Sequence Generation & Dialogue Evaluation

Ryan Lowe

McGill University

Outline

Part I: Sequence generation

- Latent variable hierarchical encoder-decoder (Serban et al., 2016)
- Actor-critic for sequence prediction (Bahdanau et al., 2016)

Part II: Dialogue evaluation

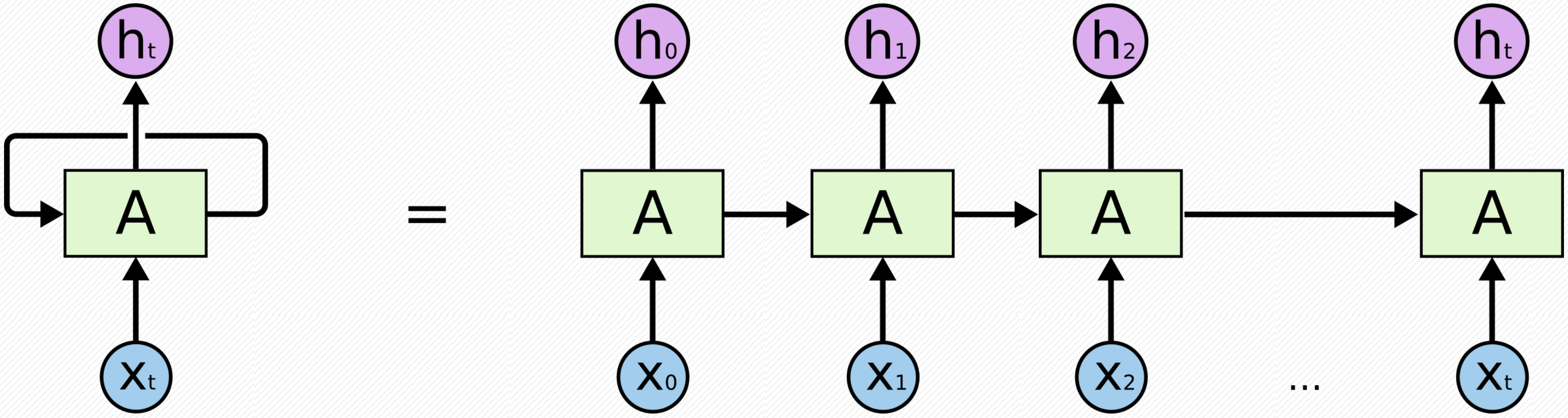
- How not to evaluate (Liu*, Lowe*, Serban*, Noseworthy* et al., 2016)
- Learning to evaluate (Lowe et al., 2016)

Part I: Sequence Generation

a) Latent variable hierarchical encoder-decoder

Recurrent neural networks

- Augment neural networks with self-loops
- Leads to the formation of a *hidden state* s_t that evolves over time: $h_t = f(W_{hh}h_{t-1} + W_{ih}x_t)$
- Used to model sequences (e.g. natural language)



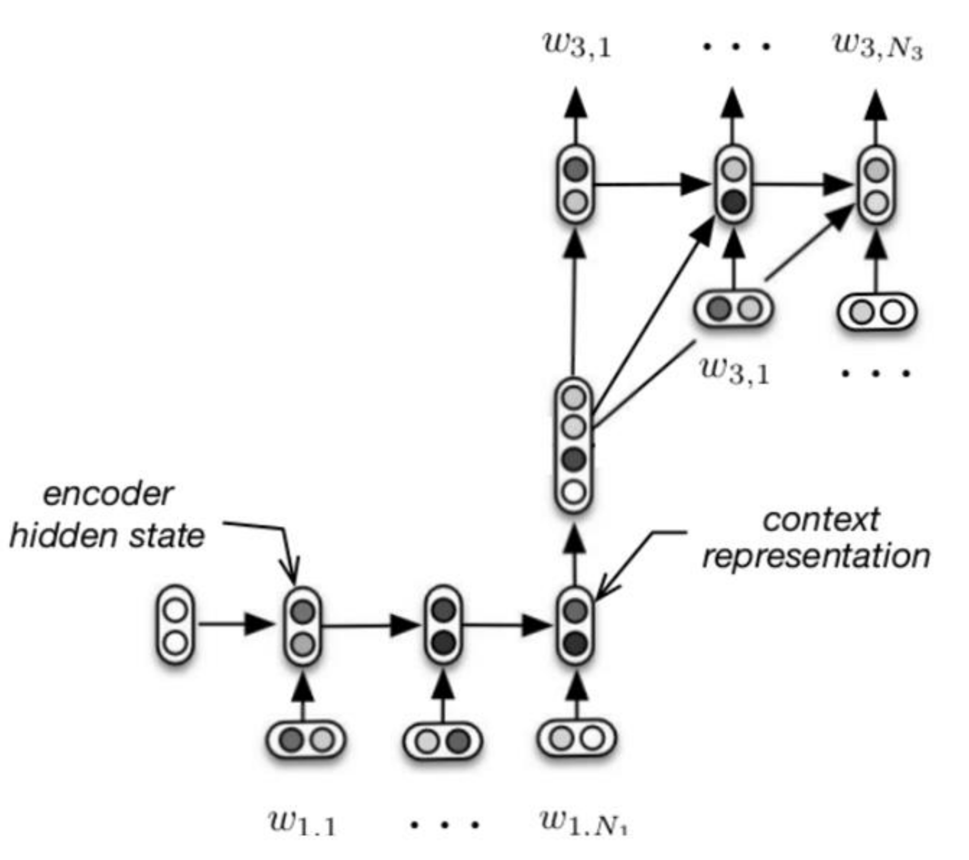
Source: colah.github.io

Sequence-to-sequence learning

- Use an RNN encoder to map an input sequence to a fixed-length vector

- Use an RNN decoder (with different parameters) to map the vector to the target sequence

(Cho et al., 2014; Sustkever et al., 2014)

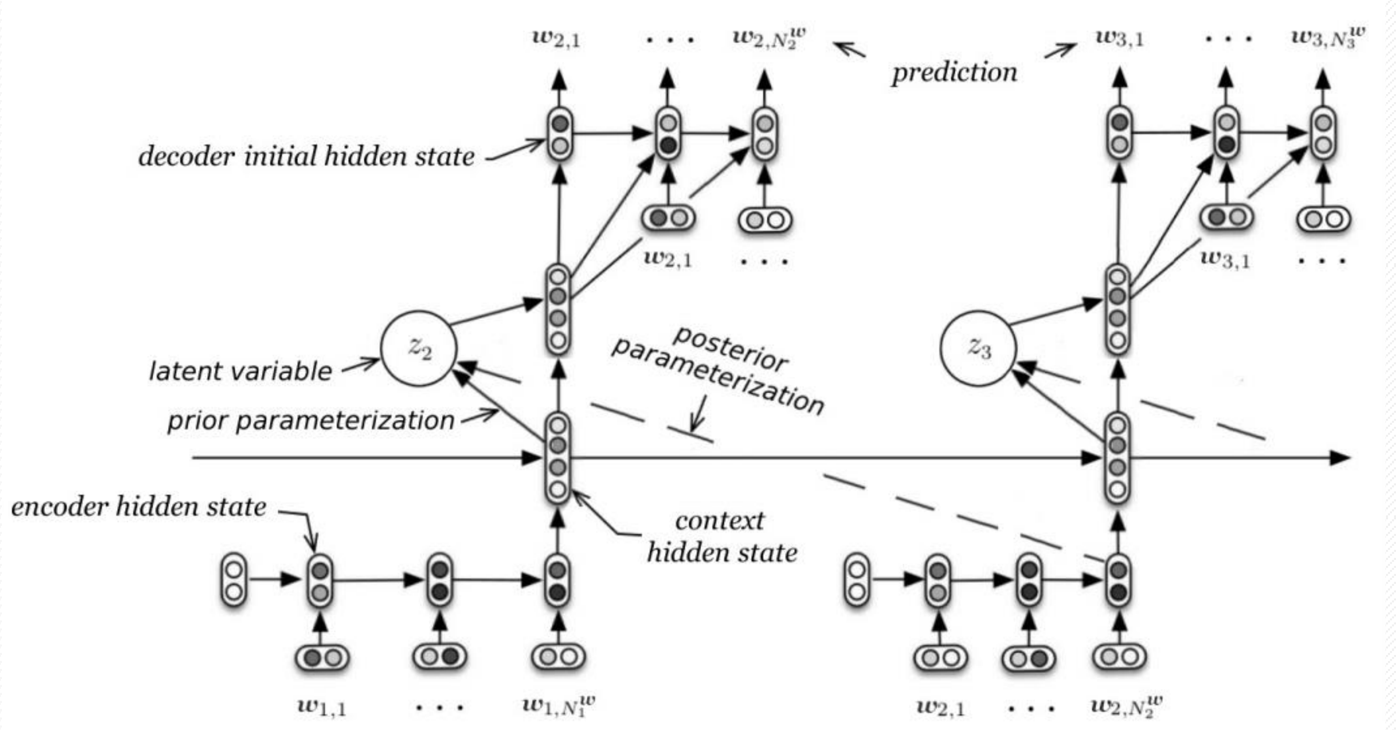


Some problems

- Strong constraint on generation process: only source of variation is at the output
- When the model lacks capacity, it is encouraged to mostly capture short-term dependencies
- Want to explicitly model variations at ‘higher level’ representations (e.g. topic, tone, sentiment, etc.)

Variational encoder-decoder (VHRED)

- Augment encoder-decoder with Gaussian latent variable *z*
- *z* can capture high-level utterance features (e.g. topic, tone)
- When generating <u>first</u> sample latent variable, <u>then</u> use it to condition generation



Serban, Sordoni, Lowe, Charlin, Pineau, Courville, Bengio. "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues." *arXiv:1605.06069*, 2016.

Variational encoder-decoder (VHRED)

- Inspired by VAE (Kingma & Welling, 2014; Rezende et al., 2014): train model with backprop using reparameterization trick
- Prior mean and variance are learned conditioned on previous utterance representation. Posterior mean and variance also conditioned on representation of target utterance.
- At training time, sample from posterior. At test time, sample from prior.
- Developed concurrently with Bowman et al. (2016)
  - Use word-dropping and KL annealing tricks

Quantitative results

Table 1: Wins, losses and ties (in %) of VHRED against baselines based on the human study (mean preferences ± 90% confidence intervals, where ∗ indicates significant differences at 90% confidence)

Opponent	Wins	Losses	Ties
Short Contexts			
VHRED vs LSTM	32.3 ±2.4	**42.5 ±2.6***	25.2 ±2.3
VHRED vs HRED	**42.0 ±2.8***	31.9 ±2.6	26.2 ±2.5
VHRED vs TF-IDF	**51.6 ±3.3***	17.9 ±2.5	30.4 ±3.0
Long Contexts			
VHRED vs LSTM	**41.9 ±2.2***	36.8 ±2.2	21.3 ±1.9
VHRED vs HRED	**41.5 ±2.8***	29.4 ±2.6	29.1 ±2.6
VHRED vs TF-IDF	**47.9 ±3.4***	11.7 ±2.2	40.3 ±3.4

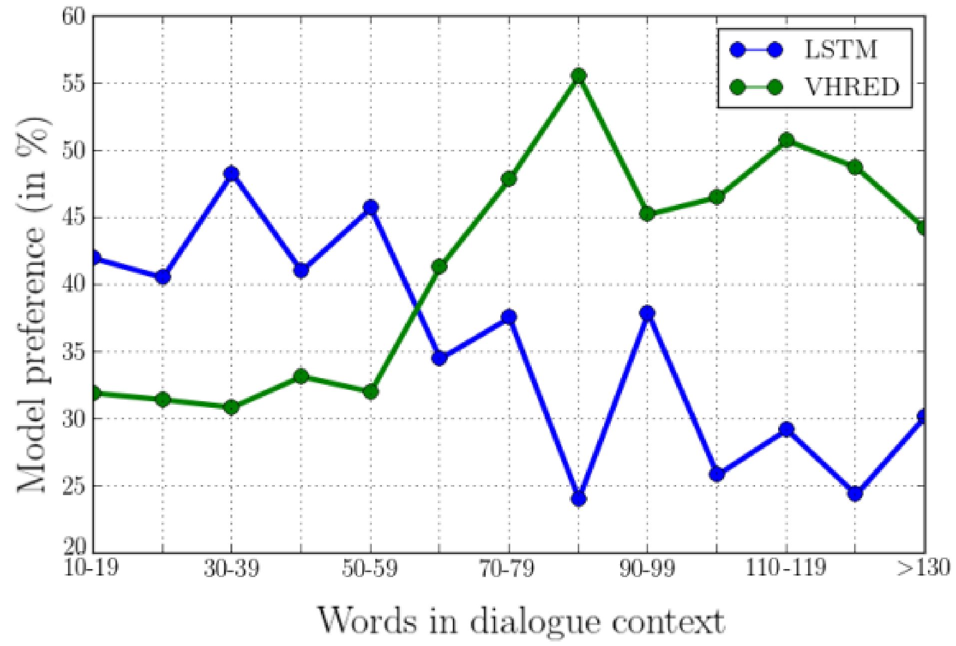


Figure 2: Human evaluator preferences for VHRED vs LSTM by context length excluding ties. For short contexts humans prefer the generic responses generated by LSTM, while for long contexts humans prefer the semantically richer responses generated by VHRED.

Cherry-picked results

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

Context	**Response**
jamaica blood!! <url> → lol. yes [name]!. [name]! → what's poppin in toronto homie	**VHRED:** nothing much. just chillin. what's good with you? **LSTM:** something!!!!!!! **HRED:** something!!!!!!
reaches for [name]! → - approaches him warily - hey there buddy. - swings him into the air - how you doin? → 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 -
my math teacher over here talkn bout nickels nd i'm tryna make big money i need to be in calculus → lol	**VHRED:** i need to go to the gym and get my hair done lol **LSTM:** lol **HRED:** how u been
i think i want an ipad2. but i already have an ibook. want vs. need. → 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. → haha doe maar p. goed bezig hoor myrthe d → 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 something

Part I: Sequence Generation

b) Actor-critic for sequence prediction

Some more problems

- Discrepancy between training and test times due to teacher forcing (conditioning next prediction on previous ground-truth output)
- Often want to maximize a task-specific score (e.g. BLEU) instead of log-likelihood

RL background

- Have states s, actions a, rewards r, policy $\pi = p(a|s)$

- Return: $R = \sum_{t=0}^{T} \gamma^t r_{t+1}$

- Value function: $V(s_t) = \mathrm{E}_{a\sim\pi}[R|s_t]$
- Action-value function: $Q(s_t, a_t) = \mathrm{E}_{a\sim\pi}[R|s_t, a_t = a]$

TD learning

- Methods for policy evaluation (i.e. calculating the value function for a policy)
- Monte Carlo learning: wait until end of the episode to observe the return R

$$V(s_t) = V(s_t) + \alpha[R - V(s_t)]$$

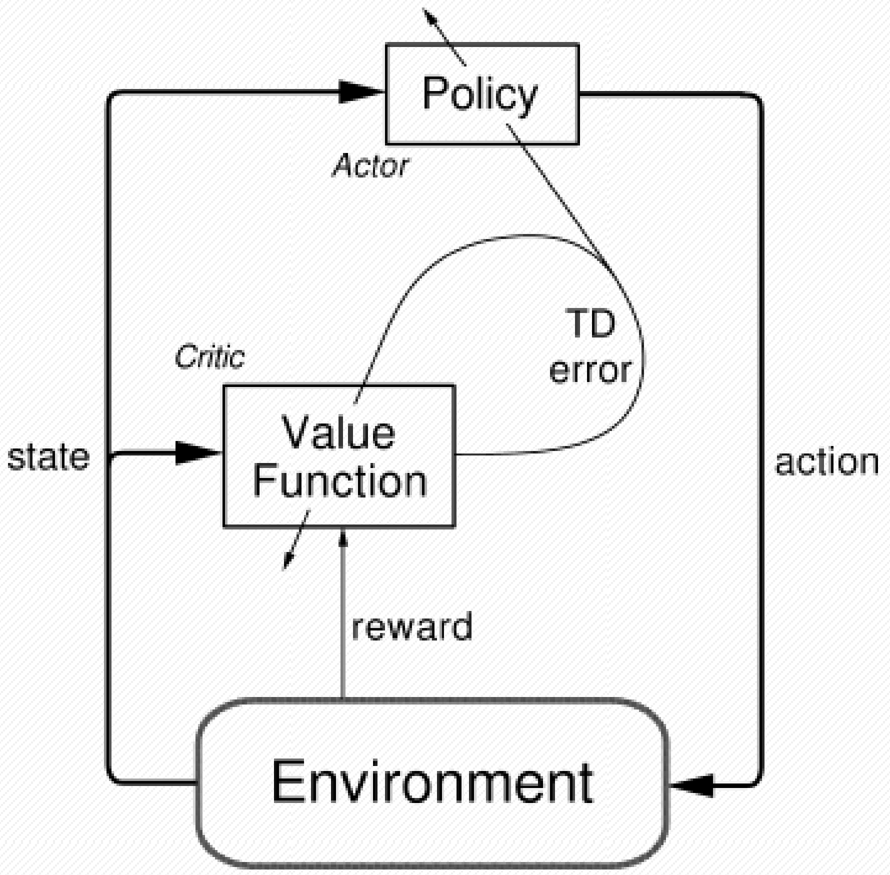
- TD(0) learning: bootstrap off your previous estimate of V

$$V(s_t) = V(s_t) + \alpha\left[\left(r_t + \gamma V(s_{t+1})\right) - V(s_t)\right]$$

- $\delta_t = \left[\left(r_t + \gamma V(s_{t+1})\right) - V(s_t)\right]$ is the TD-error

Actor-critic

- Have a parametrized value function Q (the critic) and policy π (the actor)
- Actor takes actions according to π, critic 'criticizes' them by computing Q-value

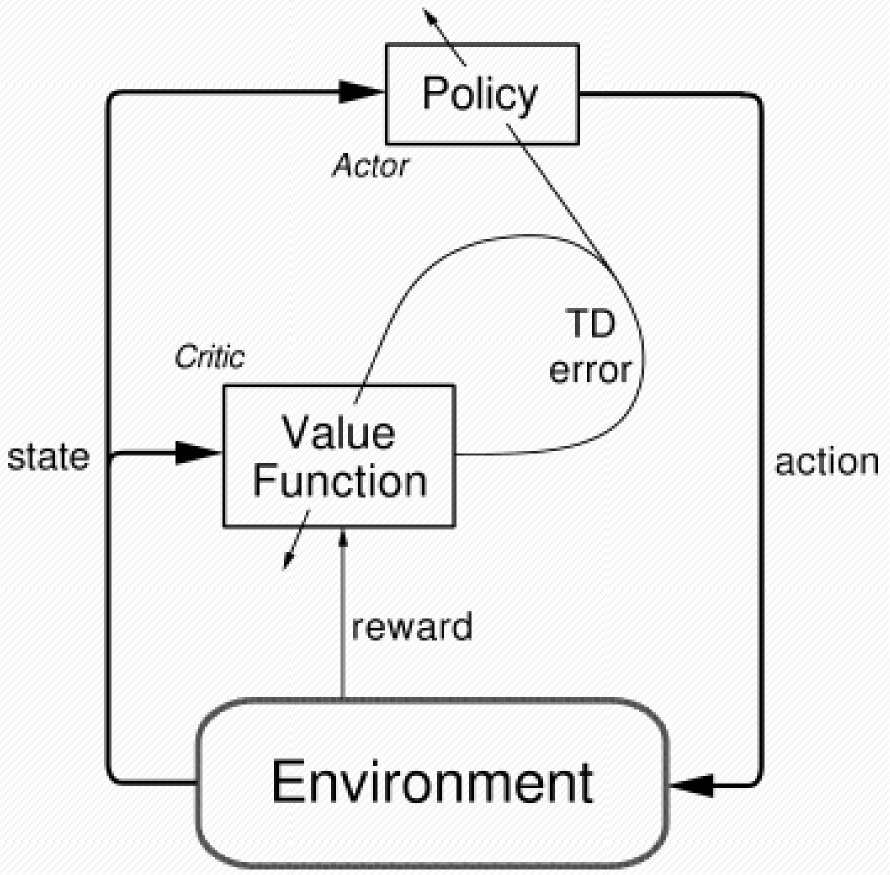


Source: Sutton & Barto (1998)

Actor-critic

- Critic usually learns with TD
- Actor learns according to the policy gradient theorem:

$$\frac{dR}{d\theta} = \mathrm{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\, Q^{\pi_\theta}(s,a)]$$



Source: Sutton & Barto (1998)

Actor-critic for sequence prediction

- Actor will be some function with parameters θ that predicts sequence one token at a time (i.e. generates 1 word at a time), *conditioned on its own previous predictions*

- Critic will be some function with parameters ϕ that computes the Q-value of decisions made by actor, which is used for learning

- Could use REINFORCE (e.g. Ranzato et al., (2015)), but this has higher variance

Actor-critic for sequence prediction

Since we are doing supervised learning, there are a couple differences to the RL case:

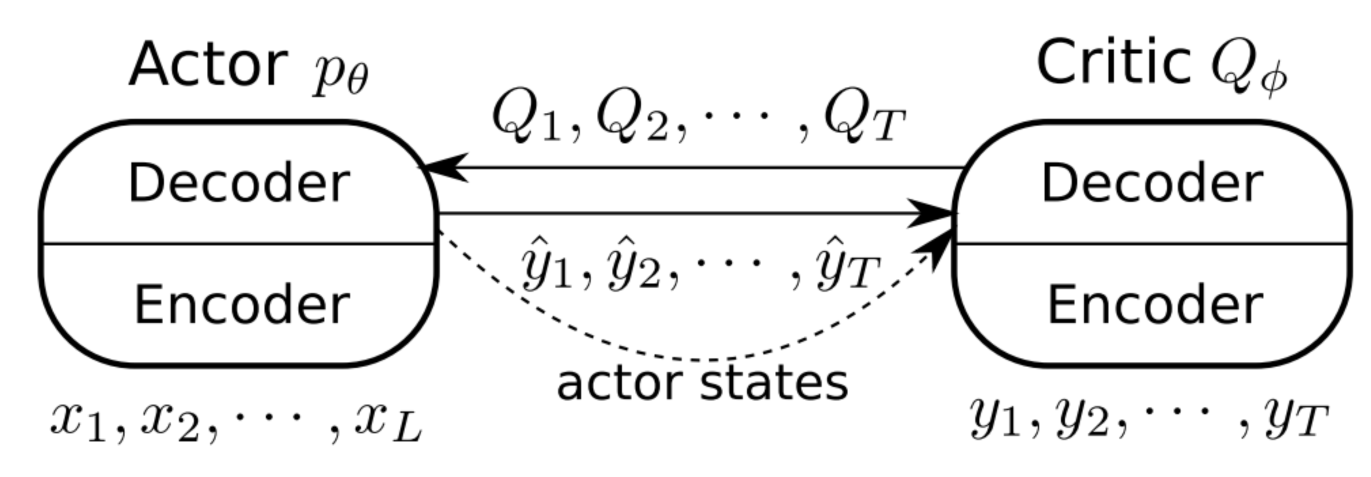
1) We can condition the critic on the actual ground-truth answer, to give a better training signal

2) Since there is a train/test split, don't use critic at test time

3) Since there is no stochastic environment, we can sum over all candidate actions to compute expectation in policy gradient thm

Deep implementation

- For actor and critic, use an RNN with 'soft-attention' (Bahdanau et al., 2015)
- Actor takes source sentence and sequence generated so far as input and predicts target sentence
- Critic takes target sentence and sequence generated so far, and computes Q-value



Tricks

- Use a target network, as in DQN
- Apply a variance penalty to the critic
- Use reward shaping to decompose final BLEU score into intermediate rewards
- Pre-train actor with log-likelihood, critic with fixed actor

Results – spelling correction

Table 1: Character error rate of different methods on the spelling correction task. In the table L is the length of input strings, η is the probability of replacing a character with a random one.

Setup	Character Error Rate		
	Log-likelihood	Actor-Critic	REINFORCE with critic
$L = 10, \eta = 0.3$	18.6	**17.2**	17.8
$L = 30, \eta = 0.3$	18.5	**17.3**	18.2
$L = 10, \eta = 0.5$	38.2	**35.9**	**35.8**
$L = 30, \eta = 0.5$	41.3	**37.0**	37.6

Results – translation

Table 2: Our machine translation results compared to the previous work by Ranzato et al. The abbreviations LL, RF, RF-C, AC stand for log-likelihood, REINFORCE, REINFORCE with critic and actor-critic training respectively. "greedy" and "beam" columns report results obtained with different decoding methods. The numbers reported with $\leq$ were approximately read from Figure 6 of Ranzato et al.

Paper	BLEU greedy	BLEU beam
Ranzato et al., LL	17.74	$\leq$ 20.3
Ranzato et al., MIXER	20.73	$\leq$ 21.9
This work, LL	19.57	21.67
This work, RF	20.64	21.45
This work, RF-C	**22.08**	**22.58**
This work, AC	21.43	22.09

Results – translation

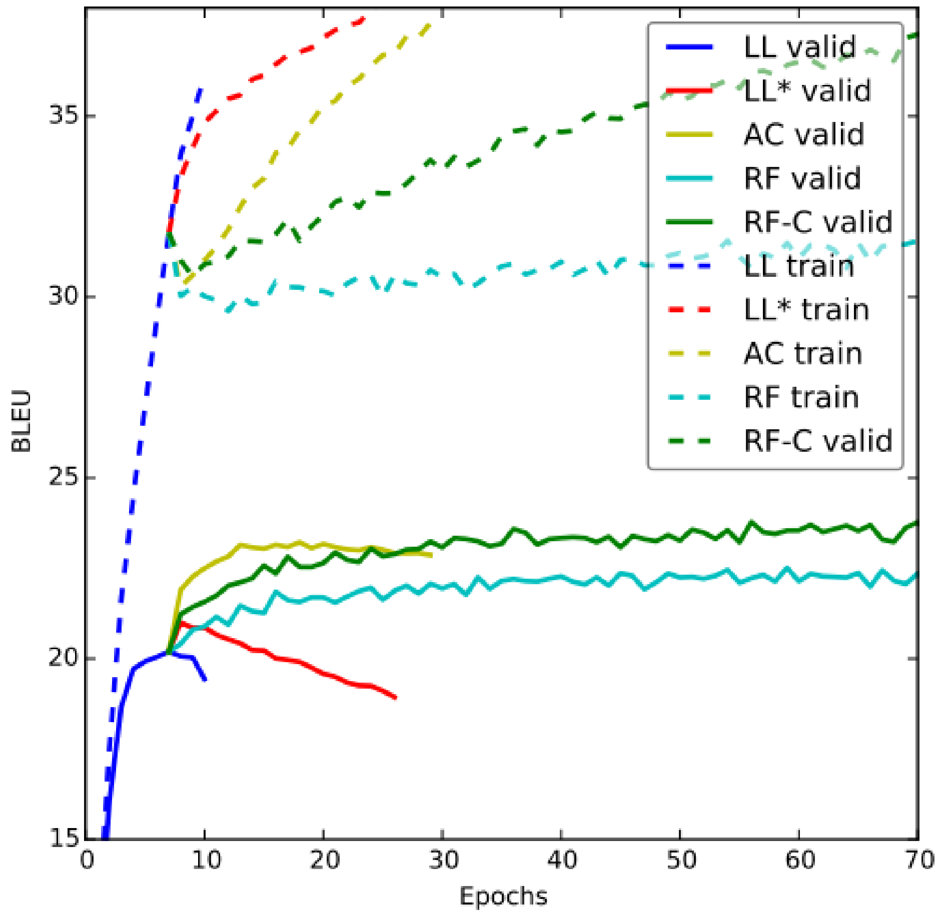


Figure 2: Progress of log-likelihood (LL), REINFORCE (RF) and actor-critic (AC) training in terms of BLEU score on the training (train) and validation (valid) datasets. LL* stands for the annealing phase of log-likelihood training. The curves start from the epoch of log-likelihood pretraining from which the parameters were initialized.

Part II: Dialogue Evaluation

Dialogue research

- Datasets for dialogue systems
    - The Ubuntu Dialogue Corpus (Lowe*, Pow* et al., 2015)
    - A survey of available corpora (Serban et al., 2015)

- Dialogue evaluation
    - How not to do it (Liu*, Lowe*, Serban*, Noseworthy* et al., 2016)
    - A slightly better way to do it (Lowe et al., 2016)

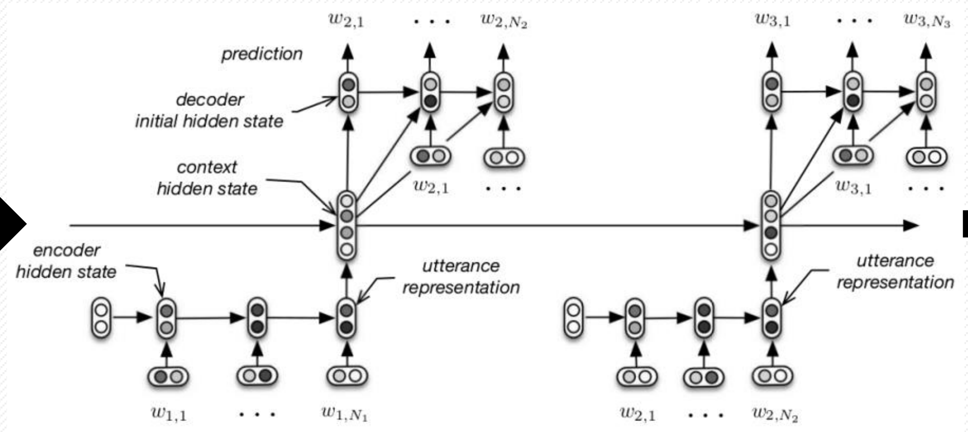
Why dialogue evaluation?

- Intelligent machines should be able to communicate with humans
- Dialogue is a great way to communicate with humans
- Hard to know if we're making progress in building dialogue models
- Particularly interested in 'non-task-oriented' setting

Comparison of ground-truth utterance

Context

Hey, want to go to the movies tonight?



Generated Response

Nah, let's do something active.

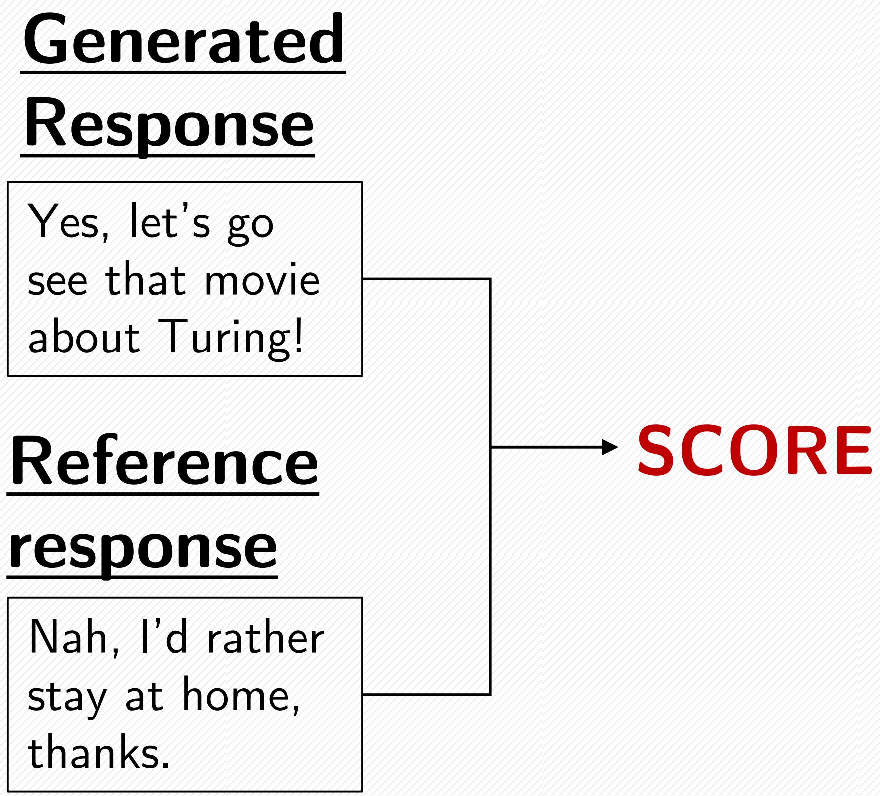
Reference response

Yeah, the film about Turing looks great!

SCORE

Comparison of ground-truth utterance

- Word-overlap metrics:
  - BLEU, METEOR, ROUGE
- Look at the number of overlapping n-grams between the generated and reference responses
- Correlate poorly with humans in dialogue



Correlation study

- Created 100 questions each for Twitter and Ubuntu datasets (20 contexts with responses from 5 'diverse models')
- 25 volunteers from CS department at McGill
- Asked to judge response quality on a scale from 1 to 5
- Compared human ratings with ratings from automatic evaluation metrics

Models for response variety

1) Randomly selected response

2) Retrieval models:
   - Response with smallest TF-IDF cosine distance
   - Response selected by Dual Encoder (DE) model

3) Generative models:
   - Hierarchical recurrent encoder-decoder (HRED)

4) Human-written response (<u>not</u> ground truth)

Goal (inter-annotator)

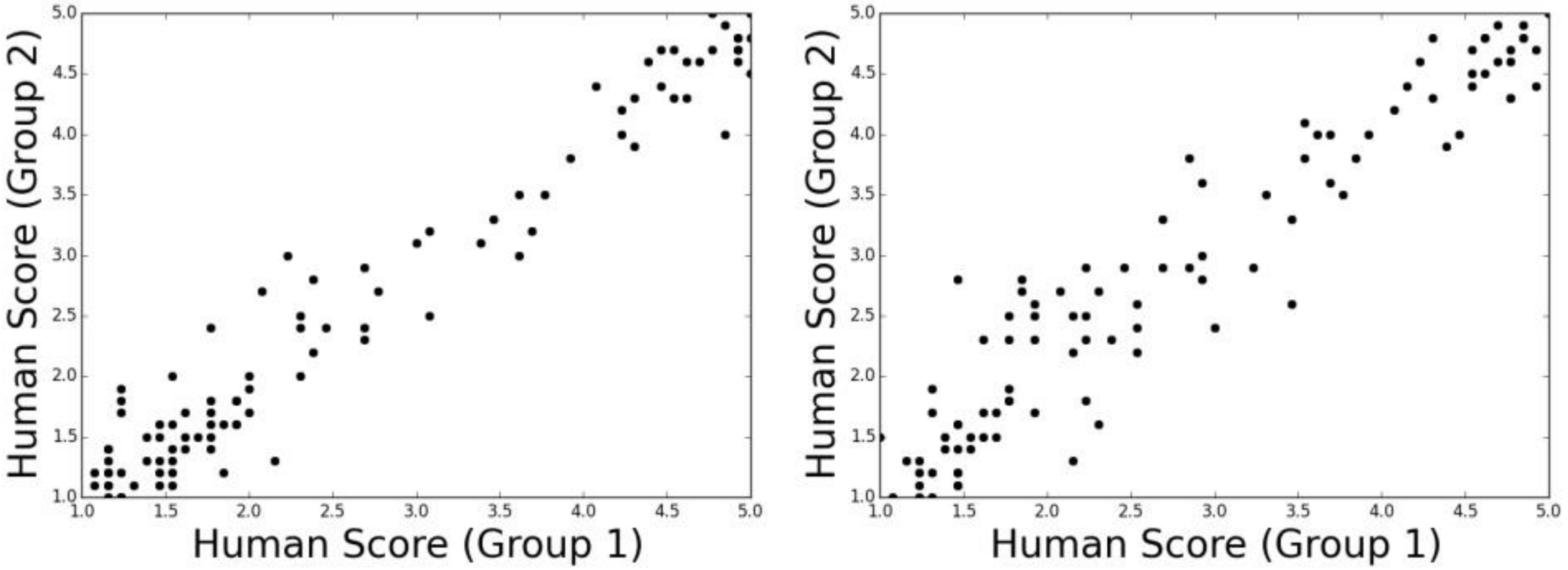
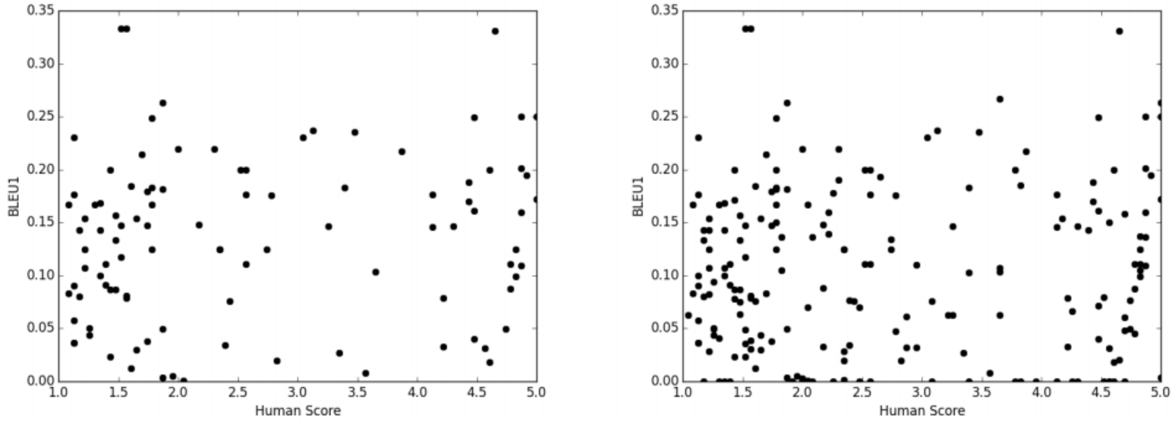
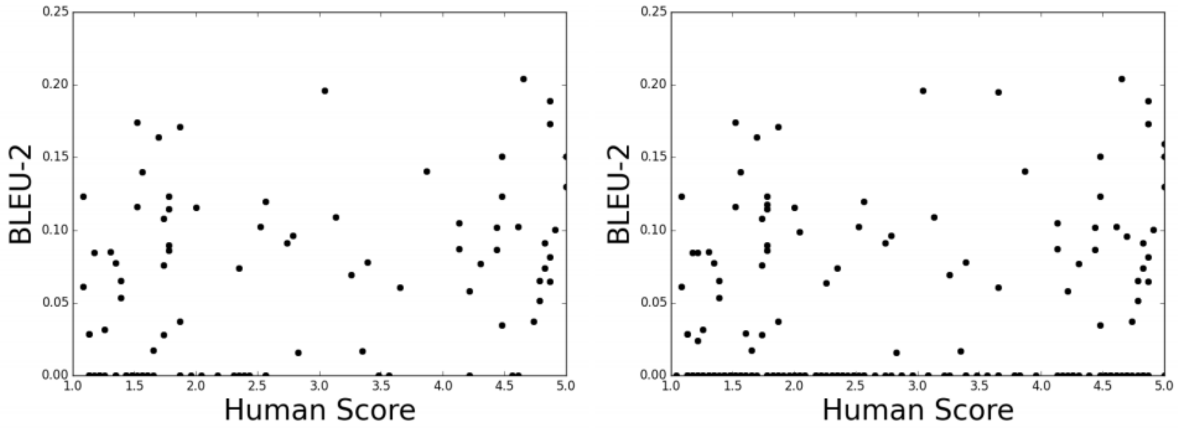


Figure 3: Scatter plots showing the correlation between two randomly chosen groups of human volunteers on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right).

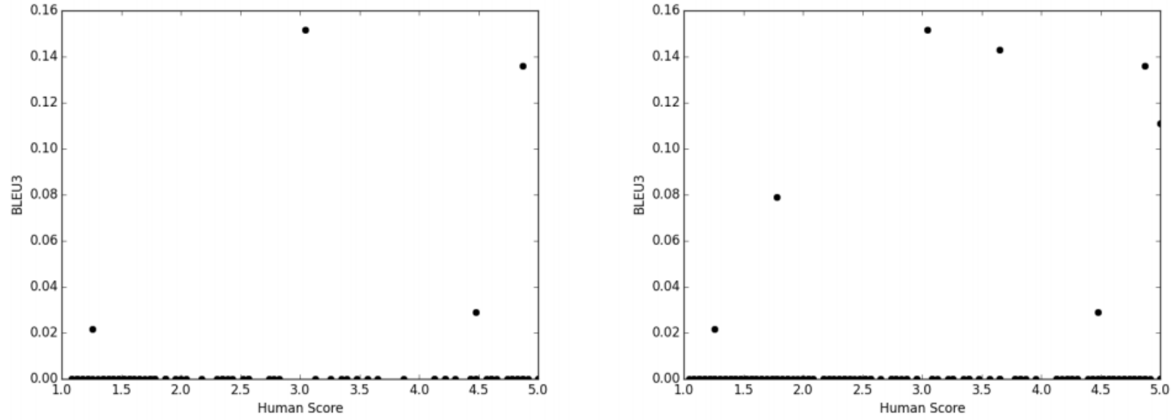
Reality (BLEU)



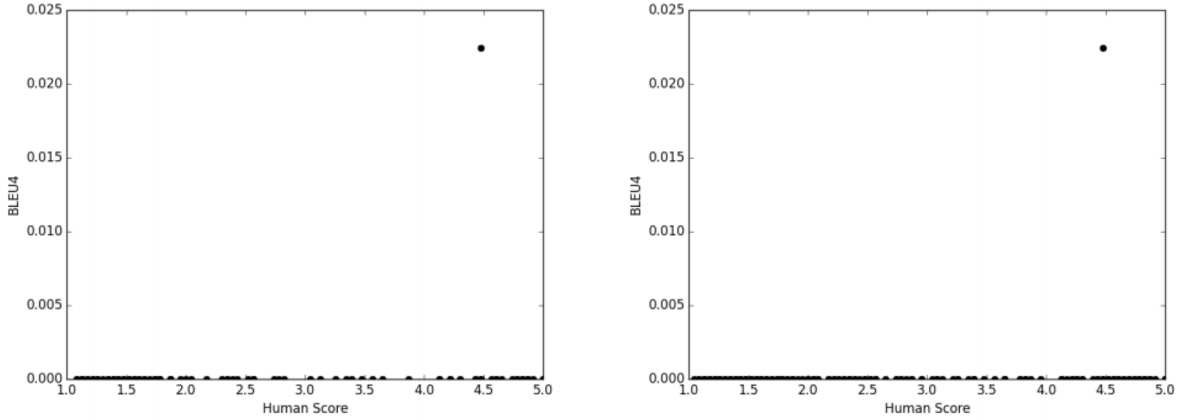
(a) BLEU-1



(b) BLEU-2

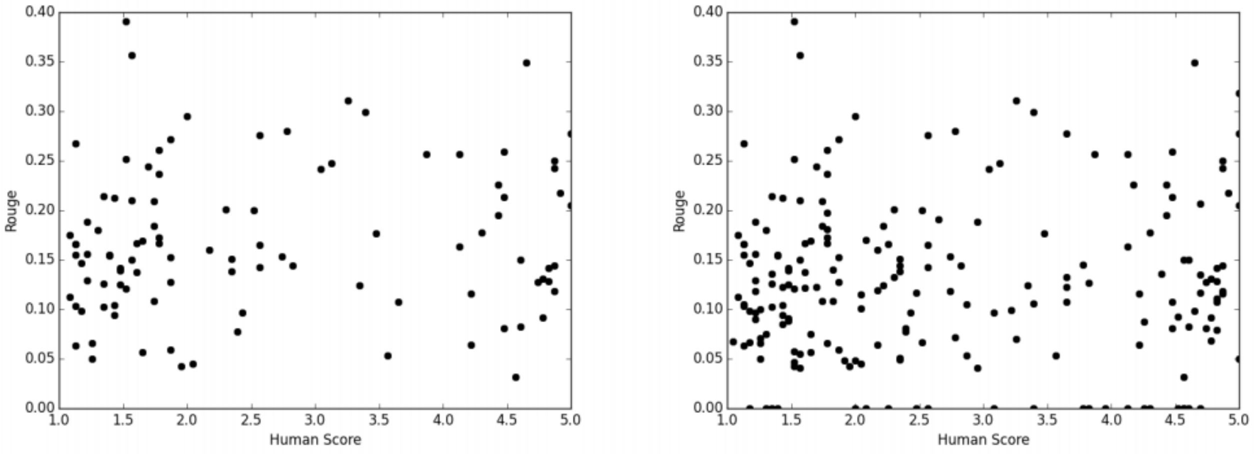


(c) BLEU-3

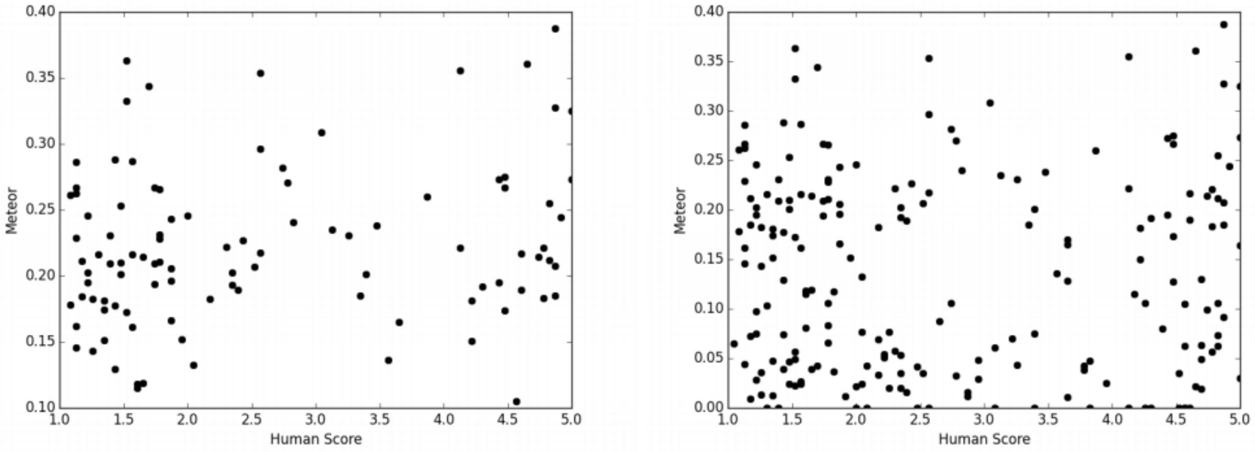


(d) BLEU-4

Reality (ROUGE & METEOR)



(a) ROUGE



(b) METEOR

Length bias

	Mean score		
	$\Delta w <= 6$ (n=47)	$\Delta w >= 6$ (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

Learning to evaluate

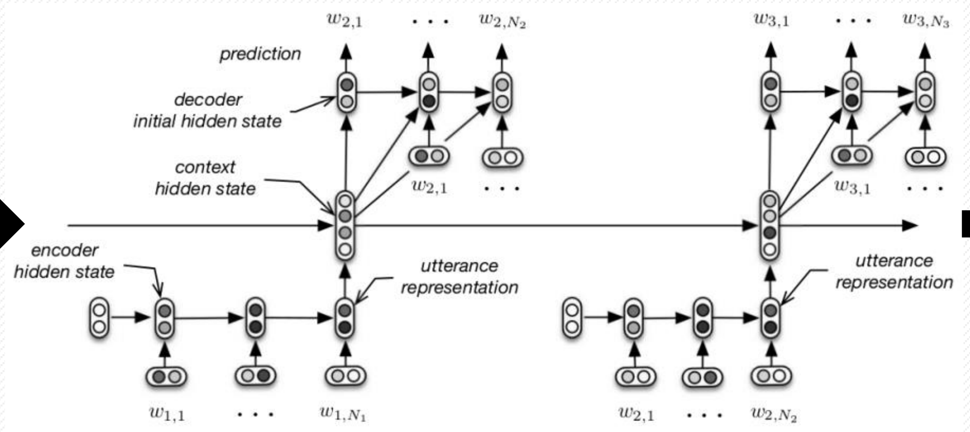
A dialogue response is probably good if it is rated highly by humans.

- Collect a labelled dataset of human scores of responses
- Build a model that learns to predict human scores of response quality (**ADEM**)
- Condition response score on the reference response and the context

Context-conditional evaluation

Context

Hey, want to go to the movies tonight?



Generated Response

Nah, let's do something active.

Reference response

Yeah, the film about Turing looks great!

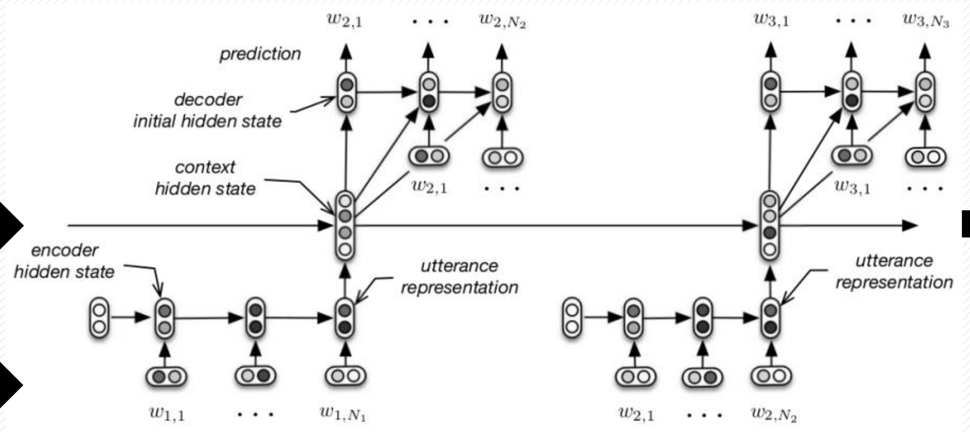
SCORE

Context-conditional evaluation

Context

~~Hey, want to go to the movies tonight?~~

Seen any good movies recently?



Generated Response

Nah, let's do something active.

Reference response

Yeah, the film about Turing looks great!

SCORE

Evaluation dataset

Conducted 2 rounds of AMT studies to get evaluation on Twitter

Study 1: ask workers to generate next sentence of a conversation

Study 2: ask workers to evaluate responses from various models (human, TFIDF, HRED, DE)

# Examples	4104
# Contexts	1026
# Training examples	2,872
# Validation examples	616
# Test examples	616
κ score (inter-annotator correlation)	0.63

ADEM

- Given: context *c*, model response *r*, reference response $\hat{r}$ (with embeddings **c**, **r**, $\hat{\mathbf{r}}$), compute score as:

$$score(c, r, \hat{r}) = (\mathbf{c}^T M \hat{\mathbf{r}} + \mathbf{r}^T N \hat{\mathbf{r}} - \alpha)/\beta$$

where *M*, *N* are parameter matrices, α, β are constants.

- Trained to minimize squared error:

$$\mathcal{L} = \sum_{i=1:K} [score(c_i, r_i, \hat{r}_i) - human\_score_i]^2 + \gamma ||\theta||_1$$

ADEM

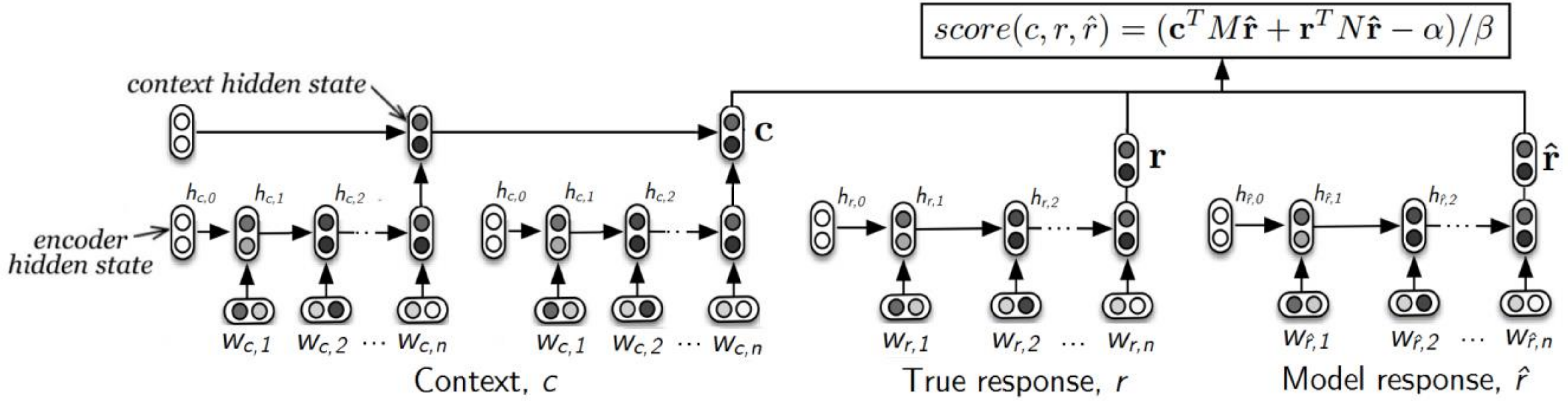


Figure 2: The ADEM model, which uses a hierarchical encoder to produce the context embedding **c**.

ADEM pre-training

- Want model that can learn from limited data (since collection is expensive)
- Pre-train RNN encoder of ADEM using VHRED

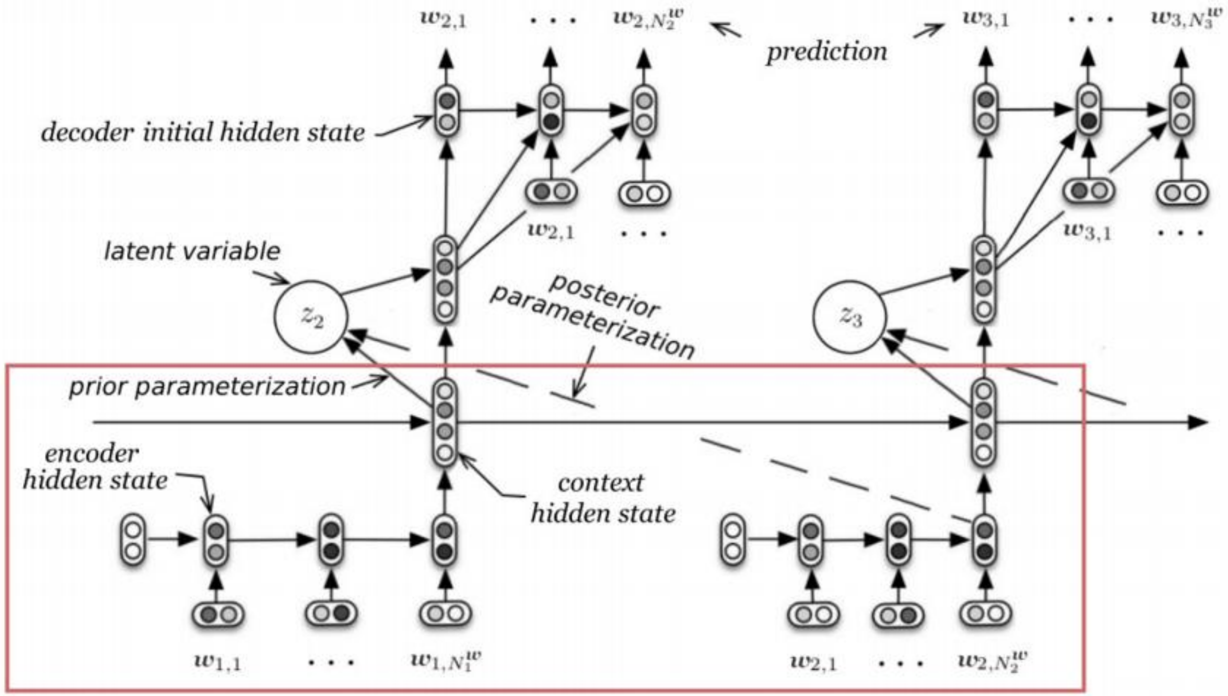
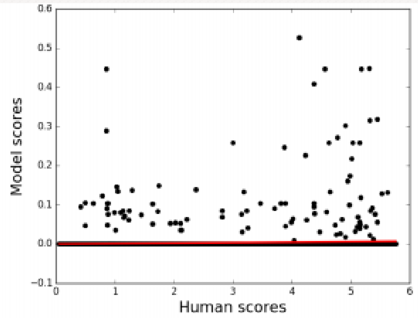


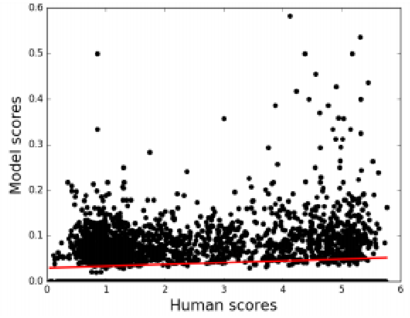
Figure 5: The VHRED model used for pre-training. The hierarchical structure of the RNN encoder is shown in the red box around the bottom half of the figure.

Utterance-level results

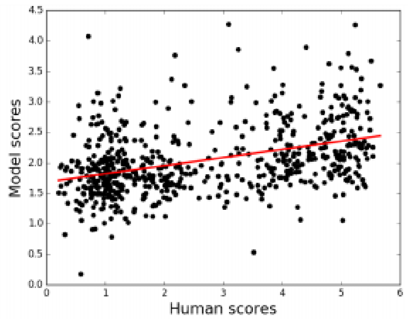
	Full dataset		Test set	
Metric	**Spearman**	**Pearson**	**Spearman**	**Pearson**
BLEU-1	0.026 (0.102)	0.055 (<0.001)	0.036 (0.413)	0.074 (0.097)
BLEU-2	0.039 (0.013)	0.081 (<0.001)	0.051 (0.254)	0.120 (<0.001)
BLEU-3	0.045 (0.004)	0.043 (0.005)	0.051 (0.248)	0.073 (0.104)
BLEU-4	0.051 (0.001)	0.025 (0.113)	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)
tweet2vec	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**	
ADEM (T2V)	0.395 (<0.001)	**0.392** (<0.001)	0.408 (<0.001)	**0.411** (<0.001)
ADEM	**0.436** (<0.001)	**0.389** (<0.001)	**0.414** (<0.001)	0.395 (<0.001)



(a) BLEU-2



(b) ROUGE



(c) ADEM

System-level results

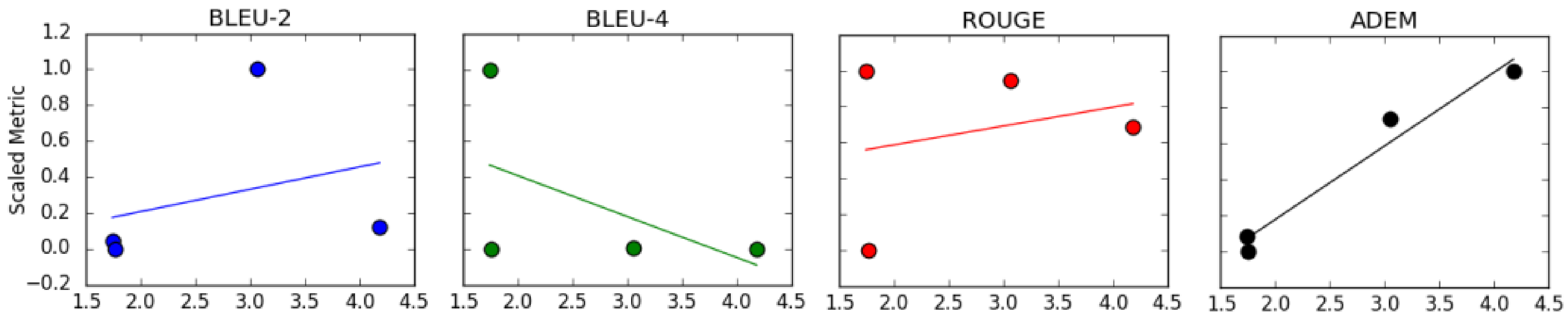


Figure 4: Scatterplots depicting the system-level correlation results for ADEM, BLEU-2, BLEU-4, and ROUGE. Each point represents the average scores for the responses from a dialogue model (TFIDF, DE, HRED, human). Human scores are shown on the horizontal axis, with normalized metric scores on the vertical axis. The ideal metric has a perfectly linear relationship.

Metric	Pearson
BLEU-1	-0.079 (0.921)
BLEU-2	0.308 (0.692)
BLEU-3	-0.537 (0.463)
BLEU-4	-0.536 (0.464)
ROUGE	0.268 (0.732)
ADEM	**0.981** (0.019)

Table 4: System-level correlation, with the p-value in brackets.

Results – generalization

	Test on full dataset		Test on removed model responses	
Data Removed	**Spearman**	**Pearson**	**Spearman**	**Pearson**
TF-IDF	0.4097 (<0.001)	0.3975 (<0.001)	0.3931 (<0.001)	0.3645 (<0.001)
Dual Encoder	0.4000 (<0.001)	0.3907 (<0.001)	0.4256 (<0.001)	0.4098 (<0.001)
HRED	0.4128 (<0.001)	0.3961 (<0.001)	0.3998 (<0.001)	0.3956 (<0.001)
Human	0.4052 (<0.001)	0.3910 (<0.001)	0.4472 (<0.001)	0.4230 (<0.001)
Average	**0.4069** (<0.001)	**0.3938** (<0.001)	**0.4164** (<0.001)	**0.3982** (<0.001)
25% at random	0.4077 (<0.001)	0.3932 (<0.001)	—	—

Table 4: Correlation for ADEM when various model responses are removed from the training set. The left two columns show performance on the entire test set, and the right two columns show performance on only responses from the dialogue model that was not seen during training.

Where does it do better?

Context	Reference response	Model response	Human score	$\lvert$BLEU-2$\rvert$ score	$\lvert$ROUGE$\rvert$ score	$\lvert$ADEM$\rvert$ score
i'd recommend <url> - or build buy an htpc and put <url> on it. → you're the some nd person this week that's recommended roku to me.	an htpc with xmbc is what i run . but i 've decked out my setup . i 've got <number> tb of data on my home server	because it's brilliant	5	1.0	1.0	4.726
imma be an auntie this weekend. i guess i have to go albany. herewego → u supposed to been here → i come off nd on. → never tell me smh	lol you sometiming	haha, anyway, how're you?	5	1.0	1.0	4.201
my son thinks she is plain. and the girl that plays her sister. seekhelp4him? → send him this. he'll thank you. <url>	you are too kind for words .	i will do	5	1.0	1.0	5.0

Table 8: Examples where both human and ADEM score the model response highly, while BLEU-2 and ROUGE do not. These examples are drawn randomly (i.e. no cherry-picking) from the examples where ADEM outperforms BLEU-2 and ROUGE (as defined in the text). ADEM is able to correctly assign high scores to short responses that have no word-overlap with the reference response. The bars around $\lvert$metric$\rvert$ indicate that the metric scores have been normalized.

Where does it do better?

- ADEM doesn't exhibit the same length bias as word overlap metrics

	Mean score		
	$\Delta w \leq 6$ (n=312)	$\Delta w > 6$ (n=304)	**p-value**
ROUGE	0.042	0.031	< 0.01
BLEU-2	0.0022	0.0007	0.23
ADEM	2.072	2.015	0.23
Human	2.671	2.698	0.83

Table 9: Effect of differences in response length on the score, Δw = absolute difference in #words between the reference response and proposed response. BLEU-1, BLEU-2, and METEOR have previously been shown to exhibit bias towards similar-length responses (Liu et al., 2016).

Potential problems

- The problem of generic responses
- Only considers single utterances, rather than a whole dialogue
- What about other aspects of dialogue quality?

Primary Collaborators

Joelle Pineau
McGill

Iulian V. Serban
U. Montreal

Mike Noseworthy
McGill

Chia-Wei Liu
McGill

Laurent Charlin
HEC Montreal

Nicolas Angelard-Gontier
McGill

Dzmitry Bahdanau
U. Montreal

Philemon Brakel
U. Montreal

Aaron Courville
U. Montreal

Yoshua Bengio
U. Montreal

References

Bahdanau, Brakel, Xu, Goyal, Lowe, Pineau, Courville, Bengio. "An Actor-Critic Algorithm for Sequence Prediction." 2016.

Bahdanau, Cho, Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." ICLR, 2015.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. "Generating sentences from a continuous space." COLING, 2016.
Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." EMNLP, 2014.

Kingma & Welling. "Auto-encoding Variational Bayes." ICLR, 2014.

Liu, Lowe, Serban, Noseworthy, Charlin, Pineau. "How NOT to Evaluate Your Dialogue System: A Study of Unsupervised Evaluation Metrics for Dialogue Response Generation." EMNLP, 2016.

Lowe, Noseworthy, Serban, Angelard-Gontier, Bengio, Pineau. "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses." 2016.

Lowe, Pow, Serban, Pineau. "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Dialogue Systems." SIGDIAL, 2015.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. "BLEU: a method for automatic evaluation of machine translation". ACL, 2002.

Ranzato, M. A., Chopra, S., Auli, M., & Zaremba, W. "Sequence level training with recurrent neural networks." ICLR, 2015.

Rezende, D. J., Mohamed, S., & Wierstra, D. "Stochastic backpropagation and approximate inference in deep generative models." ICML, 2014.

Serban, Lowe, Charlin, Pineau. "A Survey of Available Corpora for Building Data-Driven Dialogue Systems." 2016.

Serban, Sordoni, Lowe, Pineau, Courville, Bengio. "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues." AAAI, 2017.

Sutskever, Vinyals, Le. "Sequence-to-sequence Learning with Neural Networks." NIPS, 2014.

Sutton & Barto. "Reinforcement Learning: An Introduction." 1998.

Thank you

Variational encoder-decoder (VHRED)

- Inspired by VAE (Kingma & Welling, 2014; Rezende et al., 2014): train model with backprop using reparameterization trick

- Prior mean and variance are learned conditioned on previous utterance representation. Posterior mean and variance also conditioned on representation of target utterance.

- At training time, sample from posterior. At test time, sample from prior.

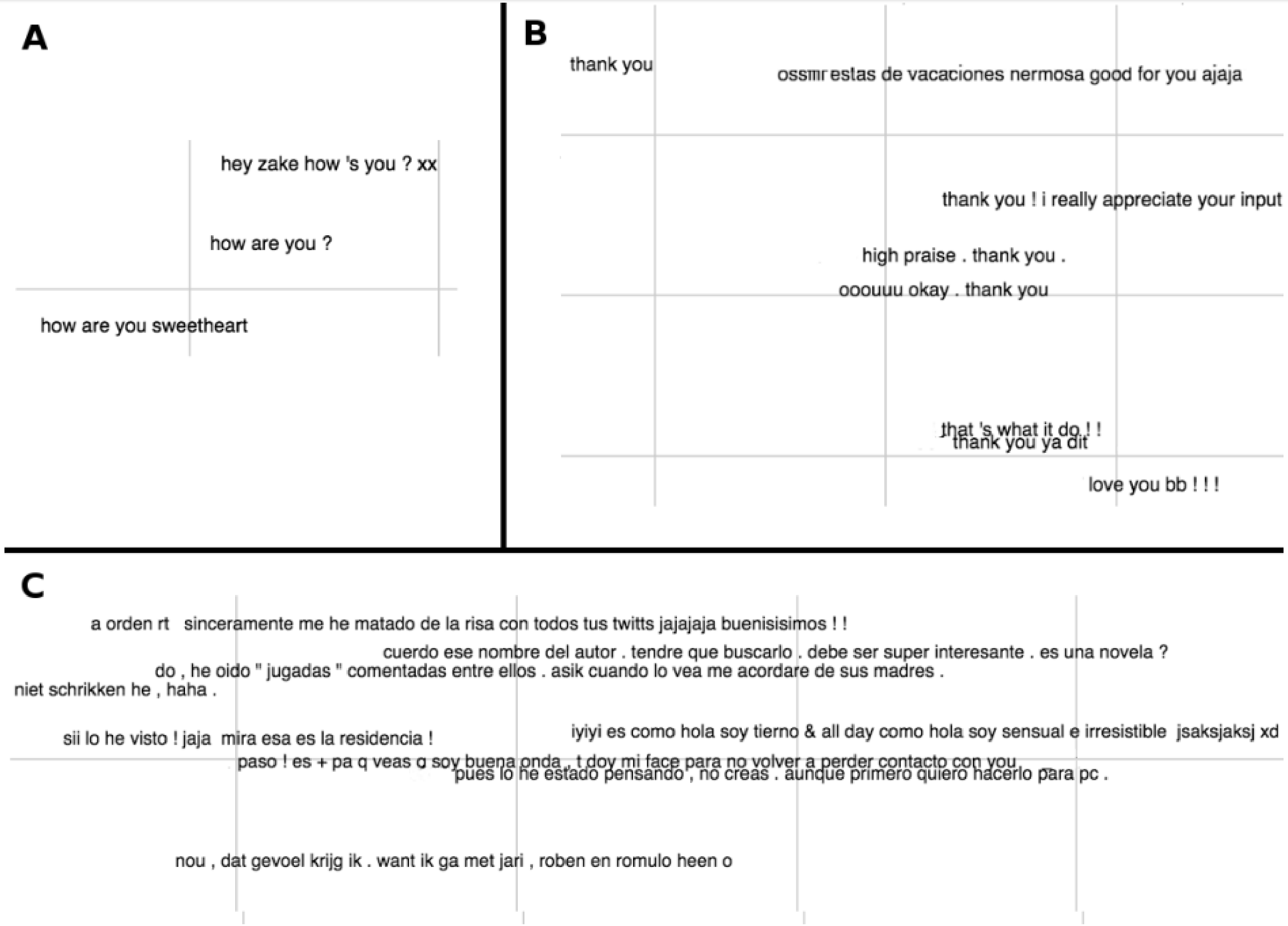
- Developed concurrently with Bowman et al. (2016)
  - Use word-dropping and KL annealing tricks

Quantitative VHRED results

Table 4: Response information content on 1-turn generation as measured by average utterance length $|U|$, word entropy $H_w = -\sum_{w \in U} p(w) \log p(w)$ and utterance entropy H_U with respect to the maximum-likelihood unigram distribution of the training corpus p.

	Twitter			Ubuntu		
Model	$\lvert U\rvert$	H_w	H_U	$\lvert U\rvert$	H_w	H_U
LSTM	11.21	6.75	75.61	4.27	6.50	27.77
HRED	11.64	6.73	78.35	**11.05**	7.53	**83.16**
VHRED	**12.29**	**6.88**	**84.56**	9.22	**7.70**	71.00
Human	20.57	8.10	166.57	18.30	8.90	162.88

VHRED results



Notation

- Let X be the input sequence, $Y = (y_1, \dots, y_T)$ be the target output sequence
- Let $\widehat{Y}_{1,\dots,t} = (\widehat{y}_1, \dots, \widehat{y}_t)$ be the sequence generated so far

- Our critic $\widehat{Q}(a; \widehat{Y}_{1,\dots,t}, Y)$ is conditioned on outputs so far $\widehat{Y}_{1,\dots,t}$, and ground-truth output Y

- Our actor $p(a; \widehat{Y}_{1,\dots,t}, X)$ is conditioned on outputs so far $\widehat{Y}_{1,\dots,t}$, and the input X

Policy Gradient for Sequence Prediction

- Denote V as the expected reward under π_θ

Proposition 1 *The gradient* $\frac{dV}{d\theta}$ *can be expressed using Q values of intermediate actions:*

$$\frac{dV}{d\theta} = \mathbb{E}_{\hat{Y} \sim p(\hat{Y})} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \frac{dp(a|\hat{Y}_{1\ldots t-1})}{d\theta} Q(a; \hat{Y}_{1\ldots t-1})$$

Algorithm

while Not Converged **do**

Receive a random example (X, Y).

Generate a sequence of actions $\hat{Y}$ from p'.

Compute targets for the critic

$$q_t = r_t(\hat{y}_t; \hat{Y}_{1\ldots t-1}, Y) + \sum_{a \in \mathcal{A}} p'(a|\hat{Y}_{1\ldots t}, X)\hat{Q}'(a; \hat{Y}_{1\ldots t}, Y)$$

Algorithm

6: Update the critic weights ϕ using the gradient

$$\frac{d}{d\phi}\left(\sum_{t=1}^{T}\left(\hat{Q}(\hat{y}_t;\hat{Y}_{1\ldots t-1},Y)-q_t\right)^2+\lambda C\right)$$

Algorithm

7: Update actor weights θ using the following gradient estimate

$$\frac{dV(X,Y)}{d\theta} = \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \frac{dp(a|\hat{Y}_{1\ldots t-1}, X)}{d\theta} \hat{Q}(a; \hat{Y}_{1\ldots t-1}, Y)$$

Tricks: target network

- Similarly to DQN, use a target network
- In particular, have both delayed actor p' and a delayed critic Q', with params θ' and ϕ', respectively
- Use this delayed values to compute target for critic:

$$q_t = r_t(\hat{y}_t; \hat{Y}_{1\ldots t-1}, Y) + \sum_{a \in \mathcal{A}} p'(a|\hat{Y}_{1\ldots t}, X)\hat{Q}'(a; \hat{Y}_{1\ldots t}, Y)$$

- After updating actor and critic, update delayed actor and critic using a linear interpolation

Tricks: variance penalty

- Problem: critic can have high variance for words that are rarely sampled

- Solution: artificially reduce values of rare actions by introducing a variance regularization term:

$$C = \sum_a \left(\hat{Q}(a; \hat{Y}_{1 \ldots t-1}) - \frac{1}{|\mathcal{A}|} \sum_b \hat{Q}(b; \hat{Y}_{1 \ldots t-1}) \right)^2 ,$$

Tricks: reward shaping

- Could train critic using all the score at the last step, but this signal is sparse
- Want to improve learning of critic (and thus the actor) by providing rewards at each time step

- If final reward is $R(\widehat{Y})$, decompose the reward into scores for all prefixes: $(R(\widehat{Y}_{1,\dots,1}), R(\widehat{Y}_{1,\dots,2}), \dots, R(\widehat{Y}_{1,\dots,T}))$
- Then the reward at time step t is:

$$r_t(\widehat{y}_t) = R(\widehat{Y}_{1\dots t}) - R(\widehat{Y}_{1\dots t-1})$$

Tricks: pre-training

- If you start off with a random actor and critic, it will take forever to learn, since the training signals would be terrible
- Instead, use pre-training: first train actor to maximize log-likelihood of correct answer
- Then, train critic by feeding samples from the (fixed) actor
- Similar to pre-training used in AlphaGo (without MC rollouts)

Experiments

Word	Words with largest $\hat{Q}$
one	and(6.623) there(6.200) but(5.967)
of	that(6.197) one(5.668) 's(5.467)
them	that(5.408) one(5.118) i(5.002)
i	that(4.796) i(4.629) ,(4.139)
want	want(5.008) i(4.160) 't(3.361)
to	to(4.729) want(3.497) going(3.396)
tell	talk(3.717) you(2.407) to(2.133)
you	about(1.209) that(0.989) talk(0.924)
about	about(0.706) .(0.660) right(0.653)
here	.(0.498) ?(0.291) –(0.285)
.	.(0.195) there(0.175) know(0.087)
$\emptyset$	.(0.168) $\emptyset$ (-0.093) ?(-0.173)

Table 3: The best 3 words according to the critic at intermediate steps of generating a translation. The numbers in parentheses are the value predictions $\hat{Q}$. The German original is “über eine davon will ich hier erzählen .” The reference translation is “and there’s one I want to talk about”.

Results

Context	Reference response	Model responses	Human score	ADEM score
photo to see my television debut go to - some. some on <url> - hehe <url> → it really was you? i thought ppl were recognizing someone who looked like you! were the oysters worth the wait?	yeah it was me . haha i 'd kinda forgotten about it it was filmed a while ago	1) i'm not sure. i just don't know what to do with it.	3	1.602
		2) you heard the horsepower productions remix of lee scratch perry's 'exercising' off his 'mighty upsetter' album?	1	1.513
		3) you wont chug a fuzzy peach navel	1	1.744
		4) they were!	5	3.274
just beat call of duty!! → want a cookie? → yes!! → come get it	im in kenmore at the moment	1) i'm gonna get a new phone some moro	1	1.848
		2) no way man.	5	4.265
		3) wow i just got a free pizza coupon! get yours before theres no more! <url>	1	0.921
		4) i'm going to go to the mall.	1	2.634
am i out of twitter jail yet? testing → yeah. i posted bail → thanks. i am a right chatter tweetbox on sundays. same happened last sunday lol	any news on meeting our user ? i go to the us on friday and i don 't want to miss anything arranged	1) i'm not sure if i'm going to be able to get it.	3	1.912
		2) good to see another mac user in the leadership ranks	4	1.417
		3) awww poor baby hope u get to feeling better soon. maybe some many work days at piedmont	2	1.123
		4) did you tweet too much?	5	2.539

Table 5: Examples of scores given by the ADEM model.