Modern Challenges in Building End-to-End Dialogue Systems

Ryan Lowe
McGill University

Primary Collaborators

Joelle Pineau
McGill

Iulian V. Serban
U. Montreal

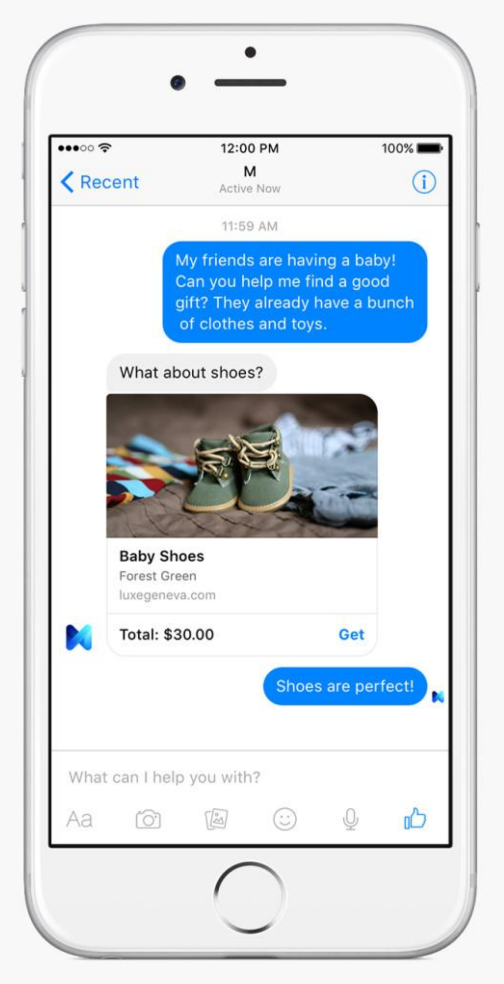
Mike Noseworthy
McGill

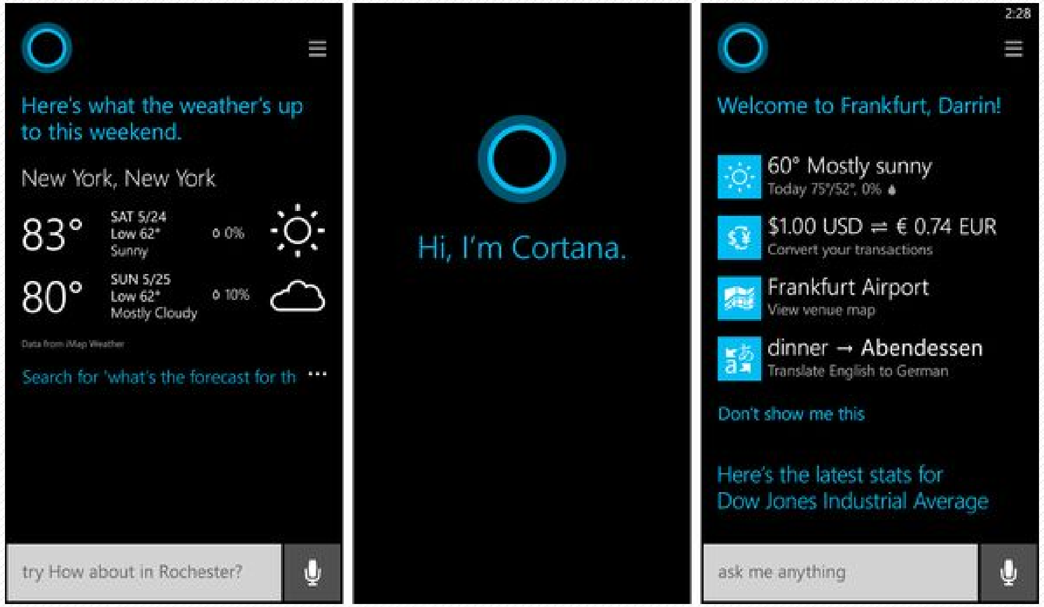
Chia-Wei Liu
McGill

Nissan Pow
McGill

Laurent Charlin
HEC Montreal

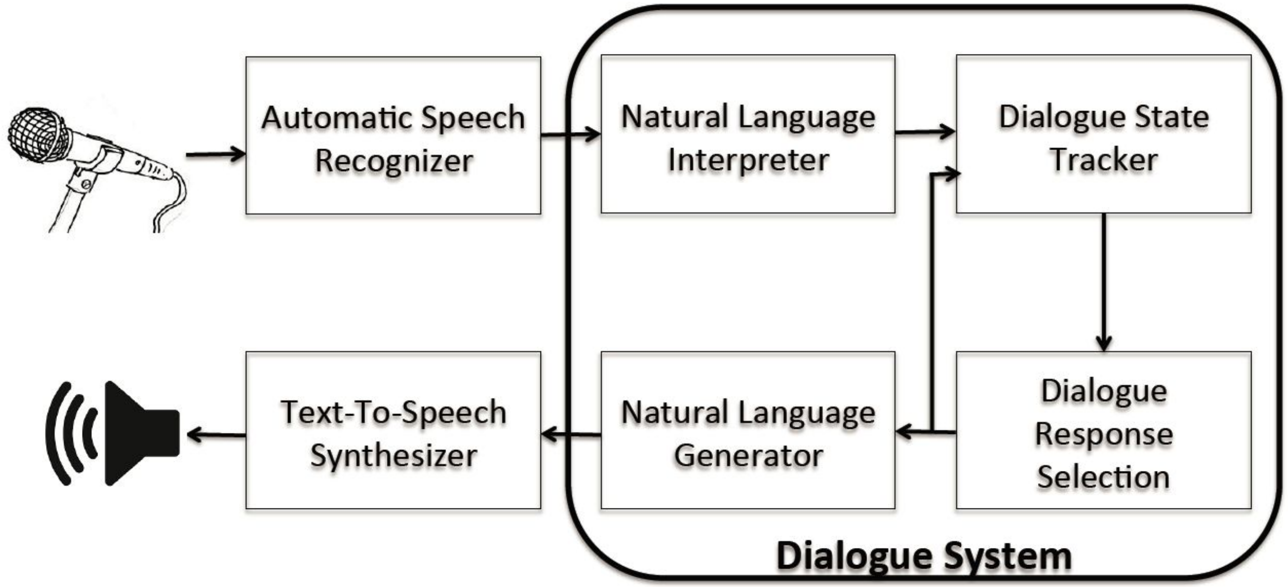
Dialogue Systems





Modular Dialogue Systems

- Traditional system consists of modules
- Each module optimized with separate objective function



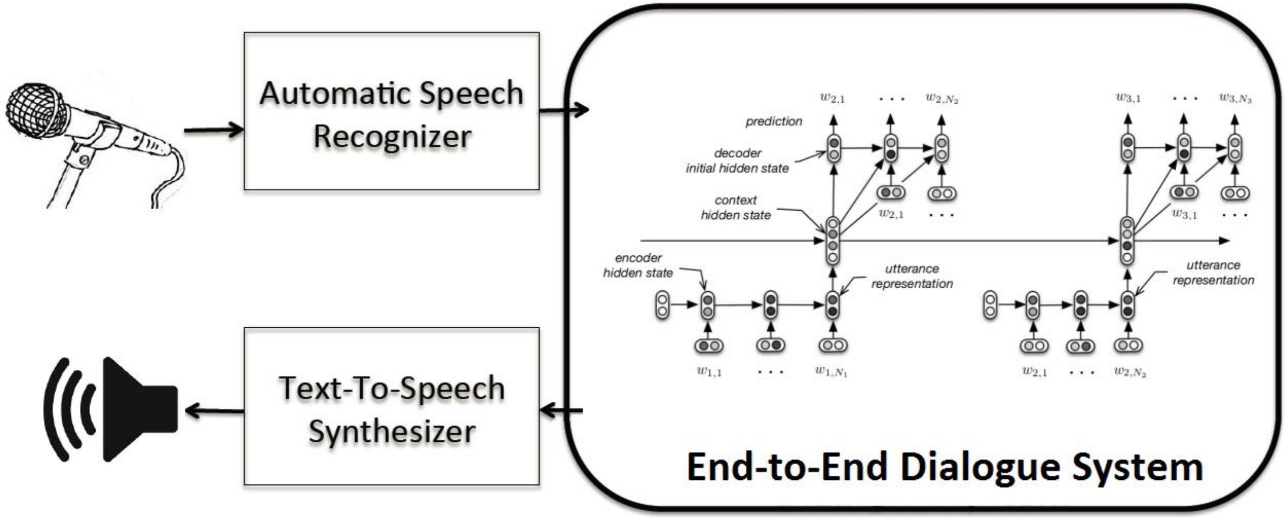
- Achieves good performance with small amounts of data

Problem: does not work well in 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)

End-to-End Dialogue Systems

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 does not scale well)

Challenge #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

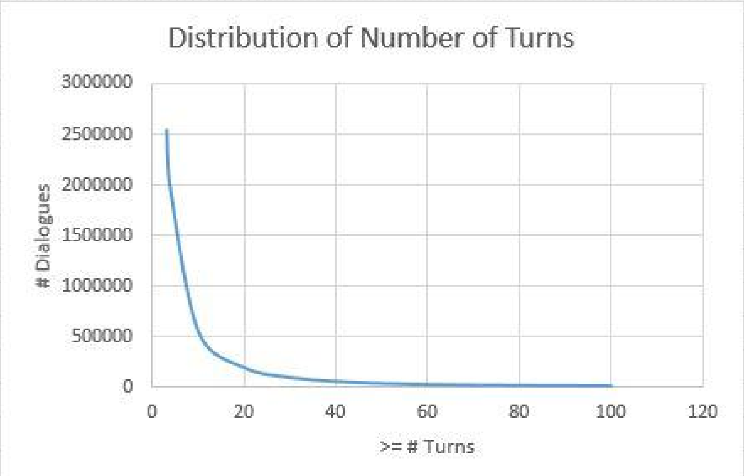
Ubuntu Dialogue Corpus




- 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	k11: 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	k11: 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.

Other Datasets

- Twitter Corpus, 850k Twitter dialogues (Ritter et al., 2011)

- Movie Dialog Dataset, 1 million Reddit dialogues (Dodge et al. 2016)

- Our survey paper covering existing datasets:

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

- **Needs more work!**

Challenge #2:
Generic Responses

The Problem of Generic Responses

- Most models trained to predict most likely next utterance given context

- But some utterances are likely given any context!

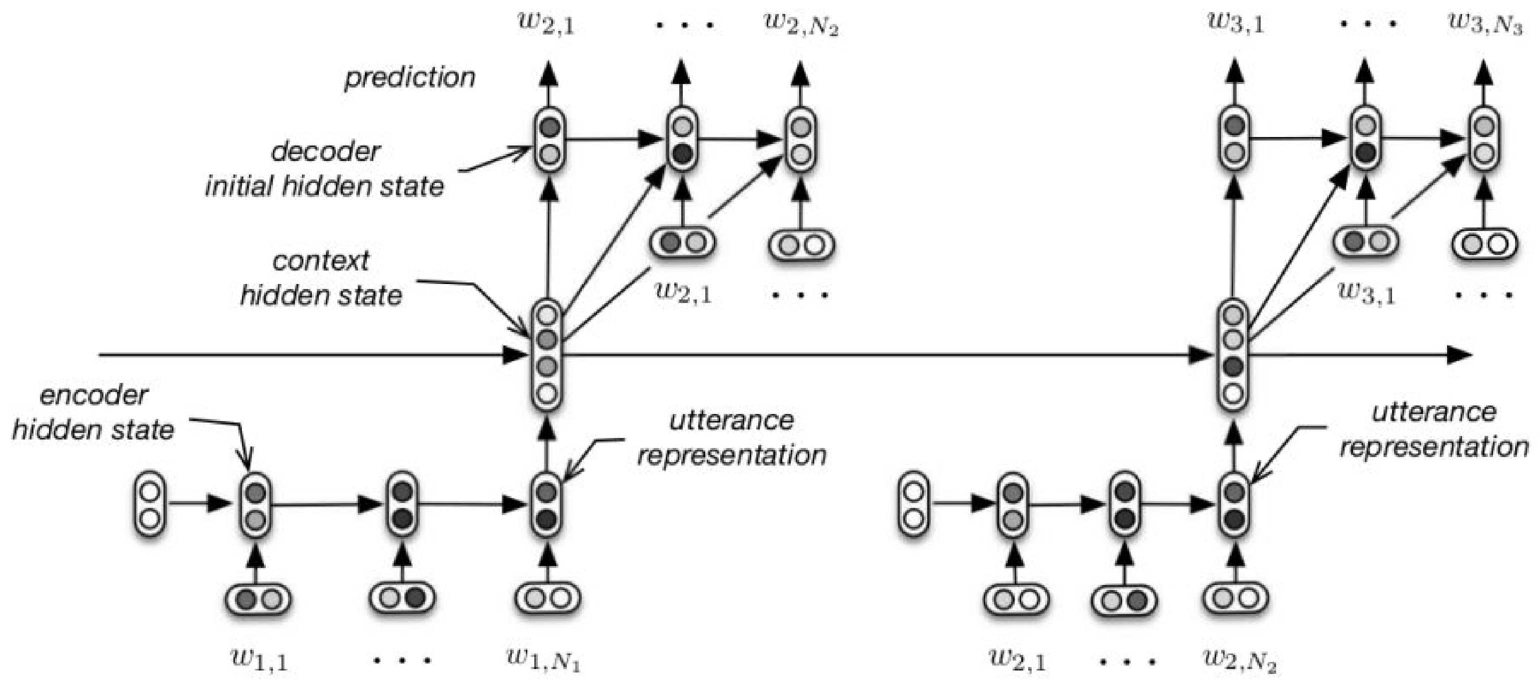
- Neural models often generate "I don't know", or "I'm not sure" to most contexts

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)

Encoder-Decoder

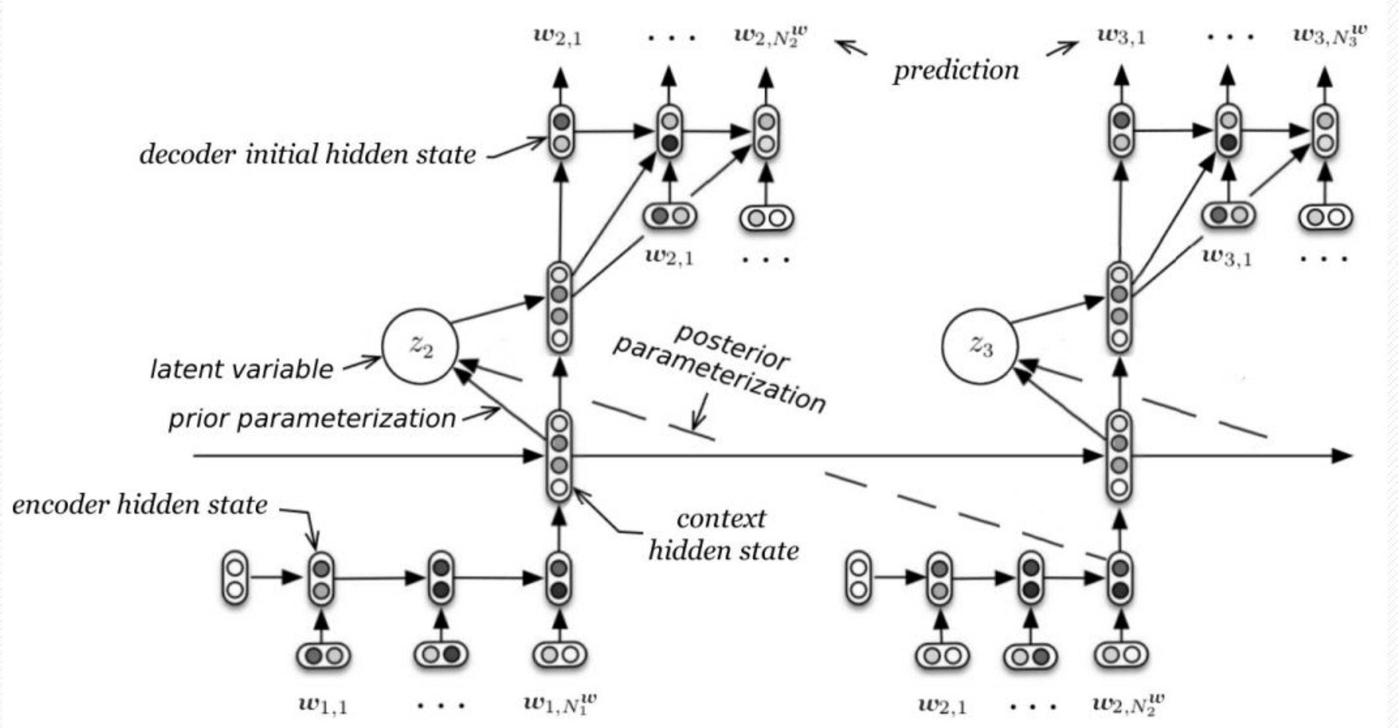
- Use RNN to encode text into fixed-length vector representation
- Use another RNN to decode representation to text
- Can make this hierarchical



Cho et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." *EMNLP* 2014.
Serban, Sordoni, Bengio, Courville, Pineau. "Building End-to-End Dialogue Systems using Generative Hierarchical Neural Network Models" *AAAI*, 2015.

Variational Encoder-Decoder (VHRED)

- Augment encoder-decoder with Gaussian latent variable
- Inspired by VAE (Kingma & Welling, 2014)
- 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)

- VHRED generates longer responses with higher entropy
- Outperforms baselines in most experiments

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)

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

Variational Encoder-Decoder (VHRED)

i love katie waissel more than like, anything in life. → i wanna c the show again! i need <number> meet her again! n i want a b-day tweet nxt week doubt that will happen x	**VHRED:** me too. i'm going to have to wait for her to come back. i'm so excited. **LSTM:** i'm sure she will. i'm sure she'll be fine. **HRED:** i'm sure she'll be fine, i'm sure she'll be fine.

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

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

Diversity-Promoting Objective

- Uses new objective: maximize the mutual information between source sentence S and target T

$$\hat{T} = \arg\max_{T} \left\{ \log p(T|S) - \lambda \log p(T) \right\}$$

- Can be considered a penalty on generic responses
- Gives slightly better results

Input: What are you doing?			
−0.86	I don't know.	−1.09	Get out of here.
−1.03	I don't know!	−1.09	I'm going home.
−1.06	Nothing.	−1.09	Oh my god!
−1.09	Get out of the way.	−1.10	I'm talking to you.
Input: what is your name?			
−0.91	I don't know.	...	
−0.92	I don't know!	−1.55	My name is Robert.
−0.92	I don't know, sir.	−1.58	My name is John.
−0.97	Oh, my god!	−1.59	My name's John.
Input: How old are you?			
−0.79	I don't know.	...	
−1.06	I'm fine.	−1.64	Twenty-five.
−1.17	I'm all right.	−1.66	Five.
−1.17	I'm not sure.	−1.71	Eight.

Li, Galley, Brockett, Gao, Dolan. "A Diversity-Promoting Objective Function for Neural Conversational Models." *arXiv:1510.03055*, 2016.

Challenge #3: Evaluation

Automatic Dialogue Evaluation

- Want a **fully automatic** way of evaluating the quality of a dialogue system

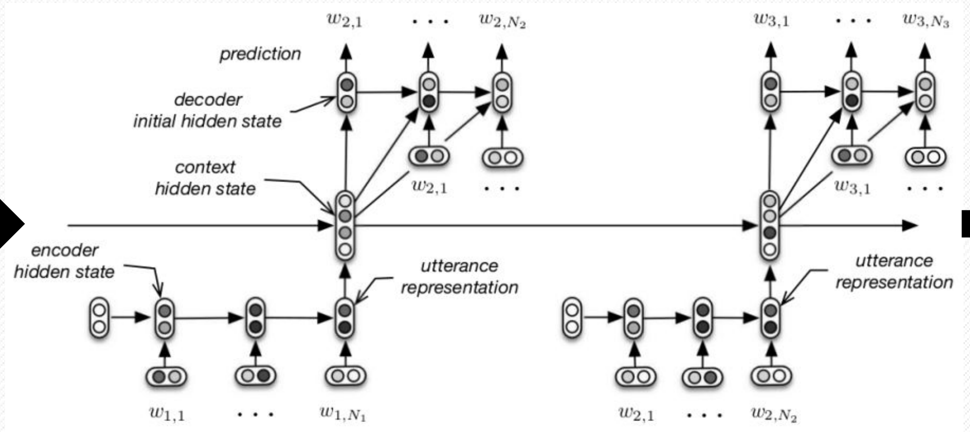
- If there is no notion of ‘task completion’, this is very hard

- Current methods compare the generated system response to the ground-truth next response

Comparison of ground-truth utterance

Context

Hey, want to go to the movies tonight?



Generated Response

Yeah, let's go see that movie about Turing!

Ground-truth response

Nah, I'd rather stay at home, thanks.

SCORE

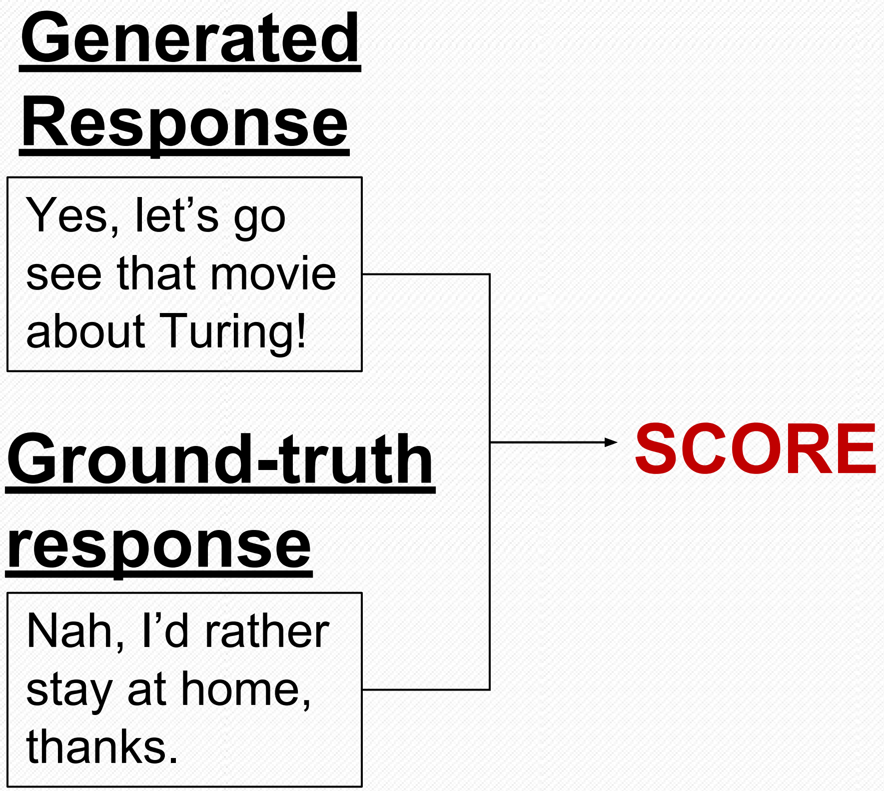
Comparison of ground-truth utterance

1) Word-overlap metrics:

- BLEU, METEOR, ROUGE

2) Word embedding-based metrics:

- Vector extrema, greedy matching, embedding average



Human 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

Liu*, Lowe*, Serban*, Noseworthy*, Charlin, Pineau. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Systems.” *EMNLP*, 2016.

Goal (inter-annotator)

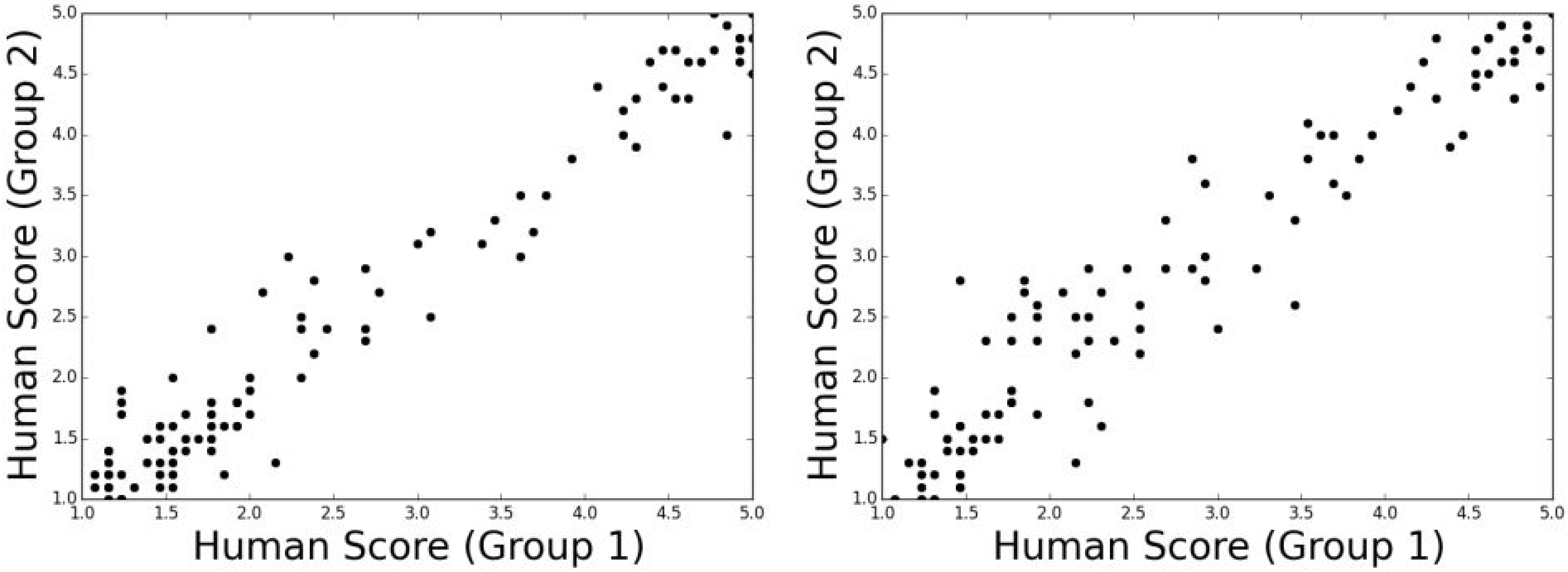
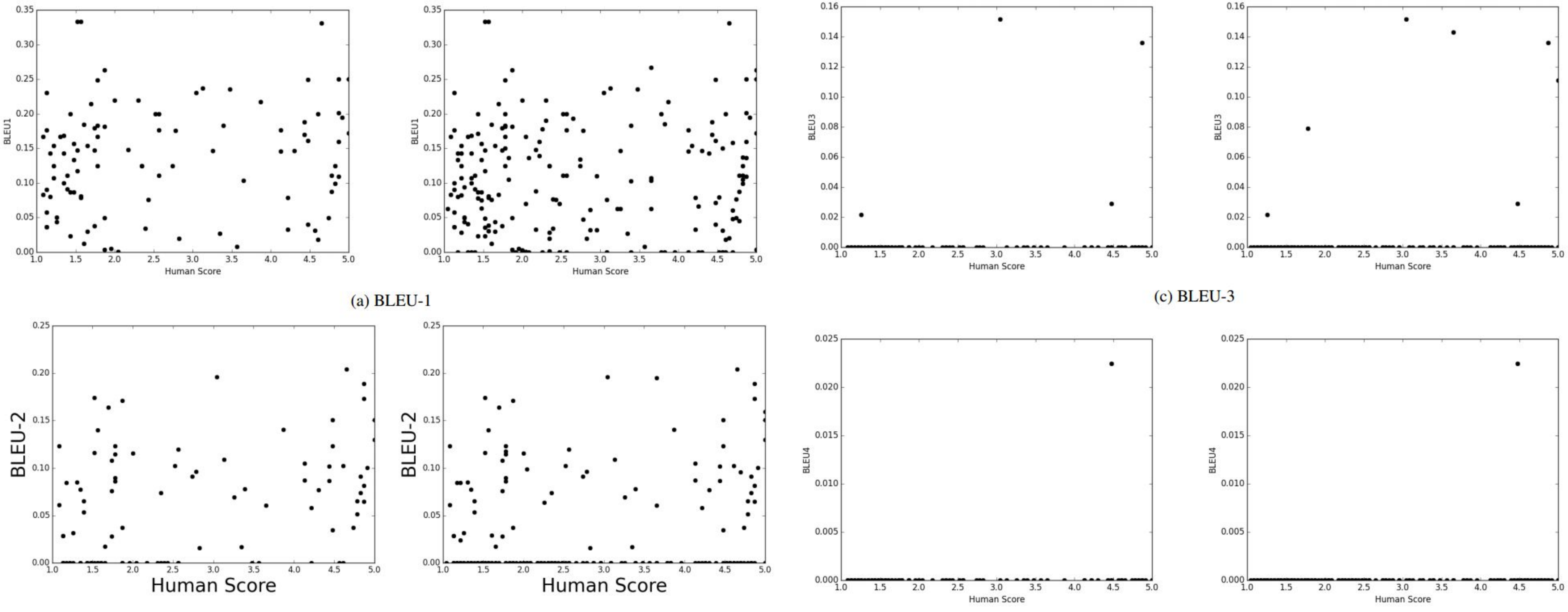


Figure 3: Scatter plots showing the correlation between two randomly chosen groups of human volunteers on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right).

Liu*, Lowe*, Serban*, Noseworthy*, Charlin, Pineau. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Systems." *EMNLP*, 2016.

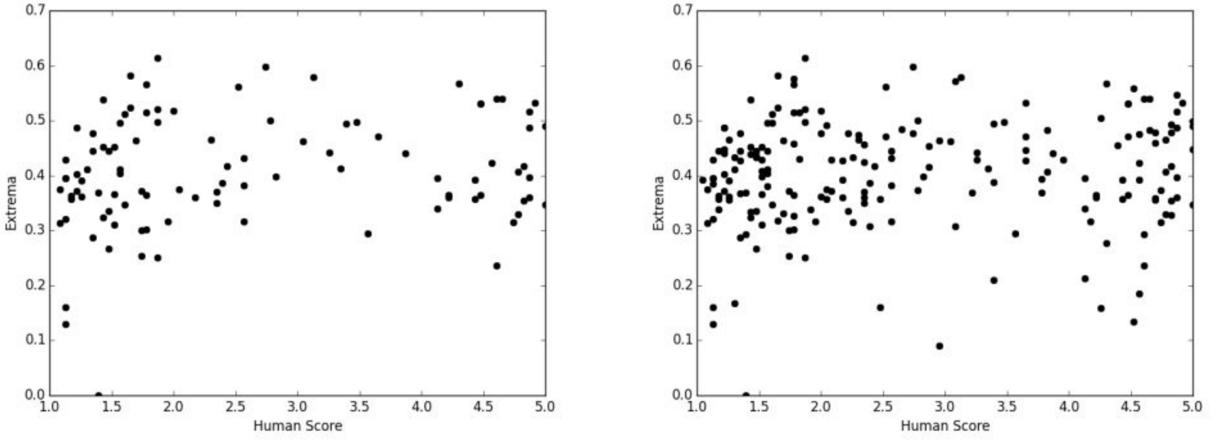
Reality (BLEU)



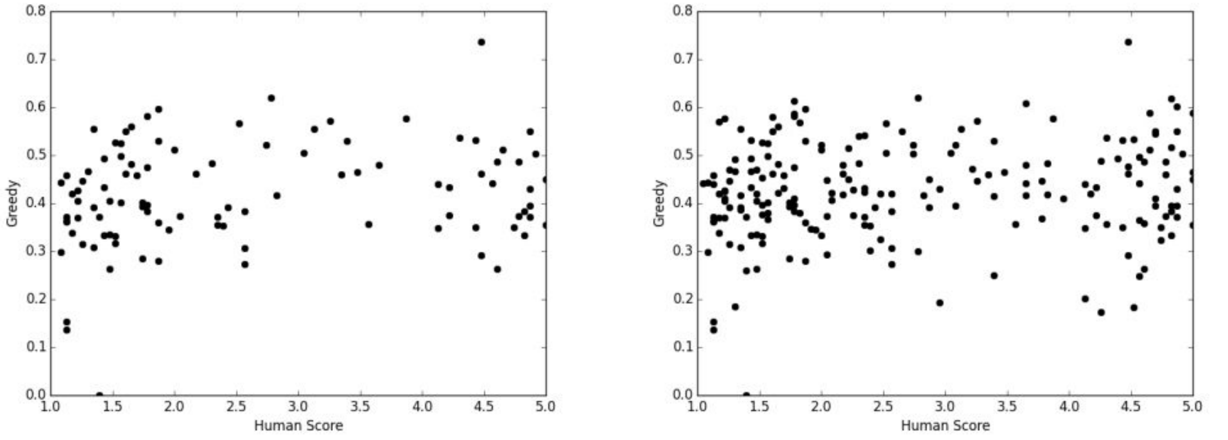
(a) BLEU-1
(b) BLEU-2
(c) BLEU-3
(d) BLEU-4

Liu*, Lowe*, Serban*, Noseworthy*, Charlin, Pineau. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Systems.” *EMNLP*,

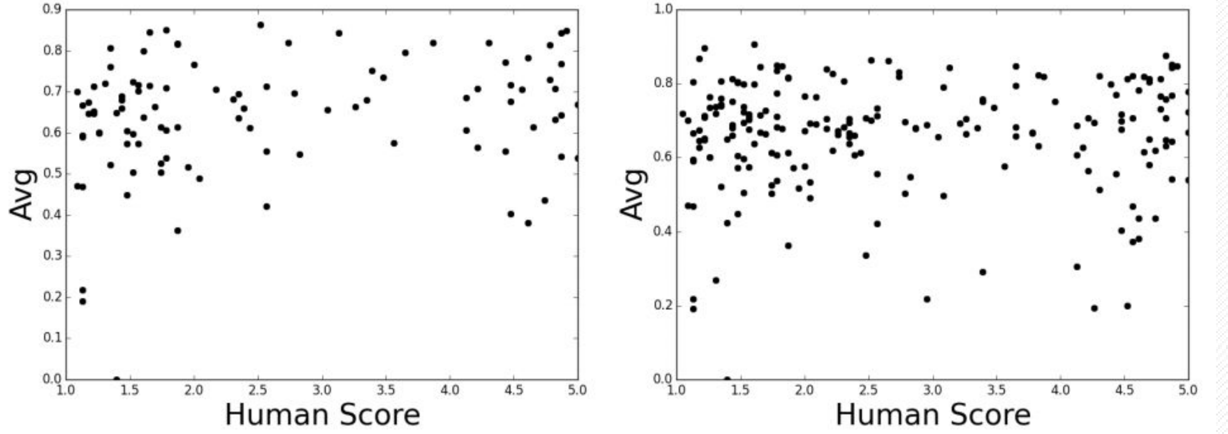
Reality (vector-based)



(a) Vector Extrema



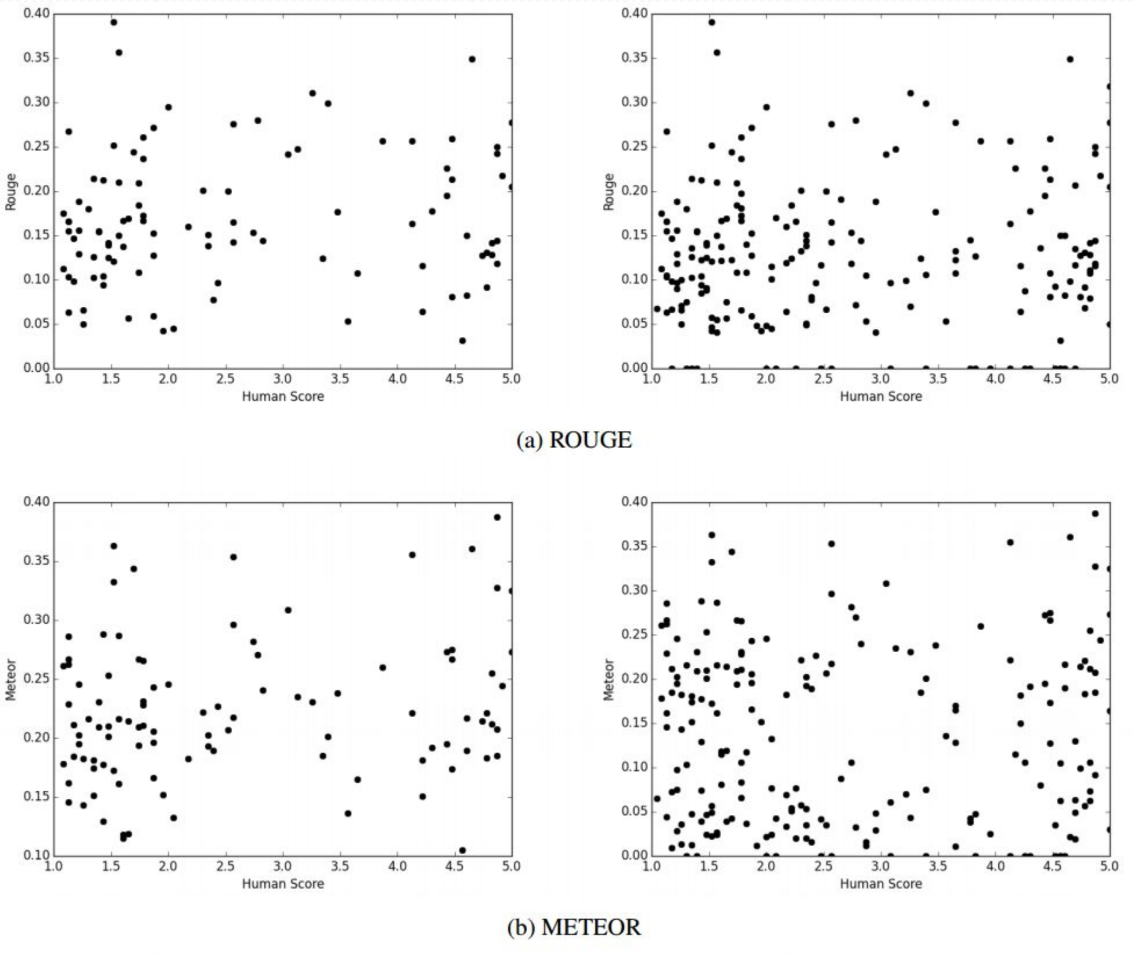
(b) Greedy Matching



(c) Vector Averaging

Liu*, Lowe*, Serban*, Noseworthy*, Charlin, Pineau. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Systems." *EMNLP*, 2016.

Reality (ROUGE & METEOR)



(a) ROUGE

(b) METEOR

Liu*, Lowe*, Serban*, Noseworthy*, Charlin, Pineau. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Systems.” *EMNLP*,

Correlation Results

	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.

Liu*, Lowe*, Serban*, Noseworthy*, Charlin, Pineau. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Systems.” *EMNLP*,

Next Utterance Classification

- Instead of evaluating model responses, can use an auxiliary task
- Have models predict next utterance in conversation from a list (multiple-choice style)
- Mitigates problem with response diversity (and many other advantages!)

Speaker A: yo .
Speaker B: damone . it 's mark .

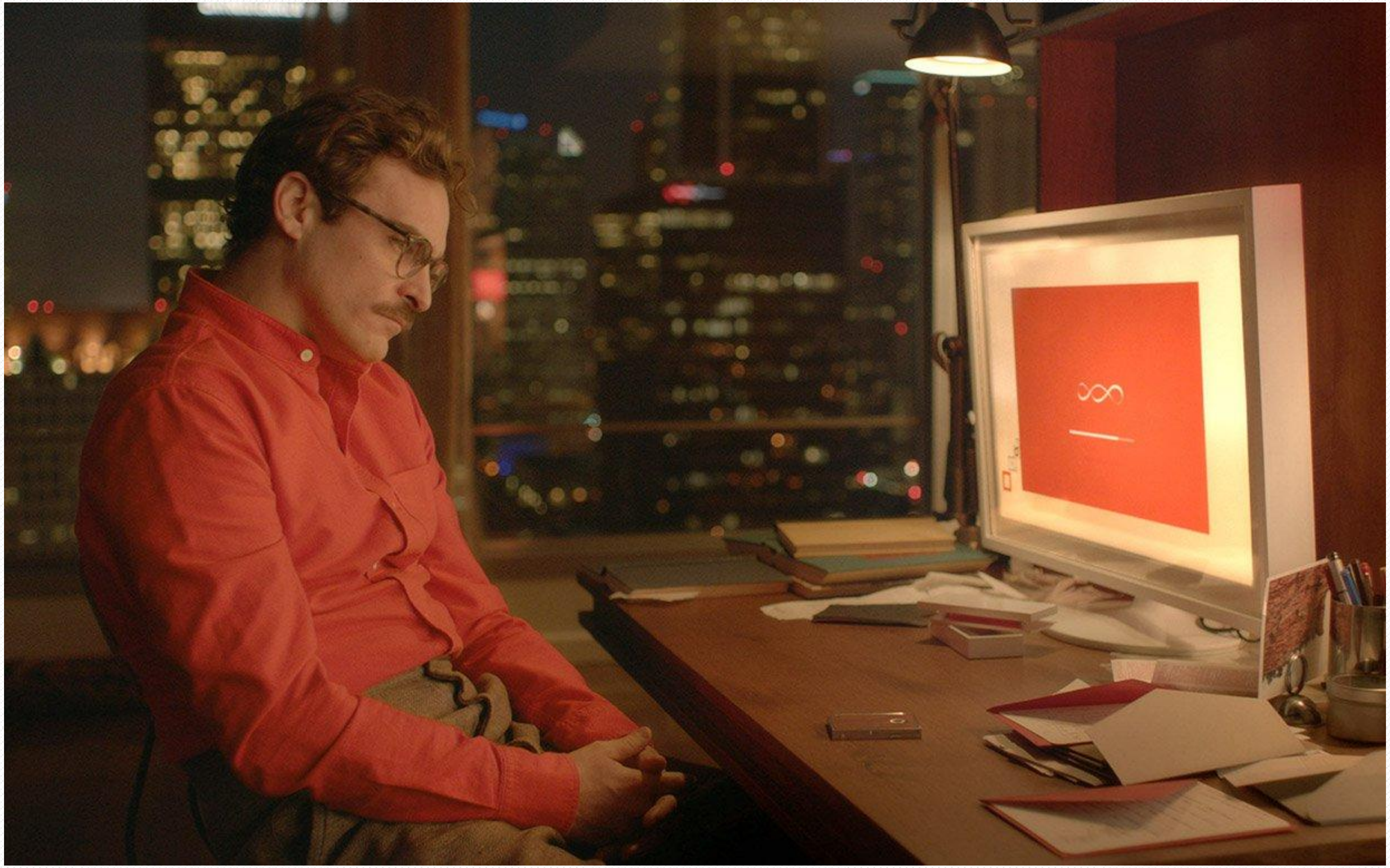
	Best Answer	Second Answer
shut up . lemme do it , red .	☐	☐
tomorrow .	☐	☐
well of course in my youth i was simply known as goldthwait .	☐	☐
sorry . that wasn 't quite what i was looking for .	☐	☐
mark . what happened to your date ?	☐	☐

Lowe, Serban, Noseworthy, Charlin, Pineau. “On the Evaluation of Dialogue Systems with Next Utterance Classification.” *SIGDIAL*, 2016.

Summary

- End-to-end systems are promising, but we have a long way to go.
- Work on collecting larger, better datasets! This is the most useful for the community!
- Don’t rely on only word-overlap metrics like BLEU! Use human evaluations (for now…)

Thank you!



References

- Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A., ... & Weston, J. (2016). Evaluating prerequisite qualities for learning end-to-end dialog systems. In *ICLR*.
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *ICLR*.
- Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2015). A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.
- Ritter, A., Cherry, C., & Dolan, W. B. (2011). Data-driven response generation in social media. In *EMNLP*.
- Shang, L., Lu, Z., & Li, H. (2015). Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.
- Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., & Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In *NAACL-HLT*.

Other curiosities

- Hard to evaluate when the proposed response has a different length than the ground-truth response

	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

Other curiosities

- Removing stop words from BLEU evaluation actually makes things worse

	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.