Dialog Systems and Visual Dialog

Sayyed Nezhadi CSC2539 Feb 2017



Picks from top categories, or try using to shop

what can i help you with

🗢 codeuridea.com

What is a Dialog System?

- A **dialog system** is a machine (computer system) with the goal of conversing with human with a logical structure.
- The communication with machine can be done through **text**, **speech**, **gesture** and so on.
- A Natural Dialog System is a form of dialog system that tries to improve usability and user satisfaction by imitating human behaviour. (Berg, 2014)
- **Turing test:** a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human.



Types of Dialog System

- Goal-oriented agents: it needs to understand the user input and complete a related task with a clear goal within a limited number of dialog turns.
 - Finite-State: Restaurant reservation, airline booking, ...
 - Active Ontalogy/Frame Based: Personal assistsant, SIRI, Alexa, Google Now
- Chatbots: general conversation with a wide scope
 - Chit-chatting
 - Entertainment
 - Examples: ELISA, ALICE, APRRY, ...



Finite-State Dialog

- A series of questions to be answered by user
- Full control of the conversation by the system
- Ignoring any unrelated answers
- Simple to build and good for simple tasks
- Only one information at a time
- Very practical but not a natural dialog



From: Dan Jurafsky slides

Active Ontology/Frame Based

- More natural conversation with mixedinitiative (Conversation initiative shifts between the user and the system)
- User can ask multiple questions or give multiple information in one sentence
- Using Frame and Slots: once all mandatory slots in a frame are filled, it will generate query to a knowledge base or external systems.
- Using Natural Language Understanding to extract slots from sentences (ML can be used).

Show me all Chinese								
LIST	CUISINE							
restaurants	in Toronto.							
LIST TYPE	LOCATION							



Some texts from: Dan Jurafsky slides

Active Ontology/Frame Based - continued



Based on a figure from Jerome Bellegarda

Example: Amazon Alexa



Example: Amazon Alexa

- Skills are voice enabled apps
- For every Intent we define as many as possible sample utterances
- Sample utterances can have **slots** in them
- Slots are categorized by slot types
- There are built-in intents to start or stop a skill or ask for help.

Slot Type "FAACODES" : AAC, AAF, AAH AAI, ...

```
Intent Schema: {
    "intent":
```

Sample Utterances:

airportInfoIntent airportInfoIntent airportInfoIntent airportInfoIntent airportInfoIntent airportInfoIntent airportInfoIntent

...

{AIRPORTCODE} airport into {AIRPORTCODE} flight delay {AIRPORTCODE} info {AIRPORTCODE} flight status {AIRPORTCODE} airport {AIRPORTCODE} flight info {AIRPORTCODE}

General Chatbots

Rule-based:

- Based on pattern matching (AIML, ChatScript, Regex, ...)
- Using mental models
- Three different memory :
 - Utterance
 - Session
 - Global

Corpus-based :

- Large corpus data
- Deep neural networks
- Information retrieval (mine conversations and retrieve similar responses)

Sample Patterns (ELIZA)

<pattern>HELLO</pattern>
<random>
How do you do. Please state your problem..
Hi. What seems to be your problem?.
</random>

<pattern>YOU ARE *</pattern>
<random>

What makes you think I am <star />?.Does it please you to believe I am <star />?.Do you sometimes wish you were <star />?.Perhaps you would like to be <star />..</random>

Some texts from: Dan Jurafsky slides

A Neural Conversation Model (O. Vinyals, Q.V. Le 2015)

- Sequence to sequence (Seq2seq) mapping using recurrent neural networks (reads the input sequence one token at a time, and predicts the output sequence, also one token at a time)
- During training, the true output sequence is given to the model



Figure 1. Using the seq2seq framework for modeling conversations.

- The model is trained to maximize the cross entropy of the correct sequence given its context
- During inference, the true output sequence is not observed, simply feed the predicted output token as input to predict the next output ("greedy" inference) or use beam search, and feed several candidates at the previous step to the next step. Selection based on the probability of the sequence

A Neural Conversation Model – sample output

- Trained on IT Helpdesk Troubleshooting dataset (Training: 30M tokens, Validation: 3M tokens)
- Trained a single layer LSTM with 1024 memory cells using stochastic gradient descent with gradient clipping. The vocabulary consists of the most common 20K words, which includes special tokens indicating turn taking and actor.

Describe your problem: I forgot my password Machine: hi , this is <NAME> from helpdesk connect , what can i help you with today ? Human: hi , i forgot my password Machine: ok great Human: can you tell me how i can recover it ? Machine: i ' II need to verify who it is at that machine Machine: can we do a hangout Machine: ? Human: yes Machine: <anonymized hangout instructions and verifications> Human: cool , i am good now Machine: how ' s this looking good ? Machine: i ' II go ahead and close this chat . Machine: take care . :) Human: bye Machine: bye

Is Neural Model Good? ③

TL;DR MICROSOFT WEB

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

by James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT

f SHARE 😏 TWEET in LINKEDIN



End-To-End Dialog Systems Using Generative Hierarchical Neural Network Models (I.V. Serban et. al. 2016)

- Three RNN modules
- Encoder RNN (encoding each subsequence into a vector)
- Context RNN

(encodes all previous sub-sequences into a vector)

• Decoder RNN

(generates the next sub-sequence)

Figure 1: Computational graph for VHRED model. Rounded boxes represent (deterministic) real-valued vectors. Variables z represent latent stochastic variables.

* The randomness injected by the variable z corresponds to higher-level decisions, like topic or sentiment of the sentence.



Table 1: Wins, losses and ties (in %) of the VHRED model against the baselines based on the human study on Twitter (mean preferences \pm 90% confidence intervals)

	Long Contexts					
Opponent	Wins	Losses	Ties	Wins	Losses	Ties
VHRED vs LSTM VHRED vs HRED	32.3 ± 2.4 42.0 \pm 2.8	${f 42.5 \pm 2.6} \\ {f 31.9 \pm 2.6}$	25.2 ± 2.3 26.2 ± 2.5	$egin{array}{r} 41.9 \pm 2.2\ 41.5 \pm 2.2 \end{array}$	$\begin{array}{ccc} 2 & 36.8 \pm 2.2 \\ 8 & 29.4 \pm 2.6 \end{array}$	21.3 ± 1.9 29.1 ± 2.6
VHRED vs TF-IDF	51.6 ± 3.3	17.9 ± 2.5	30.4 ± 3.0	47.9 ± 3.4	4 11.7 \pm 2.2	40.3 ± 3.4

Table 2: Twitter examples for the neural network models. The \rightarrow token indicates a change of turn.

Context	Response
jamaica blood!! $\langle url \rangle \rightarrow lol.$ yes [name]!. [name]! \rightarrow what's poppin in toronto homie	VHRED: nothing much. just chillin. what's good with you? LSTM: something!!!!!! HRED: something!!!!!
whhhhhhyyyy is it suddenly cold in tallahassee today?? this is florida! how can i dream of the beach when my fingers are icey cold?? \rightarrow sittin on deck in sun looking at lake travis in austin! enjoying day b4 birthday	VHRED: the sun is shining!!! LSTM: i know!!!!!!!! HRED: something!!!!!!
reaches for [name]! \rightarrow - approaches him warily - hey there buddy swings him into the air - how you doin? \rightarrow tilts his head and your wariness and reaches for your face [name].	 VHRED: - looks up at you and smiles - you're so cute kisses your cheek - LSTM: - smiles - i'm fine thank you. how are you? HRED: - grins, looking at you as i look at you -
i think i want an ipad2. but i already have an ibook. want vs. need. \rightarrow if you have an ibook. then you deserve an ipad!	VHRED: that's what i'm thinking. i'm going to buy a new one.LSTM: i want one!!!!HRED: i don't know what to do with it. i don't know what to do.
ah echt? dan anrtwoord ik volgende keer via twitter als je me smst. \rightarrow haha doe maar p. goed bezig hoor myrthe d \rightarrow goed bezig? die snap ik even niet p	 VHRED: haha, ik heb er geen zin in de something, maar ik heb er geen something meer something LSTM: haha something something something HRED: something something something something something something something

End-to-End Goal-Oriented Dialog (A. Bordes et. al 2017)

- The workhorse of traditional dialog systems is **slot-filling**
- End-to-end dialog systems, usually based on neural networks, shown promising performance in **non goal-oriented** chit-chat settings, where they were trained to predict the next utterance in social media and forum threads
- Conducting goal-oriented dialog requires skills that go beyond language modeling, e.g., asking questions to clearly define a user request, querying Knowledge Bases (KBs), interpreting results from queries to display options to users or completing a transaction
- The paper shows: end-to-end dialog system based on Memory Networks can reach promising, yet imperfect, performance and learn to perform non-trivial operations

End-to-End Goal-Oriented Dialog

Goal-oriented dialog tasks:

- A user (in green) chats with a bot (in blue) to book a table at a restaurant. Models must predict bot utterances and API calls (in dark red). Task 1 tests the capacity of interpreting a request and asking the right questions to issue an API call.
- Task 2 checks the ability to modify an API call.
- Task 3 and 4 test the capacity of using outputs from an API call (in light red) to propose options (sorted by rating) and to provide extra-information.
- Task 5 combines everything.



End-to-end Memory Network (S. Sukhabaatar 2015)



Figure 1: (a): A single layer version of our model. (b): A three layer version of our model. In practice, we can constrain several of the embedding matrices to be the same (see Section 2.2).

End-to-End Goal-Oriented Dialog - results

Jynnic lie (ge	ncratcujt	lataset					
	~						
Task	Rule-based	TF-IDF	Match	Nearest	Supervised	Memory	Networks
	Systems	no type	+ type	Neighbor	Embeddings	no match type	+ match type
T1: Issuing API calls	100 (100)	5.6 (0)	22.4(0)	55.1 (0)	100 (100)	99.9 (99.6)	100 (100)
T2: Updating API calls	100 (100)	3.4 (0)	16.4(0)	68.3 (0)	68.4 (0)	100 (100)	98.3 (83.9)
T3: Displaying options	100 (100)	8.0 (0)	8.0 (0)	58.8 (0)	64.9 (0)	74.9 (2.0)	74.9 (0)
T4: Providing information	100 (100)	9.5 (0)	17.8(0)	28.6 (0)	57.2 (0)	59.5 (3.0)	100 (100)
T5: Full dialogs	100 (100)	4.6 (0)	8.1 (0)	57.1 (0)	75.4 (0)	96.1 (49.4)	93.4 (19.7)
T1(OOV): Issuing API calls	100 (100)	5.8 (0)	22.4(0)	44.1 (0)	60.0 (0)	72.3 (0)	96.5 (82.7)
T2(OOV): Updating API calls	100 (100)	3.5 (0)	16.8(0)	68.3 (0)	68.3 (0)	78.9 (0)	94.5 (48.4)
T3(OOV): Displaying options	100 (100)	8.3 (0)	8.3 (0)	58.8 (0)	65.0 (0)	74.4 (0)	75.2 (0)
T4(OOV): Providing inform.	100 (100)	9.8 (0)	17.2(0)	28.6 (0)	57.0 (0)	57.6 (0)	100 (100)
T5(OOV): Full dialogs	100 (100)	4.6 (0)	9.0 (0)	48.4 (0)	58.2 (0)	65.5 (0)	77.7 (0)
T6: Dialog state tracking 2	33.3 (0)	1.6 (0)	1.6 (0)	21.9 (0)	22.6 (0)	41.1 (0)	41.0 (0)
Concierge ^(*)	n/a	1.1 (0.2)	n/a	13.4 (0.5)	14.6 (0.5)	16.7 (1.2)	n/a ^(†)

Data extracted from a real online concierge service performing restaurant booking

Synthetic (generated) dataset

Visual Dialog (A. Das et. al. 2016)

Computer Vision and Artificial Intelligence Trends:

- Image classification
- Scene recognition
- Object detection
- Learning to play video games
- Image and video QA

What's Next?

• Visual Dialog: Ability to hold a meaningful dialog with humans in natural language about visual content



Visual Dialog – Potential Applications

 Aiding visually impaired users in understanding their surroundings or social media content

AI: 'John just uploaded a picture from his vacation in Hawaii', **Human:** 'Great, is he at the beach?', **AI:** 'No, on a mountain'

 Aiding analysts in making decisions based on large quantities of surveillance data

Human: 'Did anyone enter this room last week?', **AI:** 'Yes, 27 instances logged on camera', **Human:** 'Were any of them carrying a black bag?'

• Interacting with an AI assistant

Human: 'Alexa – can you see the baby in the baby monitor?', AI: 'Yes, I can', Human: 'Is he sleeping or playing?'

• Robotics applications (e.g. search and rescue mission)

Human: 'Is there smoke in any room around you?', AI: 'Yes, in one room', Human: 'Go there and look for people'

Visual Dialog vs. Dialog System

- Visual Dialog Task (visual analogue of the Turing Test): given an image I, a history of a dialog consisting of a sequence of question-answer pairs, and a natural language follow-up question, the task for the machine is to answer the question in free-form natural language.
- Visual Dialog is more specific than a **general chatbot** because the dialog is about a specific image.
- Visual Dialog is not geared toward a specific goal (similar to goaldriven dialog systems). Therefore slot-filling methods won't work.

Visual Dialog vs. VQA



VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ? A: One

Captioning

Two people are in a wheelchair and one is holding a racket.

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

- Q: What are their genders ?
- A: One male and one female
- Q: Which one is holding a racket ?
- A: The woman



Visual Dialog

- Q: What is the gender of the one in the white shirt ?
- A: She is a woman
- Q: What is she doing ?

Session Variables

- A: Playing a Wii game
- Q: Is that a man to her right
- A: No, it's a woman

Visual Dialog Dataset – Data Collection

- Good data for this task should include dialogs that have:
 - Temporal continuity
 - Grounding in the image
 - Mimic natural 'conversational' exchanges
- Collected visual dialog data on images from the Common Objects in Context (COCO) dataset, which contains multiple objects in everyday scenes.
- Free form, open-ended natural language questions collected via two workers chatting on Amazon Mechanical Turk (AMT) real-time
 - The '**questioner**' sees only a single line of text describing an image (caption from COCO); the image remains hidden to the questioner.
 - Their task is to ask questions about this hidden image so as to 'imagine the scene better'
 - The '**answerer**' sees the image and the caption. Their task is to answer the questions asked by their chat partner.

Visual Dialog Dataset – Data Collection



(a) What the 'questioner' sees.

(b) What the 'answerer' sees.

(c) Example dialog from our VisDial dataset.

Figure 3: Collecting visually-grounded dialog data on Amazon Mechanical Turk via a live chat interface where one person is assigned the role of 'questioner' and the second person is the 'answerer'. We show the first two questions being collected via the interface as Turkers interact with each other in Fig. 3a and Fig. 3b. Remaining questions are shown in Fig. 3c.

Visual Dialog Dataset – Analysis

- One dialog (10 question-answer pairs) on 68k images from COCO (58k train and 10k val), or a total of 680,000 QA pairs
- More natural conversation comparing to other image QA datasets because the questioner doesn't see the image (no visual priming bias)
- Higher mean-length of answers (3.1 words) and less **binary answers** (e.g. 'Yes', 'No')
- **Coreference in dialog:** 38% of questions, 22% of answers, and nearly all (99%) dialogs contain at least one pronoun
- Temporal Continuity in Dialog Topics: based on human evaluation on smaples, across 10 rounds, VisDial question have 4:55 +- 0:17 topics on average, confirming that these are not independent questions

Visual Dialog Dataset – Analysis

	# QA	# Images	Q Length	A Length	A Length > 2	Тор-1000 А	Human Accuracy
DAQUAR [31]	12,468	1,447	11.5 ± 2.4	1.2 ± 0.5	3.4%	96.4%	-
Visual Madlibs [60]	56,468	9,688	4.9 ± 2.4	2.8 ± 2.0	47.4%	57.9%	-
COCO-QA [41]	117,684	69,172	8.7 ± 2.7	1.0 ± 0	0.0%	100%	-
Baidu [12]	316,193	316,193	-	-	-	-	-
VQA [4]	614,163	204,721	6.2 ± 2.0	1.1 ± 0.4	3.8%	82.7%	\checkmark
Visual7W [62]	327,939	47,300	6.9 ± 2.4	2.0 ± 1.4	27.6%	63.5%	\checkmark
VisDial (Ours)	680,000	68,000	4.64 ± 0	3.08 ± 0	42.2%	58.1%	\checkmark

Table 2: Comparison of existing image question answering datasets with VisDial



Visual Dialog - Evaluation Protocol

- Evaluate individual responses independently at each round (t = 1, 2, ..., 10) in a retrieval or multiple-choice setup
- The model is evaluated on retrieval metrics:
 - rank of human response
 - recall@k, i.e. existence of the human response in top-k ranked responses
 - mean reciprocal rank (MRR) of the human response
- **Candidate Answers:** groundtruth, answers to 50 similar questions, 30 most popular answers, 19 random answers

$$I = \text{image} \qquad \text{INPUT}$$

$$H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \dots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})_{H_{t-1}}$$

$$Q_t = \text{question} \qquad \text{OUTPUT}$$
sorting of 100 candidate answers
$$\{A_t^{(1)}, \dots, A_t^{(100)}\}$$

Neural Visual Dialog Models

- Experimented with the encoder-decoder combinations
- Encoders: convert inputs (I, H, Q_t) into a joint representation
- In all cases, we represent I via the I2-normalized activations from the penultimate layer of VGG-16
- For each encoder E, we experiment with all possible ablated versions: $E(Q_t)$, $E(Q_t, I)$, $E(Q_t, H)$, $E(Q_t, I, H)$
- **Decoders:** rank candidate answers based on the joint representation from encoders
- Generative (LSTM) and Discriminative (Softmax)

Visual Dialog - Encoders

• Late Fusion (LF) Encoder:

- Treat H as a long string with the entire history $(H_0, ..., H_{t-1})$ concatenated.
- Q_t and H are separately encoded with 2 different LSTMs
- individual representations of participating inputs (I, H, Q_t) are concatenated and linearly transformed to a desired size of joint representation.

• Hierarchical Recurrent Encoder (HRE):

• Similar architecture as 'Hierarchical Latent Variable Encoder-Decoder Model'

• Memory Network (MN) Encoder:

- Encode $\mathbf{Q}_t~$ with an LSTM to get a 512 -d vector
- encode each previous round of history (H₀,...,H_{t-1}) with another LSTM to get a t x 512 matrix.
- Compute inner product of question vector with each history vector to get scores over previous rounds, which are fed to a softmax to get attention-overhistory probabilities.

Visual Dialog – Late Fusion Encoder



Visual Dialog – Hierarchical Recurrent Encoder



encoded by another LSTM with shared weights



Visual Dialog - Decoders

- Generative (LSTM):
 - Encoded vector is set as the initial state of the LSTM language model
 - Maximizes the log-likelihood of the ground truth answer sequence given its corresponding encoded representation (trained end-to-end)
 - Uses the model's loglikelihood scores and rank candidate answers

• Discriminative (softmax):

- Computes dot product similarity between the input encoding and an LSTM encoding of each of the answer options
- The dot products are fed into a softmax to compute the posterior probability over the options
- Maximizes the logliklihood of the correct options and options are simply ranked based on their posterior probabilities.

Visual Dialog – Generative Decoder



- Encoded vector is set as the initial state of the LSTM language model
- Maximizes the log-likelihood of the ground truth answer sequence given its corresponding encoded representation (trained end-to-end)
- Uses the model's loglikelihood scores and rank candidate answers

Visual Dialog – Discriminative Decoder



- Computes dot product similarity between the input encoding and an LSTM encoding of each of the answer options
- The dot products are fed into a softmax to compute the posterior probability over the options
- Maximizes the logliklihood of the correct options and options are simply ranked based on their posterior probabilities

Visual Dialog – Experiment Results

• Dialogs • 50K train 8K validation 10K test					-									
• Dialogs. Jok train, ok valuation, tok test								Model	MRR	R@1	R@5	R@10	Mean	
	Higher Better			Lowe	Better			(LF-Q-D	0.482	34.29	63.42	74.31	8.87
						LF-QH-D	0.505	36.21	66.56	77.31	7.89			
							Ve		LF-QI-D	0.502	35.76	66.59	77.61	7.72
	Model	MRR	R@1	R@5	R@10	Mean	nati		LF-QIH-D	0.511	36.72	67.46	78.30	7.63
Baseline	Answer prior	0.311	10.85	30.1/	11 28	31 56	Discrimi	-	HRE-QH-D	0.489	34.74	64.25	75.40	8.32
		0.202	19.05	<i>JJ</i> .1 4	40.00	20.00			HRE-QIH-D	0.502	36.26	65.67	77.05	7.79
	NN-Q	0.392	30.54	46.99	49.98	30.88			HREA-QIH-D	0.508	36.76	66.54	77.75	7.59
	NN-QI	0.385	29.71	46.57	49.86	30.90		-	\overline{MN}	0.524	36.84	67.78	78.92	7.25
(LF-Q-G	0.403	29.74	50.10	56.32	24.06		l	MN-QIH-D	0.529	37.33	68.47	79.54	7.03
	LF-QH-G	0.425	32.49	51.56	57.80	23.11	VQA }	C	SAN1-OI-D	0.506	36.21	67.08	78.16	7.74
Ð	LF-QI-G	0.437	34.06	52.50	58.89	22.31		{	HieCoAtt-QI-D	0.509	35.54	66.79	77.94	7.68
tiv	LF-QIH-G	0.430	33.27	51.96	58.09	23.04	r							
era /	HRE-QH-G	0.430	32.84	52.36	58.64	22.59		Human Accuracies						
Gen	HRE-QIH-G	0.442	34.37	53.40	59.74	21.75	_ ((Human-Q	0.441	25.10	67.37	-	4.19
	HREA-QIH-G	0.442	34.47	53.43	59.73	21.83	Human		Human-QH	0.485	30.31	70.53	-	3.91
	$\overline{MN}-\overline{QH}-\overline{G}$	0.434	33.12	53.14	59.61	22.14			Human-QI	0.619	46.12	82.54	-	2.92
l	MN-QIH-G	0.443	34.62	53.74	60.18	21.69		l	Human-QIH	0.635	48.03	83.76	-	2.83

Visual Dialog - Conclusions

- Demonstrated the first visual chatbot.
- The results and analysis indicates that there is significant scope for improvement, the authors believe this task can serve as a testbed for measuring progress towards visual intelligence.

Potential Improvements:

- Using a model to generate responses rather than ranking candidate answers
- Include language features (e.g. part-of-speech) as the inputs
- Extend it to videos

THANKS