Dialogue State Tracking with Convolutional Semantic Taggers

Mandy Korpusik, Jim Glass
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA, USA

ICASSP, Brighton, UK
May 16, 2019
Motivation: Spoken Diet Tracking

*Coco Nutritionist* lets you record what you ate with everyday spoken natural language.
Welcome back, Mandy!

Yesterday was low in these nutrients, so consider taking supplements:

potass. (4387 / 4700mg)
vitamin B12 (0 / 1mcg)

What are you having for breakfast?
Motivation: Nutrition Multi-turn Dialogue

Nutrition question answering

– *is grilled chicken or red meat better?*
– *What should I eat for dinner?*
– *What is a healthy breakfast*
– *Which cereal is best to keep you satisfied?*
– *How many calories in ## of food item*
– *Is milk healthy? …*

Personalized food recommendation

(Kopusik et al., CBRecSys, 2016)
Motivation: Nutrition Multi-turn Dialogue

Nutrition question answering

– *is grilled chicken or red meat better?*
– *What should I eat for dinner?*
– *What is a healthy breakfast*
– *Which cereal is best to keep you satisfied?*
– *How many calories in ## of food item*
– *Is milk healthy? …*

Personalized food recommendation

(Korusik et al., CBRecSys, 2016)
Overview

• Motivation: Nutrition

• Introduction

• Our work in 3 state tracking challenges:
  – DSTC7
  – DSTC6
  – DSTC2

• Conclusion
Introduction: Spoken Dialogue Systems

can you book a table for two in bombay in a cheap price range

Intent Detection

Semantic Tagging

can you book a table for two in bombay in a cheap price range

Database Retrieval

Restaurant Knowledge Base
Introduction: Spoken Dialogue Systems

Our Goal: Develop ability to do dialogue state tracking.
can you book a table for **two** in **bombay** in a **cheap** price range

**Dialogue State**

- cuisine: None
- location: bombay
- number: 2
- price: cheap
- atmosphere: None

System: any preference on a type of cuisine

**Dialogue State**

- cuisine: **indian**
- location: bombay
- number: 2
- price: cheap
- atmosphere: None
Dialogue State Tracking Challenges (DSTC)

- DSTC1 (2013): human-computer bus timetables
- DSTC2 and 3 (2014): human-computer restaurant info
- DSTC5 (2016): multilingual tourist info
- DSTC6 (2017): 3 tracks, end-to-end learning
- DSTC7 (2019): 3 tracks (response selection, generation, and audio-visual)
Dialogue State Tracking Challenges (DSTC)

• DSTC1 (2013): human-computer bus timetables
• DSTC2 and 3 (2014): human-computer restaurant info
• DSTC5 (2016): multilingual tourist info
• DSTC6 (2017): 3 tracks, end-to-end learning
• DSTC7 (2019): 3 tracks (response selection, generation, and audio-visual)
Student-Advisor Partial Dialogue:

**ADVISOR** / Hi! What can I help you with?
**STUDENT** / Hello! I’m trying to schedule classes for next semester. Can you help me?
**STUDENT** / Hardware has been an interest of mine.
**STUDENT** / But I don’t want too hard of classes.
**ADVISOR** / So are you interested in pursuing Electrical or Computer Engineering?
**STUDENT** / I’m undecided.
**STUDENT** / I enjoy programming but enjoy hardware a little more.
**ADVISOR** / Computer Engineering consists of both programming and hardware.
**ADVISOR** / I think it will be a great fit for you.
**STUDENT** / Awesome, I think that’s some good advice.
**STUDENT** / What classes should I take to become a Computer Engineer?
**ADVISOR** / You haven’t taken EECS 203, 280, and 270, so it may be in your best interest to take one or two of those classes next semester.
**STUDENT** / Ok. Which of those is in the morning. I like morning classes.
DSTC7: Convolutional Neural Encoder

Binary Verification: 1 (Match) / 0 (Not)

Meanpool + Sigmoid

Dropout + Batch Norm

Dot Product

Conv + ReLU

Maxpooling

0 0 0 ... 481 is the early morning and is quite similar to EECS381, so you might want to skip it.

Candidate System Response

0 0 0 ... What time does the course occur? I like afternoon classes and will find something else if it's scheduled too early.

EECS351 is after lunch. The others are before.
EECS481 is from nine to ten thirty and EECS 492 is from ten thirty to twelve.

Previous Utterances
Goal: select the best system response.

Hi!

Hello, what can I help you with today?

I'd like to book a table for six people in an expensive price range with British food.

I'm on it!

Where should it be?

In London.

Ok, let me look into some options for you.

Actually I would prefer for four.

Sure. Is there anything else to update?

No.

Ok let me look into some options for you.

api_call(British, London, Six, Expensive)

Task 1
Issuing API calls

Task 2
Updating API calls

(Bordes & Weston, 2016)
Task 1: API Call

Dialogue State
- cuisine: indian
- location: bombay
- number: 2
- price: cheap
- atmosphere: casual

api_call indian bombay two cheap casual

Restaurant Knowledge Base
Task 2: Updating API Call

Dialogue State

- cuisine: indian
- location: bombay
- number: 2
- price: cheap
- atmosphere: casual

Actually there are four of us

Dialogue State

- cuisine: indian
- location: bombay
- number: 4
- price: cheap
- atmosphere: None
DSTC6 Data

• 10,000 simulated training dialogues per task
• KB of restaurants
  – 10 cuisines
  – 10 locations
  – 3 price ranges
  – 4 party sizes
DSTC6: Related Work

- 2 challenge participants achieved 100% on all tasks:
  - Extended Hybrid Code Networks for DSTC6 (*Ham et al., 2017*)
  - Modeling Conversations to Learn Responding Policies (*Bai et al., 2017*)
DSTC6: Related Work

- 2 challenge participants achieved 100% on all tasks:
  - Extended Hybrid Code Networks for DSTC6 (Ham et al., 2017)
  - Modeling Conversations to Learn Responding Policies (Bai et al., 2017)
DSTC6: Our Binary CNN Baseline

Binary Verification: 1 (Match) / 0 (Not)

Meanpool + Sigmoid

Dot Product

Conv + ReLU

Maxpooling

... 0 0 0 ... good morning

... hello what can i help you with

... let's do moderate price range, and keep

... expensive price range for another day

... 0 0 0 ... ok let me look into

... some options for you
DSTC6: Our Full CNN Architecture

Approach:

1. Select action template with CNN.
2. Populate action template with CNN-predicted semantic tags.
DSTC6: Our Full CNN Architecture

Approach:
1. Select action template with CNN.
2. Populate action template with CNN-predicted semantic tags.
Step 1: Semantic Tagging

Problem: prior state-of-the-art Conditional Random Field (CRF) model requires hand-crafted features.

Solution: use a neural network to automatically learn features during training.
Step 1: Semantic Tagging

can you book a table for two in bombay in a cheap price range

- Number
- Location
- Price
Step 1: Generating Tagging Data

api_call indian bombay two cheap casual

cuisine: indian
location: bombay
number: 2
price: cheap
atmosphere: casual

can you book a table for two in bombay in a cheap price range i'm looking for a casual atmosphere ...
Step 1: Semantic Tagging Results

<table>
<thead>
<tr>
<th>Semantic Tag</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuisine</td>
<td>100</td>
<td>96.9</td>
<td>98.4</td>
</tr>
<tr>
<td>Location</td>
<td>100</td>
<td>95.9</td>
<td>97.9</td>
</tr>
<tr>
<td>Number</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Price</td>
<td>96.9</td>
<td>96.5</td>
<td>96.7</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>All</td>
<td>99.8</td>
<td>99.8</td>
<td>99.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Filter</th>
<th>Top-3 Highest Activation Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>french, spanish, italian</td>
</tr>
<tr>
<td>52</td>
<td>two, six, four</td>
</tr>
<tr>
<td>63</td>
<td>bombay, london, paris</td>
</tr>
</tbody>
</table>

expensive is tempting but cheap may be more reasonable

let me check if london or bombay would work
Step 2: Action Template Selection

<table>
<thead>
<tr>
<th>Action Template</th>
<th>Ranked Actions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>ok let me look into some options for you</td>
<td>1) request_api_slot</td>
</tr>
<tr>
<td>api_call</td>
<td>2) ok let me look into some options for you</td>
</tr>
<tr>
<td>i’m on it</td>
<td>ok let me look into some options for you</td>
</tr>
<tr>
<td>hello what can i help you with today</td>
<td>...</td>
</tr>
<tr>
<td>sure is there anything else to update</td>
<td>...</td>
</tr>
<tr>
<td>you’re welcome</td>
<td>Maxpool + Softmax</td>
</tr>
<tr>
<td>what do you think of this option:</td>
<td>Conv + ReLU</td>
</tr>
<tr>
<td>great let me do the reservation</td>
<td>0 0 0 ... good morning</td>
</tr>
<tr>
<td>sure let me find another option for you</td>
<td>hello what can i help you with</td>
</tr>
<tr>
<td>here it is</td>
<td>...</td>
</tr>
<tr>
<td>whenever you’re ready</td>
<td>let's do moderate price range, and keep</td>
</tr>
<tr>
<td>the option was</td>
<td>expensive price range for another day</td>
</tr>
<tr>
<td>i am sorry i don’t have an answer to that question</td>
<td>...</td>
</tr>
<tr>
<td>is there anything i can help you with</td>
<td>...</td>
</tr>
<tr>
<td>request_api_slot</td>
<td>...</td>
</tr>
</tbody>
</table>
Step 3: Final Response Generation

1) Action mask
   api_call: masked out if any slots are still unspecified.
   request_api_slot: masked out if all slots are specified.

2) Use dialogue state
   api_call: populate slots with values in dialogue state.
   request_api_slot: select the next slot missing a value.

<table>
<thead>
<tr>
<th>Slot</th>
<th>System Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuisine</td>
<td>any preference on a type of cuisine</td>
</tr>
<tr>
<td>Location</td>
<td>where should it be</td>
</tr>
<tr>
<td>Number</td>
<td>how many people would be in your party</td>
</tr>
<tr>
<td>Price</td>
<td>which price range are you looking for</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>are you looking for a specific atmosphere</td>
</tr>
</tbody>
</table>

Dialogue State
- cuisine: indian
- location: bombay
- number: 2
- price: cheap
- atmosphere: casual

Dialogue State
- cuisine: None
- location: bombay
- number: 2
- price: cheap
- atmosphere: None
DSTC6: Test Results

100% precision on both tasks

<table>
<thead>
<tr>
<th>Model</th>
<th>Task 1 P@1</th>
<th>Task 1 P@2</th>
<th>Task 1 P@5</th>
<th>Task 2 P@1</th>
<th>Task 2 P@2</th>
<th>Task 2 P@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>10.2</td>
<td>20.4</td>
<td>50.9</td>
<td>0.95</td>
<td>19.5</td>
<td>46.7</td>
</tr>
<tr>
<td>TFIDF</td>
<td>21.0</td>
<td>29.9</td>
<td>52.2</td>
<td>36.7</td>
<td>47.4</td>
<td>66.9</td>
</tr>
<tr>
<td>SVM</td>
<td>81.3</td>
<td>81.6</td>
<td>83.0</td>
<td>74.5</td>
<td>76.4</td>
<td>78.9</td>
</tr>
<tr>
<td>LSTM</td>
<td>84.3</td>
<td>90.6</td>
<td>98.5</td>
<td>77.8</td>
<td>84.0</td>
<td>97.8</td>
</tr>
<tr>
<td>Hier. LSTM</td>
<td>88.6</td>
<td>94.1</td>
<td>99.9</td>
<td>81.7</td>
<td>92.6</td>
<td>100</td>
</tr>
<tr>
<td>Bai et al.</td>
<td>99.8</td>
<td>100</td>
<td>100</td>
<td>99.7</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Ham et al.</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Binary CNN</td>
<td>78.9</td>
<td>88.9</td>
<td>99.7</td>
<td>69.0</td>
<td>79.3</td>
<td>99.6</td>
</tr>
<tr>
<td>Our Model</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
From Simulated to Real Data: WOZ 2.0

Task: predict all the user’s requested and informable slots at each turn in a restaurant booking dialogue.

<table>
<thead>
<tr>
<th>Slot</th>
<th>Type</th>
<th>Num Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>Informable, Requestable</td>
<td>75</td>
</tr>
<tr>
<td>Area</td>
<td>Informable, Requestable</td>
<td>7</td>
</tr>
<tr>
<td>Price range</td>
<td>Informable, Requestable</td>
<td>4</td>
</tr>
<tr>
<td>Name</td>
<td>Requestable</td>
<td>N/A</td>
</tr>
<tr>
<td>Address</td>
<td>Requestable</td>
<td>N/A</td>
</tr>
<tr>
<td>Phone</td>
<td>Requestable</td>
<td>N/A</td>
</tr>
<tr>
<td>Postcode</td>
<td>Requestable</td>
<td>N/A</td>
</tr>
<tr>
<td>Signature</td>
<td>Requestable</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(User: Is there any place here in the centre that serves corsica food? 
food = corsica; area = centre)

(System: What price range are you looking for? 
User: Any price range will do. 
food = corsica; area = centre; price = dontcare)

(System: There are no restaurants available matching your criteria. Would you like to try a different area, price range, or food type? 
User: Are there any restaurants in the centre that serves North American type of food? 
food = north_american; area = centre; price = dontcare)

(System: Yes. The gourmet burger kitchen serves north american food and is located in the centre part of the city. Would you like their location? 
User: Can I get the phone number? 
food = north_american; area = centre; price = dontcare; requested = phone)

(System: The phone number for gourmet burger kitchen is 01223 312598. 
User: Thank you. Good bye. 
food = north_american; area = centre; price = dontcare)
From Simulated to Real Data: WOZ 2.0

Our Goal: Show our NN models generalize to real dialogues.

Requestable Slots: 1) $P(\text{phone}) > 0.5$
2) address

0 0 0 ... Would you like their location? Can I get the phone number?
From Simulated to Real Data: WOZ 2.0

Separately trained informable slot models.

**Food Slots:**
1) $P(\text{corsica}) > 0.5$
2) indian

**Area Slots:**
1) $P(\text{centre}) > 0.5$
2) north

**Name Slots:**
1) $P(\text{None})$
2) curry prince

**Price Slots:**
1) $P(\text{None})$
2) dontcare

0 0 0 ... Is there any place here in the centre that serves corsica food?
CNN is competitive with state-of-the-art, without requiring semantic dictionaries or pre-trained word vectors.
Conclusion

Demonstrated our neural network models’ ability to do dialogue state tracking in several domains.

Future Work:

• Experiment on the remaining DSTC6 subtasks.
• Jointly train tagger and action selector as end-to-end model.
• Automatically learn action mask by adding a feature to action selector model indicating whether all slots have values.
• Apply these techniques to the nutrition domain!