Statistical Dialogue Systems
NPFL099 Statistické Dialogové systémy

12. Multimodal Systems
(+a few words about domain adaptation)

Ondřej Dušek & Vojtěch Hudeček
http://ufal.cz/npfl099
9. 1. 2020
Multimodal Dialogue Systems

• adding more modalities to voice/text
  • input:
    • touch
    • drawing
    • gaze, gestures, facial expressions
    • voice pitch/tone
    • image
  • output:
    • graphics
    • gaze, gestures, facial expressions, body movement
• either traditional/modular and mostly rule-based systems, or very experimental (not much use in practice)
Standard Multimodal DS Schema

• basically the same as voice/text DSs
• adding multiple input modules
  • for multiple modalities
  • each with its own NLU-like interpretation
  • interpretations are merged
• multiple output modules
  • each with its own generation
  • dialogue manager output is split
• typically ready-made off-the-shelf modules
  • it’s too complex/costly to build these custom
Smart Devices

• Phones, wearables, smart speakers with a display
  • incl. Google Assistant, Alexa & Siri
  • admittedly not so much dialogue, more of commands
  • cloud-based operation for most

• Input
  • touch: active & passive gestures (touch/accelerometer)
    • “raise to speak”
    • rarely visually sensing gestures
    • doesn’t support gaze

• Output
  • graphics: card interface
  • generation functions rule-based/low-level

https://www.wearable.com/android-wear/how-to-use-voice-commands-on-android-wear
https://www.cnet.com/reviews/amazon-echo-spot-review/
“Classical” Multimodal Systems

• closed-domain task-oriented dialogue systems
• map-based: town information with map input & output
  • touch / pen – drawing, map display
    • reacting to zooming, area selection
    • handwriting recognition (as alternative input)
  • similar to Google Assistant, but more interactive
• in-car: voice & button control
• custom architectures
  • off-the-shelf modules
  • rule-based touch input processing

what about here?

S: I found 3 albums by The Beatles in your collection
<shows listing on screen>
U: Play the third one.
U: Which songs are on this one?
<selects an album from listing on screen>

(Johnston et al., 2002) https://www.aclweb.org/anthology/P02-1048/
(Johnston et al., 2014) https://www.aclweb.org/anthology/W14-4335
(Becker et al., 2006) https://www.aclweb.org/anthology/P06-4015
Virtual Agents

• character face/full body
  • on screen or 3D projected (FurHat)

• a lot more outputs
  • full motion video – facial expressions, gaze, gestures, body movement
    • a lot of it “automatic”, designed to look natural/match what’s said

• additional inputs – gaze & facial expression
  • checking user engagement/sentiment

• dialogue management mostly rule-based
  • retrieval with non-linguistic inputs (Virtual Humans/SimSensei)
  • limited-domain custom rules (FurHat)

• tutoring/training, healthcare

https://vhtoolkit.ict.usc.edu/

https://youtu.be/ejczMs6b1Q4

(Al Moubayed et al., 2012)
(Rushforth et al., 2009)
(DeVault et al., 2014)

https://doi.org/10.1007/978-3-642-34584-5_9
https://doi.org/10.1007/978-3-642-04380-2_82
https://dl.acm.org/doi/10.5555/2615731.2617415
Robots

• similar to virtual agents, but with actual hardware
  • different user’s perception
    • body gestures more prominent
    • touching the robot is possible
  • situated deployment – need to track user engagement
    • is the user still talking to the robot?
  • hardware limitations
    • mostly no facial expr./gaze output, some sensors missing etc.

• off-the-shelf robots (Nao, Pepper)
  • built-in & additional sensors (e.g. Kinect)
  • custom rule-based gesture generation
  • controlled via a computer (not autonomous)

• “receptionist” – directions, information

(Bohus et al., 2014)  https://dl.acm.org/doi/10.5555/2615731.2615835
Multi-party Dialogue

- Relevant for both virtual agents & robots
  - supported by most previously mentioned projects
- How to handle multiple counterparts?
  - users or other robots/virtual agents
- gaze/engagement/speech detection
  - who’s speaking/looking etc.
- rules for multiple counterparts
  - switching gaze to address them
    - here, 3D is better than 2D (otherwise gaze ambiguous)
  - telling one to wait for another
- customer service, information

https://youtu.be/oOp4XP_ziMw
http://www.danbohus.com/
(Foster et al., 2012)
(Bohus et al., 2014)
(Skantze & Al Moubayed, 2012)
http://dl.acm.org/citation.cfm?doid=2388676.2388680
http://dl.acm.org/doi/10.5555/2615731.2615835
https://doi.org/10.1145/2388676.2388698

Interaction 1
(Socially inappropriate)

| One person, A, approaches the bar and turns towards the bartender |
| Robot (to A): How can I help you? |
| A: A pint of cider, please. |
| A second person, B, approaches the bar and turns towards the bartender |
| Robot (to B): How can I help you? |
| B: I’d like a pint of beer. |
| Robot: (Serves B) |

Interaction 2
(Socially appropriate)

| Robot (to A): How can I help you? |
| A: A pint of cider, please. |
| Robot (to B): One moment, please. |
| Robot: (Serves A) |
| Robot (to B): Thanks for waiting. |
| How can I help you? |
| B: I’d like a pint of beer. |
| Robot: (Serves B) |
Specific uses

• Air traffic controller training – radar as a modality
  • multiple agents/systems representing pilots
  • radar charting each agent’s behavior
• single ASR, many TTSs
  • varied accents
• all rule-based
  • very limited domain
  • bearings, flight levels

NPFL099 L12 2019

End-to-end Multimodal

• recent, experimental

• enhancing end-to-end DS architectures with image input
  • no video input
  • no avatars, facial expressions, gestures etc.
  • not much graphics output either

• also using off-the-shelf components
  • especially for image recognition – ready-made convolutional architectures
  • textual parts based on known architectures (HRED, MemNN etc.)

• mostly just end-to-end prediction
  • pretrained image recognition parts are kept fixed, no end-to-end training
Pretrained convolutional nets

• Data: ImageNet Challenge
  • >1M images, 1000 classes
  • just classify the object in the image
  • CNNs are way better than anything that came before them

• AlexNet – 1\textsuperscript{st} deep CNN
  • 5 conv layers, ReLU activations, max pooling & 3 dense layers

• VGGNet – improvement
  • more layers, smaller CNN kernels (3x3, 2x2 pooling with stride 2)
    • reduces # of parameters, same function

---

https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaecccc96
Pretrained CNNs

• **ResNet** – residual networks
  • trying to simplify the mappings found by CNNs
    • with regular CNNs, deeper might not be better (vanishing gradient problem)
  • “shortcuts”: adding identity / linear projection to convolutions
    • learning a **residual** CNN mapping (“what projection can’t handle”)
  • allows much deeper networks – alleviates vanishing gradients

• **Inception** – more CNN kernels in parallel
  • for detecting different-sized object features
  • 1x1 depth reductions, depth-wise concatenations
  • better results with shallower networks

https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96
(He et al., 2016) https://arxiv.org/abs/1512.03385
(Szegedy et al., 2015) http://arxiv.org/abs/1409.4842
Pretrained CNNs

• Faster R-CNN
  • object detection – harder task
  • detecting boxes (regions) for multiple objects in image

• Pipeline:
  • Region prediction network (detect salient boxes)
  • Region-of-interest pooling (consolidate features)
  • Region-based CNN (classify)
Region prediction

- pretrained VGG as feature extraction
  - features for each of the anchor points (regularly spaced in the image)
- for each anchor point, predict:
  - anchor base size & h/w ratio (e.g. 64-128-256px, 0.5/1/1.5)
  - $p$(this is object) & $p$(this is background)
  - anchor $\Delta x$, $\Delta y$, $\Delta h$, $\Delta w$
  - all of this via convolutions 😊
- trained using object/non-object anchors
- overlapping predictions unified

R-CNN classification

• basically the same as image classification (given region)
  • with one more box coordinates fix

• sharing VGG features from RPN
  • this makes it much faster (only the pooling & prediction layers are new)


**Task**: have a meaningful dialogue about an image
- close to visual QA: **human asks, system responds**
- but VD is multi-turn & human doesn’t see the image (just a caption)
  - follow-up questions possible – coreference
  - people are not primed by the image when asking questions

- not much realistic purpose other than to test the models
- dataset of 10-turn dialogues on 120k images
  - collected via crowdsourcing
  - connecting 2 people live to talk about an image
- shared challenges

[Image: A man and woman on bicycles are looking at a map.
Person A (1): where are they located
Person B (1): in city
Person A (2): are they on road
Person B (2): sidewalk next to 1
Person A (3): any vehicles
Person B (3): 1 in background
Person A (4): any other people
Person B (4): no
Person A (5): what color bikes
Person B (5): 1 silver and 1 yellow
Person A (6): do they look old or new
Person B (6): new bikes
Person A (7): any buildings
Person B (7): yes
Person A (8): what color
Person B (8): brick
Person A (9): are they tall or short
Person B (9): I can’t see enough of them to tell
Person A (10): do they look like couple
Person B (10): they are]

Caption: A man and woman on bicycles are looking at a map.

[Image: A sink and toilet in a small room.
Q3: can you see anything else ?
A3: there is a shelf with items on it
Q4: is anyone in the room ?
A4: nobody is in the room
Q5: can you see on the outside ?
A5: no, it is only inside
Q6: what color is the sink ?
A6: the sink is white
Q7: is the room clean ?
A7: it is very clean
Q8: is the toilet facing the sink ?
A8: yes the toilet is facing the sink
Q9: can you see a door ?
A9: yes, I can see the door
Q10 what color is the door ?
A10 the door is tan colored]
Base Visual Dialogue Models

(Das et al., 2017) [http://arxiv.org/abs/1611.08669](http://arxiv.org/abs/1611.08669)

**Hierarchical Recurrent Encoder**

- Each turn
- $E_t$: encoding
- $R_t$: attention over $H$
- LSTM

**Late Fusion**

- Basic encoders for everything
- Simple projection to initialize decoder
- LSTM decoder (same for all)

**Memory Network**

- Memory (1 hop only)
- Fully-connected layer
- Weighted sum
- $t \times 512$: attention over history

**CNN output**

- Current user input
- $I$, $Q_t$
- $H^c$

### Internal Attention Over History

- $Q_t$:
  - $t$ rounds of history (concatenated)
  - $Q_t = \{Q_1, Q_2, ..., Q_t\}$
Visual Dialogue Evaluation

• BLEU etc. possible but not used here
• IR setup used instead
  • system given ground-truth dialogue history + user input & 100 candidate answers to score/rank

• IR metrics:
  • ground-truth response rank (average)
  • recall@k (% cases where ground-truth is included in top k)
  • mean reciprocal rank: $\frac{1}{\text{ground truth rank}}$ (1 if ground truth is first, 0.5 if second etc.)
  • normalized discounted cumulative gain
    • for multiple acceptable answers out of the 100 candidates
    • DCG: $\sum_{i=1}^{100} \frac{c_i \text{ relevant?}}{\log_2(i+1)}$, normalize by highest possible DCG (all good answers on top)

• problem: images only give modest gain over text-only models

https://visualdialog.org/challenge/2019#evaluation
https://en.wikipedia.org/wiki/Discounted_cumulative_gain
RL for Visual Dialogue

• human questioner replaced by “Q-bot”
  • Q-bot needs to guess an image

• using HRE architecture for both
  • Q-bot produces a VGG-like image representation
  • both trained via REINFORCE
  • same reward for both: lowered difference of Q-bot’s predicted representation to the ground-truth image
    • per turn

• curriculum learning
  • start with fully supervised
  • use RL only for last k turns
  • increase k

• combination with supervised works best

(Das et al., 2017)
http://arxiv.org/abs/1703.06585
Guess What

• guessing one of the objects in an image
  • GuessWhat data (150k guessing dialogues)
• 3 models:
  • question generation – LSTM
    • running through all previous questions
    • conditioned on VGG image features & previous replies
  • “oracle” – reply generation (Y/N/NA)
    • feed-forward from LSTM question encoding
      + object category + object size & position in image
  • guesser – select object from list of candidates
    • dot product & softmax over last LSTM generator state
      + candidate objects categories & sizes/positions
    • triggered at the end of the dialogue

(Strub et al., 2017)
https://www.ijcai.org/proceedings/2017/385
Guess What

- training via RL
  - optimize the sequence of questions to find the object
  - states = each word
    - <?> = answer retrieved, <stop> = end of dialogue, guesser applied
  - simple 0/1 reward (guesser found the correct object at the end or not)
- supervised pretraining
- only question generator trained using RL (REINFORCE)
  - learns to stop after 4.1 questions on average
    - without penalty
    - guesser might be less accurate for long dialogues

<table>
<thead>
<tr>
<th></th>
<th>New Objects</th>
<th>New Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suppressed</td>
<td>39.2% ± 0.2</td>
<td>38.0% ± 0.1</td>
</tr>
<tr>
<td>Supervised</td>
<td>40.7% ± 0.1</td>
<td>39.4% ± 0.3</td>
</tr>
<tr>
<td>REINFORCE</td>
<td>46.1% ± 0.0</td>
<td>44.8% ± 0.2</td>
</tr>
<tr>
<td>REINFORCE</td>
<td>53.3% ± 0.3</td>
<td>52.3% ± 0.2</td>
</tr>
<tr>
<td>REINFORCE</td>
<td>49.5% ± 0.0</td>
<td>48.5% ± 0.2</td>
</tr>
<tr>
<td>REINFORCE</td>
<td>44.9% ± 0.1</td>
<td>45.8% ± 0.1</td>
</tr>
<tr>
<td>Human</td>
<td>84.4%</td>
<td>81.4%</td>
</tr>
<tr>
<td>Human with Guesser</td>
<td>63.8%</td>
<td>65.1%</td>
</tr>
<tr>
<td>Random</td>
<td>18.1%</td>
<td>20.2%</td>
</tr>
</tbody>
</table>
Image Chat

(Shuster et al, 2018)
http://arxiv.org/abs/1811.00945

• Open chat about an image
  • no particular task
  • specific personality traits of both participants
• Crowdsourced data
  • ~200k dialogues, 3 turns per dialogue (A-B-A)
  • A & B have predefined personalities
• Evaluation: recall@1 (out of 100 candidates)
Image Chat

- Pretrained input components
  - concatenating ResNet & Faster R-CNN image features
  - Transformer encoders pretrained on Reddit
    - 1 encoder for the dialogue context
    - 1 encoder for candidate response

- personality embedding

- multimodal combiner
  - concatenate all inputs except candidate & self-attend (transformer)
    - using 3 transformers, then summing them (~ensemble)

- dot product with candidate encoding for ranking (ranking only!)

- the same model can also do captioning and visual QA
Chat with Facial Expressions

• Data collected from movie clips
  • selected scenes with single face on screen
  • automatically annotated FACS facial features
    • movements of 18 muscles + 3 head position features
    • 48k utterances (=pretraining needed)
• facial features clustered & synced with words
  • 200 facial expression “templates”
• HRED-like setup
  • parallel encoders – words & facial in sync (+ turn-level encoders)
  • parallel decoders + fine-grained micro-gesture generator (frame-level LSTM)
  • training: supervised, RL (word F1 as reward), GAN (human/machine discriminator)
• avatar with external ASR/TTS

Shopping Dialogues

- a more practical task: fashion e-shop
  - both user & system can show images
  - user can reference/talk about images
- semi-automatic data (150k dialogues)
  - 84 predefined intents
  - generated dialogue flows
    - based on predefined state automata over intents
  - utterances for each intent written by annotators
  - backed by a large manually curated DB of fashion items with images (1M items)
    - taxonomy (man > apparel > lower body > trousers)
    - (generated) celebrity profiles/endorsements

(Saha et al., 2018)
http://arxiv.org/abs/1704.00200
Shopping Dialogues

- Models similar to visual dialogue
  - variants of multimodal HRED
  - VGG image input
- image input
  - turn-level
  - concatenated with utterance
    - seems to work better (fewer turns)
- text/image responses
  - shared encoder
  - text generation (word-by-word)
  - image ranking (needs rough retrieval)
    - so far just “select 1 out of 5”

(Saha et al., 2018) http://arxiv.org/abs/1704.00200
(Agarwal et al., 2018) http://aclweb.org/anthology/W18-6514
Domain Adaptation

• pretraining
  • BERT, but also any other model
  • weight sharing: copy weights for similar slots in target domain

• delexicalization
  • assuming your domains are similar (e.g. TVs → PCs)

• pseudo in-domain data selection
  • find data similar to your domain in the source domain

• forcing shared latent space (see few-shot end-to-end models)

• multi-task training
  • your task in source domain & different task in target domain

• partial handcrafting (see Hybrid Code Networks)
Summary

• “traditional” multimodal systems, with components
  • combination of off-the-shelf components
    • parallels for ASR/NLU & NLG/TTS in I/O modalities
  • dialogue typically quite simple
  • modalities: static graphics / touch / gaze / facial expr. / avatars / robots
  • often support multi-party dialogue

• end-to-end multimodal systems
  • mostly experimental, based on HRED with pretrained CNNs
    • VGG, ResNet, Inception (just image classification), Faster R-CNN (+object detection)
  • visual dialogue: questions & answers about an image
  • guessing: finding an object in image
  • image chat: open-domain, based on image
  • task-oriented: shopping dialogue with product images
Thanks

Contact us:

odusek@ufal.mff.cuni.cz
hudecek@ufal.mff.cuni.cz
or on Slack

Get the slides here:

http://ufal.cz/npfl099

References/Inspiration/Further:

• Volha Pethukova’s course (Uni Saarland):
  https://www.lsv.uni-saarland.de/multimodal-dialogue-systems-summer-2019/
• McTear et al. (2016): The Conversational Interface – Talking to Smart Devices
• papers referenced on slides
Exam

• Written test, ~10 questions
  • 60% = pass (C), 73%+ = B, 88%+ = A
  • points might be adapted based on your overall performance
  • expected 1 hr, but you’ll be given at least 2hrs (no pressure on time)

• Covering all lectures

• Question format
  • you’ll need to write stuff on your own (not a-b-c-d)
  • explanation of terms/concepts
    • no exact formulas needed (if needed, they might be provided)
    • but you should know the principles of how stuff works
  • relationships between concepts (“what’s the difference between X & Y”)
  • “how would you build X”
  • focused on “important” stuff – see summaries at the end of each lecture