Knowledge-aware Multimodal Dialogue Systems

Lizi Liao¹, Yunshan Ma¹, Xiangnan He¹, Richang Hong², Tat-Seng Chua¹

¹National University of Singapore, ²Hefei University of Technology

24 October 2018
Any similar one in blue?

How to match with it?

Is there any such restaurant nearby?

Is there any shop selling this nearby?
Early 1990s

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

Early 2000s

Multi-modal systems
e.g., Microsoft MiPad, Pocket PC
mainly focused on multimodal interface

TV Voice Search
e.g., Bing on Xbox

Early 2010s

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

2017

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”

Virtual Personal Assistants

DARPA
CAI Project

Material:
http://deepdialogue.miulab.tw

Apple Siri
(2011)

Google Now (2012)

Google Assistant (2016)

Microsoft Cortana
(2014)

Amazon Alexa/Echo (2014)

Facebook M & Bot (2016)

Google Home (2016)
Challenges

1. Understanding semantics from text and image

Hi, what can I do for you?

Show some similar dresses in blue color.

Hi

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Found some blue dresses like these.

1. Understanding semantics from text and image

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?

Hi, what can I do for you?
Challenges

1. Understanding semantics from text and image

2. Incorporating domain knowledge

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 stiletto?

Yes, it is a good match.
Challenges

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

Hi, what can I do for you?

Show some similar dresses in blue color.

Sorry, could not find anything similar.
System Overview

- **Hierarchical RNN**

  + **3 core components**

  ![Diagram](image)

  1. **Taxonomy-based Visual Semantic Learning**
  2. **Incorporation of Domain Knowledge**
  3. **Reinforcement Learning**

  - **User input**
  - **Context hidden state** $h_t$
  - **Utterance** $u_t$
  - **Response**
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

More accurate concept prediction
2. Incorporating Domain Knowledge

- **Incorporate Knowledge by Multimodal Knowledge Memory Network**

Knowledge triplets:
- \( g_1 = \{ \text{blue skater dress, match with, silver stilettos} \} \)
- \( g_2 = \{ \text{blue, match with, silver} \} \)
- \( g_3 = \{ \text{skater dress, match with, stilettos} \} \)

I like this one, what kind of shoe will it go well with it?

Multimodal Encoder

\[ h_t \]

Multimodal Knowledge Memory

\[ \Sigma \]

Domain knowledge base

> 300 K triplets

Decoder

\[ h_t + s \]

It goes well with silver stilettos
3. Training with Reinforcement Signals

- Improve dialogue flow via reinforcement signals in **two stages training**

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**

**Rewards**

- **Text response**
  \[ R(h, r) = \text{BLEU score} \]

- **Image response**
  \[ R(h, r) = \text{sim}(I, I^+) - \text{sim}(I, I^-) \]

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

---

**Text Response**

- BLEU
  - MHRED: 0.45
  - MHRED+TK: 0.5
  - MHRED+TK+EK: 0.6
  - MHRED+TK+RL: 0.7
  - KMD: 0.75

**Image Response**

- R@1
  - MHRED: 0.38
  - MHRED+TK: 0.4
  - MHRED+TK+EK: 0.5
  - MHRED+TK+RL: 0.55
  - KMD: 0.65

---

- © Dataset: 150 K conversation sessions, 1.05 M products, avg. 4 images each
Experiments

Sample responses

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

**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?

**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