ACM Multimedia 2018

Knowledge-aware Multimodal Dialogue Systems

Lizi Liao[1], Yunshan Ma[1], Xiangnan He[1], Richang Hong[2], Tat-Seng Chua[1]

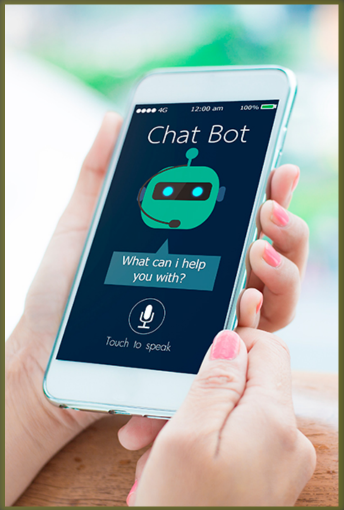
[1]National University of Singapore, [2]Hefei University of Technology

24 October 2018

Why Multimodal Dialogue?

Any similar one in blue?

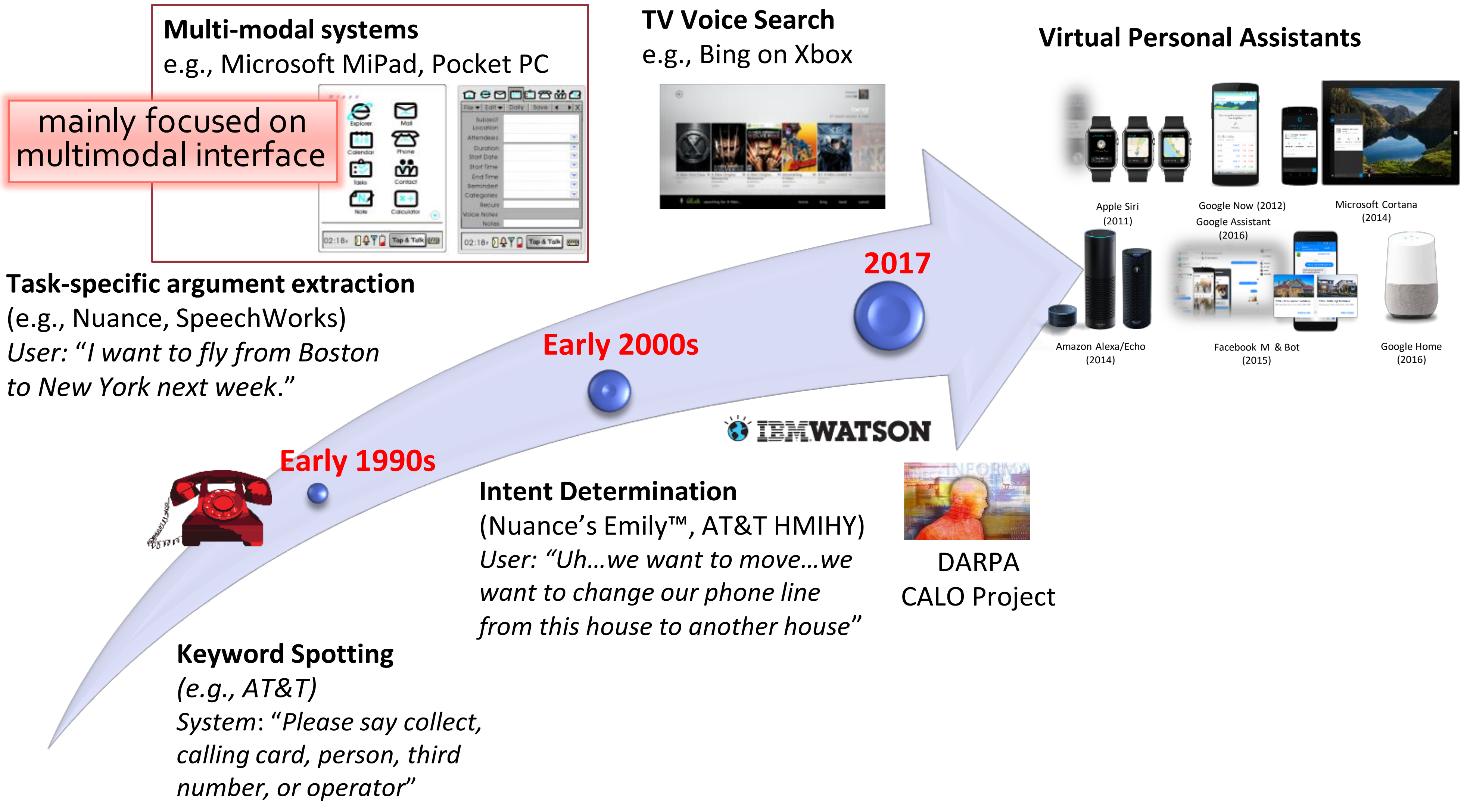
How to match with it?



Is there any such restaurant nearby?

Is there any shop selling this nearby?

Evolution of Dialogue Systems



Keyword Spotting
(e.g., AT&T)
System: *"Please say collect, calling card, person, third number, or operator"*

Early 1990s

Task-specific argument extraction
(e.g., Nuance, SpeechWorks)
User: "I want to fly from Boston to New York next week."

Multi-modal systems
e.g., Microsoft MiPad, Pocket PC

mainly focused on multimodal interface

Early 2000s

Intent Determination
(Nuance's Emily™, AT&T HMIHY)
User: "Uh…we want to move…we want to change our phone line from this house to another house"

TV Voice Search
e.g., Bing on Xbox

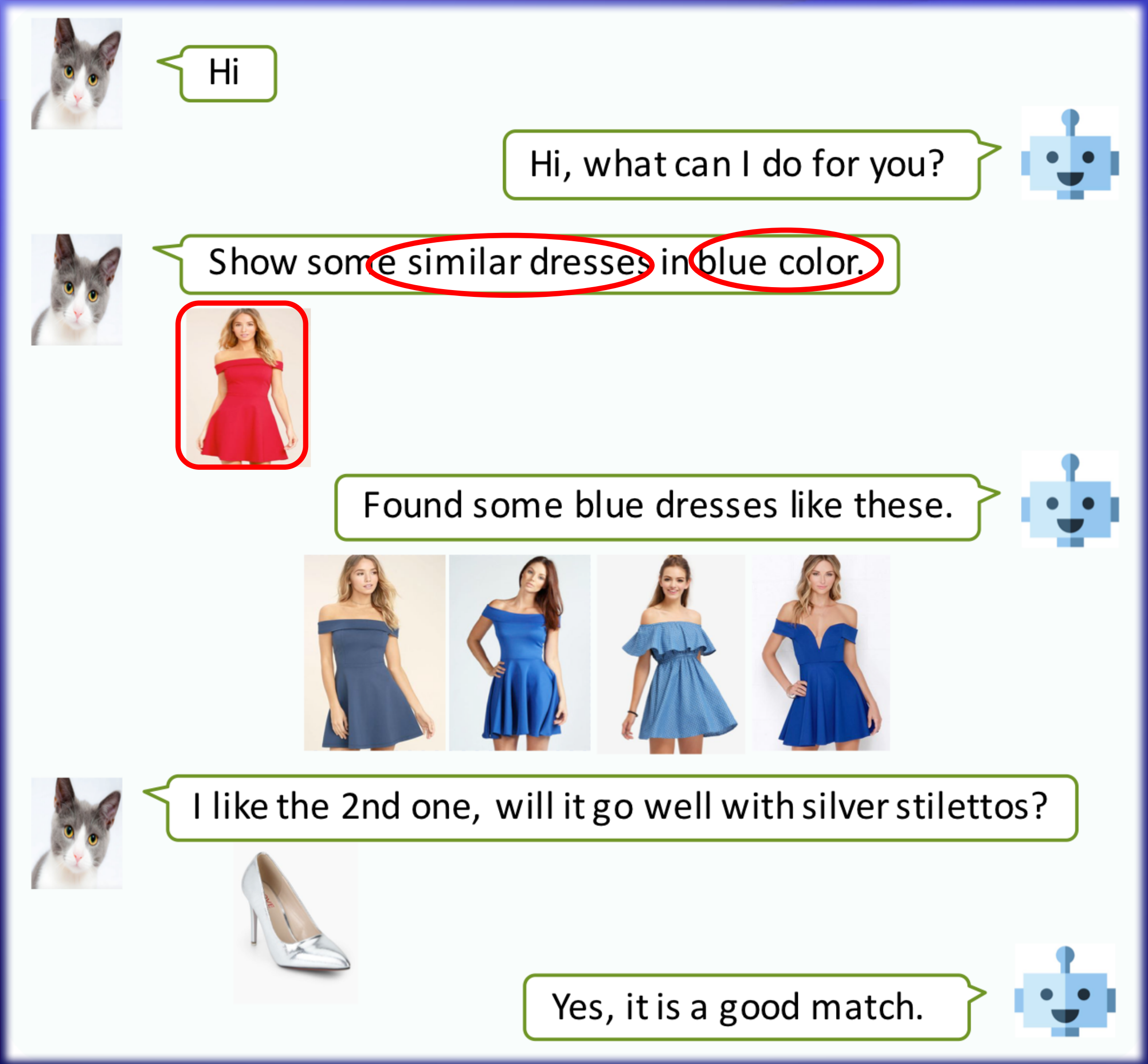
IBM WATSON

DARPA
CALO Project

2017

Virtual Personal Assistants

Challenges



Hi

Hi, what can I do for you?

Show some similar dresses in blue color.

Found some blue dresses like these.

I like the 2nd one, will it go well with silver stilettos?

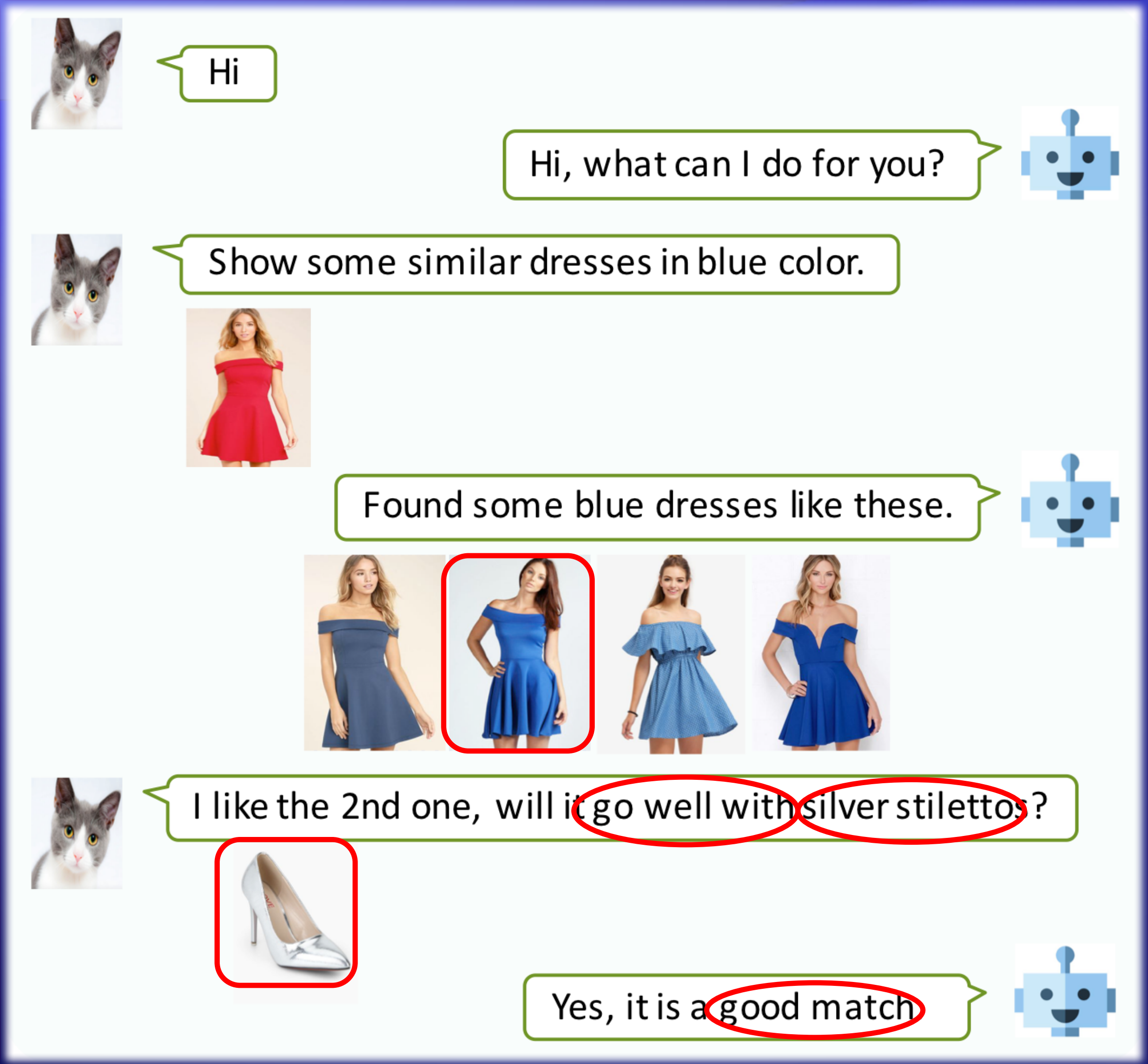
Yes, it is a good match.

1 Understanding semantics from text and image

2

3

Challenges



Hi

Hi, what can I do for you?

Show some similar dresses in blue color.

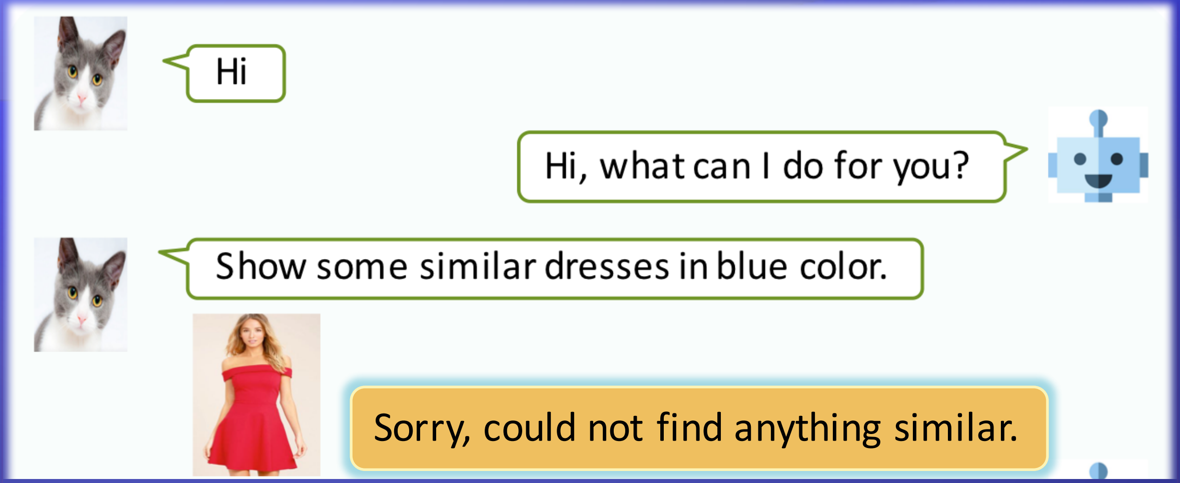
Found some blue dresses like these.

I like the 2nd one, will it go well with silver stilettos?

Yes, it is a good match

1. Understanding semantics from text and image
2. Incorporating domain knowledge
3.

Challenges



Hi

Hi, what can I do for you?

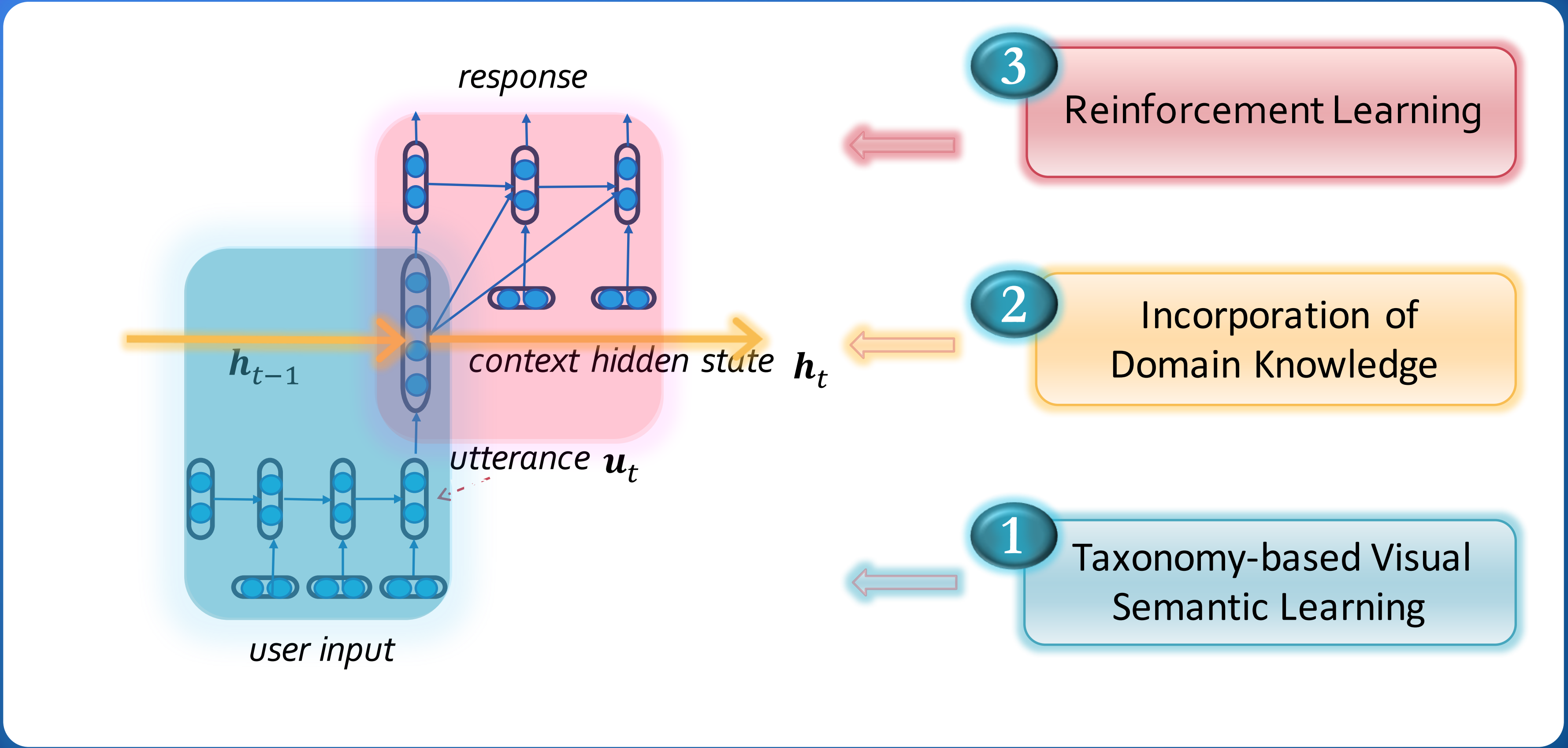
Show some similar dresses in blue color.

Sorry, could not find anything similar.

1. Understanding semantics from text and image
2. Incorporating domain knowledge
3. Improving Dialogue flow

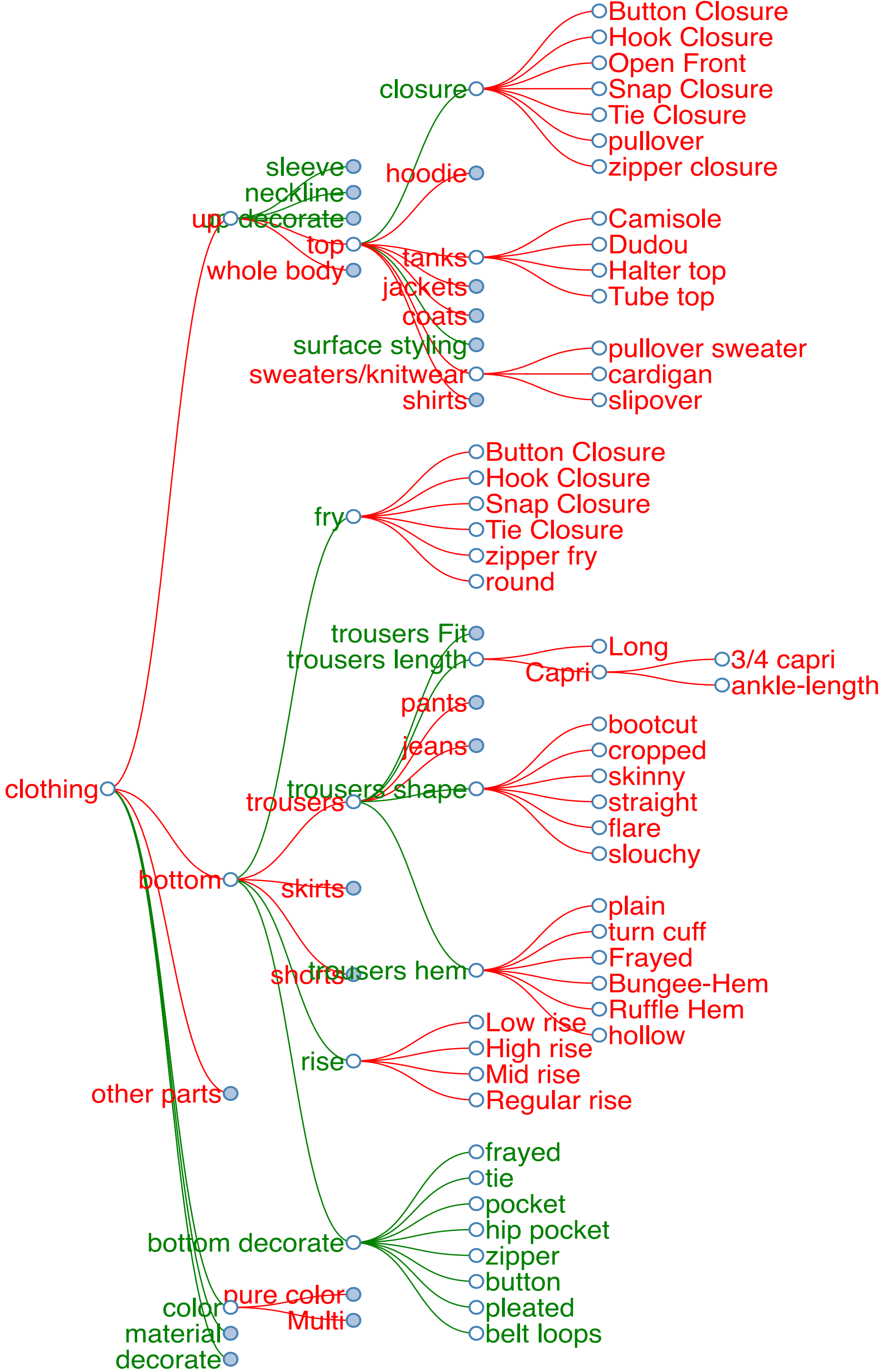
System Overview

- Hierarchical RNN **+ 3 core components**


response
h_{t-1}
context hidden state h_t
utterance u_t
user input
3 Reinforcement Learning
2 Incorporation of Domain Knowledge
1 Taxonomy-based Visual Semantic Learning

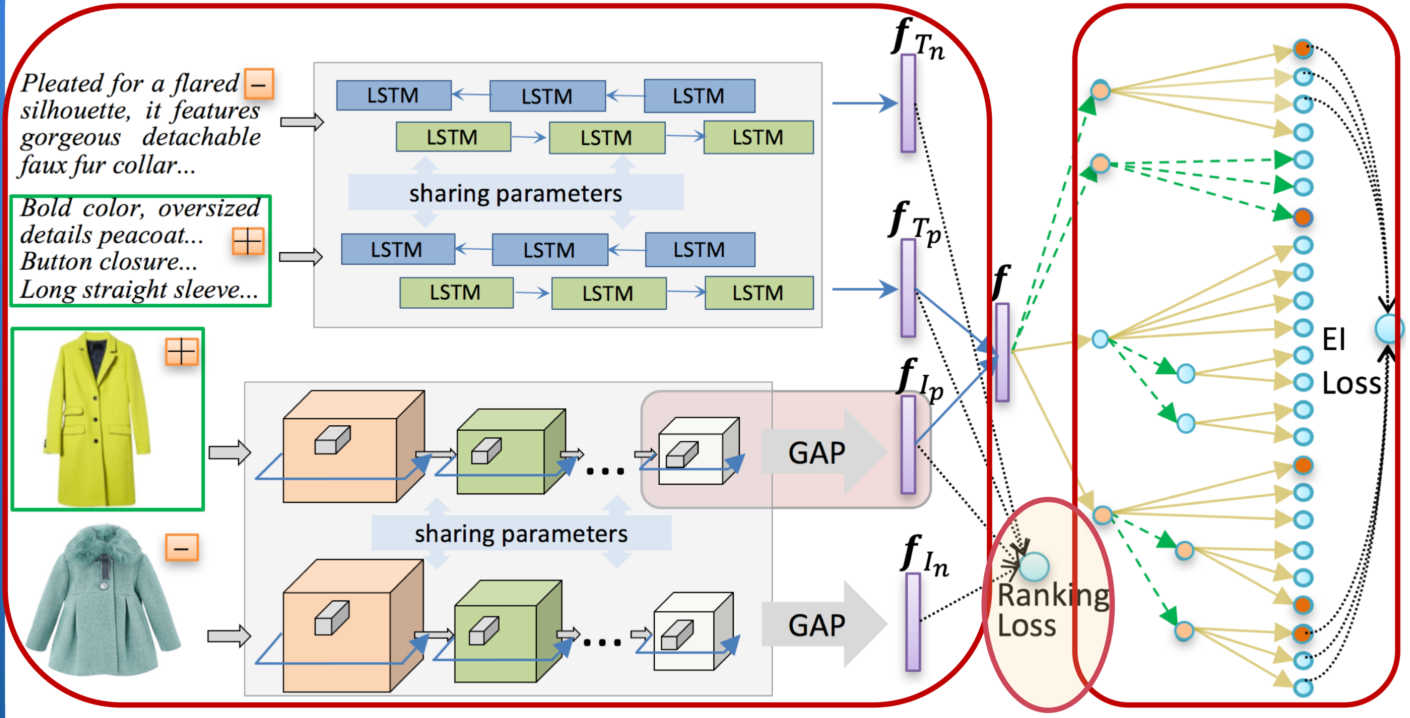
1. Learning Taxonomy-based V

- Human perception of product organization and product similarity
  - General to specific
  - Exclusive and Independent relationships (EI)



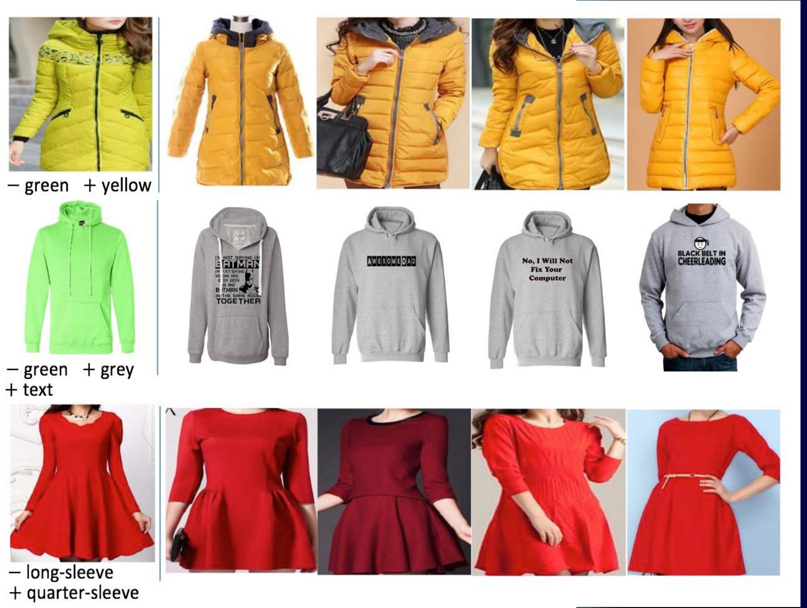
1. Learning Taxonomy-based Visual Semantics

- Map images and text into a joint visual semantic space
- Leverage EI tree taxonomy to guide fashion concepts learning



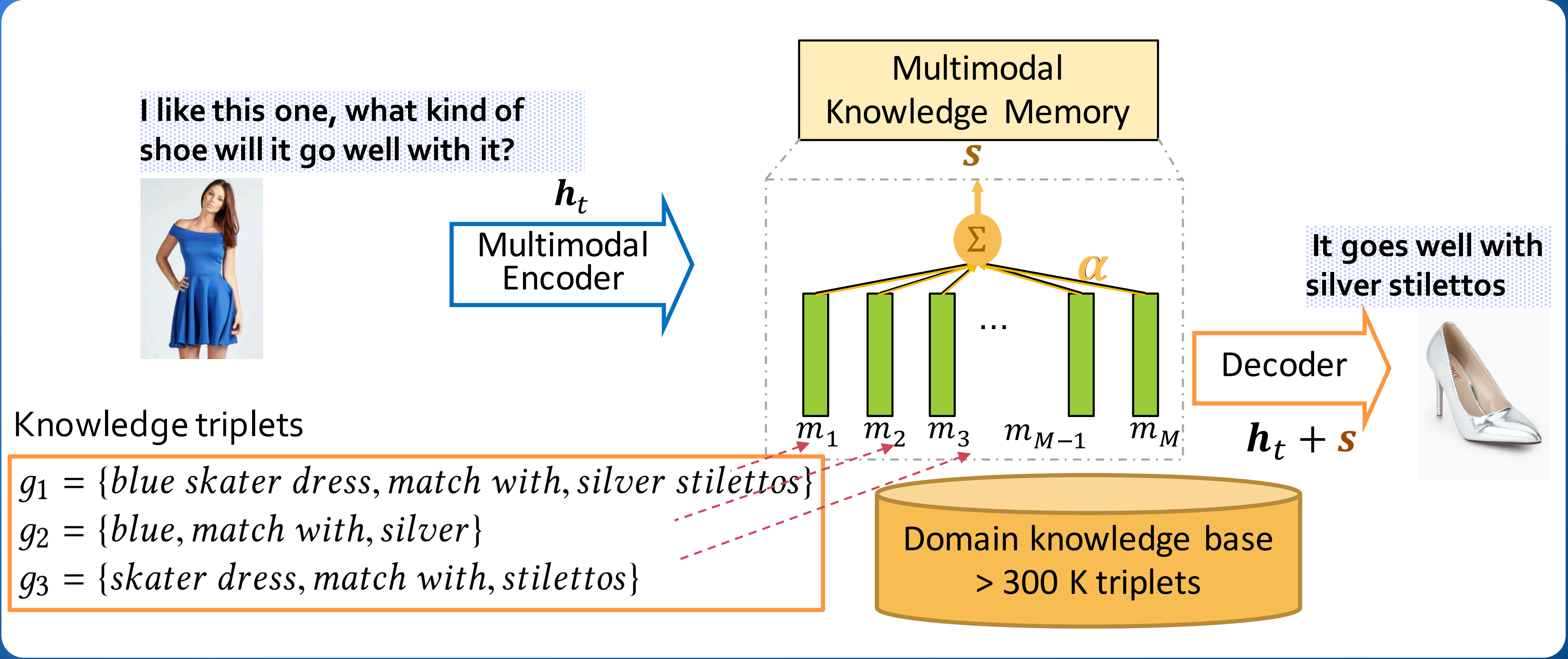
Pleated for a flared silhouette, it features gorgeous detachable faux fur collar...
–
Bold color, oversized details peacoat... Button closure... Long straight sleeve...
+
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
sharing parameters
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
+
–
GAP
sharing parameters
GAP
f_{T_n}
f_{T_p}
f_{I_p}
f_{I_n}
f
EI Loss
Ranking Loss

More accurate concept prediction



2. Incorporating Domain Knowledge

- Incorporate Knowledge by Multimodal Knowledge Memory Network



Knowledge triplets

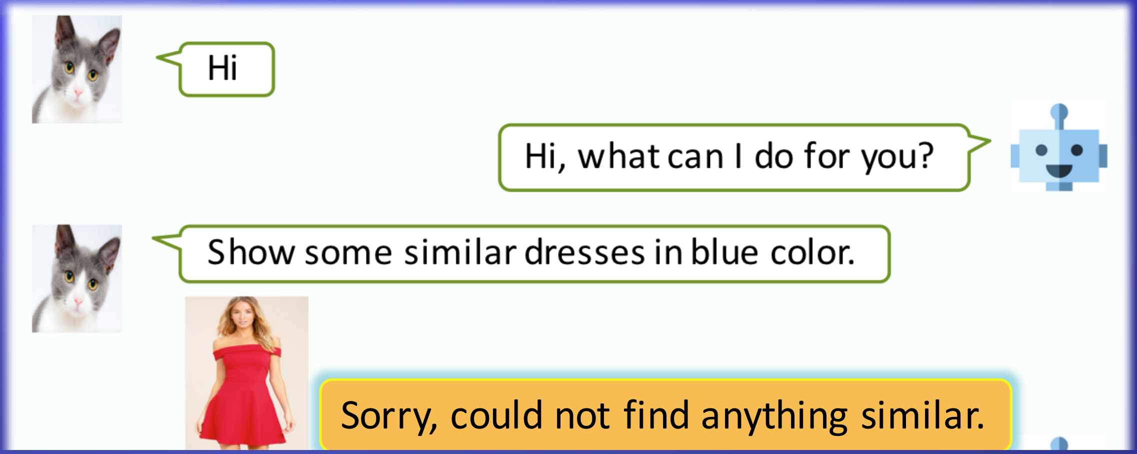
$$g_1 = \{blue\ skater\ dress, match\ with, silver\ stilettos\}$$

$$g_2 = \{blue, match\ with, silver\}$$

$$g_3 = \{skater\ dress, match\ with, stilettos\}$$

3. Training with Reinforcement Signals

- Improve dialogue flow via reinforcement signals in two stages training



Hi

Hi, what can I do for you?

Show some similar dresses in blue color.

Sorry, could not find anything similar.

1 Predict a generated target utterance given the dialogue context in a **supervised** fashion

2 Initialized the policy model using the model trained during the first stage, start **fine-tune**

3. Training with Reinforcement Signals

- Improve dialogue flow via reinforcement signals in two stages training

- Text response

$$R(h,r) = BLEU\ score$$

- Image response

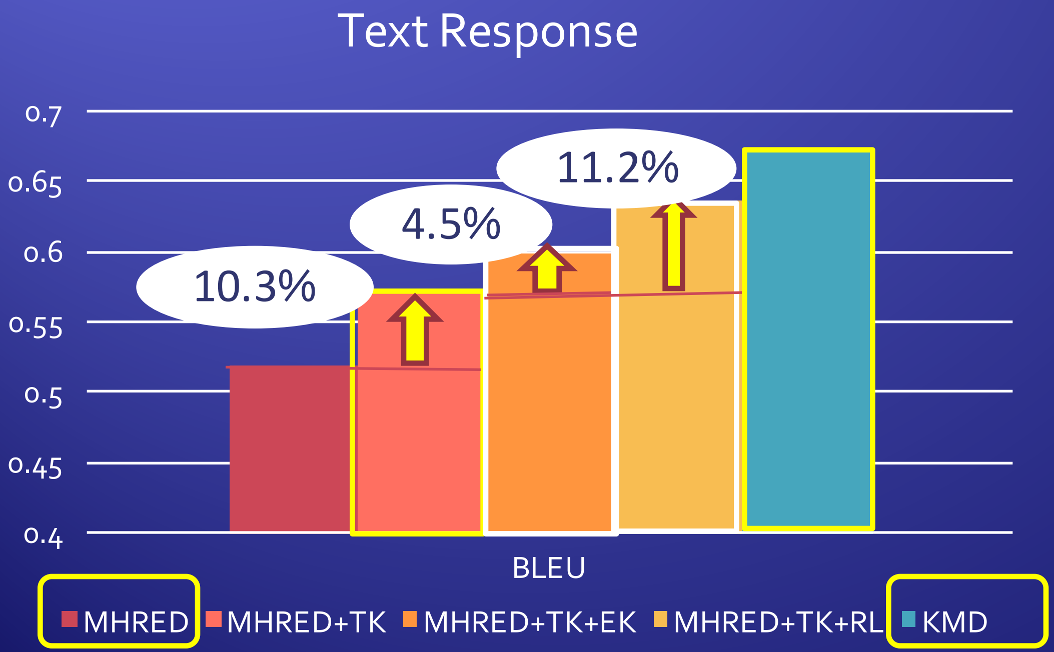
$$R(h,r) = sim(\boldsymbol{I},\boldsymbol{I}^{+}) - sim(\boldsymbol{I},\boldsymbol{I}^{-})$$

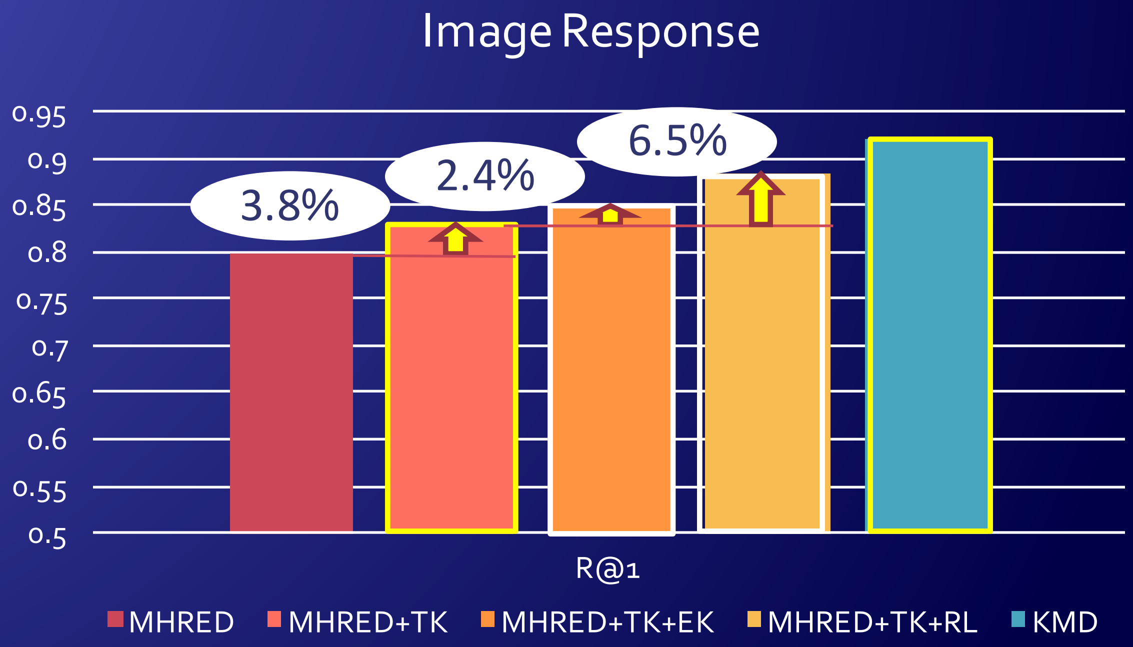
1 Predict a generated target utterance given the dialogue context in a **supervised** fashion

2 Initialized the policy model using the model trained during the first stage, start **fine-tune**

Experiments

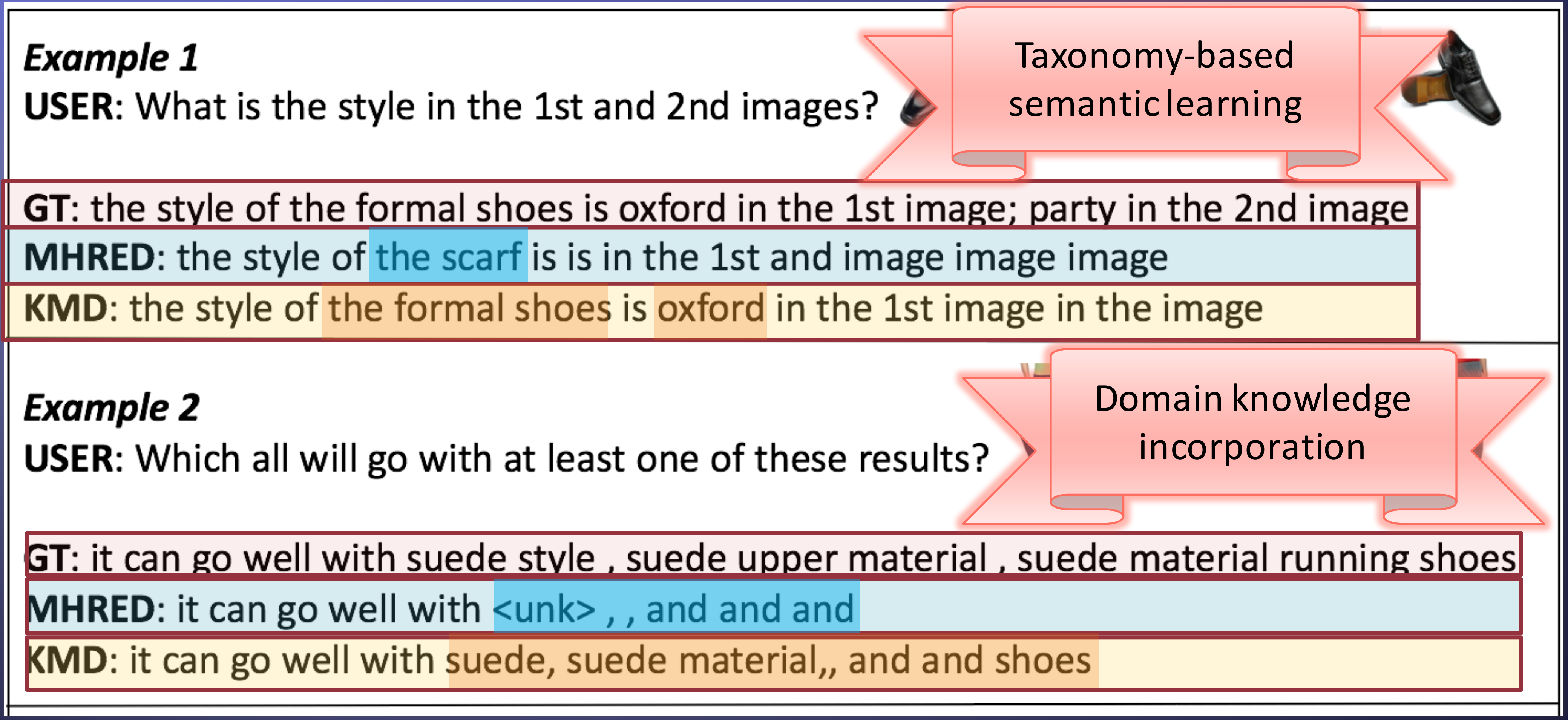
- **Dataset:** 150 K conversation sessions, 1.05 M products, avg. 4 images each
  - + TK ◆ learns more informative representations for fashion products
  - + EK ◆ generates responses not only based on conversation context but also on domain knowledge
  - + RL ◆ fine-tunes the backbone network and optimize the BLEU score or image similarity as rewards





Experiments

- Sample responses



Example 1

USER: What is the style in the 1st and 2nd images?

Taxonomy-based semantic learning

GT: the style of the formal shoes is oxford in the 1st image; party in the 2nd image

MHRED: the style of the scarf is is in the 1st and image image image

KMD: the style of the formal shoes is oxford in the 1st image in the image

Example 2

USER: Which all will go with at least one of these results?

Domain knowledge incorporation

GT: it can go well with suede style , suede upper material , suede material running shoes

MHRED: it can go well with <unk> , , and and and

KMD: it can go well with suede, suede material,, and and shoes

Conclusion and Future Work

- Multimodal Dialogue Systems
  - Offer an effective way for information seeking
  - Provide a general scheme for dialogue systems with in-depth visual understanding
  - Emphasize domain knowledge incorporation for enhancing bot intelligence

- Future Work
  - Maintain and update the domain knowledge base
  - Generalize to other domains such as travel, healthcare
  - Analyze dialogue acts to increase interpretability of dialogue flow control
  - Start procedural knowledge learning for performing tasks such as nudging customers

Thank You

Q & A