Lecture 13

Multimodal Virtual Assistants

By Giovanni Campagna, Jackie Yang, and Monica Lam
Goals of this Lecture

• Multimodal: Today’s software is GUI Based
  • How do we combine the best of GUI and Voice?
• RUSS: Rapid Universal Support System
  • Think out of the box: A “meta” agent that reads instructions
  • Technology: Learn how to use semantic parsing
• DoItHere
  • Think out of the box: Mobile devices are NOT small computers
  • Technology: Natural combination of voice and touch
Most Common Chatbots

• Customer service
  • The agents are often using the same web interface
  • They can do what consumers do
  • But people like to call – why?

• Call agent written with dialogue trees + intent classification
  • Trained for each task
Webtask Instructions in NL

Customer Service Web Page

Help instructions

Help & Customer Service

Gifts, Gift Cards, and Registries › Gift Cards ›

Redeem a Gift Card

When you redeem an Amazon.com Gift Card or gift voucher to your account, the funds are stored in Your Account and will automatically apply to your next eligible order.

To redeem a gift card:

1. Find the claim code.
2. Go to Redeem a Gift Card.
3. Enter your claim code and select Apply to Your Balance.

Grounding Open-Domain Instructions to Automate Web Support Tasks
Nancy Xu, Sam Masling, Michael Du, Giovanni Campagna, Larry Heck, James Landay, Monica S Lam
Many Other Examples

Can we train one universal assistant that reads arbitrary instructions interacts with users fills in the form
Webtask Instructions in NL

Customer Service Web Page

Help instructions

1. Ask the user for the claim code
2. Go to Redeem a Gift Card.
3. Enter the claim code and select **Apply to Your Balance**.

Agent instructions

1. Ask the user for the claim code
2. Go to Redeem a Gift Card.
3. Enter the claim code and select **Apply to Your Balance**.

Quiz: How?
RUSS (Rapid universal support system)

Natural Language instructions are parsed to an intermediate DSL (ThingTalk)

that can be executed interactively in a DOM.
(1) Semantic Parsing: 6 Actions for Most Web Instructions

<table>
<thead>
<tr>
<th>Agent Action</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>@goto(url)</td>
<td>Navigate to the given URL</td>
</tr>
<tr>
<td>@enter(element_id, dict_key)</td>
<td>Find the closest match to the given dictionary key and enter its value in the given input element</td>
</tr>
<tr>
<td>@click(element_id)</td>
<td>Click on the given element</td>
</tr>
<tr>
<td>@read(element_id)</td>
<td>Read the content of the given element to the user</td>
</tr>
<tr>
<td>@say(message)</td>
<td>Read the given message to the user</td>
</tr>
<tr>
<td>@ask(dict_key)</td>
<td>Ask the user for the value of a dictionary key</td>
</tr>
</tbody>
</table>

**element_id**: element ID in the DOM (Document Object Model)
Hardest Part: From NL description for human to operation using ElementIds

Given a web page:

**DOM:**

```plaintext
element_id: 1, type = "body"
element_id: 2, type = "h1", text = "Your Orders"
element_id: 3, type = "form"

... 
element_id: 48, type = "label", text = "order number"
element_id: 49, type = "input"

... 
```

**From:**

**Instruction:** “Enter the user’s order number in the text field that says order number”

**To:**

**Action:** `@enter(text = order_number, element = 49)`
Two Steps: Add Intermediate Function: Retrieve

**DOM:**
- element_id: 1, type = "body"
- element_id: 2, type = "h1", text = "Your Orders"
- element_id: 3, type = "form"
  
  ... element_id: 48, type = "label", text = "order number"
  element_id: 49, type = "input"
  ...

**Instruction:** “Enter the user’s order number in the text field that says order number”

**Semantic Parser (1)**

**ThingTalk:**
\[ \text{@retrievel(description} = \text{"order number"}, \text{type} = \text{input}) \Rightarrow \text{@enter}(\text{text} = \text{order number}. \text{element} = \text{id}) \]

**Grounding Model (2)**

**Action:** @enter(text = order_number, element = 49)

**Grounding Function**

<table>
<thead>
<tr>
<th>Grounding Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>@retrievel(descr, type, loc, above, below, right_of, left_of) : element_id</td>
<td>Retrieves the elements matching the descriptors, returns an element_id.</td>
</tr>
</tbody>
</table>

formal description independent of the DOM
Design of Retrieve

Another Example

- Enter the user’s email (clause 3)
- in the text field under (clause 2)
- email address or username (clause 1)

Design Rationale

- Compositional
- Formal representation of NL
- Independent of the DOM

Quiz: What is the benefit?
Grounding Model: ThingTalk -> DOM Element

@retrieve(
  description="order number",
  type=input
)

Features
  type=input
  description="order number"

DOM:
  element_id: 1, type = "body"
  element_id: 2, type = "h1", text = "Your Orders"
  element_id: 3, type = "form"

  element_id: 48, type = "label", text = "order number"
  element_id: 49, type = "input"

filter by type and location

encode in SentenceBERT and find max cosine similarities of embedding with embedding of input text description

element id = 49
Semantic Parser: Language -> ThingTalk

{“@click”: 0.03, “@enter”: 0.08, “description”: 0.06, “color”: 0.01, …}

[CLS] Enter username in input field
Natural Language Web Instruction

@retrieve (description = “username

Attention
BERT

Feed Forward

Pointer Switch

LSTM
Embedding

Pool
Training

• Template-based synthesis method to generate sample instruction -> ThingTalk pairs.

• Total: 1.5M training samples from roughly 840 distinct templates.
Evaluation: The RUSS Evaluation Dataset

- 80 tasks
- 741 instructions
- 22 help centers
Entity Extraction

• Entity Strings → Placeholder tokens for parsing
• Example:
  Pre-parsing
    https://www.amazon.com/gp/help/customer/display.html?nodeId=G5D4TA7NBKQT7GW2 → “URL”
  Post-parsing
    “URL” → https://www.amazon.com/gp/help/customer/display.html?nodeId=G5D4TA7NBKQT7GW2
## Accuracy on Semantic Parsing

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RUSS (1.5M training parses)</strong></td>
<td>87.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ablations</th>
<th>Accuracy (dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RUSS (1.5M training parses)</strong></td>
<td>88.2%</td>
</tr>
<tr>
<td>– entity extraction</td>
<td>77.6%</td>
</tr>
<tr>
<td>– 1M training parses, entity extraction</td>
<td>70.0%</td>
</tr>
</tbody>
</table>
Accuracy on Grounded Instructions

<table>
<thead>
<tr>
<th>Model</th>
<th>Grounding Acc (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUSS</td>
<td>63.6%</td>
</tr>
<tr>
<td>End-to-End Baseline</td>
<td>51.1%</td>
</tr>
<tr>
<td>PhraseNode</td>
<td>46.5%</td>
</tr>
</tbody>
</table>
End-To-End Evaluation: User Study

# 1 Redeem Amazon Gift Card
# 2 Get Pinterest Ad Account Number
# 3 Log out of all Spotify accounts
# 4 Create new Walmart account
# 5 Send Google feedback
Summary of Russ

• A meta-agent
  • Trained once, run on any web service instructions
  • Read instructions → answer calls immediately
• Design of ThingTalk
  • Separate problem into semantic parsing + grounding model
• Performance
  • 76.6% end-to-end accuracy
  • 69% of users prefer RUSS over following web instructions
• 1 quarter project!
DoThisHere
Multimodal Interaction to Improve Cross-Application Tasks on Mobile Devices

Jackie (Junrui) Yang
Monica S. Lam
James A. Landay
Users do many cross-app tasks on their smartphones

- 31.8% of smartphone usage sessions involved two or more apps [1]
- Our survey on MTurk (N=75) shows eight common categories of cross-app tasks

Users do many cross-app tasks on their smartphones
Current interfaces offer limited support for information transfer.

For **simple information**, users have to read info and remember the result.

For **complex information**, users have to go back and forth between apps.

Hard to get it right with voice.
DoThisHere introduces: Query, Do, and Keep

For **simple information**, user can use **Query** and **Do** to transfer information

For **complex information**, user can keep part of the interface for reference
DoThisHere introduces: Query, Do, and Keep

For **simple information**, user can use **Query** and **Do** to transfer information.

For **complex information**, user can keep part of the interface for reference.

Insert my duo code here.

Show me the review on IMDB.

Keep this.
DoThisHere can be portable to other platforms

DoThisHere is based on existing mobile infrastructure
DoThisHere can be portable to other platforms

DoThisHere is based on existing mobile infrastructure

Mitigates the “cold start problem”
DoThisHere can support majority of cross-apps tasks

- Among the cross-app tasks we collected, DoThisHere can help on 95% of the unsupported tasks
DoThisHere may help reduce users’ cognitive load

- Among the cross-app tasks we collected, DoThisHere can help on 95% of the unsupported tasks.

- Our user study (N=12) indicated that DoThisHere may help on reducing user’s cognitive load (NASA-TLX).

Task:
1: Keep
2, 3: Query
4: Do

*: p< 0.05, **: p < 0.01
How can we better support **complex data** transfer?

How can we **support more** than what’s currently supported in voice?
ReactGenie: toolkit for multimodal interaction apps

Jackie (Junrui) Yang, Hung Bui, Karina Li, Shuning Zhang, Monica Lam, James Landay
The benefit of multimodal interaction is beyond cross-app tasks
ReactGenie tries to solve two problems of DoThisHere

- Lack of skills
- Lack of complex types on screen

10+ actions in the app vs. 2 actions via voice

More than just text!
ReactGenie is based on ReactNative

State
- Restaurant
- OrderFoods
- Reserve
- Cuisine
- Address

Food
- Name
- Type
- Price

Account
- Email
- Name
- ProfilePicture

UI
- Login Page
- Restaurant Page
- Food Detail Page
- Shopping Cart
- Account Page
ReactGenie is based on ReactNative

Automatically generates skills

Annotation + GenieLamp

State

Restaurant
OrderFoods
Reserve
Cuisine
Address

Food
Name
Type
Price

Account
Email
Name
ProfilePicture

Annotation + Mapping

UI

Login Page
Restaurant Page
Food Detail Page
Shopping Cart
Account Page

Automatically maps data and UI
GenieLamp (Genie Large lAnguage Model Parser)

Code + Annotation

// Here are all the function that we have

class Restaurant {
    string name;
    string address;
    ...

    // Examples:
    user: get me the best restaurant in Palo Alto
    parsed: Restaurant.all().matching(field: .address, value: "Palo Alto")
    ...

    // Current User
    user: order the same burger that I ordered at McDonald last time
    parsed: Order.current.addFoods(foods: Order.all().matching(field: ...

GPT-3 Prompt

Interpreter

Live Demo
Future work

- Working on from UI annotations to mapping between data and UI
- Building demo apps
  - Food ordering app
  - Messaging app
- Chart/Visualization library