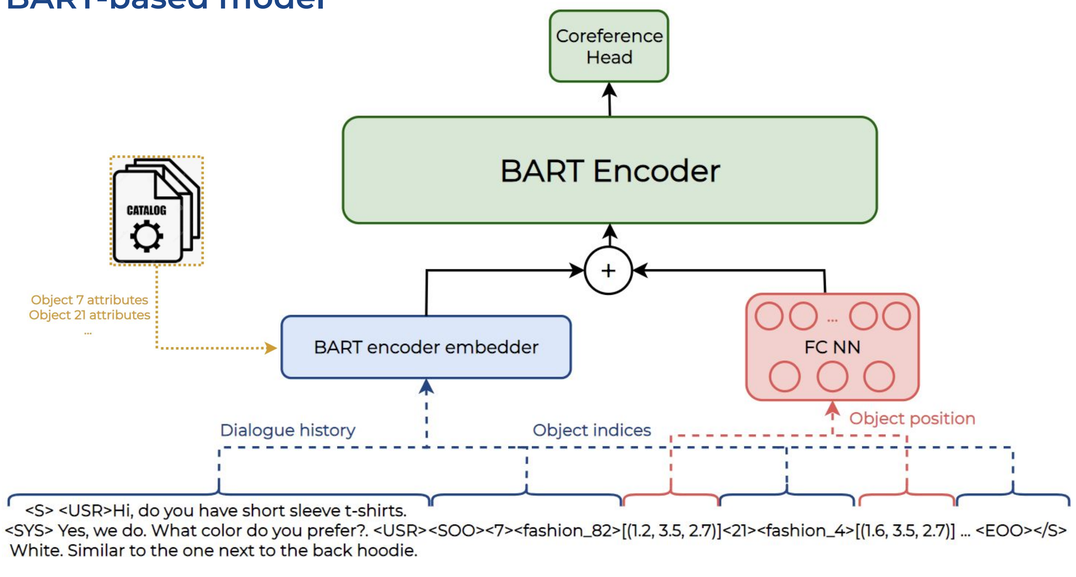
SOTA MMCR Systems

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.



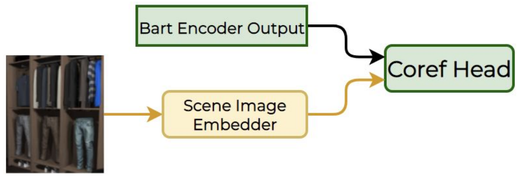
BART-based model



- 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.



- 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 (New York Uni. Shanghai)	67.4.%
BART-based (KAIST & Samsung Research)	74.3%
BART 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

Project carried out under the supervision of Svetlana Stoyanchev together with Rama Doddipatla and Simon Keizer from the Dialog group at Toshiba Europe Ltd., in collaboration with Kate Knill from the Engineering Department of the University of Cambridge.