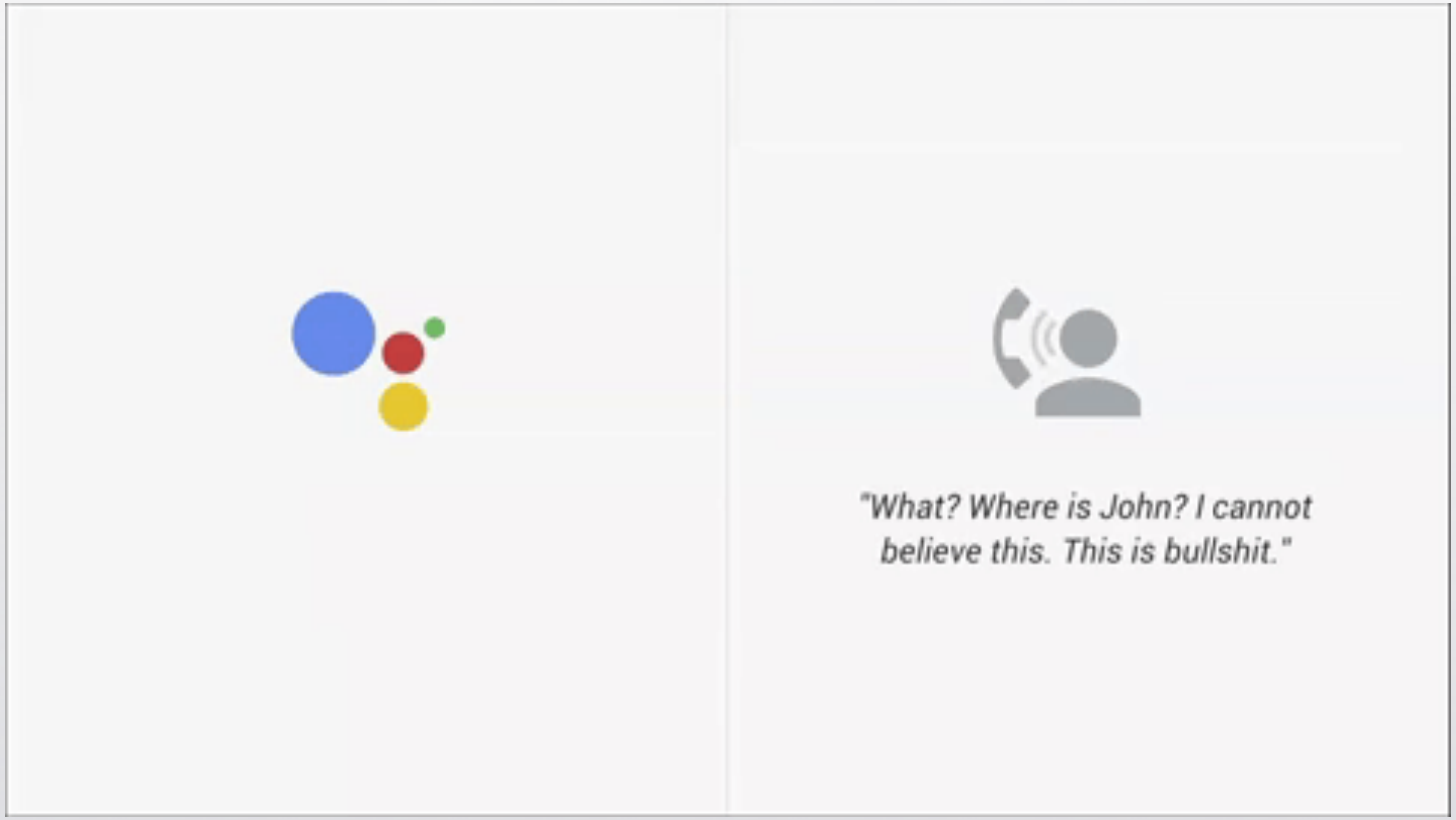
Deep Learning in Open-Domain Dialogue Systems

Yanju Chen

Google Duplex: A Task-Driven Bot



https://www.youtube.com/watch?time_continue=5&v=qB9sYGZJdbs

When a bot meets ... another bot ...

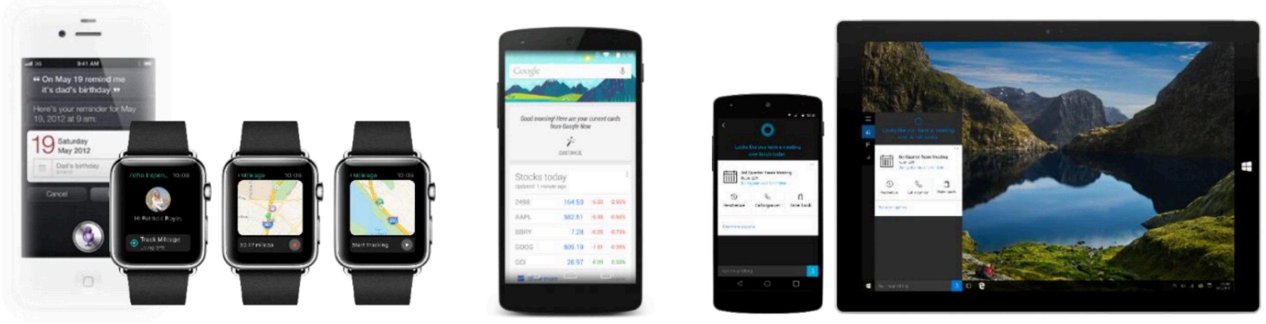
JaviAir
@Javi

Google Duplex: "Hi! Uhm... I'd like to make a dinner reservation for 3."
Restaurant: "Sure! What time would you like?"
Google Duplex: "it's, uhh... for tomorrow May the 11th at NULL POINTER EXCEPTION."
Restaurant: "Internal Exception: Invalid parameter not satisfying: time".

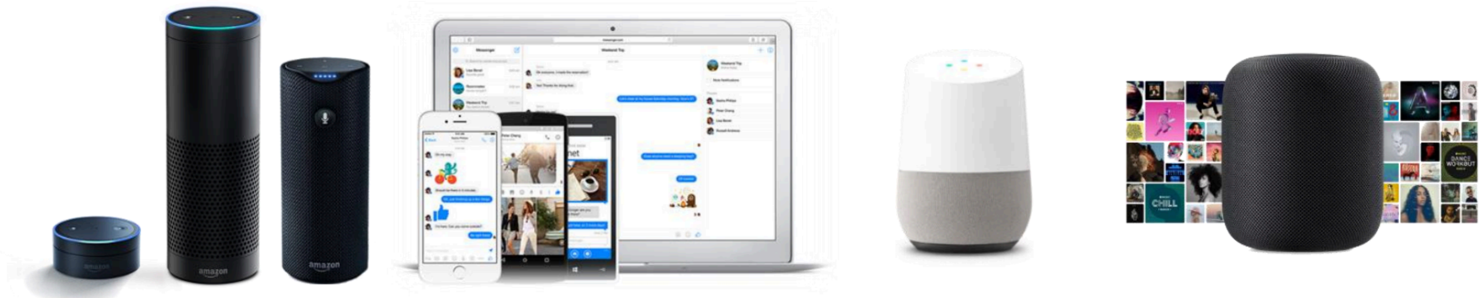
10:03 AM - May 10, 2018 · Berkeley, CA

10.9K 4,742 people are talking about this

Incorporated Dialogue Systems

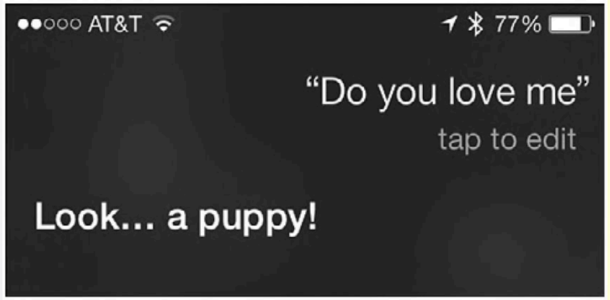


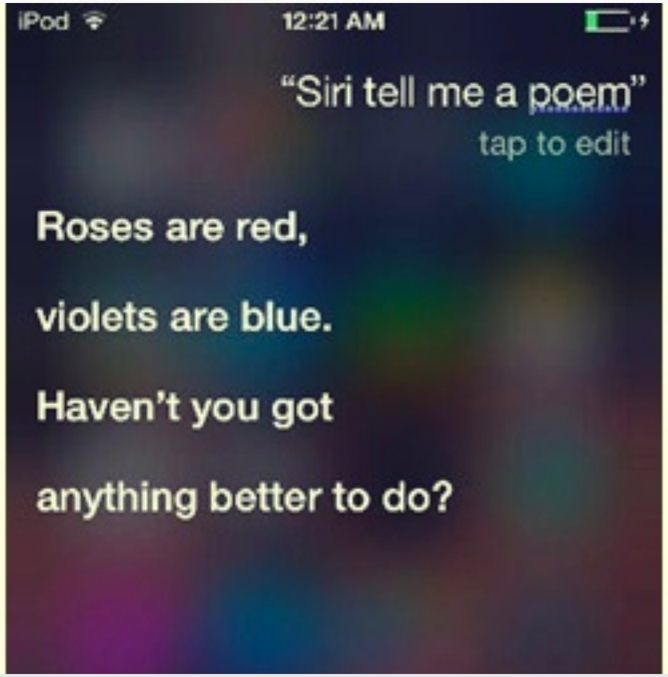
Apple Siri (2011) | Google Now (2012) Google Assistant (2016) | Microsoft Cortana (2014)



Amazon Alexa/Echo (2014) | Facebook M & Bot (2015) | Google Home (2016) | Apple HomePod (2017)

A Chit-Chat Bot

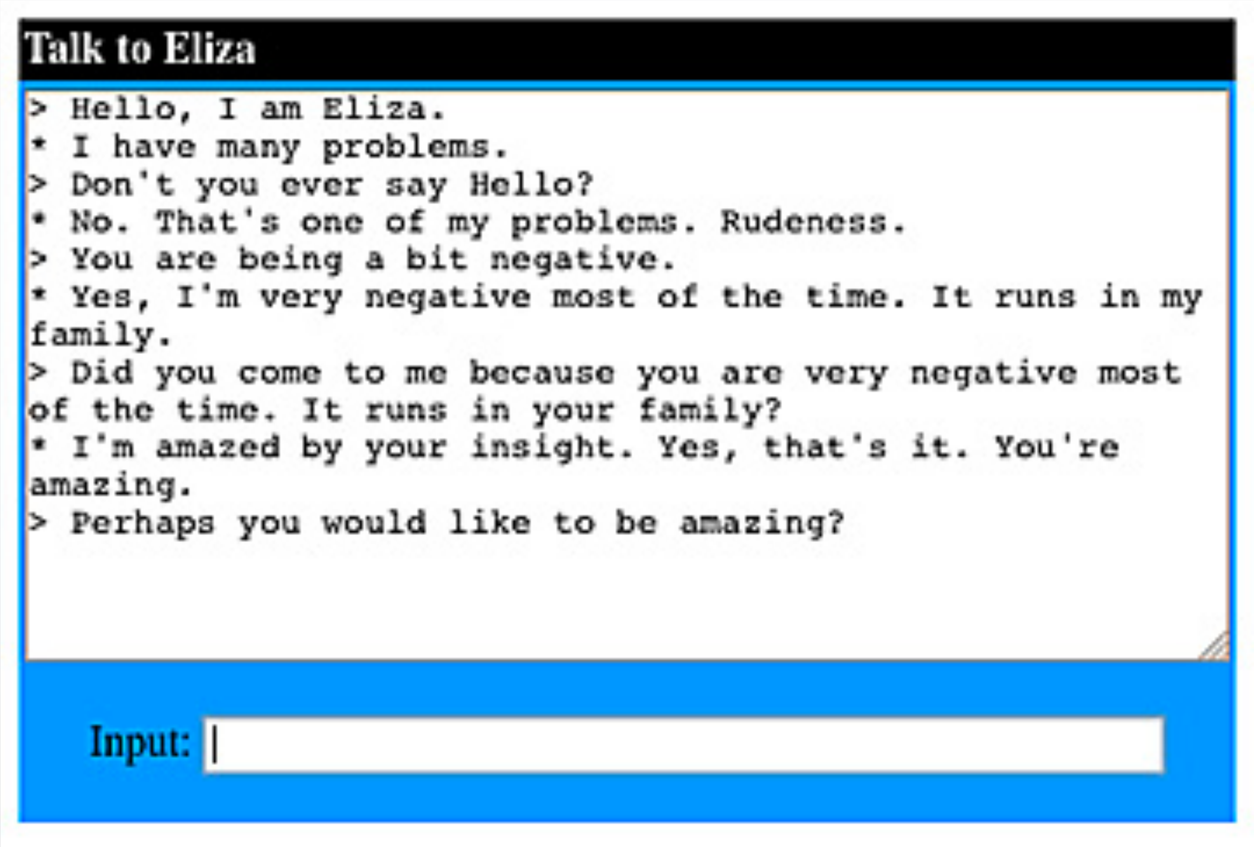




Why chit-chat bot? What's the meaning of it?

- entertainment value
- new interaction design
- scientific usage
- decrease human workload
- important part of task bot

ELIZA (1966): Psychological Therapist



Weizenbaum, “ELIZA – A computer program for the study of natural language communication between man and machine”, Communications of ACM 1966
http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm

Two Branches of Dialogue Systems

task-driven dialogue system	data-driven dialogue system (today's topic)
Task-Oriented Bot	**Chit-Chat Bot**

Task-Oriented Bot (task-driven dialogue system)

- Personal assistant, helps users achieve a certain task
- Combination of rules and statistical components
  - POMDP for spoken dialog systems (Williams and Young, 2007)
  - End-to-end trainable task-oriented dialogue system (Wen et al., 2016)
  - End-to-end reinforcement learning dialogue system (Li et al., 2017; Zhao and Eskenazi, 2016)

Chit-Chat Bot (data-driven dialogue system — today's topic)

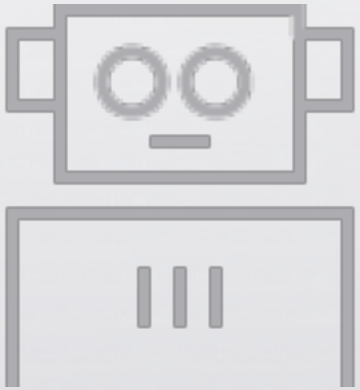
- No specific goal, focus on natural responses
- Using variants of seq2seq model
  - A neural conversation model (Vinyals and Le, 2015)
  - Reinforcement learning for dialogue generation (Li et al., 2016)
  - Conversational contextual cues for response ranking (Al-Rfou et al., 2016)

- Challenges for Chit-Chat Bots:
  - **understand** what you ask
  - generate **coherent** and **meaningful** responses
    - domain knowledge, discourse knowledge, world knowledge
  - responses should be **consistent** and **interactive**

Chen and Gao, "Open-Domain Neural Dialogue Systems", IJCNLP 2017
Li, "Deep Learning in Open Domain Dialogue Generation", Stanford University 2017

Outline

- PART I. Open-Domain Dialogue Systems
- PART II. Open-Domain Dialogue Evaluations



Open-Domain Dialogue Systems

Problem Formalization (Generative Model)

- Single-Turn Dialogue
  - (message, response) – (m, r)
- Multi-Turn Dialogue
  - (context, message, response) – (c, m, r)

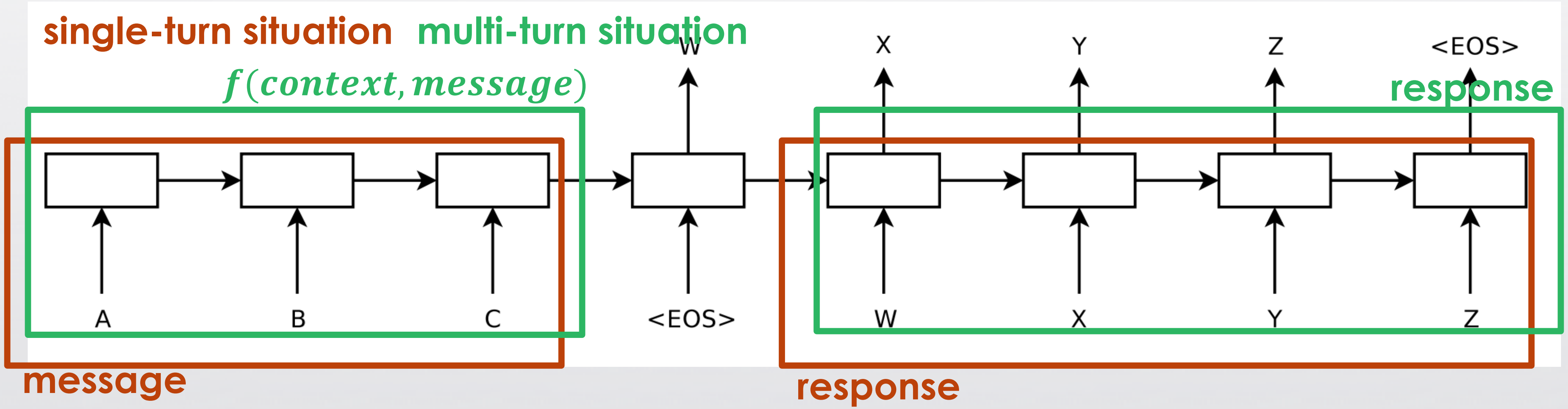
A: I'm worried about something.
B: What's that?
A: Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.
B: That's annoying, but nothing to worry about. *Just breathe deeply when you feel yourself getting upset*.
A: Ok, I'll try that.
B: Is there anything else bothering you?
A: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.
B: Do you have any other plans this weekend?
A: I'm supposed to work on a paper that'd due on Monday.
B: *Try not to take on more than you can handle*.
A: You're right. I probably should just work on my paper. Thanks!

- Goal of Generative Dialogue Model
  - to generate entirely new sentences that are unseen in the training set

Sordoni et al., "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses", ACL 2015

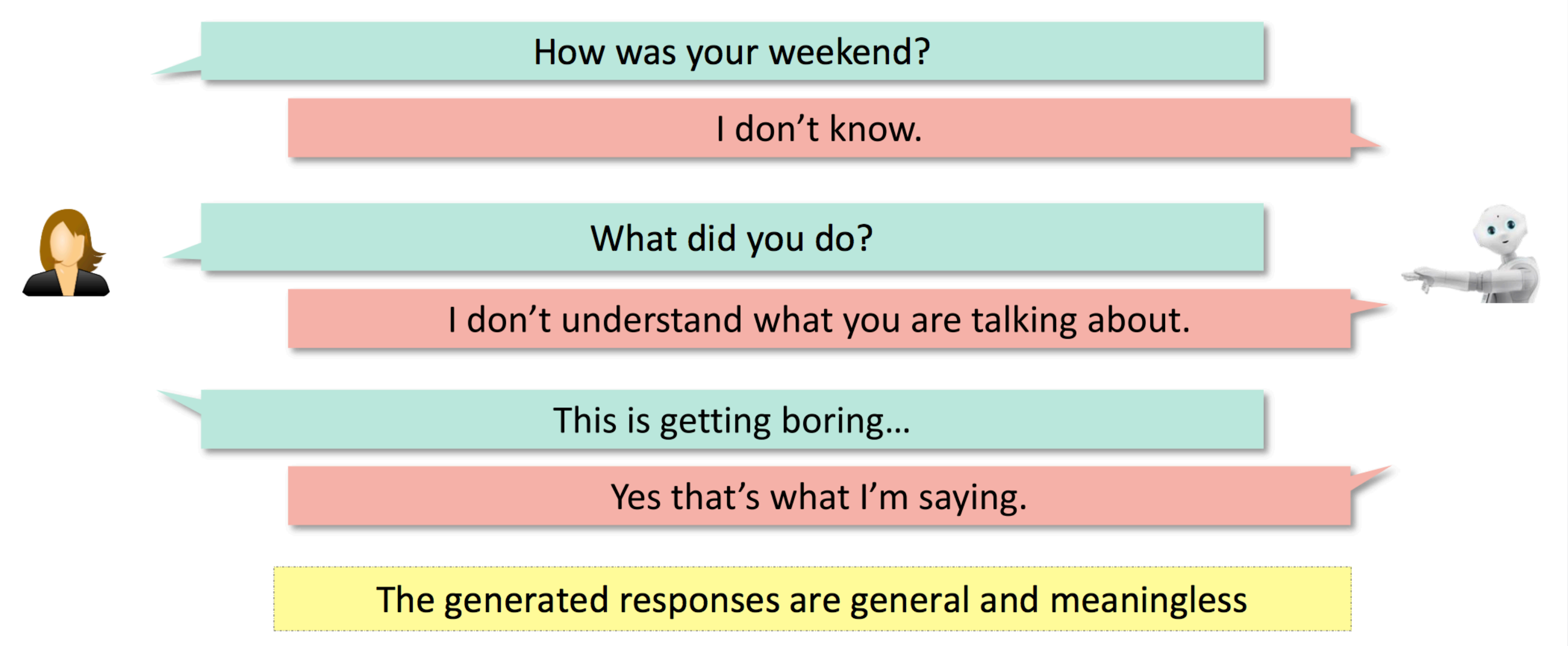
Sequence-To-Sequence Model (seq2seq)

- Vanilla Sequence-To-Sequence Model



- How about **multi-turn** situation? (context, message, response)
  - wrap (context, message) into a function and transform to a new sequence?

The Blandness Problem (Response Diversity)



- <u>**Not active or engaging at all!**</u>
- Maybe we should pay attention to:
  - how to capture dialogue **topics**
  - how to make it **human-like**

How to Capture Dialogue Topics

Figure 1: Example of three consecutive utterances occurring between two Twitter users A and B.

- In fact, there are early works on dialogue topic capturing using deep learning, even before SEQ2SEQ.

Context-Sensitive Generation

- Motivation:
  - **<u>explicitly</u>** consider and model dialogue context

($\boldsymbol{s}$) how are you ?

- Methods: **extends** the **Recurrent Language Model (RLM)**
  - given sentence $s = s_1, \ldots, s_T$, to estimate:

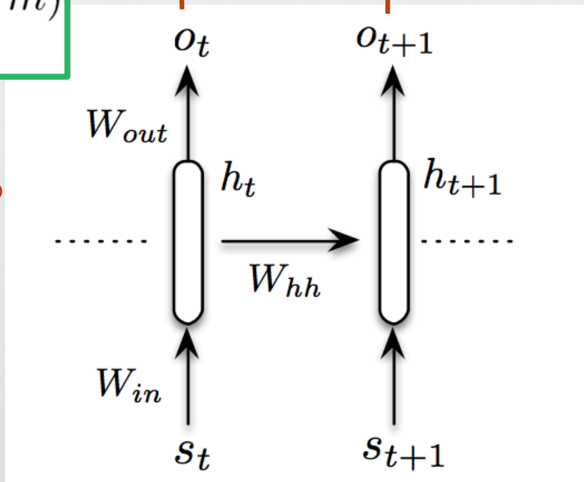
$$p(s) = \prod_{t=1}^{T} p(s_t | s_1, \ldots, s_{t-1}) \rightarrow \boxed{p(r|c,m) = \prod_{t=1}^{T} p(r_t | r_1, \ldots, r_{t-1}, c, m)}$$

probability of a natural language sentence s

What?? No sequence generated? How does this work?

$p(s_{t+2} | s_1, \ldots, s_{t+1})$

$p(s_{t+1} | s_1, \ldots, s_t)$



- Dialogue Generation before SEQ2SEQ
  - complex systems generate candidate responses
  - use features to re-rank candidate responses
  - **RLM** provides a feature for a candidate response

Sordoni et al., "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses", ACL 2015
Mikolov et al., "Recurrent Neural Network Based Language Model", INTERSPEECH 2010

Context-Sensitive Models: RLMT

- **Tripled** Language Model (RLMT) - Baseline
  - **<u>concatenate the triple</u>** (context, message, response): $s = [c; m; r]$

$(\boldsymbol{RLM})\ \boldsymbol{s} =$ ok good luck !
$(\boldsymbol{RLMT})\ \boldsymbol{s} =$ because of your game ? yeah i 'm on my way now . ok good luck !

Figure 1: Example of three consecutive utterances occurring between two Twitter users A and B.

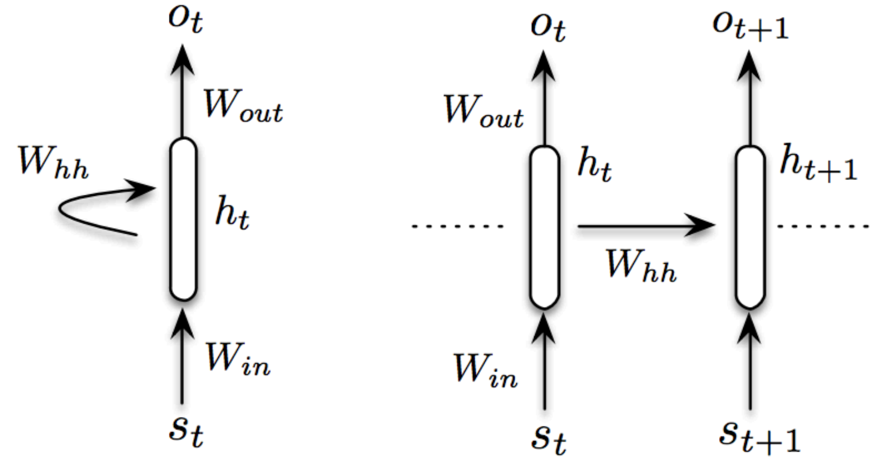


Figure 2: Compact representation of an RLM (left) and unrolled representation for two time steps (right).

context too long
computation cost

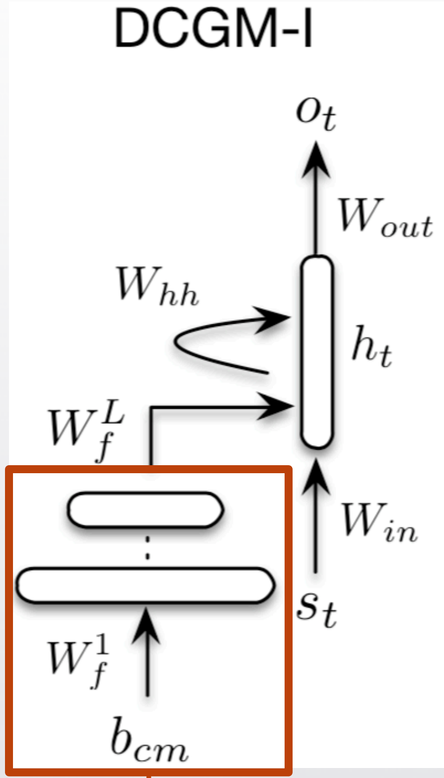
Context-Sensitive Models: DCGM-I

- Dynamic-Context Generative Models I (DCGM-I)
  - model **word occurrences** in context
  - $b_{cm} \in \mathbb{R}^V$: bag-of-words representation
  - k_L: context-message encoding
  - adding context vector as additional bias to RLM:

$$h_t = \sigma(h_{t-1}^\top W_{hh} + \boxed{k_L} + s_t^\top W_{in})$$

additional bias

do not distinguish between c and m
m and r have stronger dependency



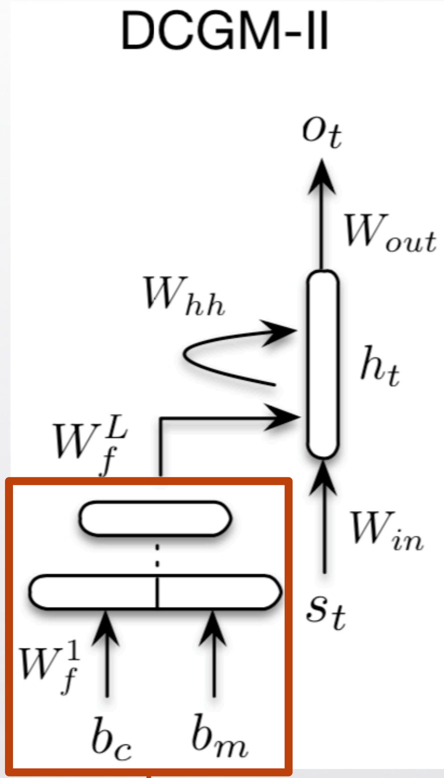
Multi Layer Perceptron

Context-Sensitive Models: DCGM-II

- Dynamic-Context Generative Models II (DCGM-II)
  - model **word occurrences**, **distinguish** context and message
  - $b_c, b_m \in \mathbb{R}^V$: bag-of-words representation
  - k_L: context-message encoding
  - adding context vector as additional bias to RLM

$$h_t = \sigma(h_{t-1}^\top W_{hh} + \boxed{k_L} + s_t^\top W_{in})$$

additional bias



Multi Layer Perceptron

Dataset & Evaluation Settings

- Dataset: selected 4,232 Twitter (c, m, r) triplets, 2,118/2,114 for train/test
- Automatic Evaluations
  - BLEU
  - METEOR
- Multi-Reference Extraction
  - **Why?** The set of reasonable responses is **vast** and **diverse**.
  - **How?** Use Information Retrieval method to **select more candidate response**.
  - Retain high-quality candidates by human evaluation.

Sordoni et al., "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses", ACL 2015

Experiment Results (Automatic Evaluation)

MT n-best	**BLEU** (%)	**METEOR** (%)	**IR n-best**	**BLEU** (%)	**METEOR** (%)
MT $_{9\ \text{feat.}}$	3.60 (*-9.5%*)	9.19 (*-0.9%*)	IR $_{2\ \text{feat.}}$	1.51 (*-55%*)	6.25 (*-22%*)
CMM $_{9\ \text{feat.}}$	3.33 (*-16%*)	9.34 (*+0.7%*)	CMM $_{9\ \text{feat.}}$	3.39 (*-0.6%*)	8.20 (*+0.6%*)
▷ MT + CMM $_{17\ \text{feat.}}$	3.98 (-)	9.28 (-)	▷ IR + CMM $_{10\ \text{feat.}}$	3.41 (-)	8.04 (-)
RLMT $_{2\ \text{feat.}}$	4.13 (*+3.7%*)	9.54 (*+2.7%*)	RLMT $_{2\ \text{feat.}}$	2.85 (*-16%*)	7.38 (*-8.2%*)
DCGM-I $_{2\ \text{feat.}}$	4.26 (*+7.0%*)	9.55 (*+2.9%*)	DCGM-I $_{2\ \text{feat.}}$	3.36 (*-1.5%*)	7.84 (*-2.5%*)
DCGM-II $_{2\ \text{feat.}}$	4.11 (*+3.3%*)	9.45 (*+1.8%*)	DCGM-II $_{2\ \text{feat.}}$	3.37 (*-1.1%*)	8.22 (*+2.3%*)
DCGM-I + CMM $_{10\ \text{feat.}}$	4.44 (*+11%*)	9.60 (*+3.5%*)	DCGM-I + CMM $_{10\ \text{feat.}}$	4.07 (*+19%*)	8.67 (*+7.8%*)
DCGM-II + CMM $_{10\ \text{feat.}}$	4.38 (*+10%*)	9.62 (*+3.5%*)	DCGM-II + CMM $_{10\ \text{feat.}}$	4.24 (*+24%*)	8.61 (*+7.1%*)

Table 3: Context-sensitive ranking results on both MT (left) and IR (right) n-best lists, $n = 1000$. The subscript $_{\text{feat.}}$ indicates the number of features of the models. The log-linear weights are estimated by running one iteration of MERT. We mark by ($\pm$%) the relative improvements with respect to the reference system (▷).

Sordoni et al., "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses", ACL 2015

Experiment Results (Human Evaluation)

System A	System B	Gain (%)	CI
HUMAN	MT+CMM	13.6*	[12.4,14.8]
DCGM-II	MT	1.9*	[0.8, 2.9]
DCGM-II+CMM	MT	3.1*	[2.0, 4.3]
DCGM-II+CMM	MT+CMM	1.5*	[0.5, 2.5]
DCGM-II	IR	5.2*	[4.0, 6.4]
DCGM-II+CMM	IR	5.3*	[4.1, 6.6]
DCGM-II+CMM	IR+CMM	2.3*	[1.2, 3.4]

Table 4: Pairwise human evaluation scores between System A and B. The first (second) set of results refer to the MT (IR) hypothesis list. The asterisk means agreement between human preference and BLEU rankings.

Sordoni et al., "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses", ACL 2015

Sample Responses from the System

(***context***) no one can help this time its just i need a support system
(***message***) well i m here to support you. whenever you need me.
(***response***) thank you, i appreciate it thank you

(***context***) nothin much, and how's the **book**?!
(***message***) its good but i'm only like halfway through cuz i don't feel like **reading**. i'm so **bored** ...
(***response***) that's good! i have the **book** but i'm **bored** too.

(***context***) thanks, your way better though
(***message***) i'm just good at expressing it that's but **trust** me you are freakin amazing and waay better
(***response***) i know, i don't **trust** you, i love you too! thank you

(***context***) today i will walk for miles. clear sky clear mind.
(***message***) love walking. it's too **hot** to walk this month. i haven't moved. **frozen**. heat freeze.
(***response***) it's too **hot** for you to get up to honey if you have **frozen** yogurt to you.

Though reasonably plausible in the context, responses are still generic or commonplace.

How to Make Dialogue Human-Like

- Several Factors:
  - Ease of Answering?
  - Information Flow? (contribute new information)
  - Semantic Coherence?
  - …
- We now use SEQ2SEQ to $\boxed{\text{generate}}$ the dialogue responses.
- What if there's a human-like model to help $\boxed{\text{discriminate}}$ all the dialogues?

Generative Adversarial Nets (GAN)

generator **discriminator**

Li et al., "Deep Reinforcement Learning for Dialogue Generation", EMNLP 2016

Design a GAN for dialogue generation

- Our Motivation:
  - Produce sequences that are **indistinguishable** from human-generated dialogue utterances.
- Problem Formalization
  - given dialogue history x: a sequence of dialogue utterances
  - to generate response $y = \{y_1, y_2, \dots, y_T\}$
- What we have:
  - Generator: SEQ2SEQ
  - Discriminator: a binary classifier $Q_+(\{x, y\})$

Li et al., "Adversarial Learning for Neural Dialogue Generation", EMNLP 2017

Training a generator: maximize the likelihood

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

objective function

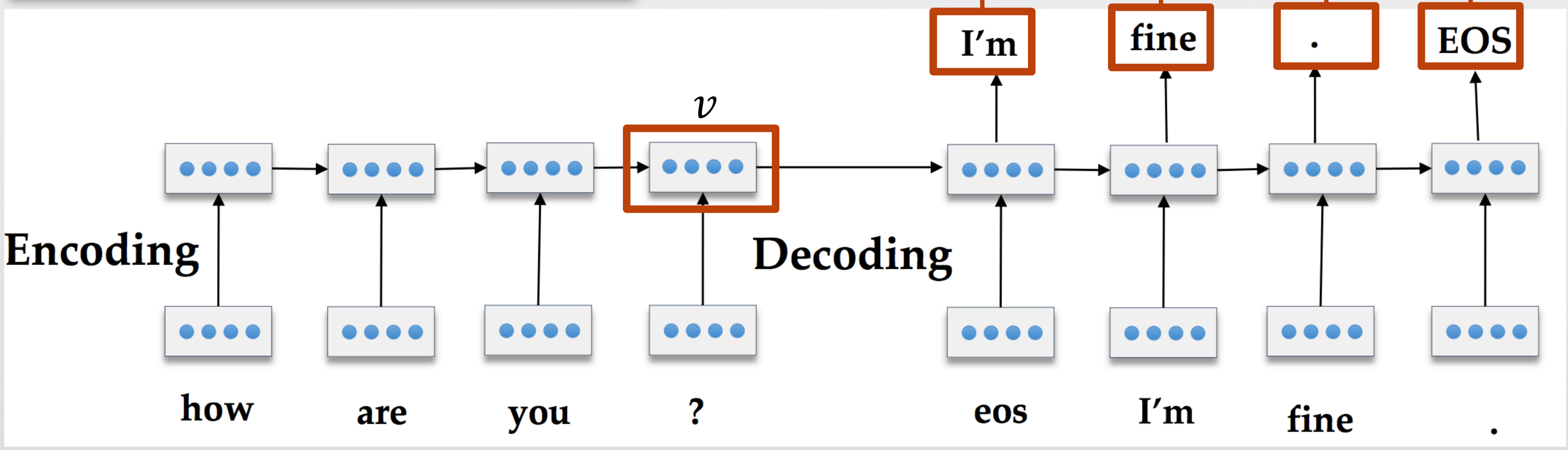
($\boldsymbol{x}$) how are you ?
($\boldsymbol{y}$) I'm fine . EOS

$p(\text{"}EOS\text{"}|v, \text{"}eos\text{"}, \text{"}I'm\text{"}, \text{"}fine\text{"}, \text{"}.\text{"})$

$p(\text{"}.\text{"}|v, \text{"}eos\text{"}, \text{"}I'm\text{"}, \text{"}fine\text{"})$

$p(\text{"}fine\text{"}|v, \text{"}eos\text{"}, \text{"}I'm\text{"})$

$p(\text{"}I'm\text{"}|v, \text{"}eos\text{"})$



Li et al., "Adversarial Learning for Neural Dialogue Generation", EMNLP 2017

Training a generator: maximize the rewards

reward

- How to use discriminator signal $\boxed{Q_{+}(\{x,y\})}$

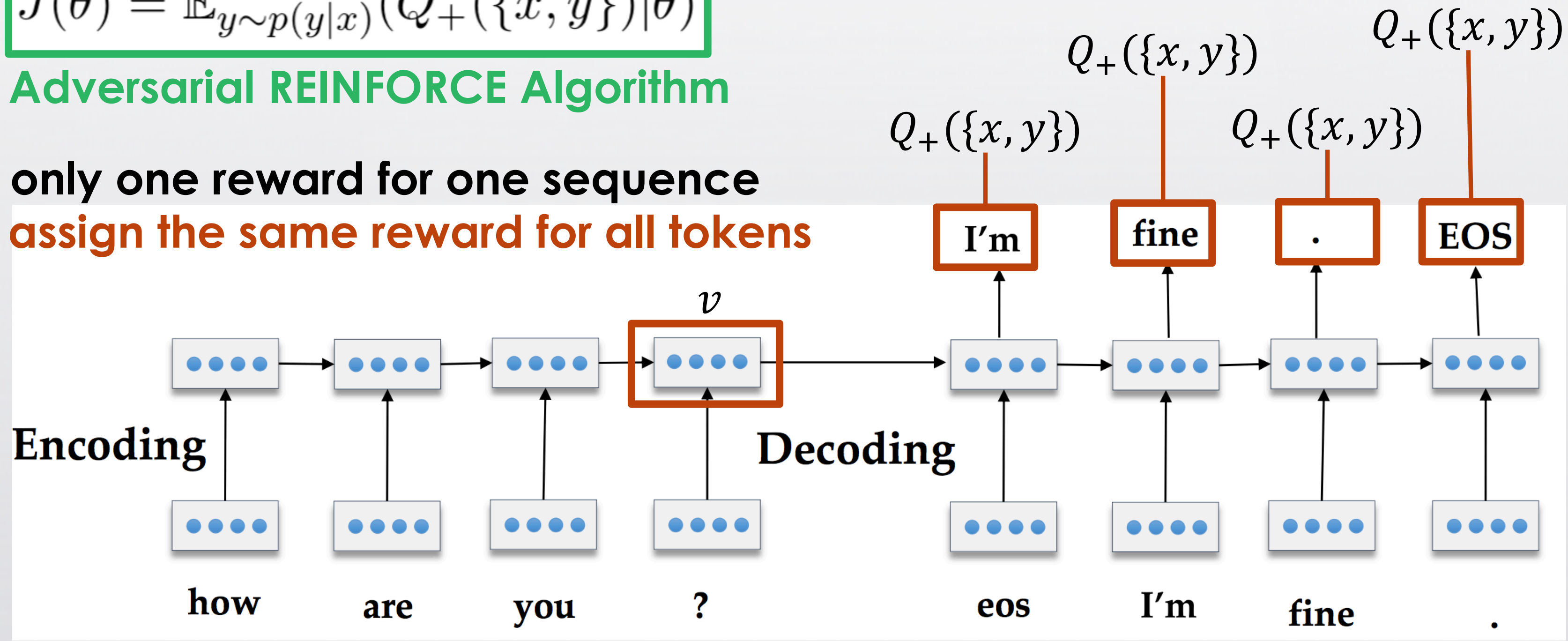
$Q_{+}(\{\text{“}how\ are\ you\ ?\text{”},\text{”}I'm\ fine\ .\ EOS\text{”}\})$

(**x**) what ‘s your name
(**y**) i am john
(**model**) i don ‘t know

$$\boxed{J(\theta)=\mathbb{E}_{y\sim p(y|x)}(Q_{+}(\{x,y\})|\theta)}$$

Adversarial REINFORCE Algorithm

only one reward for one sequence
assign the same reward for all tokens



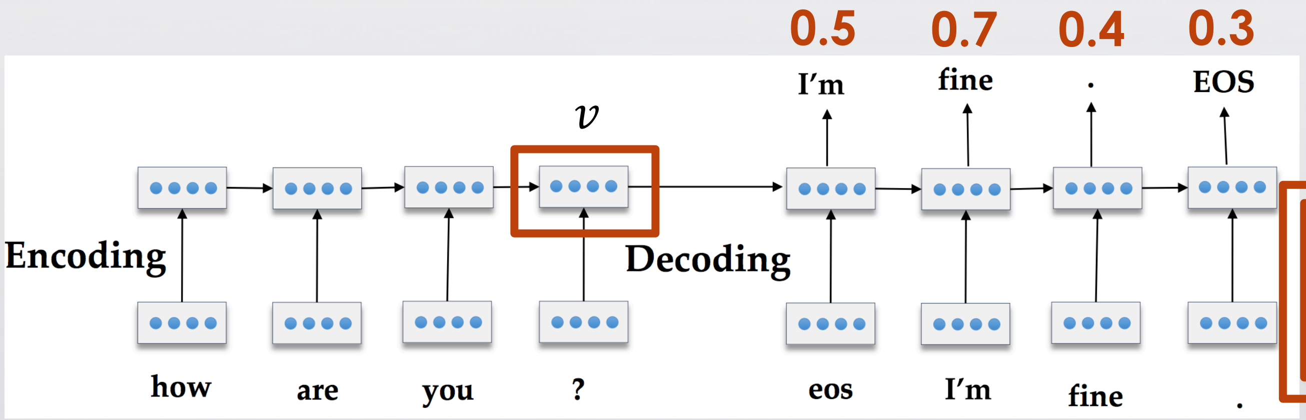
Li et al., “Adversarial Learning for Neural Dialogue Generation”, EMNLP 2017

Reward for Every Generation Step (REGS)

- The Idea **(Monte Carlo Search)**
  - estimate the quality of current token by **rolling out** complete sentence N times

$$J(\theta) = \sum_t \mathbb{E}_{y_t \sim p(y_t|x, Y_{1:t-1})}(Q_+(x, Y_t)|\theta)$$

REGS Monte Carlo



($\boldsymbol{x}$) what 's your name
($\boldsymbol{y}$) i am john
(***model***) i don 't know

average:0.5

i am jack	0.9	i don 't know	0
i am john	1	i am leaving	0.1
i am joe	0.9	i am exhausted	0.1

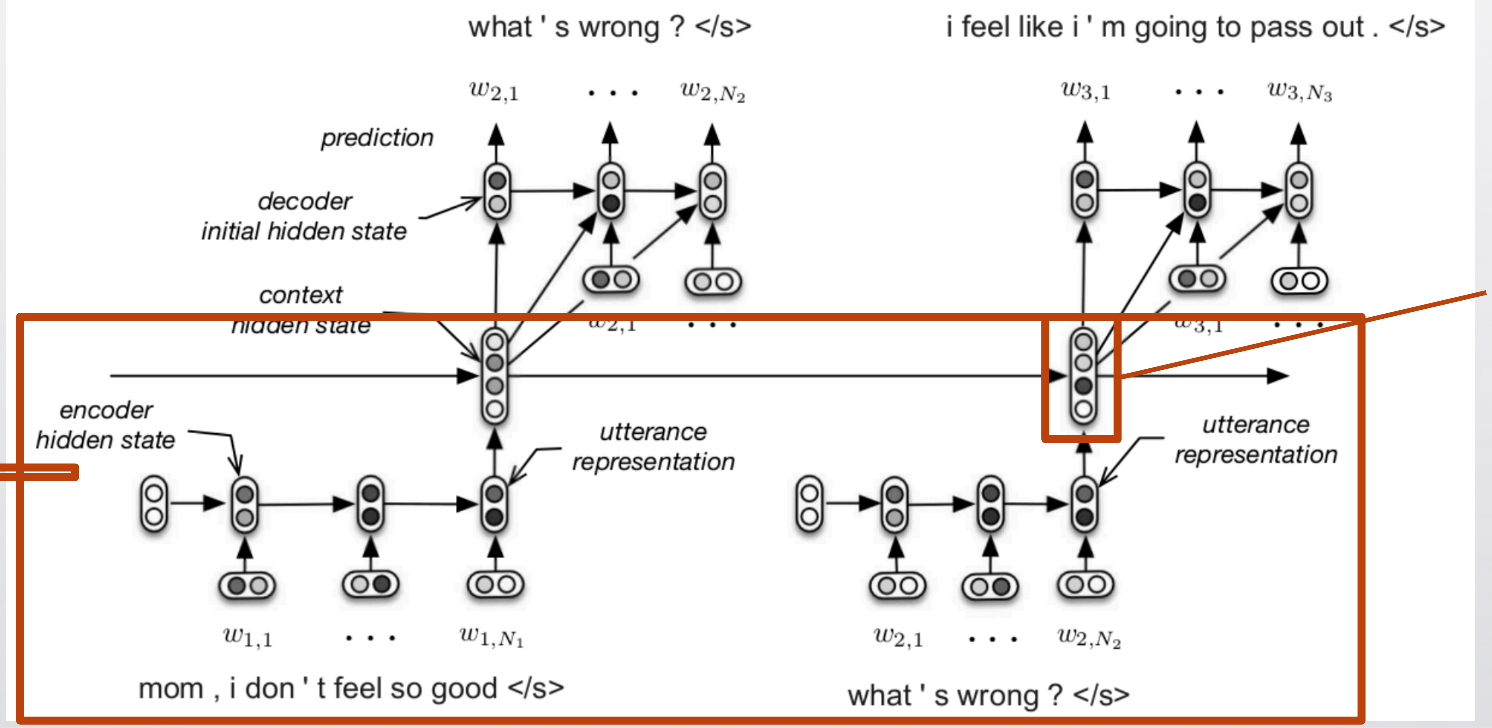
($\boldsymbol{x}$) what 's your name
($\boldsymbol{y}$) i am john
(***model***) i don 't know

average:0.6

i don 't understand	0.5
i don 't want to answer	0.8
i don 't have a name	0.7

Designing a discriminator

- What we want the discriminator to be:
  - binary classifier: $Q_+(\{x, y\})$
  - context aware: consider history x and response y



sentence level

word level

dialogue encoding

Li et al., "Adversarial Learning for Neural Dialogue Generation", EMNLP 2017
Serban et al., "Building end-to-end dialogue systems using generative hierarchical neural network models", AAAI 2016
Li et al., "A hierarchical neural autoencoder for paragraphs and documents", ACL 2015

The Adversarial Training Algorithm

For number of training iterations **do**
. **For** i=1,D-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Update D using (X, Y) as positive examples and $(X, \hat{Y})$ as negative examples.
. **End**
.
. **For** i=1,G-steps **do**
. Sample (X,Y) from real data
. Sample $\hat{Y} \sim G(\cdot|X)$
. Compute Reward r for $(X, \hat{Y})$ using D.
. Update G on $(X, \hat{Y})$ using reward r
. Teacher-Forcing: Update G on (X, Y)
. **End**
End

- Dataset: OpenSubtitles (22.27M sentence fragments)
- Pre-Train: Generator and Discriminator
- Negative Samples for Discriminator
  - beam search with mutual information reranking (Li et al., 2016)
  - sampling
- Other Tricks

Add Teacher Forcing phase to stabilize the training.

Li et al., "Adversarial Learning for Neural Dialogue Generation", EMNLP 2017

Sample Responses & Human Evaluations

(***source***) tell me ... how **long** have you had this falling **sickness** ?
(***MLE***) i 'm not a **doctor**.
(***Sample***) well everything you did was totally untrue .
(***REINFORCE***) i don 't know how **long** it 's been .
(***REGS Monte Carlo***) A few **months**, I guess .

(***source***) so i had the **doctors** test **sammy** 's response to conditioning .
(***MLE***) **sammy** wrote the **test sammy** wrote the **test** .
(***Sample***) objects pick objects up objects objects objects objects objects
(***REINFORCE***) i 'm not sure that 's relevant .
(***REGS Monte Carlo***) so he took the **pills** .

(***source***) the **man** who does that is not **afraid** to die .
(***MLE***) he 's not **afraid** of the dark .
(***Sample***) even the program is the only truth
(***REINFORCE***) he 's also not **afraid** of me .
(***REGS Monte Carlo***) he 's such a **brave man** !

Setting	adver-win	adver-lose	tie
single-turn	0.62	0.18	0.20
multi-turn	0.72	0.10	0.18

Table 4: The gain from the proposed adversarial model over the mutual information system based on pairwise human judgments.

Li et al., "Adversarial Learning for Neural Dialogue Generation", EMNLP 2017

Open-Domain Dialogue Evaluations

The Problems of Dialogue Evaluations

- NLP tasks have their own **automatic** evaluation metrics:
  - Machine Translation: BLEU, METEOR
  - Summarization: ROUGE
  - Open-Domain Dialogue Generation: ???

biased and correlate with human poorly on dialogue evaluation

- Challenges in dialogue evaluation:
  - diversity of valid responses
- Then how to evaluate a dialogue?
  - ... except for human evaluation

Context of Conversation
Speaker A: Hey John, what do you want to do tonight?
Speaker B: Why don't we go see a movie?
Potential Responses
Response 1: Nah, I hate that stuff, let's do something active.
Response 2: Oh sure! Heard the film about Turing is out!

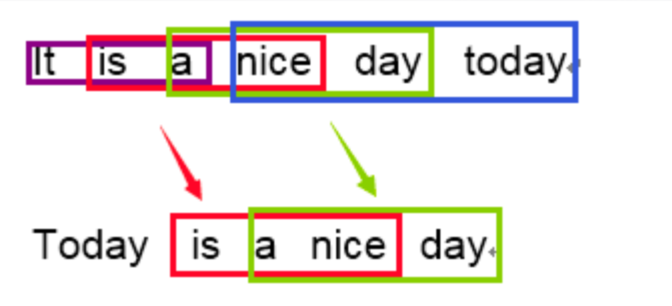
responses do not share any words

Liu et al., "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation", EMNLP 2016

Word Overlap-Based Metrics

- BLEU (Papineni et al., 2002)

unigram 5/6



trigram 2/4

- METEOR (Banerjee and Lavie, 2005)

i am attending a class

i am listening to a lecture

unigram alignment based on different rules

- ROUGE-L (Lin, 2004): Longest Common Subsequence
  - n-gram f-measure

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

It **is a nice day** today

Today **is a nice day**

Liu et al., "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation", EMNLP 2016

Embedding-Based Metrics

- Greedy Matching (word-level cosine similarity)

i found homework hard

i have difficulties finishing my homework

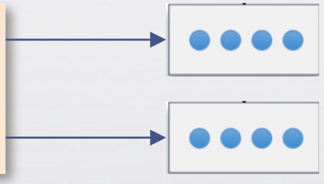
i found homework hard

i have difficulties finishing my homework

- Embedding Average (sentence-level cosine similarity)

i found homework hard

i have difficulties finishing my homework



- Vector Extrema (Forgues et al., 2014) (sentence-level)

	I	found	homework	hard	<sentence>
d1					hard
d2					found
d3					homework
d4					found
d5					I
d6					homework
d7					homework

Evaluations on Dialogue Models

	Spearman	p-value	**Pearson**	p-value
BLEU-1	0.1580	0.12	0.2074	0.038
BLEU-2	0.2030	0.043	0.1300	0.20

Table 4: Correlation between BLEU metric and human judgements after removing stopwords and punctuation for the Twitter dataset.

	Mean score $\Delta w <= 6$ (n=47)	Mean score $\Delta w >= 6$ (n=53)	p-value
BLEU-1	0.1724	0.1009	< 0.01
BLEU-2	0.0744	0.04176	< 0.01
Average	0.6587	0.6246	0.25
METEOR	0.2386	0.2073	< 0.01
Human	2.66	2.57	0.73

Table 5: Effect of differences in response length for the Twitter dataset, Δw = absolute difference in #words between a ground truth response and proposed response

Evaluations on Dialogue Models

	Twitter				Ubuntu			
Metric	**Spearman**	p-value	**Pearson**	p-value	**Spearman**	p-value	**Pearson**	p-value
Greedy	0.2119	0.034	0.1994	0.047	0.05276	0.6	0.02049	0.84
Average	0.2259	0.024	0.1971	0.049	-0.1387	0.17	-0.1631	0.10
Extrema	0.2103	0.036	0.1842	0.067	0.09243	0.36	-0.002903	0.98
METEOR	0.1887	0.06	0.1927	0.055	0.06314	0.53	0.1419	0.16
BLEU-1	0.1665	0.098	0.1288	0.2	-0.02552	0.8	0.01929	0.85
BLEU-2	0.3576	< 0.01	0.3874	< 0.01	0.03819	0.71	0.0586	0.56
BLEU-3	0.3423	< 0.01	0.1443	0.15	0.0878	0.38	0.1116	0.27
BLEU-4	0.3417	< 0.01	0.1392	0.17	0.1218	0.23	0.1132	0.26
ROUGE	0.1235	0.22	0.09714	0.34	0.05405	0.5933	0.06401	0.53
Human	0.9476	< 0.01	1.0	0.0	0.9550	< 0.01	1.0	0.0

Table 3: Correlation between each metric and human judgements for each response. Correlations shown in the human row result from randomly dividing human judges into two groups.

What’s wrong with the evaluation metrics? (on Open-Domain Dialogue)

- Automatic Metrics:
  - correlate very weakly with human judgement
  - incapable of considering the semantic similarity between responses
- Human Evaluations:
  - too expensive
  - time-consuming for every model specification

Evaluation of generative approaches	
Automatic	**Manual**
N-gram diversity (Li et al. 2016b); **BLEU** (Li et al. 2016a, Sordoni et al. 2015), **DeltaBLEU** (Galley et al. 2015); **Length metrics** (Mou et al. 2016, Li et al. 2016b); **Perplexity** (Vinyals & Le, 2015); **ROUGE** (Gu et al., 2016); **METEOR** (Sordoni et al., 2015); **Embedding-based metrics** (Serban et al. 2016b, Serban et al. 2017)	**Pairwise comparison with rule-based system** (Vinyals & Le, 2015); - **between models** (Li et al. 2016b, Wen et al. 2016, Serban et al. 2016b); **next utterance rating** (Sordoni et al. 2015) ; **5 turn 3rd party rating** (Li et al., 2016b)
+++ fast, uncostly, scalable, easily reproducible --- non-correlated with human evaluation	+++ test specific quality, representative --- costly, non-reproducible, possibly biased

Table 1: This table offers an overview on what automatic and human measures have been used for the quality evaluation of response generation by unsupervised dialogue systems. Expanded version of Helen Hastie (NIPS 2016) with evaluation of evaluation by Antoine Bordes (NIPS 2016).

Lowe et al., ”Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses”, ACL 2017

Learning to Evaluate Dialogue Responses

- Motivations:
  - **train** an **automatic dialogue evaluation model (ADEM)** to **predict human scores** and can:
    - capture **semantic similarity** beyond word overlap statistics
    - exploit both the **context** and **reference** responses

Context of Conversation Speaker A: Hey, what do you want to do tonight? Speaker B: Why don't we go see a movie?
Model Response Nah, let's do something active.
Reference Response Yeah, the film about Turing looks great!

Figure 1: Example where word-overlap scores fail for dialogue evaluation; although the model response is reasonable, it has no words in common with the reference response, and thus would be given low scores by metrics such as BLEU.

Lowe et al., "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses", ACL 2017

Automatic Dialogue Evaluation Model (ADEM)

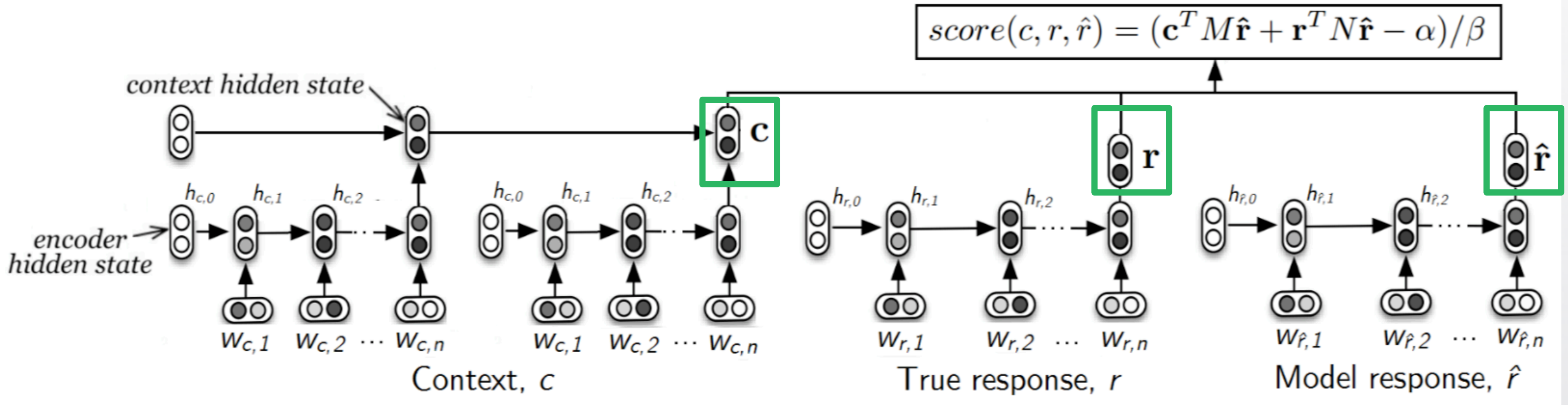


Figure 2: The ADEM model, which uses a hierarchical encoder to produce the context embedding $\mathbf{c}$.

- $M, N \in \mathbb{R}^n$: linear projection (without activation)

$$\mathcal{L} = \sum_{i=1:K} [score(c_i, r_i, \hat{r}_i) - human_i]^2 + \gamma ||\theta||_2$$

Utterance-Level Correlations

	Full dataset		Test set	
Metric	Spearman	Pearson	Spearman	Pearson
BLEU-2	0.039 (0.013)	0.081 (<0.001)	0.051 (0.254)	0.120 (<0.001)
BLEU-4	0.051 (0.001)	0.025 (0.113)	0.063 (0.156)	0.073 (0.103)
ROUGE	0.062 (<0.001)	0.114 (<0.001)	0.096 (0.031)	0.147 (<0.001)
METEOR	0.021 (0.189)	0.022 (0.165)	0.013 (0.745)	0.021 (0.601)
T2V	0.140 (<0.001)	0.141 (<0.001)	0.140 (<0.001)	0.141 (<0.001)
VHRED	-0.035 (0.062)	-0.030 (0.106)	-0.091 (0.023)	-0.010 (0.805)
	Validation set		**Test set**	
C-ADEM	0.338 (<0.001)	0.355 (<0.001)	0.366 (<0.001)	0.363 (<0.001)
R-ADEM	0.404 (<0.001)	0.404 (<0.001)	0.352 (<0.001)	0.360 (<0.001)
ADEM (T2V)	0.252 (<0.001)	0.265 (<0.001)	0.280 (<0.001)	0.287 (<0.001)
ADEM	**0.410** (<0.001)	**0.418** (<0.001)	**0.428** (<0.001)	**0.436** (<0.001)

Lowe et al., "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses", ACL 2017

Utterance-Level Correlations

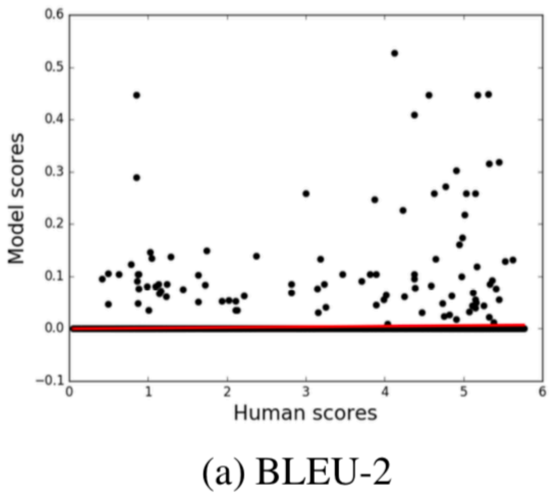


Figure 4: Scatter plot showing model against human scores, for BLEU-2 and ROUGE on the full dataset, and ADEM on the test set. We add Gaussian noise drawn from $\mathcal{N}(0, 0.3)$ to the integer human scores to better visualize the density of points, at the expense of appearing less correlated.

Lowe et al., "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses", ACL 2017

Summary

Summary

- Dialogue systems remain a challenging topic
  - diversity problem
  - context aware generation
  - higher level human-like generation
- Open-domain dialogue evaluation remains an open problem
  - extended reference
  - adversarial evaluation