SUMMARY

- Last week:
  - BART-based model: deleted all heads but coref head and obtained similar performance (74%)
  - BART-based model: included “object descriptions” using **two attributes** (brand and price) and the performance increased a little bit, around **75.5 %**.
  - BART-based model: included larger “object descriptions” using **four attributes** (brand, price, color and type) and boosted the performance to **76.1%**.
  - We decided to investigate how to improve **scene-independent models** (do not use Obj IDs) and how to boost the performance using more **visual information**. It seems that **further studying the UNITER model** seems the way to go.
  - Adding scene image embeddings to the BART-based model can also be investigated.
- This week so far:
  - Replicated UNITER model that uses previously mentioned objects (72.62%) and the one that uses prev_mentioned + attention bias (+ 74%).
  - First draft of the poster.
- To do this week (plan):
  - Finish the poster (before Wednesday).
  - Check the performance of the UNITER model if we remove Object IDs.
  - Poster presentation
  - Anything else?

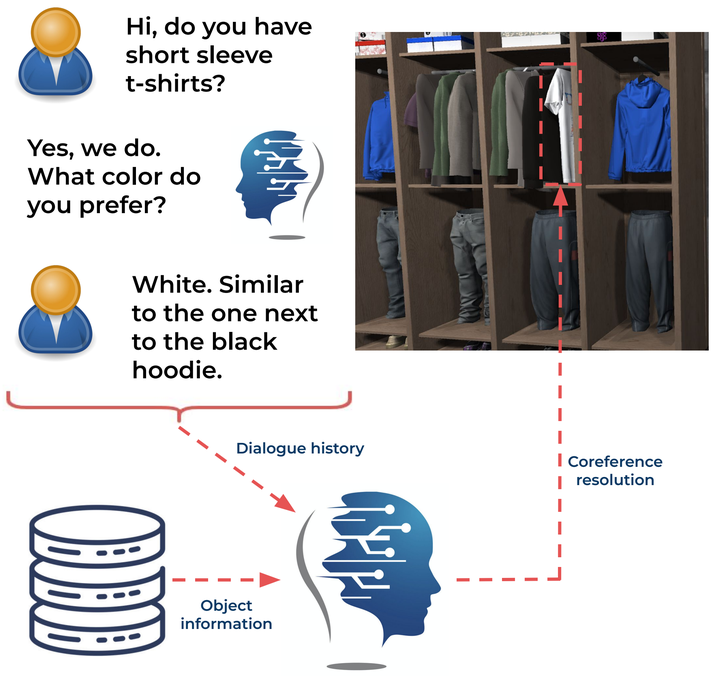
Multimodal Coreference Resolution

Alejandro Santorum Varela
Svetlana Stoyanchev, Kate Knill

UNIVERSITY OF CAMBRIDGE

TOSHIBA

Introduction

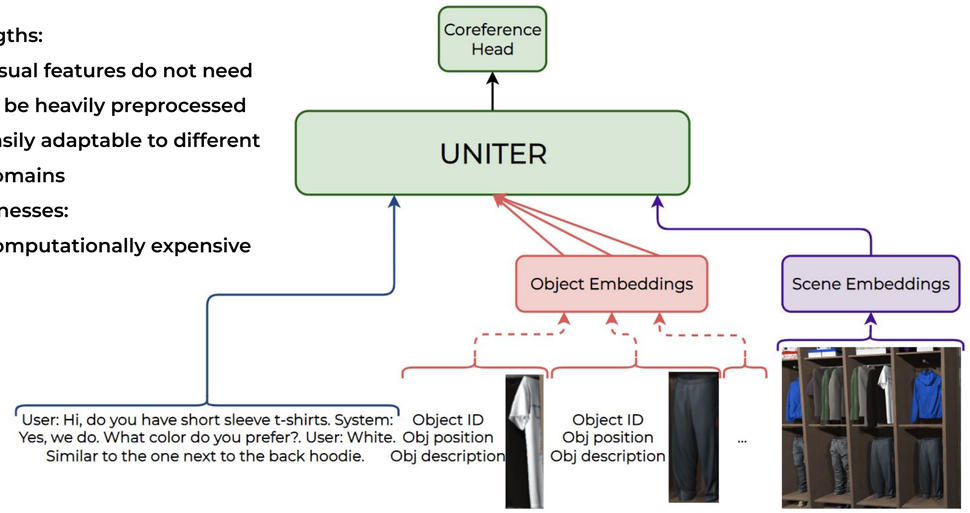


- Applications:
  - Virtual assistant
  - In-site interpreter
- SIMMC2 dataset published by Facebook Research is used for investigation.
- DSTC10 challenge partially focused on multimodal coreference resolution.

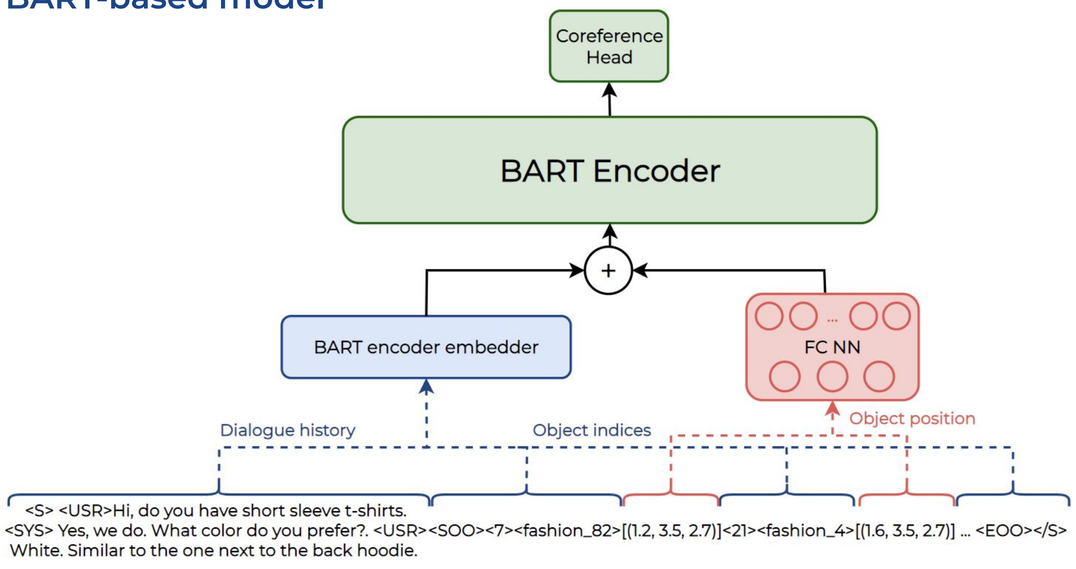
SOTA MMCR Systems

UNITER-based model

- Strengths:
  - Visual features do not need to be heavily preprocessed
  - Easily adaptable to different domains
- Weaknesses:
  - Computationally expensive



BART-based model



- Strengths:
  - One of the best solutions for coreference resolution task.
  - Winner of DSTC10 on this task.
- Weaknesses:
  - Bad at handling objects not seen in training.
  - Strong dependent on natural language.

Proposed improvements

- Include object descriptions in the input of the BART-based model.
- Suppress object IDs in UNITER-based model to make it scene-independent.
- Provide image embeddings to improve the Coreference head of the BART-based model.

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%
Our improved system	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