NPFL099 Statistical Dialogue Systems

11. Multimodal Systems
(+some notes on domain adaptation)

http://ufal.cz/npfl099

Ondřej Dušek, Vojtěch Hudeček & Zdeněk Kasner

12. 12. 2022
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

https://www.lsv.uni-saarland.de/multimodal-dialogue-systems-summer-2019/
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

(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

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>
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
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

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

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** – 1st 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-7baaaecccc96
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)


Using Transformers

(Carion et al., 2020)

• Object detection: CNN + Transformer
  • trained end-to-end
  • predicts all objects at once
    • max. $k$ objects: class + box (or “null”)
  • set-based loss (any order allowed)

• Classification: Transformer-only
  • split image into 16x16 patches
    • flatten (256xRGB=768-dim)
    • pass through a single “embedding” matrix
    • feed into standard transformer
  • adding positional embeddings
  • special “start” token
    • attention here used in a feed-forward classifier

(Dosovitskiy et al., 2020)
**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
- not very deep dialogue: history only needed in ~11% cases

(Agarwal et al., 2020)
Base Visual Dialogue Models

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

Hierarchical Recurrent Encoder

Late fusion

- simple projection to initialize decoder
- LSTM decoder (same for all)
- basic encoders for everything
- VGG (same for all)

Memory Network

- memory (1 hop only)
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
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
- trained jointly with RL

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

better model: (Suglia et al., 2020) https://www.aclweb.org/anthology/2020.coling-main.95
Shopping Dialogues

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

- 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

SHOPPER: Hello
AGENT: Hello, is there anything i can help you with today?
SHOPPER: show me some espadrilles in beige/blue within $200.
AGENT: Sure. Let me just quickly browse through my catalogue

AGENT:
SHOPPER: List more in the upper material as in the 5th image and style as the 3rd and the 5th

AGENT:
SHOPPER: Can you list images of the 1st result from all orientations?

AGENT:
SHOPPER: And the 3rd one?

AGENT:

<table>
<thead>
<tr>
<th>Id</th>
<th>Automata State Type</th>
<th>State Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>greeting</td>
<td>Shopper greets</td>
</tr>
<tr>
<td>2</td>
<td>self-info</td>
<td>Shopper gives information about himself/herself</td>
</tr>
<tr>
<td>3</td>
<td>give-criteria</td>
<td>Describes his requirements</td>
</tr>
<tr>
<td>4</td>
<td>show-image</td>
<td>Agent shows relevant responses for the Shopper's query</td>
</tr>
<tr>
<td>5</td>
<td>give-image-description</td>
<td>Agent generates short description of the product, using visual and catalog information</td>
</tr>
<tr>
<td>6</td>
<td>Like/Dislike specific items / overall items, show-more</td>
<td>Shopper expresses negative or positive preference specifically towards one or more items previously or currently shown, or a overall general preference towards all the items and optionally shows a new image to possibly modify his requirements and wants to see more</td>
</tr>
<tr>
<td>7</td>
<td>show-orientation</td>
<td>Shopper wants to see an item from different orientations</td>
</tr>
<tr>
<td>8</td>
<td>show-similar</td>
<td>Shopper wants to see similar to a particular item</td>
</tr>
</tbody>
</table>
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](http://arxiv.org/abs/1704.00200)
(Agarwal et al., 2018) [http://aclweb.org/anthology/W18-6514](http://aclweb.org/anthology/W18-6514)
Using Images to boost NLU

- Grounding all words to images – “vokenization”
  - images + captions, nearest neighbor search → assigning an image to each word – voken
  - train a vokenizer (LM assigning images to tokens) on this
  - apply it to vokenize large training data for BERT

- Finetuning BERT with:
  - masked language modelling (as usual)
  - predicting a voken for all words (masked or not)
    - classification – mimicking the vokenizer

- Further finetuning for a NLU task
  - better performance when vokens were used

(Tan and Bansal, 2020)
Situated Tasks (Amazon SimBot Challenge)

- Home tasks – commander & follower
  - find best trajectory of actions to complete task
  - include dialogue, ask if unsure

- EMMA – integrated architecture for this
  - core: Transformer LM (+sparse attention)
  - “visual tokens” to represent video input
  - explicit actions/words on the output
  - pretraining on many vision & language tasks
    - captioning, image QA, location, relations

(Suglia et al., 2022) https://aclanthology.org/2022.sigdial-1.62/
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)
• “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
  • task-oriented: shopping dialogue with product images
  • situated tasks: discussing & executing household actions
Contact us:
https://ufaldsg.slack.com/
{odusek,hudecek,kasner}@ufal.mff.cuni.cz
Skype/Meet/Zoom (by agreement)

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

Next week:
last lecture & labs
• In-person written test, 10 questions covering lectures, 10 points each
  • 50% on homework assignments needed to do the test
  • counts for 75% of the grade, 25% comes from homework assignments
  • grades: 1 = 87%+, 2 = 74%+, 3 = 60%+ (for the weighted combo)
  • expected 1 hr, but you’ll be given at least 2hrs (no pressure on time)

• Question type: 2-3 sentences to answer
  • 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
  • list of possible questions to be published soon (by Dec 31)