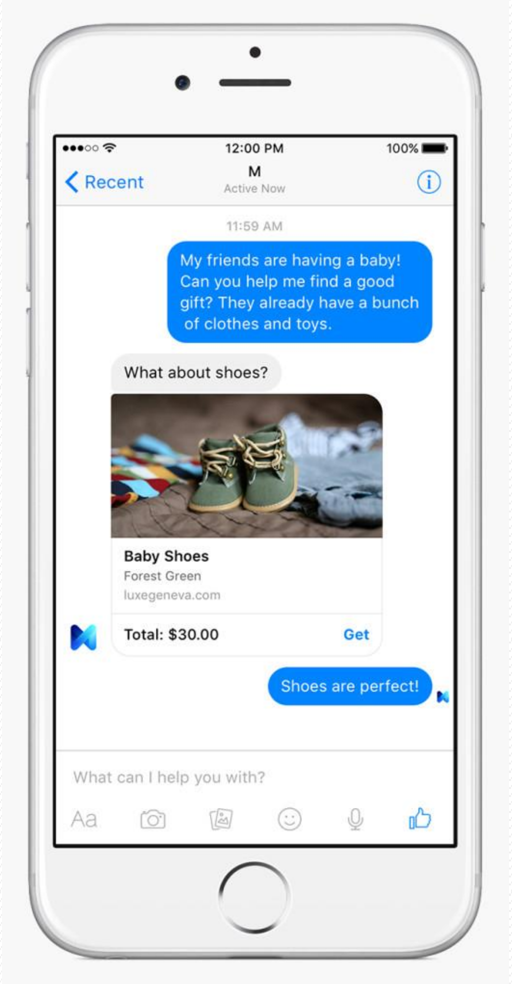
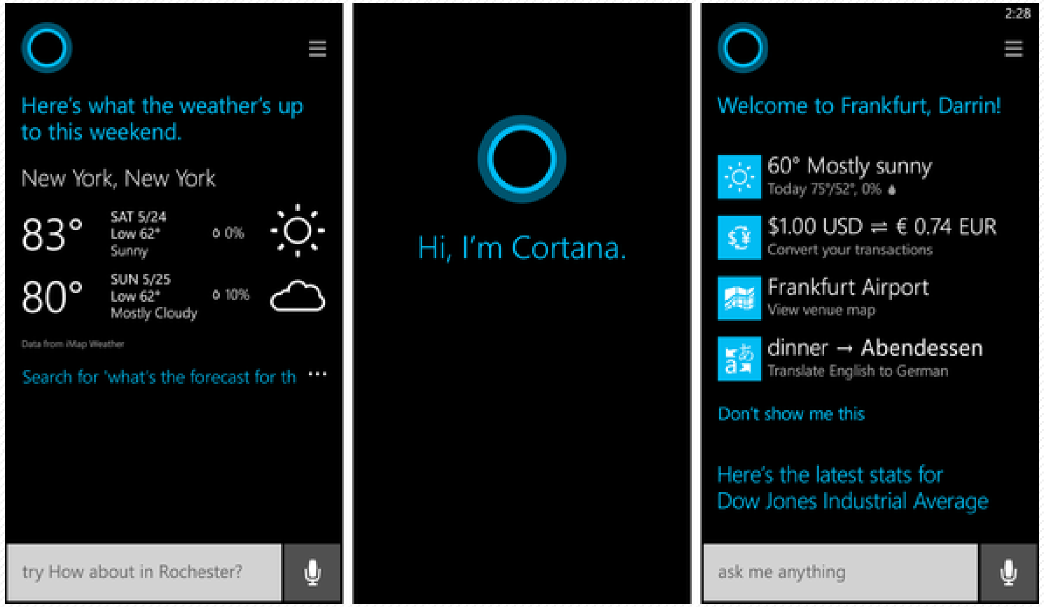
The Problem(s) with Neural Chatbots

Ryan Lowe

McGill University, OpenAI

Dialogue systems





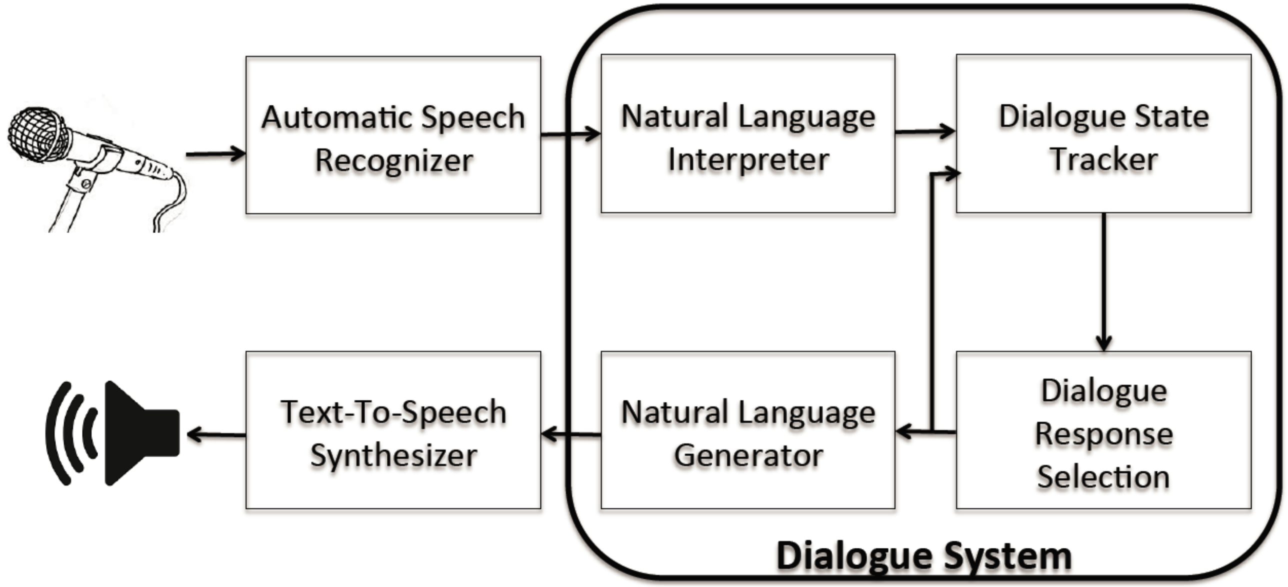
Why work on dialogue systems?

- Many commercial applications
- Creating a ‘general-purpose communicating agent’
  - An agent that can communicate with humans on many topics, to exchange knowledge and complete a variety of tasks in its environment.
- Language is a natural communication interface between humans and machines

Modular dialogue systems

- Traditional system consists of modules

- Each module optimized with separate objective function

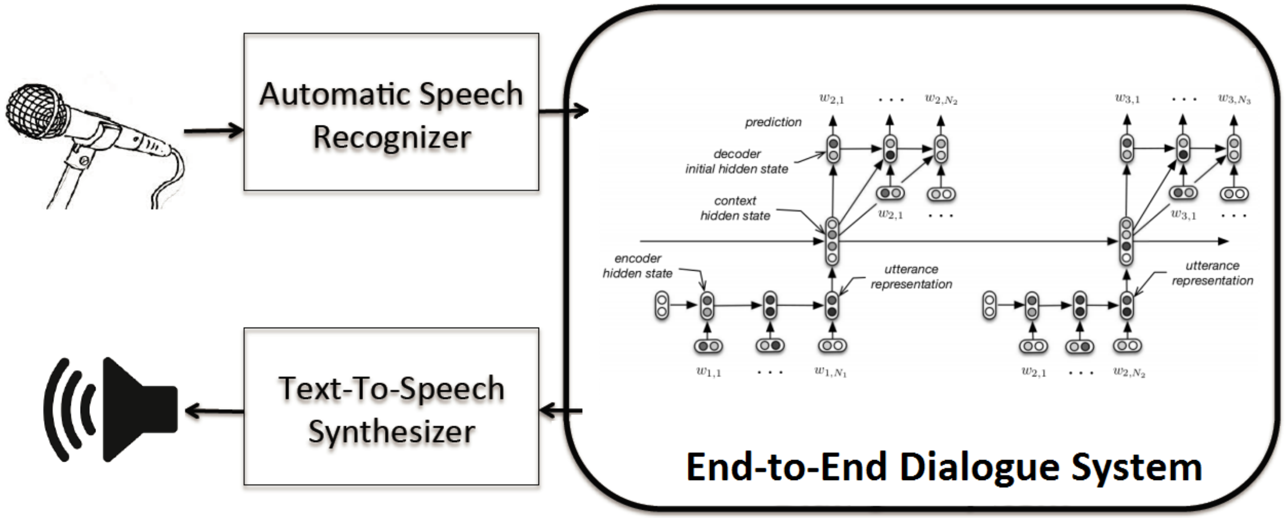


- Achieves fairly good performance with small amounts of data

Problem: require supervised data for each module – does not scale well to general domains!

End-to-end dialogue systems

- A single model trained directly on conversational data



- Uses a single objective function, usually maximum likelihood on next response

- Significant recent work using neural networks to predict the next response. (Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015)

Why might this work?

Advantages of end-to-end systems:

1) Does not require feature engineering (only architecture engineering).

2) Can be transferred to different domains.

3) Does not require supervised data for each module!

(collecting this data is difficult at large scales)

Early results – Vinyals & Le (2015)

Machine: *hi*
Human: *hello*
Machine: *could you please let me know what are the operating systems you are using ?*
Human: *i am using linux*
Machine: *ok it was in the ssl vpn on which you stated it was broken*
Machine: *please check if you are logged in to corp network*
Machine: *what is the error that you are running please*
Human: *i am seeing an error related to vpn*
Machine: *what is the error message that you are getting when connecting to vpn using network connect ?*
Human: *connection refused or something like that*
Machine: *may i know the version of network connect you connect ?*

Human: *i am not sure i know that*
Machine: *is the network connect prompting for any user certificate*
Machine: *?*
Human: *yes*
Machine: *are you at home or at work at the moment ?*
Machine: *office ?*
Human: *i am at home*
Machine: *try this*
Machine: *goto <URL> page and enter your name and password*
Human: *done*
Machine: *try logging in with <NAME> and check if its working fine now*
Human: *yes , now it works !*
Machine: *great*

Current results (VHRED)

<**first_speaker**> what do you mean?

<**second_speaker**> what are you talking about?

<**first_speaker**> i 'm not a big fan of cats , dogs.

What went wrong?

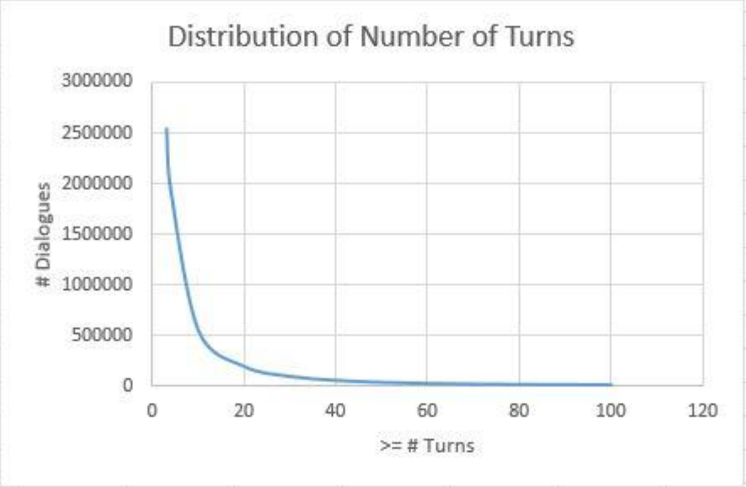
Problem #1: Data

Dialogue datasets

- Building general-purpose dialogue systems requires lots of data
- The best datasets are proprietary
- We need **large** (>500k dialogues), **open-source** datasets to make progress

Ubuntu Dialogue Corpus

- Large dataset of ~1 million tech support dialogues
- Scraped from Ubuntu IRC channel
- 2-person dialogues extracted from chat stream



ubuntuaddicted	what's my ip?	[02:59]
DF3D2	k1l: so I reinstalled fglrx manually, and startx just keeps saying "no protocol specified"	[02:59]
nahtnam	ubuntuaddicted: Are you in europe?	[03:00]
xtpeeps	Anyone can introduce me some interest channel of irc:p THX	[03:00]
timwis	hey guys, just did a fresh install on a Lenovo yoga to Pro, and I'm getting Wi-Fi is disabled by hardware switch. Any idea how to resolve?	[03:01]
DF3D2	k1l: and time out in locking the .Xauthority file	[03:01]
Bashing-om	DF3D2: Before you rebooted, did you do -> sudo amdconfig --initial <- ??	[03:01]
timwis	this article suggests I modify ideapad-laptop.c but it doesn't seem to exist on the filesystem http://billauer.co.il/blog/2014/08/linux-ubuntu-yoga-hardware-blocked-wireless-lan/	[03:01]
xangua	!alis | xtpeeps	[03:01]
ubottu	xtpeeps: alis is a services bot that can help you find channels. Read "/msg alis help list" . For more help or questions relating to alis, please join #freenode. Example usage: /msg alis list #ubuntu* or /msg alis list *http*	[03:01]
DF3D2	Bashing-om: yes	[03:01]
ubuntuaddicted	nahtnam, no, why?	[03:01]
DF3D2	Bashing-om: I also did rm -r ~/.Xauthority as I saw suggested on the web, didn't help	[03:02]
cfhowlett	timwis, yep. only took me 3 years to learn. hit the windows wifi switch but experiment with combinations: ctrl F2 does it on my DELL in ubuntu. In windows: f2	[03:02]
cfhowlett	timwis, ctrl, alt, shift and super keys are all candidates	[03:03]
timwis	that article actually suggests that with the Lenovo laptops there's a problem beyond that	[03:04]
timwis	what is the super key?	[03:04]
cryptodan	the windows key	[03:04]
cfhowlett	timwis, aka "windows" key	[03:04]
timwis	ah! super indeed	[03:04]
somsip	timwis: windows key, or mod key, between left ctrl and left alt usually	[03:04]

Sender	Recipient	Utterance
Old		I dont run graphical ubuntu, I run ubuntu server.
bur[n]er	Old	you can use ”ps ax” and ”kill (PID#)”
kuja	Taru	Haha sucker.
Taru	Kuja	?
kuja	Taru	Anyways, you made the changes right?
Taru	Kuja	Yes.
kuja	Taru	Then from the terminal type: sudo apt-get update
Taru	Kuja	I did.

Lowe*, Pow*, Serban, Pineau. “The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems.” *SIGDIAL*, 2015.

Ubuntu Dialogue Corpus

Pros:

- Hard
- Large
- Open-source
- Related to many real-world technical problems

Cons:

- Too hard?
- Not perfectly disentangled
- Requires external knowledge to solve
- Ideally suited for task-oriented setting, but no reward signal in dataset

Large-scale dialogue datasets

- Ubuntu Dialogue Corpus (Lowe et al., 2015)
- Twitter Corpus (Ritter et al., 2011)
- Movie Dialog Dataset (Dodge et al. 2016)
- Reddit
- ...

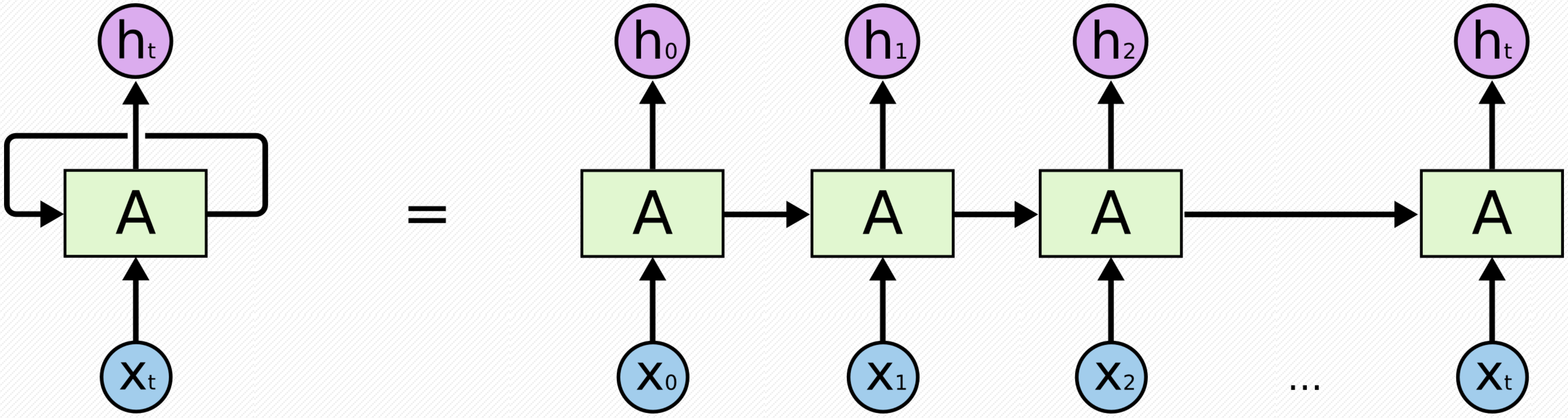
Survey paper covering existing datasets:

Serban, Lowe, Charlin, Pineau. "A Survey of Available Corpora for Building Data-Driven Dialogue Systems." *arXiv:1512.05742*, 2015.

Problem #2: Model Architecture

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)



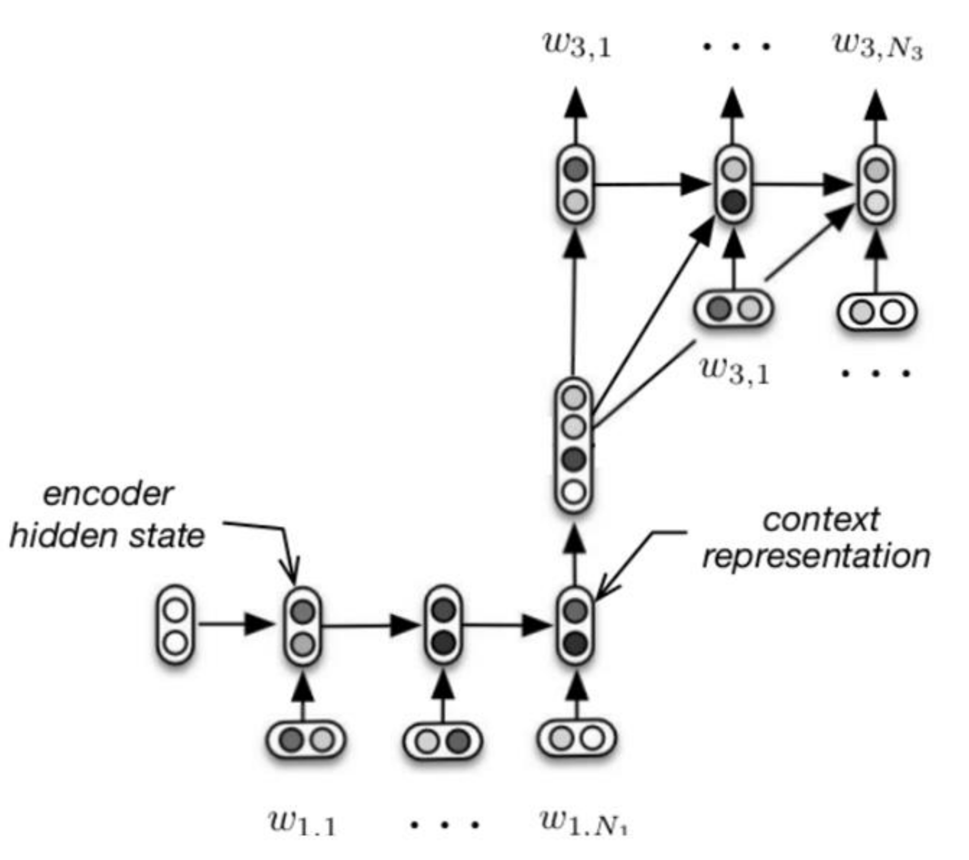
Source: colah.github.io

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)



Main goal

Build models with right inductive biases to effectively represent dialogue data

Judge model quality by quality of generated responses

Some problems: generic responses

- Most models trained to predict most likely next utterance given context
- But some utterances are likely given any context!
- Neural models often generate "I don't know", or "I'm not sure" to most contexts

Input: What are you doing?		
-0.86	I don't know.	–
-1.03	I don't know!	–
-1.06	Nothing.	–
-1.09	Get out of the way.	–
Input: what is your name?		
-0.91	I don't know.	...
-0.92	I don't know!	–
-0.92	I don't know, sir.	–
-0.97	Oh, my god!	–
Input: How old are you?		
-0.79	I don't know.	...
-1.06	I'm fine.	–
-1.17	I'm all right.	–
-1.17	I'm not sure.	–

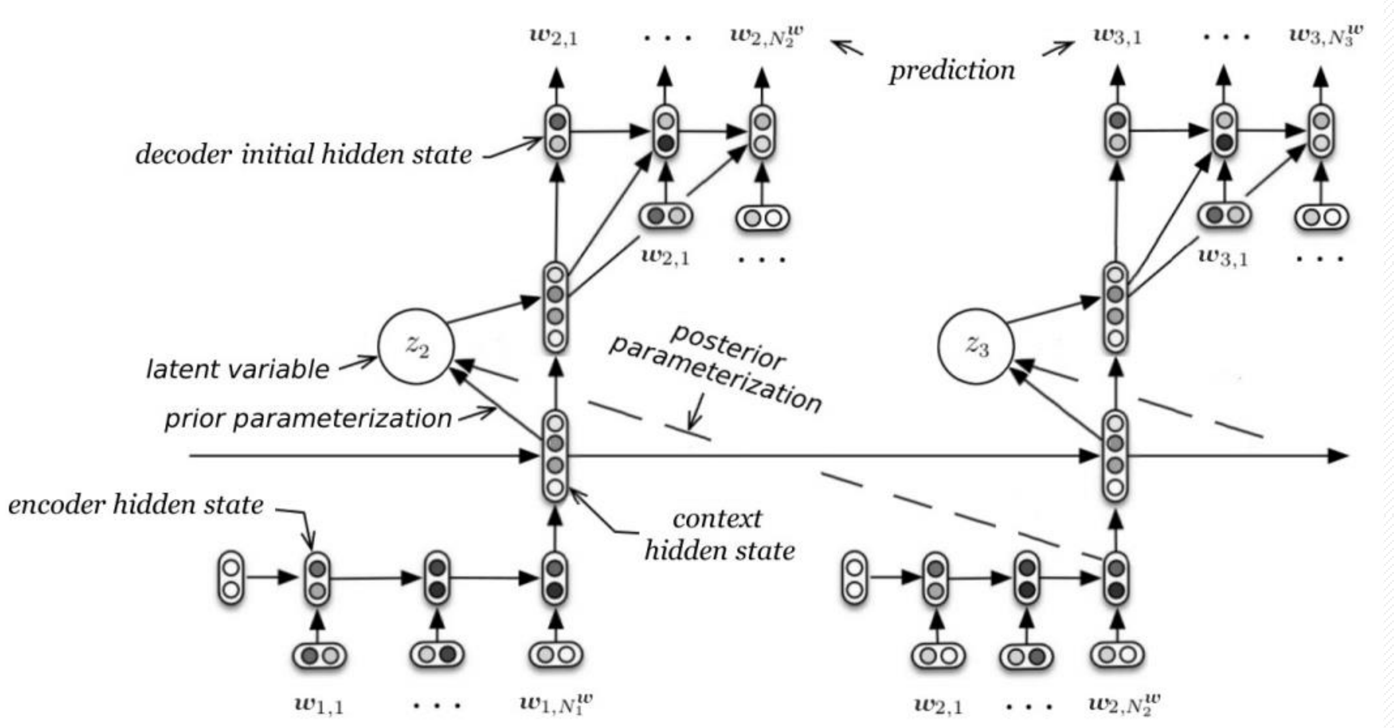
(Li et al., 2016)

More 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 HRED with Gaussian latent variable *z*
- *z* can capture high-level utterance features (e.g. topic, tone)
- When generating first sample latent variable, then 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	$\mathbf{42.5 \pm 2.6}^*$	25.2 ± 2.3
VHRED vs HRED	$\mathbf{42.0 \pm 2.8}^*$	31.9 ± 2.6	26.2 ± 2.5
VHRED vs TF-IDF	$\mathbf{51.6 \pm 3.3}^*$	17.9 ± 2.5	30.4 ± 3.0
Long Contexts			
VHRED vs LSTM	$\mathbf{41.9 \pm 2.2}^*$	36.8 ± 2.2	21.3 ± 1.9
VHRED vs HRED	$\mathbf{41.5 \pm 2.8}^*$	29.4 ± 2.6	29.1 ± 2.6
VHRED vs TF-IDF	$\mathbf{47.9 \pm 3.4}^*$	11.7 ± 2.2	40.3 ± 3.4

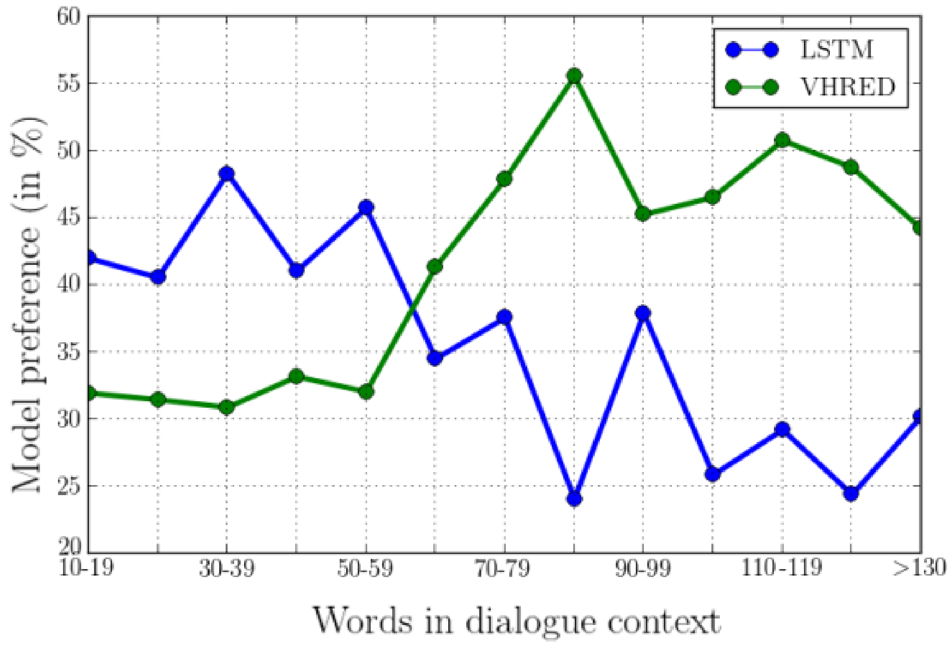


Figure 2: Human evaluator preferences for VHRED vs LSTM by context length excluding ties. For short contexts humans prefer the generic responses generated by LSTM, while for long contexts humans prefer the semantically richer responses generated by VHRED.

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

Future work

- Many interesting areas to be investigated:
  - Modifying the loss function
  - Adversarial training
  - Reinforcement learning
  - Learning from human interaction
  - ...

Problem #3: Evaluation

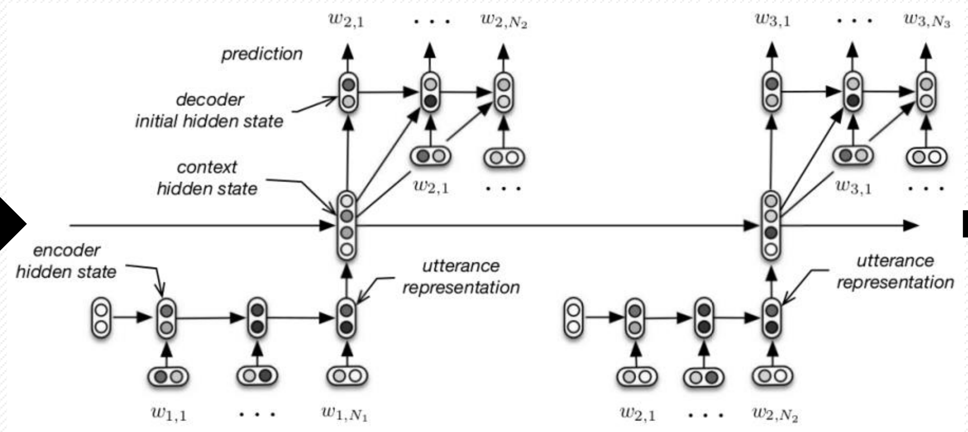
Dialogue evaluation

- Hard to know if we're making progress in building dialogue models
- Important to define – wrong metrics can lead to spurious research
- Human evaluation is effective, but slow and expensive – want to have an automatic evaluation metric
- **Lack of reliable metrics means researchers only compare to their own previously implemented models**

Comparison of ground-truth utterance

Context

Hey, want to go to the movies tonight?



Generated Response

Nah, let's do something active.

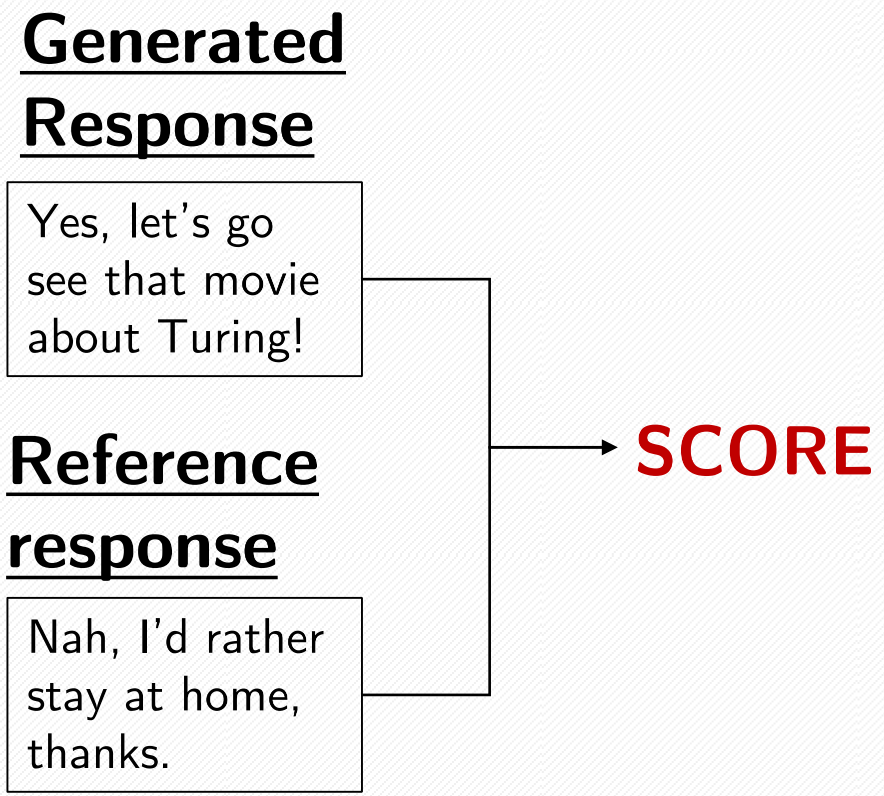
Reference response

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)

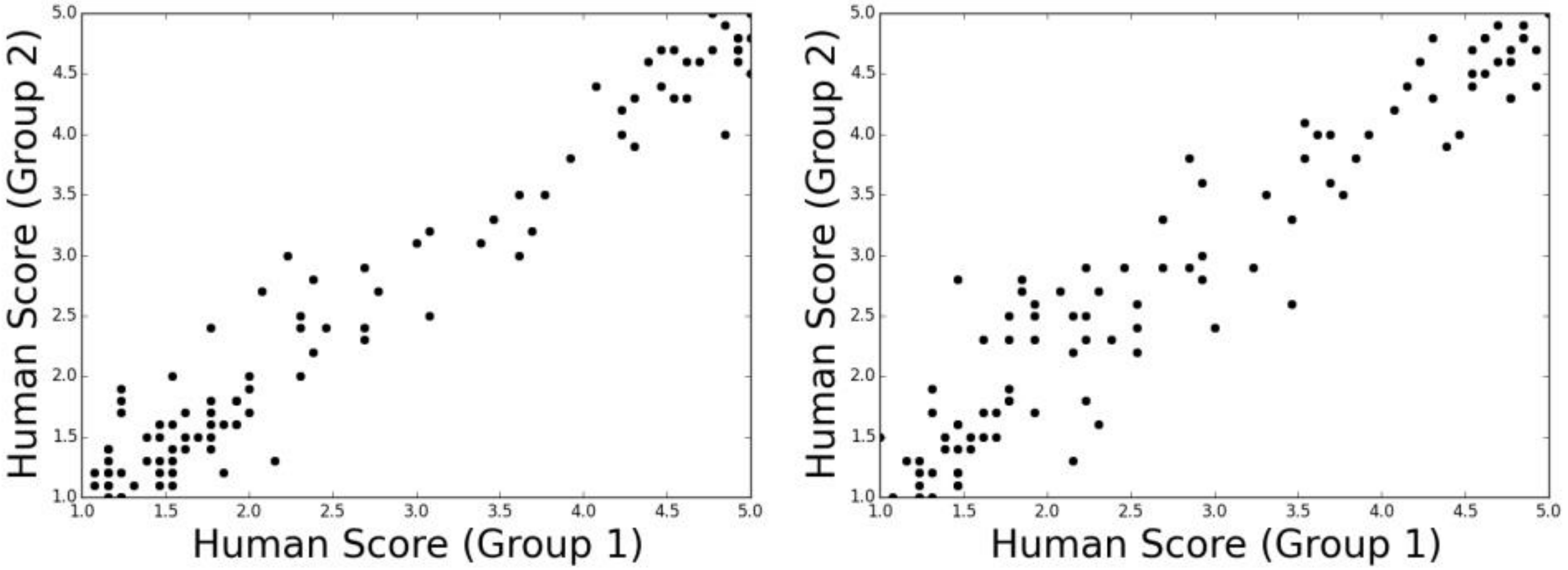
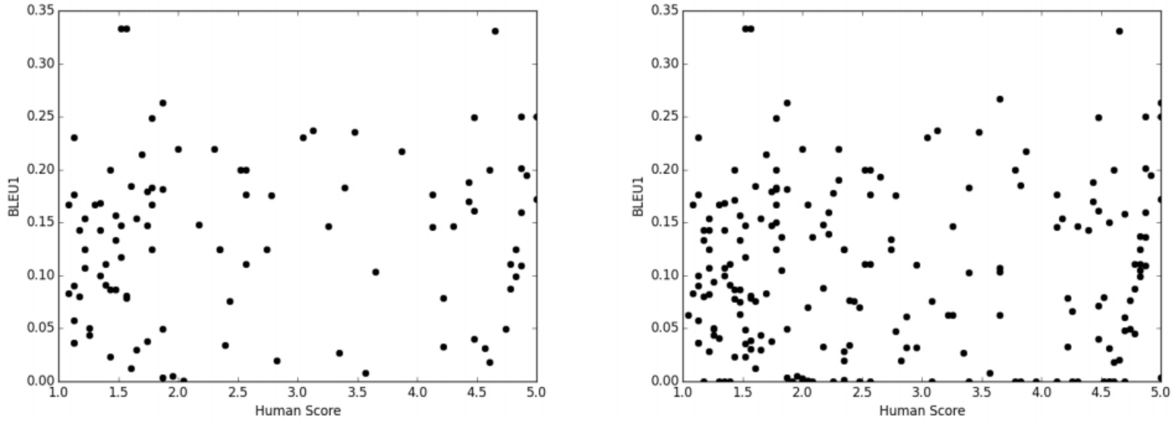
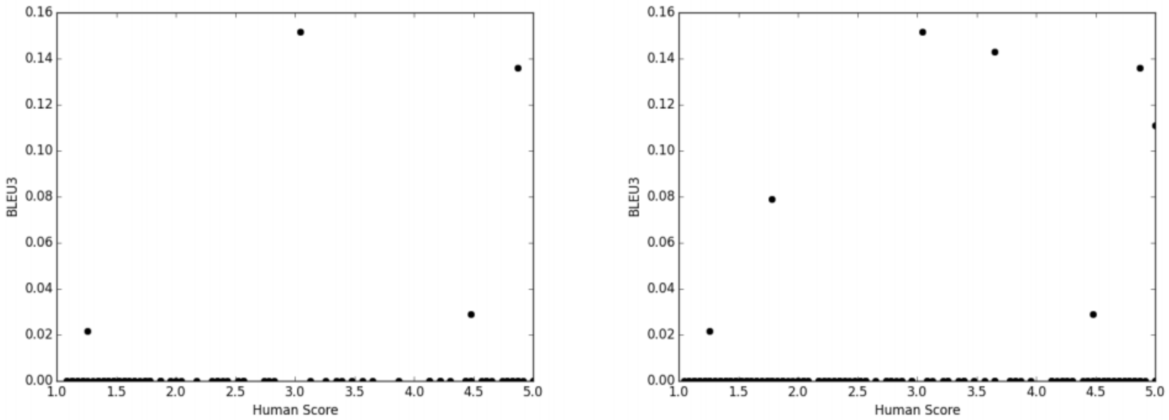


Figure 3: Scatter plots showing the correlation between two randomly chosen groups of human volunteers on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right).

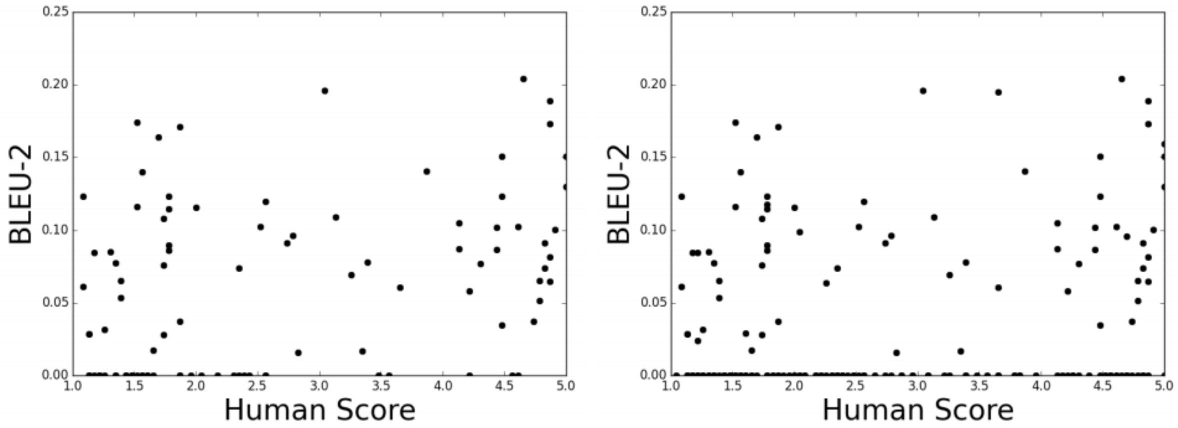
Reality (BLEU)



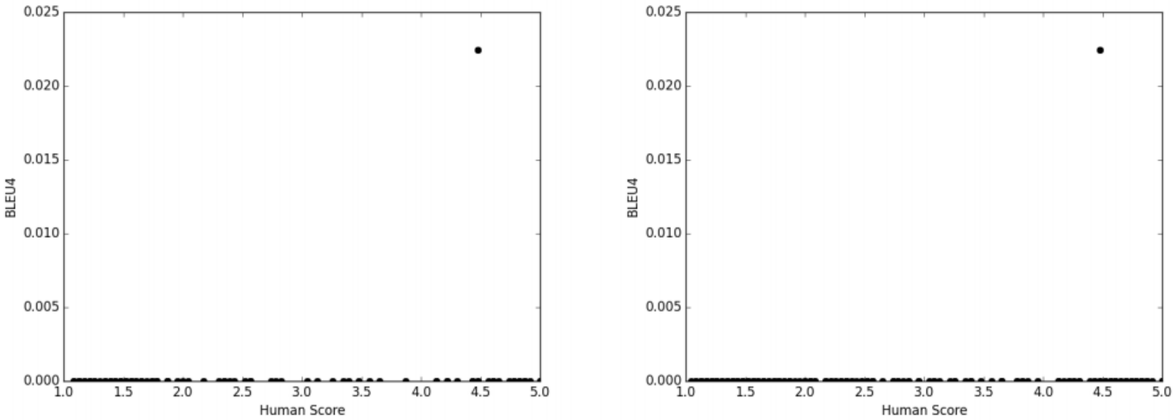
(a) BLEU-1



(c) BLEU-3

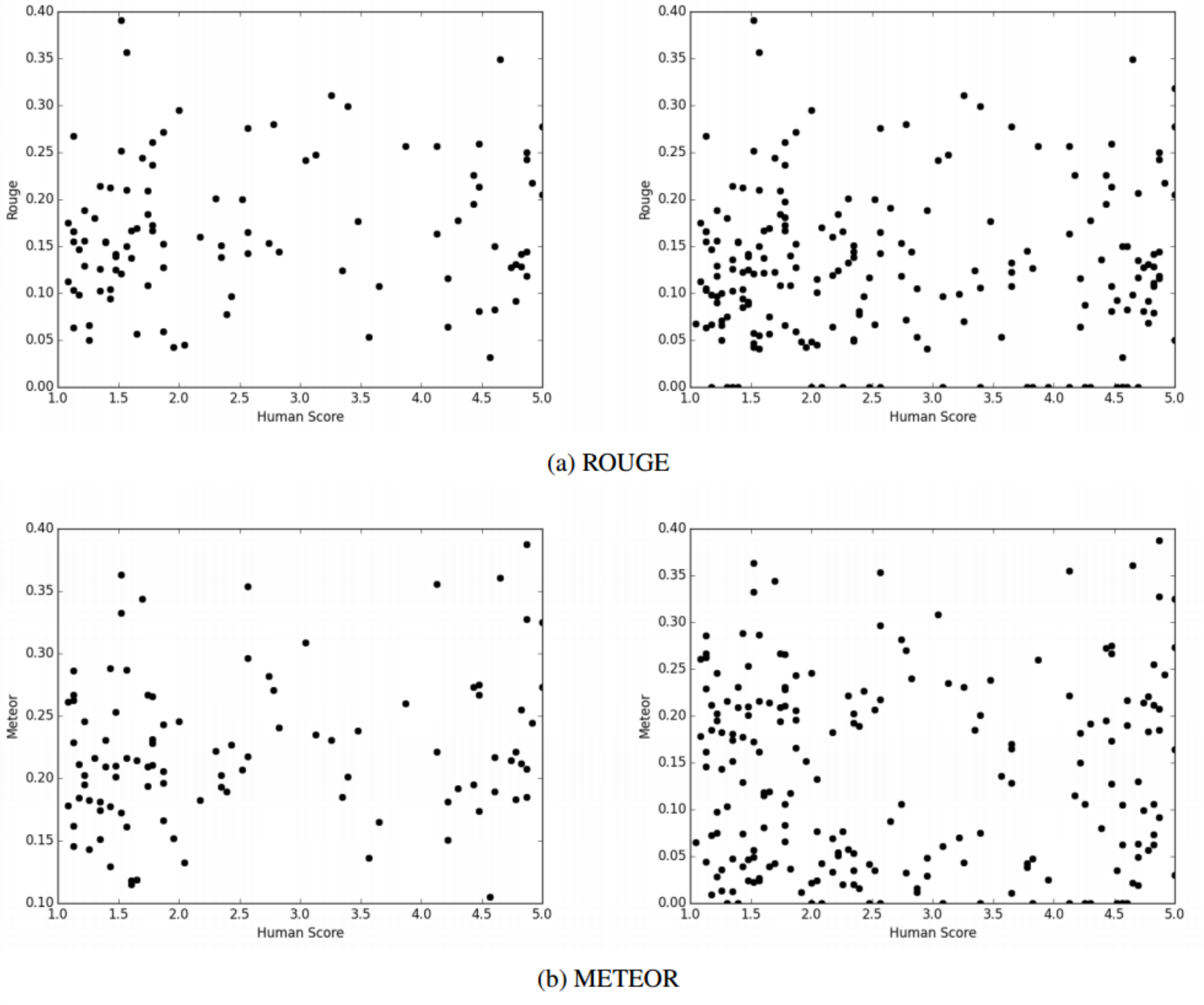


(b) BLEU-2



(d) BLEU-4

Reality (ROUGE & METEOR)



(a) ROUGE

(b) METEOR

Correlation Results

Original paper (Liu et al., 2016):

	Twitter			
Metric	**Spearman**	p-value	**Pearson**	p-value
Greedy	0.2119	0.034	0.1994	0.047
Average	0.2259	0.024	0.1971	0.049
Extrema	0.2103	0.036	0.1842	0.067
METEOR	0.1887	0.06	0.1927	0.055
BLEU-1	0.1665	0.098	0.1288	0.2
BLEU-2	0.3576	< 0.01	0.3874	< 0.01
BLEU-3	0.3423	< 0.01	0.1443	0.15
BLEU-4	0.3417	< 0.01	0.1392	0.17
ROUGE	0.1235	0.22	0.09714	0.34
Human	0.9476	< 0.01	1.0	0.0

After removing pre-processing artifacts (<speaker> token):

Metric	**Spearman**	**Pearson**
BLEU-1	-0.026 (0.80)	0.016 (0.87)
BLEU-2	0.065 (0.52)	0.080 (0.43)
BLEU-3	0.139 (0.17)	0.088 (0.39)
BLEU-4	0.139 (0.17)	0.092 (0.36)
ROUGE	-0.083 (0.41)	-0.010 (0.92)

Word-overlap metrics are poor substitute for human evaluations

Learning to evaluate

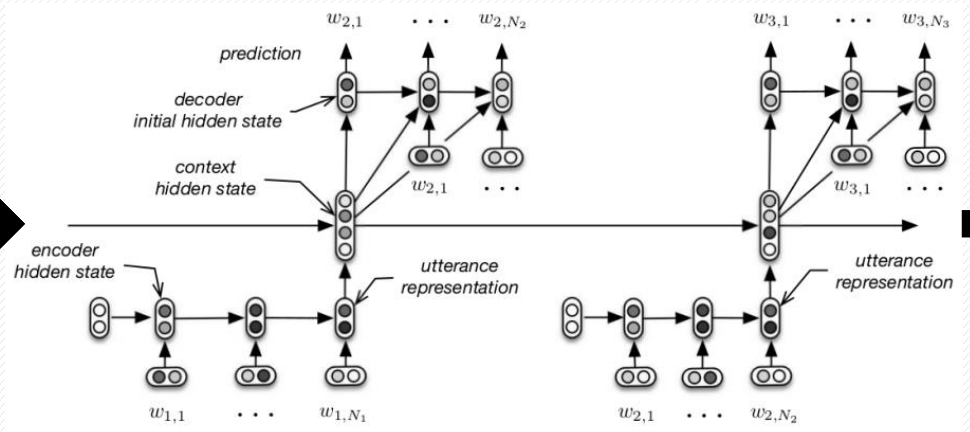
A dialogue response is probably good if it is rated highly by humans.

- 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 <u>and</u> 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!

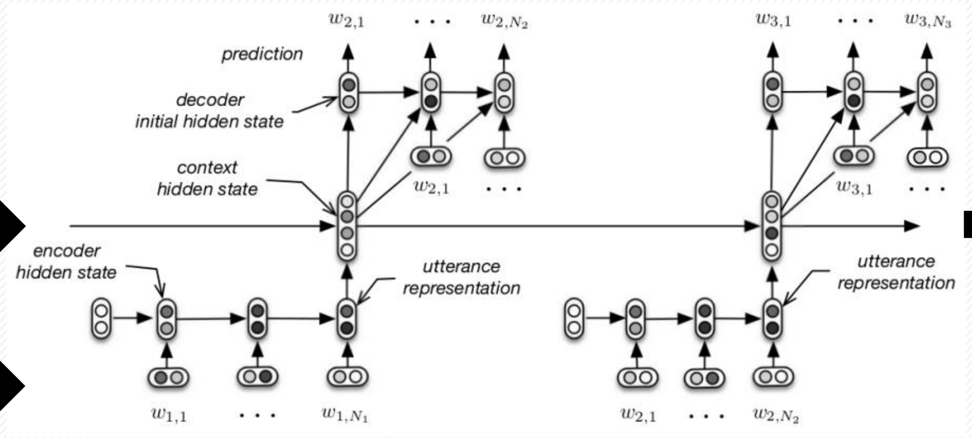
SCORE

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

Dialogue response score should also depend on context!

Evaluation dataset

Conducted 2 rounds of AMT studies to get evaluation on Twitter

<u>Study 1:</u> ask workers to generate next sentence of a conversation

<u>Study 2:</u> 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

Evaluation dataset

- Our simplifying assumption is that dialogue response quality measured by 'appropriateness'

- In our experiments, other measures ('topicality', 'informativeness', etc.) either had little inter-annotator agreement, or correlated strongly with 'appropriateness'

Measurement	κ **score**
Overall	0.63
Topicality	0.57
Informativeness	0.31
Background	0.05

Table 1: Median κ inter-annotator agreement scores for various questions asked in the survey.

ADEM

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

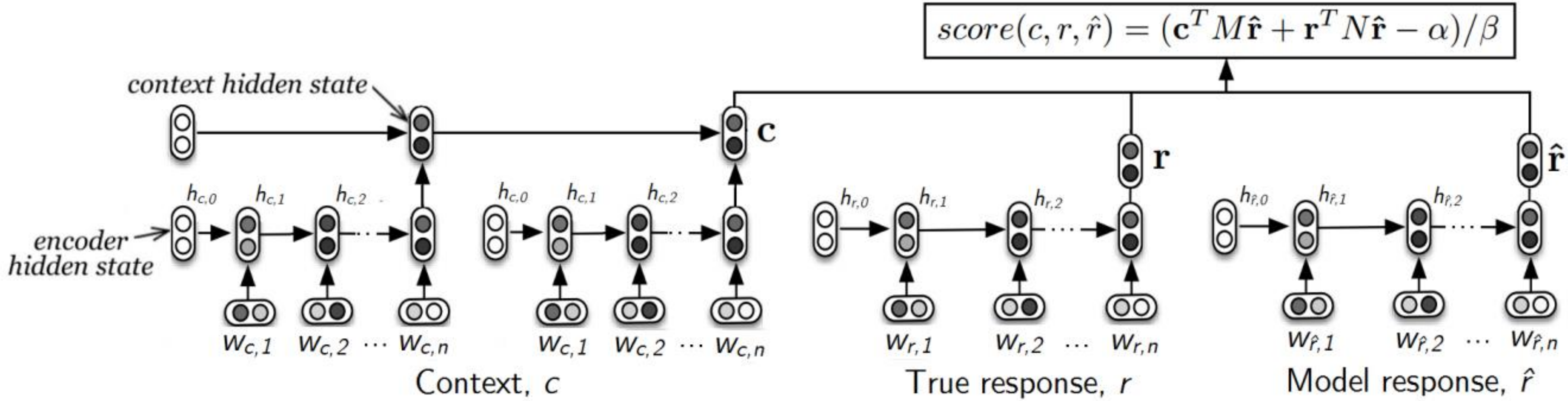


Figure 2: The ADEM model, which uses a hierarchical encoder to produce the context embedding **c**.

ADEM pre-training

- Want model that can learn from limited data (since collection is expensive)

- Pre-train RNN encoder of ADEM using VHRED

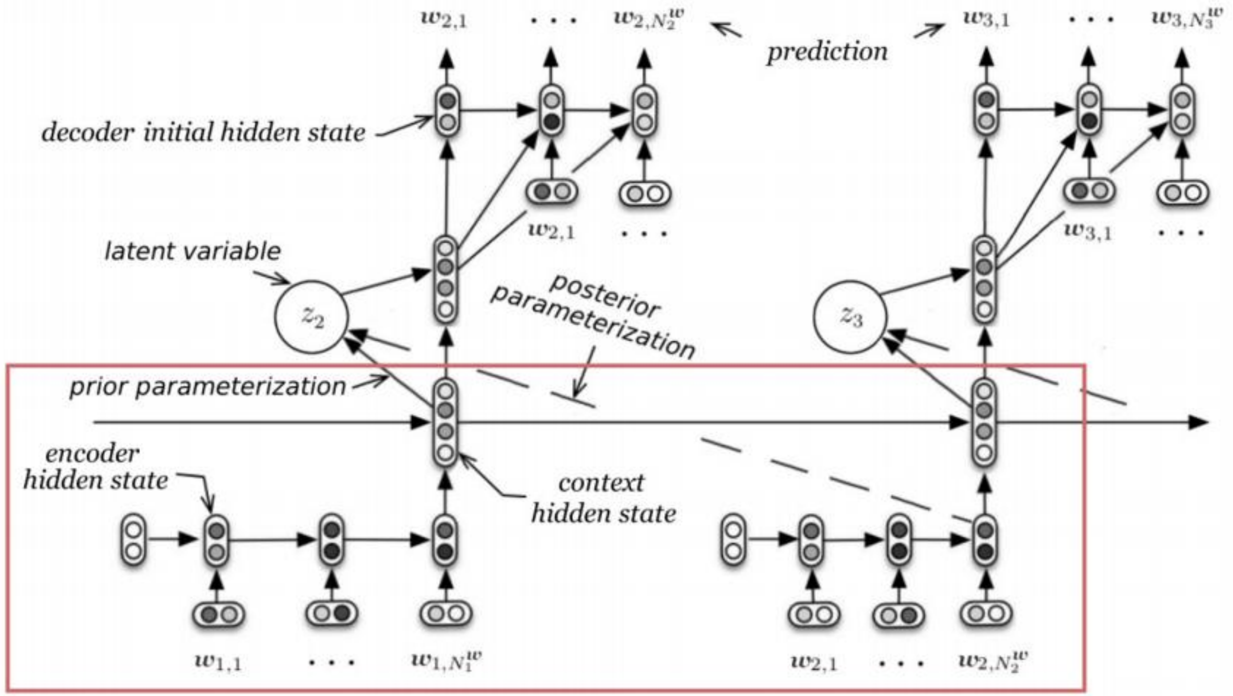


Figure 5: The VHRED model used for pre-training. The hierarchical structure of the RNN encoder is shown in the red box around the bottom half of the figure.

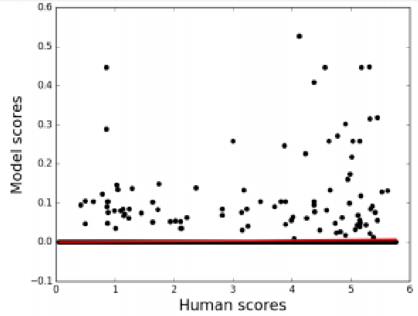
Length correlation

Problem: humans favour shorter responses, and ADEM can trivially use this for better performance (length gets 0.27 correlation with human score)

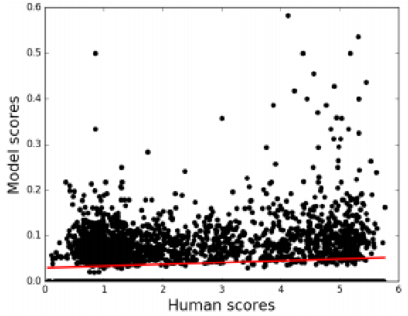
Solution: bin training set examples by length, re-weight samples such that each length bin has same average score

Utterance-level results

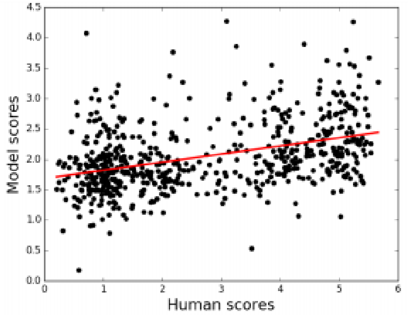
	Full dataset		Test set	
Metric	**Spearman**	**Pearson**	**Spearman**	**Pearson**
BLEU-2	0.039 (0.013)	0.081 (<0.001)	0.051 (0.254)	0.120 (<0.001)
BLEU-4	0.051 (0.001)	0.025 (0.113)	0.063 (0.156)	0.073 (0.103)
ROUGE	0.062 (<0.001)	0.114 (<0.001)	0.096 (0.031)	0.147 (<0.001)
METEOR	0.021 (0.189)	0.022 (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)



(a) BLEU-2

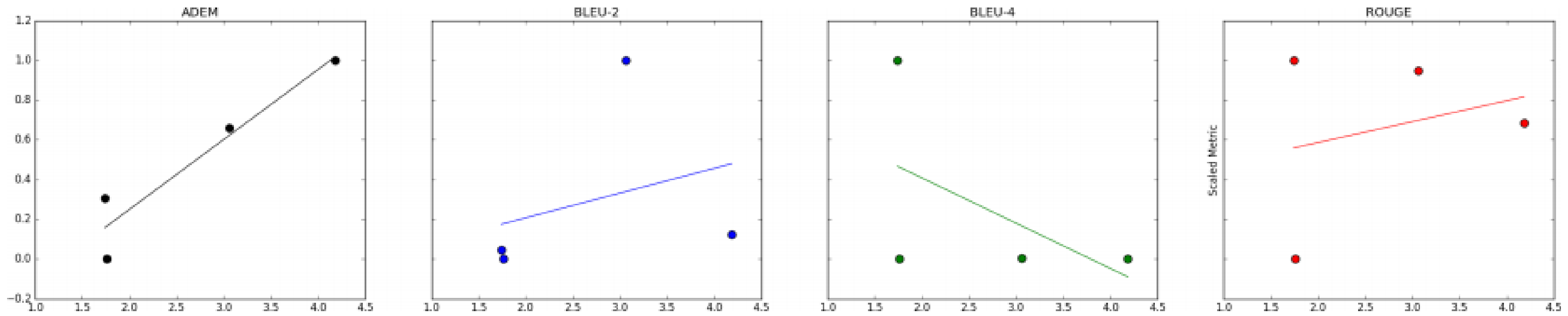


(b) ROUGE



(c) ADEM

System-level results



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	**0.954** (0.046)

Figure 4: Scatterplots depicting the system-level correlation results for ADEM, BLEU-2, BLEU-4,and ROUGE on the test set. 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.

Results – generalization

	Test on full dataset		Test on removed model responses	
Data Removed	**Spearman**	**Pearson**	**Spearman**	**Pearson**
TF-IDF	0.406 (<0.001)	0.409 (<0.001)	0.186 (0.021)	0.196 (0.015)
Dual Encoder	0.364 (<0.001)	0.373 (<0.001)	0.026 (0.749)	0.027 (0.736)
HRED	0.393 (<0.001)	0.396 (<0.001)	0.151 (0.060)	0.148 (<0.070)
Human	0.292 (<0.001)	0.298 (<0.001)	0.216 (<0.010)	0.148 (<0.070)
Average	**0.364**	**0.369**	**0.145**	**0.130**
25% at random	0.378 (<0.001)	0.384 (<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 responses only from the dialogue model not seen during training. The last row (25% at random) corresponds to the ADEM model trained on all model responses, but with the same amount of training data as the model above (i.e. 25% less data than the full training set).

How useful is this?

- Moderately. Need to collect more data for better generalization
- Only considers single utterances, rather than a whole dialogue
- What about other aspects of dialogue quality?

Adversarial evaluation

- Rather than imitating human scores, train a model to distinguish between real and generated responses (Kannan et al, 2016; Li et al., 2017)
- Similar to discriminator in a GAN
- Combines well with ADEM – want dialogue responses that are appropriate, and similar to human responses

Model	Accuracy (%)
HRED	99.28
VHRED	97.87
Reference	97.27
Average	**98.14**

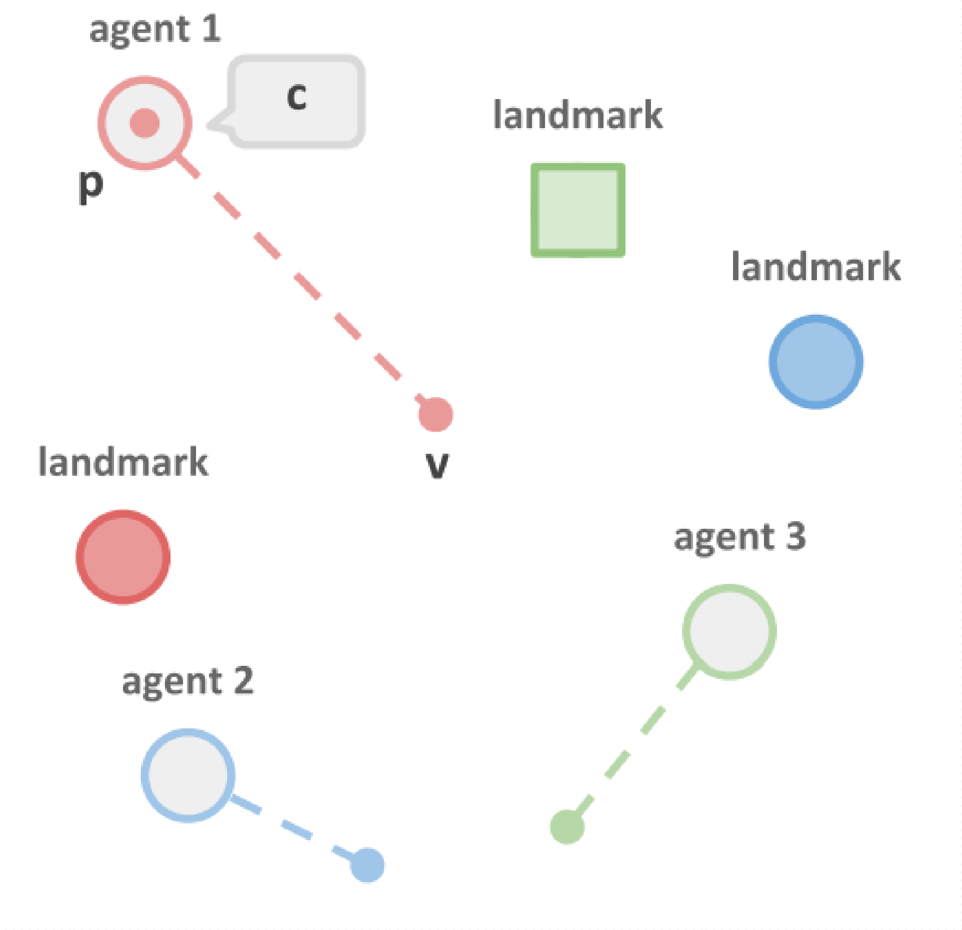
Table 6: Performance of the CAE model in terms of accuracy of predicting y.

Problem #4: Entire Premise?

Learning from static datasets

- Will training solely from static datasets lead to a 'general-purpose communicating agent'?
- Probably not. In this setting, we are primarily learning the statistical structure of language
- But we also want to learn the function of language, and ground the learned language in the agent's observations
- An alternative approach: have simulated agents in physical environments learn to communicate to solve tasks in that environment (Gauthier & Mordatch, 2016)

Multi-agent language learning



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


Nissan Pow
McGill


Aaron Courville
U. Montreal


Yoshua Bengio
U. Montreal

References

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.

Kannan, Vinyals. "Adversarial Evaluation of Dialogue Models." NIPS Workshop on Adversarial Training, 2016.

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

Li, Monroe, Shi, Ritter, Jurafsky. "Adversarial Learning for Neural Dialogue Generation." 2017.

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.

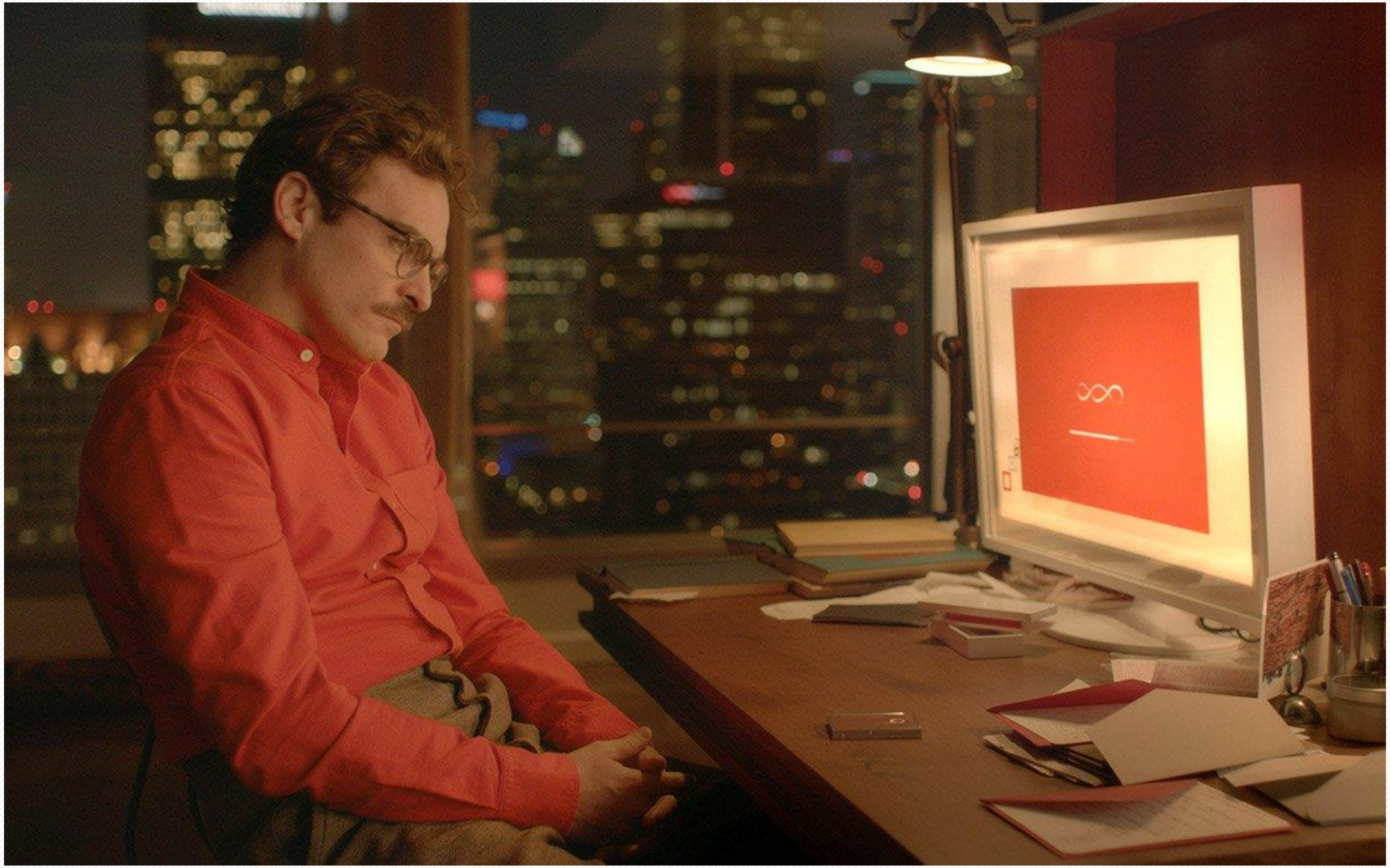
Serban, Lowe, Charlin, Pineau. "A Survey of Available Corpora for Building Data-Driven Dialogue Systems." 2016.

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!

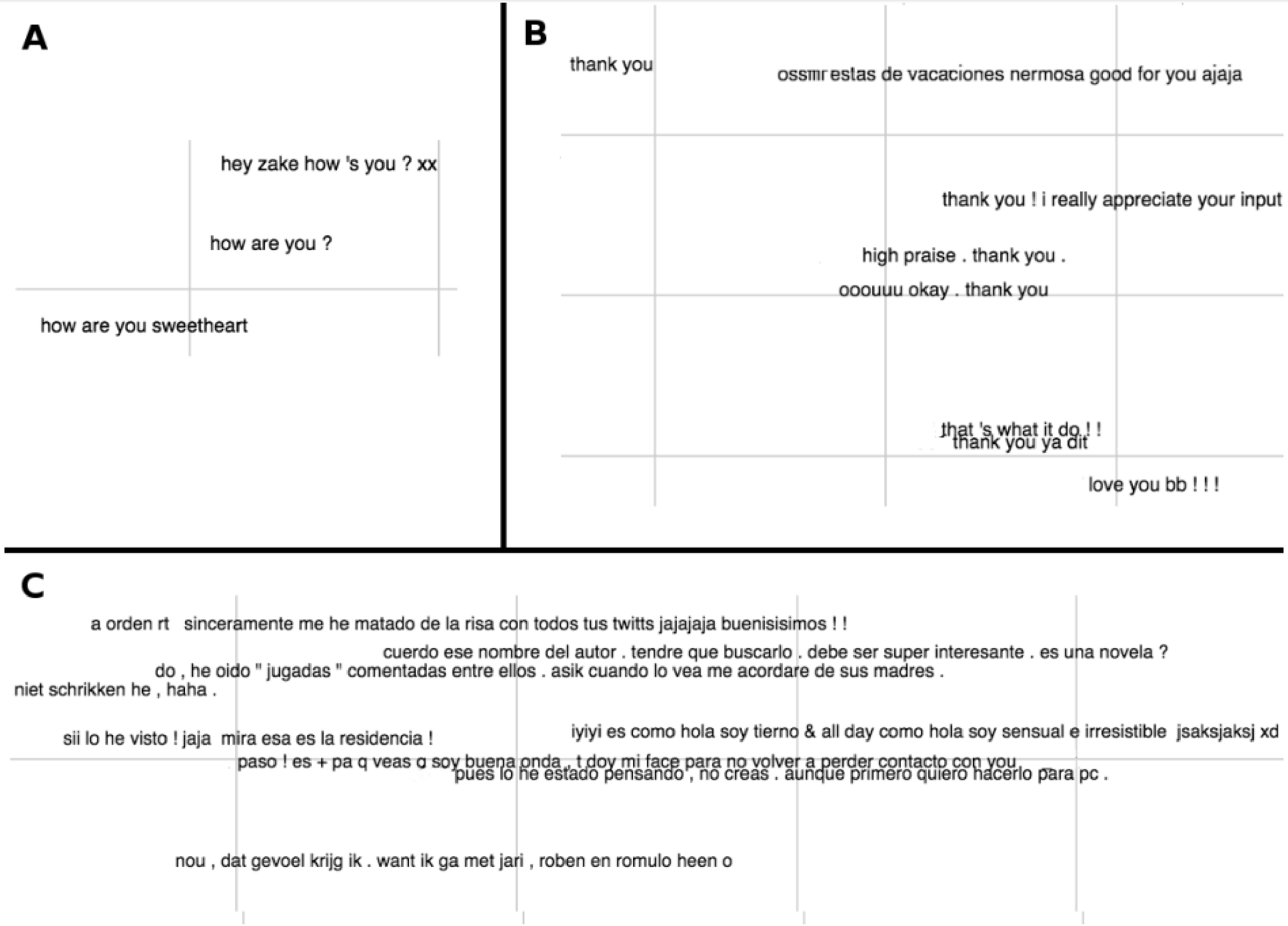


Quantitative VHRED results

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



Length bias of word overlap metrics

	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

Where does ADEM 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 |metric| indicate that the metric scores have been normalized.