Neural Dialog Systems

Or Biran*

orb@cs.columbia.edu

11/20/2019

* elemental cognition

Real world systems

Virtual personal assistants

Chatbots

Commercial assistants

Real world systems

Virtual personal assistants: Alexa, Cortana, Siri, Alisa, Conversica, Google

Chatbots

Commercial assistants

Real world systems

Virtual personal assistants: Alexa, Cortana, Siri, Alisa, Conversica, Google

Chatbots: XiaoIce (小冰) – aka Rinna, ~~Tay,~~ Zo

Commercial assistants

Real world systems

Virtual personal assistants: Alexa, Cortana, Siri, Alisa, Conversica, Google

Chatbots: XiaoIce (小冰) – aka Rinna, ~~Tay,~~ Zo

Commercial assistants: WeChat (e-commerce), WhatsApp (flights), websites

Types of dialog systems

	Chatbots	Task-driven
Retrieval		
Generation		

Types of dialog systems

	Chatbots	Task-driven
Retrieval		
Generation		

Types of dialog systems

	Chatbots	Task-driven
Retrieval		
Generation		

Types of dialog systems

	Chatbots	Task-driven
Retrieval		
Generation		
Hybrid		

Types of dialog systems

	Chatbots	Task-driven	Collaborative learning
Retrieval			
Generation			
Hybrid			

Types of dialog systems

Component architecture

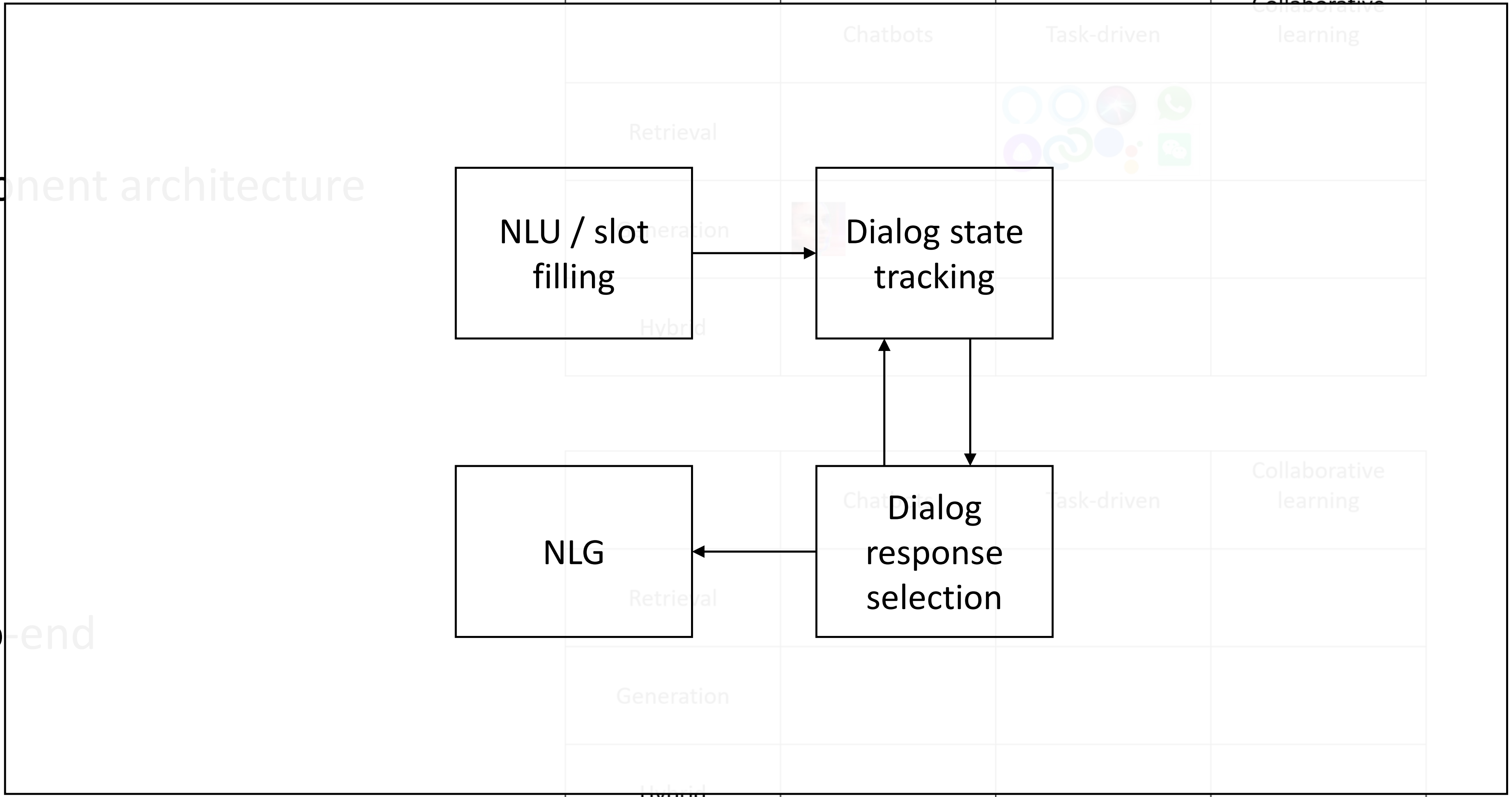
	Chatbots	Task-driven	Collaborative learning
Retrieval			
Generation			
Hybrid			

End-to-end

	Chatbots	Task-driven	Collaborative learning
Retrieval			
Generation			
Hybrid			

Types of dialog systems

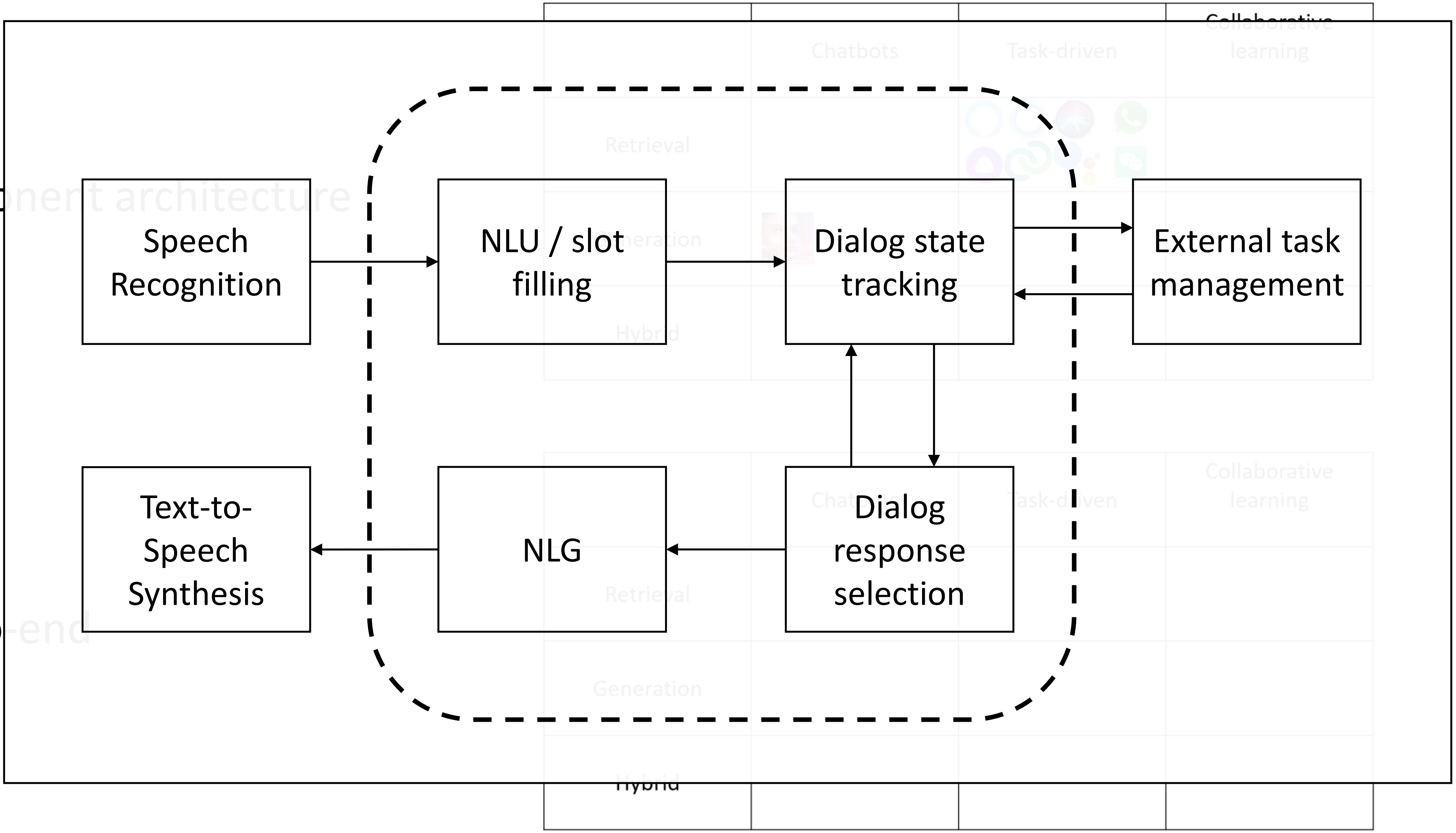
Component architecture



End-to-end

Types of dialog systems

Component architecture



End-to-end

Types of dialog systems

Component architecture

	Chatbots	Task-driven	Collaborative learning
Retrieval			
Generation			
Hybrid			

End-to-end

	Chatbots	Task-driven	Collaborative learning
Retrieval			
Generation			
Hybrid			

Dialog challenges

Multi-turn generation

Dialog challenges

Multi-turn generation

Response diversity

Machine translation

- Semantics fully determined; some lexical diversity

Summarization

- Semantics superset determined

Question answering

- Semantics unknown, but right answer is fully determined

Dialog

- Many right answers!

Dialog challenges

Multi-turn generation

- Response diversity
- Theory of mind
  - Speaker intent

H:	Please book me a flight to Boston tomorrow.
M:	How about one that leaves at 9am?
H:	No, I need to sleep.
M:	How about one that leaves at 8am?

Dialog challenges

Multi-turn generation

- Response diversity
- Theory of mind
  - Speaker intent
  - Emotional state

H:	Please book me a flight to Boston tomorrow.
M:	How about one that leaves at 9am?
H:	No, I need to sleep.
M:	How about one that leaves at 8am?
H:	NO!!!
M:	7am?

Dialog challenges

Multi-turn generation

- Response diversity
- Theory of mind
  - Speaker intent
  - Emotional state
- Pragmatics
  - Speech acts / implicature

H:	Do you know what time the meeting is?
M:	Yes.

H:	I can’t remember the name of the service you recommended.
M:	It’s ok. Humans often don’t remember things.

Dialog challenges

Multi-turn generation

Response diversity

Theory of mind

Speaker intent

Emotional state

Pragmatics

Speech acts / implicature

H:	Do you know what time the meeting is?
M:	Yes.

H:	I can’t remember the name of the service you recommended.
M:	It’s ok. Humans often don’t remember things.

Grice’s maxims

- Quality (say true things)
- Quantity (don’t say too much / too little)
- Relevance (be relevant)
- Manner (express things clearly)

Dialog challenges

Multi-turn generation

- Response diversity
- Theory of mind
  - Speaker intent
  - Emotional state
- Pragmatics
  - Speech acts / implicature
  - Prosody

I’m not flying to Boston

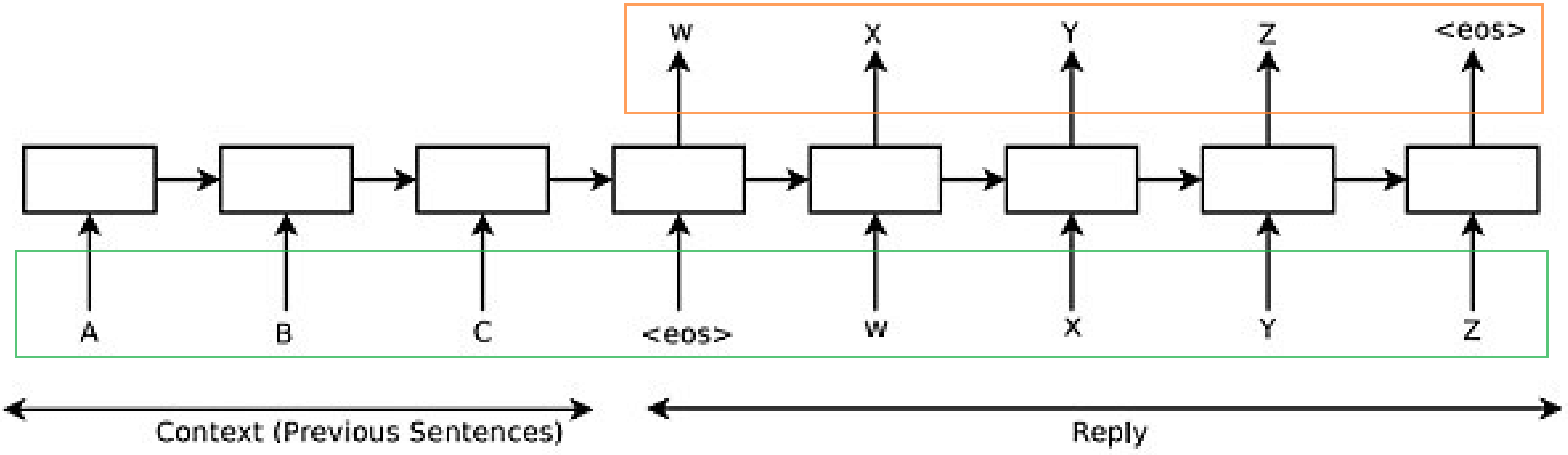
Dialog challenges

Multi-turn generation

- Response diversity
- Theory of mind
  - Speaker intent
  - Emotional state
- Pragmatics
  - Speech acts / implicature
  - Prosody
  - Turn taking
- Personality

End-to-end neural dialog systems

- Basic idea: reuse MT encoder-decoder model
  - RNNs with attention (2015)



End-to-end neural dialog systems

- Basic idea: reuse MT encoder-decoder model
  - RNNs with attention (2015)
- 3 papers within 2 months!

	Shang et al., 2015	Sordoni et al., 2015	Vinyals and Le, 2015
Input (concatenated)	Last message	Last two messages	Entire conversation
Language	Chinese	English	English
Domain(s)	Weibo (microblogs)	Twitter	Ubuntu IT helpdesk + OpenSubtitles (movie conversations)
Extras		Crafted features extracted from input (BOW, length diff, lexical matches…)	

End-to-end neural dialog systems

- Basic idea: reuse MT encoder-decoder model
  - RNNs with attention (2015)
- 3 papers within 2 months!
- Conclusions:
  - Outputs usually plausible in context
  - Generalizes to unseen questions and topics

H:	hello!
M:	hello!
H:	how are you?
M:	I’m good.
H:	what’s your name?
M:	I’m Julia .
H:	when were you born?
M:	July 20th.
H:	what year were you born?
M:	1977.
H:	where are you from?
M:	I’m out in the boonies.
H:	are you a man or a woman?
M:	I’m a woman.
H:	why are we here?
M:	I’m not sure.
H:	okay, bye!
M:	bye.
H:	see you later.
M:	bye.

H: who is skywalker?	H: is sky blue or black?
M: he is a hero .	M: blue.
H: who is bill clinton ?	H: does a cat have a tail?
M: he’s a billionaire.	M: yes.

Examples from Vinyals and Le, 2015

End-to-end neural dialog systems

- Basic idea: reuse MT encoder-decoder model
  - RNNs with attention (2015)
- 3 papers within 2 months!
- Conclusions:
  - Outputs usually plausible in context
  - Generalizes to unseen questions and topics
  - Tend to be generic and repetitive ("safe response")
  - No consistent personality
  - Long outputs can go off-topic or become incoherent

I:	There are always 8 players at the Italian restricted area. Unbelievable!
O:	Unbelievable!

Shang et al., 2015

H:	what is your job?
M:	I'm a lawyer.
H:	what do you do?
M:	I'm a doctor.

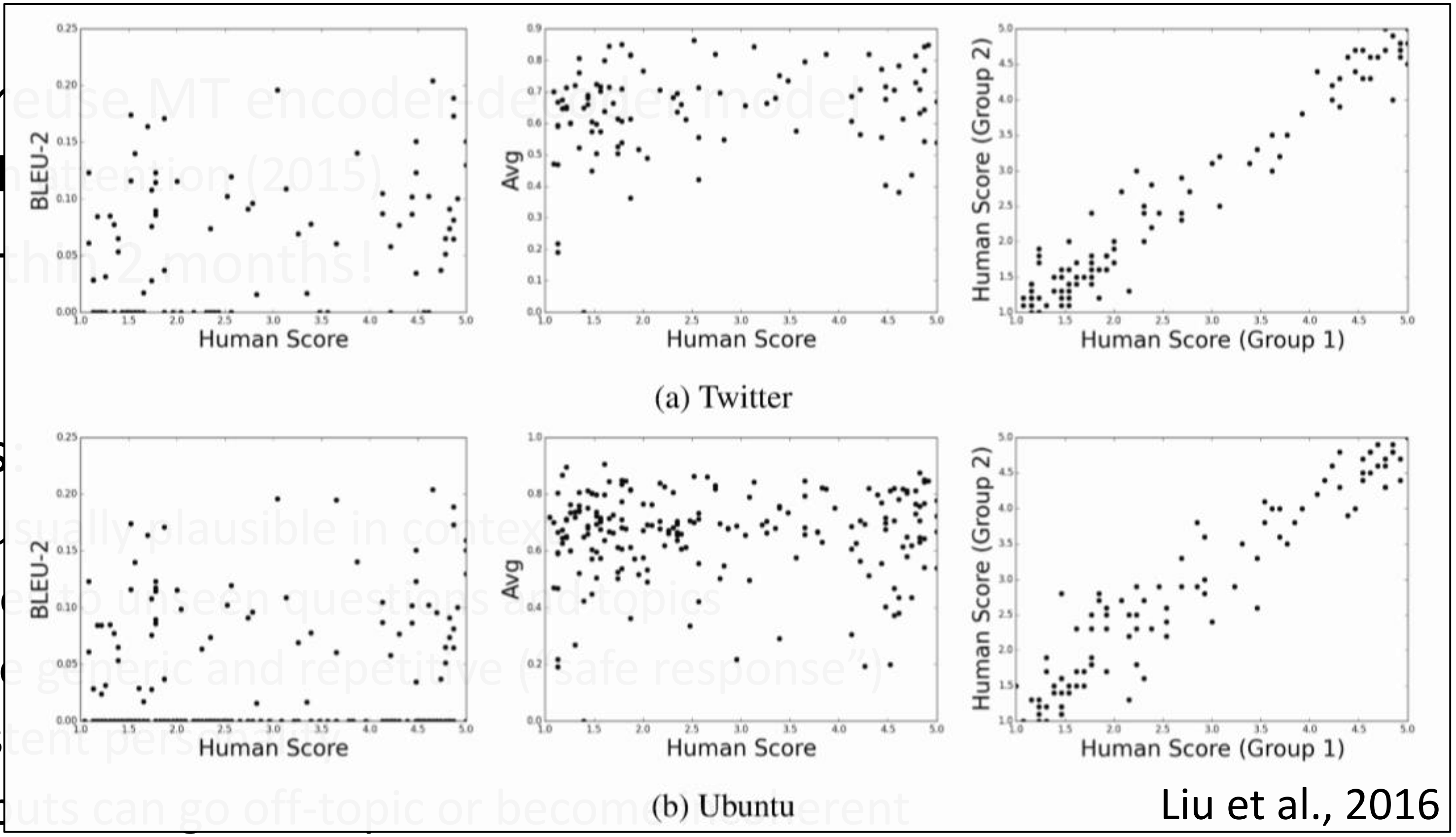
Vinyals and Le., 2015

T-2:	today i will walk for miles. clear sky clear mind.
T-1:	love walking. it's too hot to walk this month. i haven't moved. frozen. heat freeze.
T:	it's too hot for you to get up to honey if you have frozen yogurt to you.

Sordoni et al., 2015

End-to-end neural dialog systems

- Basic idea: reuse MT encoder-decoder model
  - RNNs with attention (2015)
- 3 papers within 2 months!



- Conclusions:
  - Outputs usually plausible in context
  - Generalize to unseen questions and topics
  - Tend to be generic and repetitive ("safe response")
  - No consistent personality
  - Long outputs can go off-topic or become incoherent
  - BLEU not useful: no correlation with human scores

Global coherence, meaningful content and "safe responses"

Two problems with seq2seq models

- Have trouble with long memory
- Tend to favor generic, non-informative output

Global coherence, meaningful content and “safe responses”

Two problems with seq2seq models

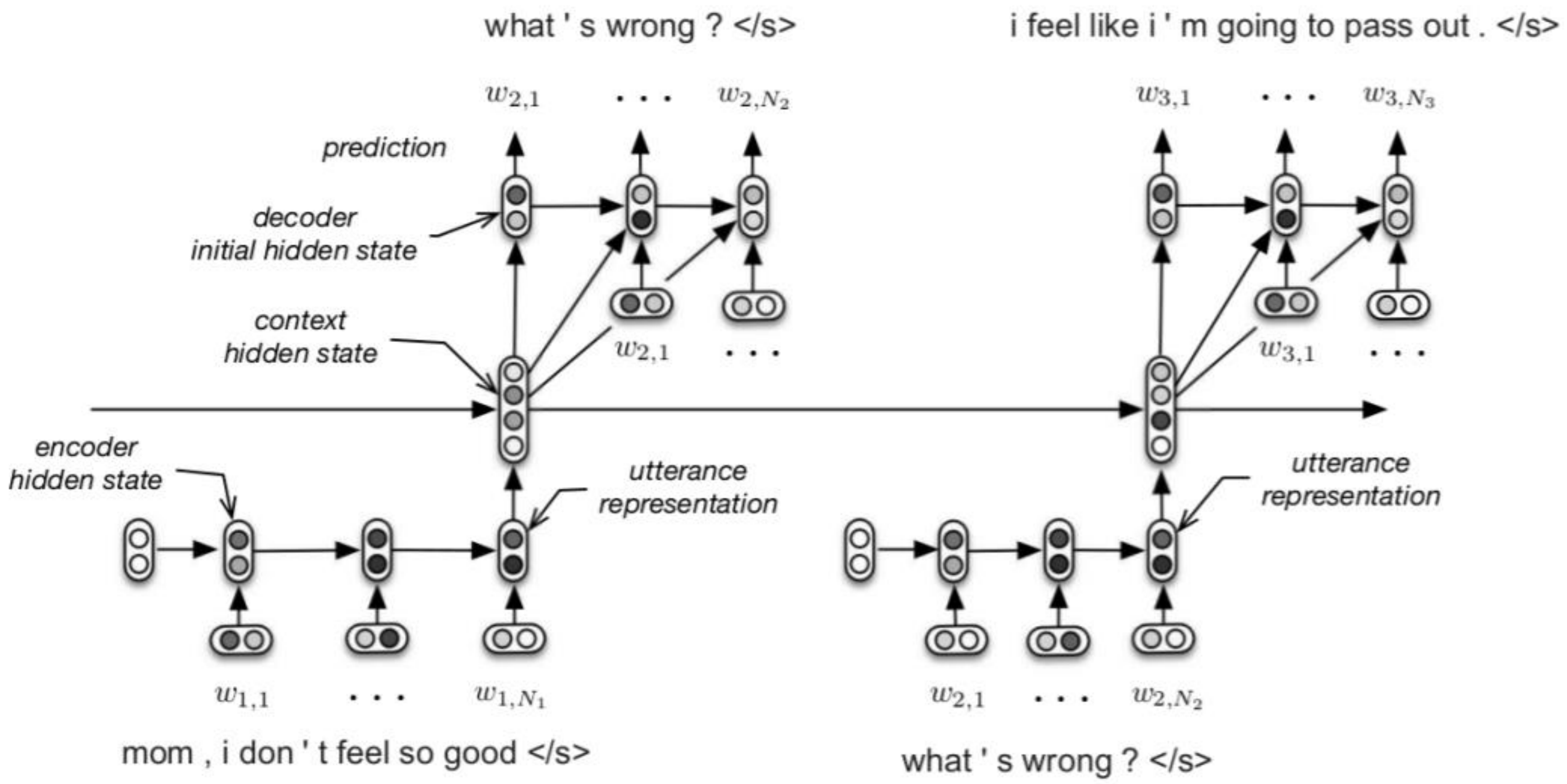
- Have trouble with long memory
- Tend to favor generic, non-informative output

Three general approaches / lines of research:

- Add features
  - Personality embeddings (Li et al., 2016b); Topics (Xing et al., 2016); Situations (Sato et al., 2017)
- Improve model architecture
  - Hierarchical encoders (Serban et al., 2016)
  - Memory networks (Bordes et al, 2016)
  - Variational AutoEncoders (Serban et al., 2017; Zhao et al., 2017; Shen et al., 2018; Park et al., 2018)
- Improve training
  - Diversity-promoting objective function (Li et al., 2015)
  - Reinforcement Learning with heuristic rewards (Li et al., 2016a); Adversarial Learning (Li et al., 2017; Xu et al., 2017; Liu and Lane, 2018)

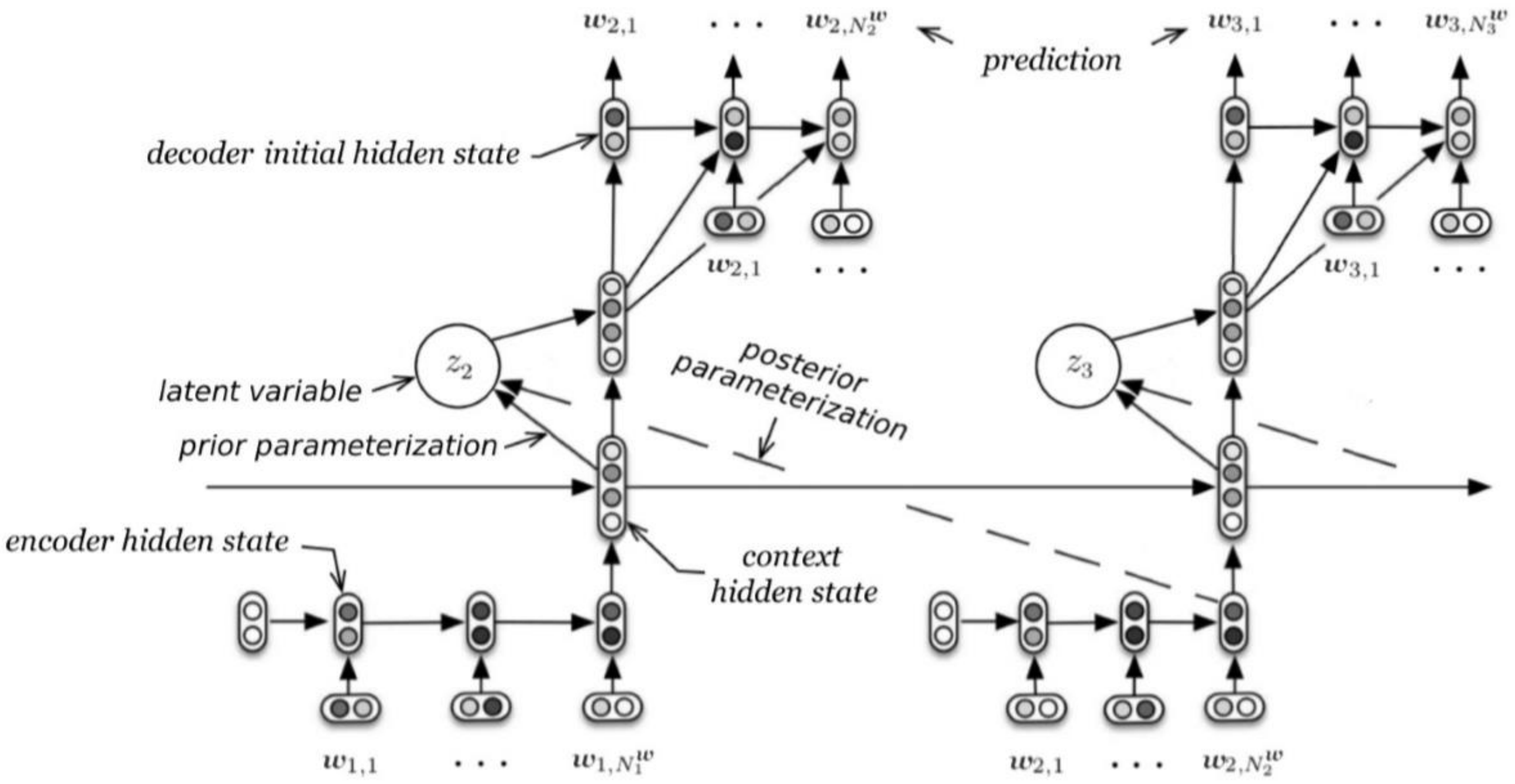
Hierarchical encoders (Serban et al., 2016)

Utterance-level hidden state



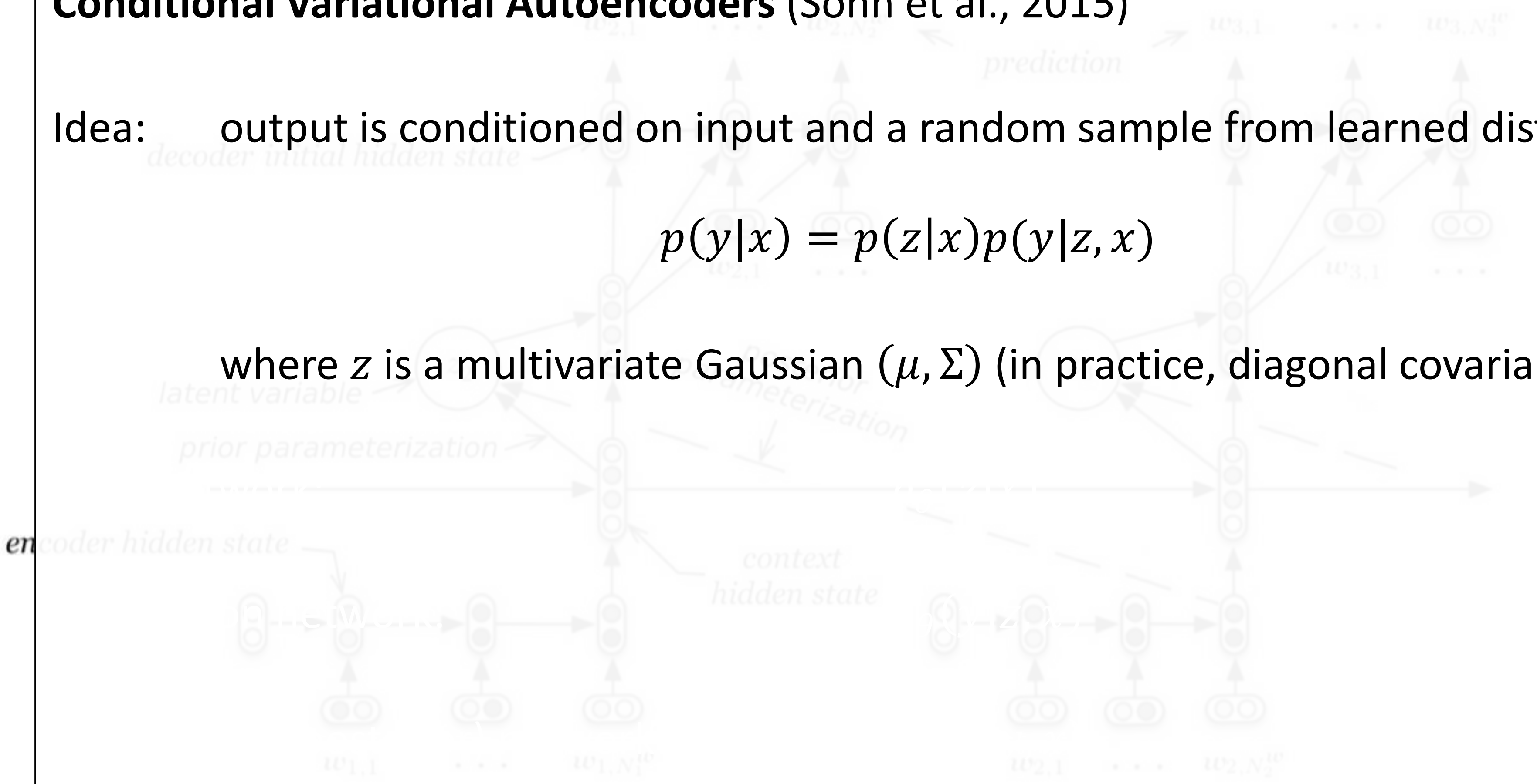
Variational hierarchical encoders (Serban et al., 2017)

Add variational latent variable



Variational hierarchical encoders (Serban et al., 2017)

Add variational latent variable



Conditional Variational Autoencoders (Sohn et al., 2015)

Idea: output is conditioned on input and a random sample from learned distribution

$$p(y|x) = p(z|x)p(y|z, x)$$

where z is a multivariate Gaussian (μ, Σ) (in practice, diagonal covariance)

Variational hierarchical encoders (Serban et al., 2017)

Add variational latent variable

Conditional Variational Autoencoders (Sohn et al., 2015)

Idea: output is conditioned on input and a random sample from learned distribution

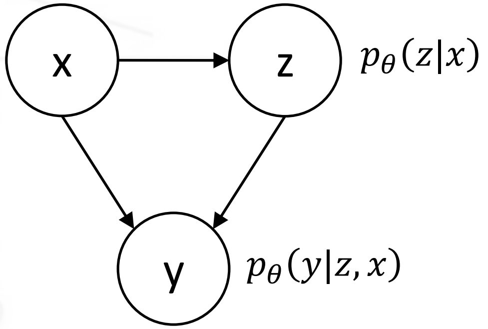
$$p(y|x) = p(z|x)p(y|z,x)$$

where z is a multivariate Gaussian (μ, Σ) (in practice, diagonal covariance)

Prior network: $p_\theta(z|x)$

Generation network: $p_\theta(y|z,x)$

Recognition (posterior) network: $q_\phi(z|y,x)$



Variational hierarchical encoders (Serban et al., 2017)

Add variational latent variable

Conditional Variational Autoencoders (Sohn et al., 2015)

Can be trained with stochastic gradient variational Bayes (Kingma and Welling, 2013)

Use the variational lower bound of the LL as objective:

$$\log p_\theta(y|x) \geq -KL\big(q_\phi(z|x,y) \| p_\theta(z|x)\big) + \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta\,(y|x,z)]$$

Variational hierarchical encoders (Serban et al., 2017)

Add variational latent variable

Conditional Variational Autoencoders (Sohn et al., 2015)

Can be trained with stochastic gradient variational Bayes (Kingma and Welling, 2013)

Use the variational lower bound of the LL as objective:

$$\log p_\theta(y|x) \geq -KL\big(q_\phi(z|x,y) \| p_\theta(z|x)\big) + \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta\,(y|x,z)]$$

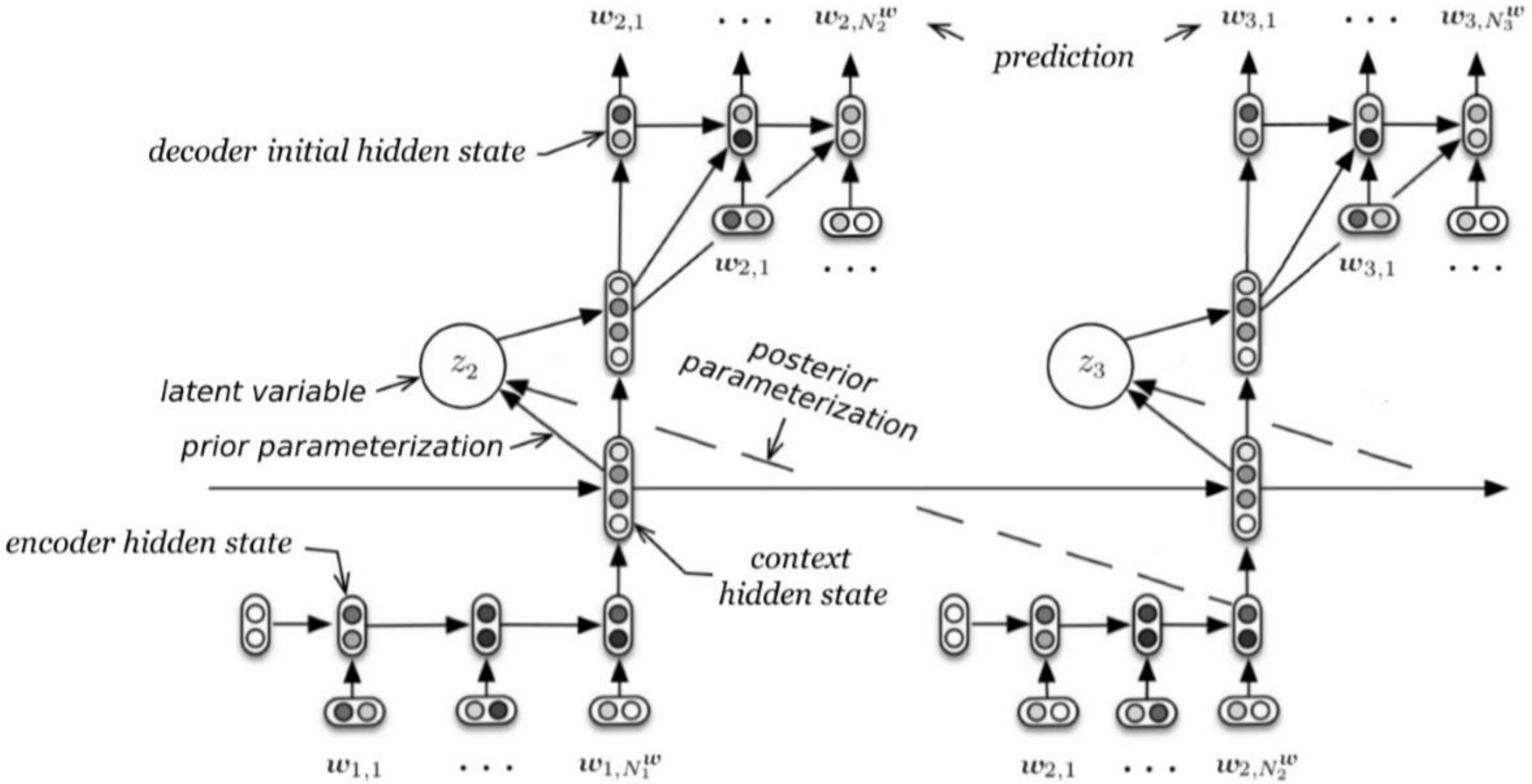
Actual objective (reparameterization trick):

$$\tilde{\mathcal{L}}_{CVAE}(x,y;\theta,\phi) = -KL\big(q_\phi(z|x,y) \| p_\theta(z|x)\big) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta\big(y|x,z^{(l)}\big)$$

$$z^{(l)} = g_\phi\big(x,y,\epsilon^l\big) = \mu + \sigma \odot \epsilon^l$$

Variational hierarchical encoders (Serban et al., 2017)

Add variational latent variable



Variational hierarchical encoders (Serban et al., 2017)

Add variational latent variable

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!!!!!!
whhhhhhyyyy is it suddenly cold in tallahassee today?? this is florida! how can i dream of the beach when my fingers are icey cold?? → sittin on deck in sun looking at lake travis in austin! enjoying day b4 birthday	**VHRED:** the sun is shining!!! **LSTM:** i know!!!!!!!!! **HRED:** something!!!!!!!
reaches for [name]! → - 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 -
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
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.
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

Vanishing latent variable problem

With enough training data, the decoder RNN learns to ignore z ☹

Proposed solutions:

- KL annealing (Bowman et al., 2016) – gradually increase weight of KL
- BOW loss (Zhao et al., 2017) – add x_{BOW} term to loss function; latent variable captures global information about target response
- Hierarchical latent variables (Park et al., 2018)
- Learn to generate prior sample with GAN instead of random sample (Gu et al., 2019)

Consistent personality

message	Where do you live now?
response	I live in Los Angeles.
message	In which city do you live now?
response	I live in Madrid.
message	In which country do you live now?
response	England, you?
message	Where were you born?
response	I was born in Canada.
message	Where are you from?
response	England, you?
message	Where did you grow up?
response	I grew up in Texas.
message	How old are you?
response	16 and you?
message	What's your age?
response	18.
message	What is your major?
response	I'm majoring in psychology
message	What did you study in college?
response	English lit.

(Li et al., 2017)

Personality embeddings (Li et al., 2017)

Learn speaker embeddings v_i with a word prediction task for all utterances of speaker i

Derive speaker-addressee embeddings $V_{i,j}$ from v_i and v_j with a learned weighted linear combination

$$V_{i,j} = \tanh\left(W_1 v_i + W_2 v_j\right)$$

$V_{i,j}$ is added as input to each step in the decoder

- Helps with consistency
- Can infer biographic information based on similar personalities
- Adequate speaker-addressee style based on similar personalities

Personality embeddings (Li et al., 2017)

Learn speaker embeddings v_i with a word prediction task for all utterances of speaker i

Derive speaker-addressee embeddings $V_{i,j}$ from v_i and v_j with a learned weighted linear combination

$$V_{i,j} = \tanh(W_1 v_i + W_2 v_j)$$

$V_{i,j}$ is added as input to each step in the decoder

- Helps with consistency
- Can infer biographical information based on similar personalities
- Adequate speaker-addressee style based on similar personalities

message	Where is your hometown?
response	I'm from England.
message	Where are you from?
response	I'm from England.
message	In which city do you live now?
response	I live in London.
message	In which country do you live now?
response	I live in England.
message	What is your major?
response	Business. You?
message	What did you study in college?
response	I did business studies.
message	How old are you?
response	I'm 18.
message	What is your age?
response	I'm 18.

Personality embeddings (Li et al., 2017)

Learn speaker embeddings v_i with a word prediction task for all utterances of speaker i

Derive speaker-addressee embeddings $V_{i,j}$ from v_i and v_j with a learned weighted linear combination

$$V_{i,j} = \tanh(W_1 v_i + W_2 v_j)$$

$V_{i,j}$ is added as input to each step in the decoder

- Helps with consistency
- Can infer biographic information based on similar personalities
- Adequate speaker-addressee style based on similar personalities

message	Do you love me?
baseline	Of course, I love you.
	addressee: Emily, speaker: Sheldon
response	Of course, I love you, Emily.
	addressee: Emily, speaker: Penny
response	Of course, I love you, Emily.
	addressee: Leonard, speaker: Penny
response	Of course, I love you, Leonard.
	addressee: Monica, speaker: Chandler
response	Of course I love you. I want to marry you.
	addressee: Rachel, speaker: Chandler
response	Of course I love you.
	addressee: Ross, speaker: Chandler
response	Of course I love you.
	addressee: Ross, speaker: Rachel
response	Of course I love you. (kisses him)
	addressee: Emily, speaker: Rachel
response	Of course I love you.

Building dialog data sets (personality)

Personality embeddings work for existing users, but we want to create personas from scratch, using language

Mazare et al., 2018

- Mine persona-context-response triples from Reddit
- Persona is set sentences generally describing user

Persona: [“I like sport”, “I work a lot”]

Context: “I love running.”

Response: “Me too! But only on weekends.”

Building dialog data sets (personality)

Personality embeddings work for existing users, but we want to create personas from scratch, using language

Mazare et al., 2018

- Mine persona-context-response triples from Reddit
- Persona is set sentences generally describing user

Zhang et al., 2018

- Turkers create personas with few sentences
- (Other) Turkers assigned personas randomly, get paired up and chat

Building dialog data sets (personality)

Personality embeddings work for existing users, but we want to create personas from scratch, using language

Mazare et al., 2018

- Mine persona-context-response triples from Reddit
- Persona is set of sentences generally describing user

Zhang et al., 2018

- Turkers create personas
- (Other) Turkers assigned personas randomly, get paired up and chat

Persona 1	**Persona 2**
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

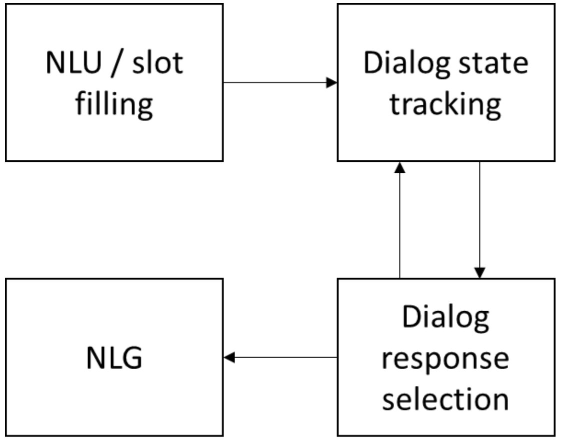
[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

End-to-end structured dialog

Wizard of Oz setting: a human pretending to be a dialog system

MultiWOZ (Budzianowski et al., 2018)

- Full length dialogs in seven task-driven domains
- Annotated with DB entries, belief state and dialog acts
- Allows large scale training of individual components



End-to-end structured dialog

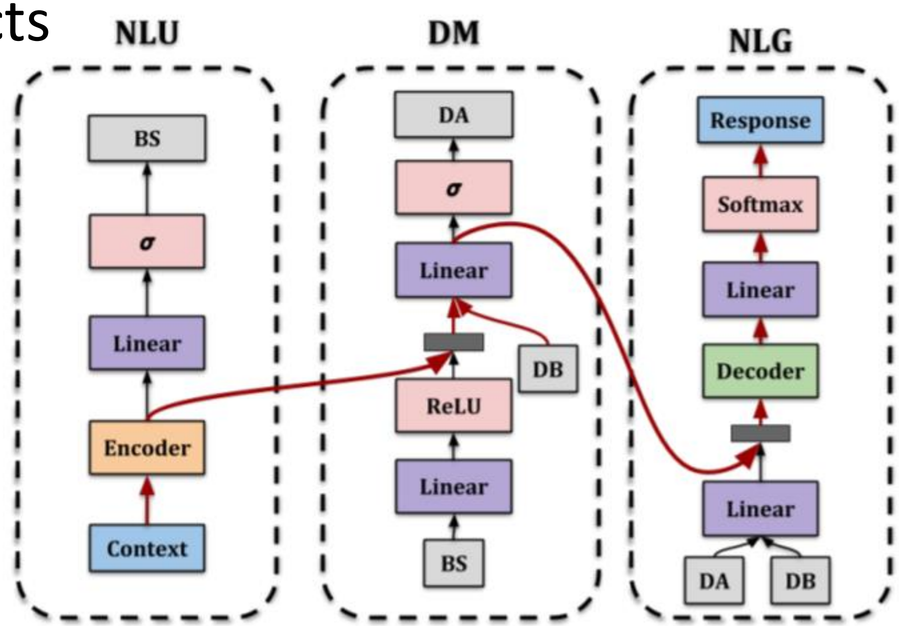
Wizard of Oz setting: a human pretending to be a dialog system

MultiWOZ (Budzianowski et al., 2018)

- Full length dialogs in seven task-driven domains
- Annotated with DB entries, belief state and dialog acts
- Allows large scale training of individual components

Structured fusion networks (Mehri et al., 2019)

- Multitask training of individual components
- End-to-end network uses pre-trained components



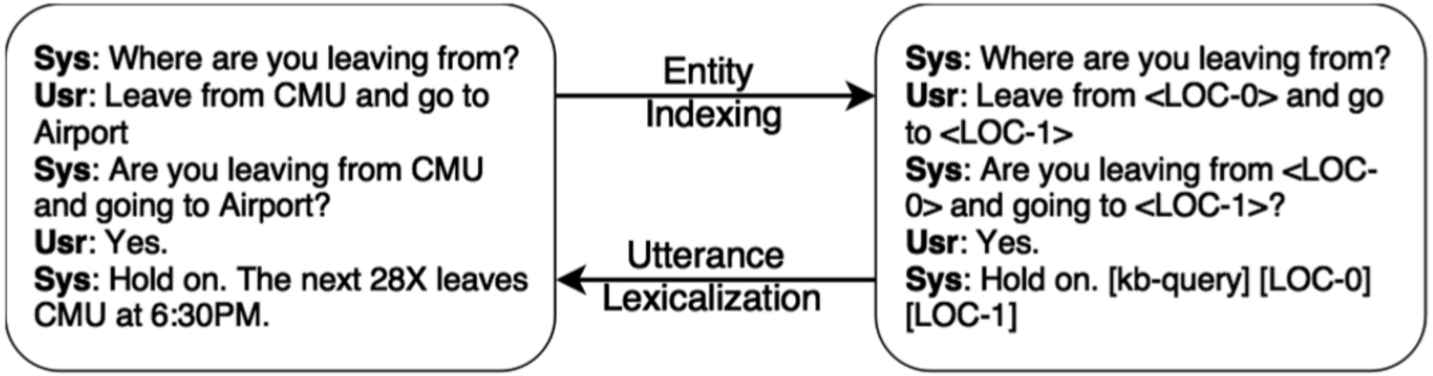
Handling OOV entities

Seq2seq models rely on a fixed vocabulary learned from the training set. Test sets typically have a similar vocabulary

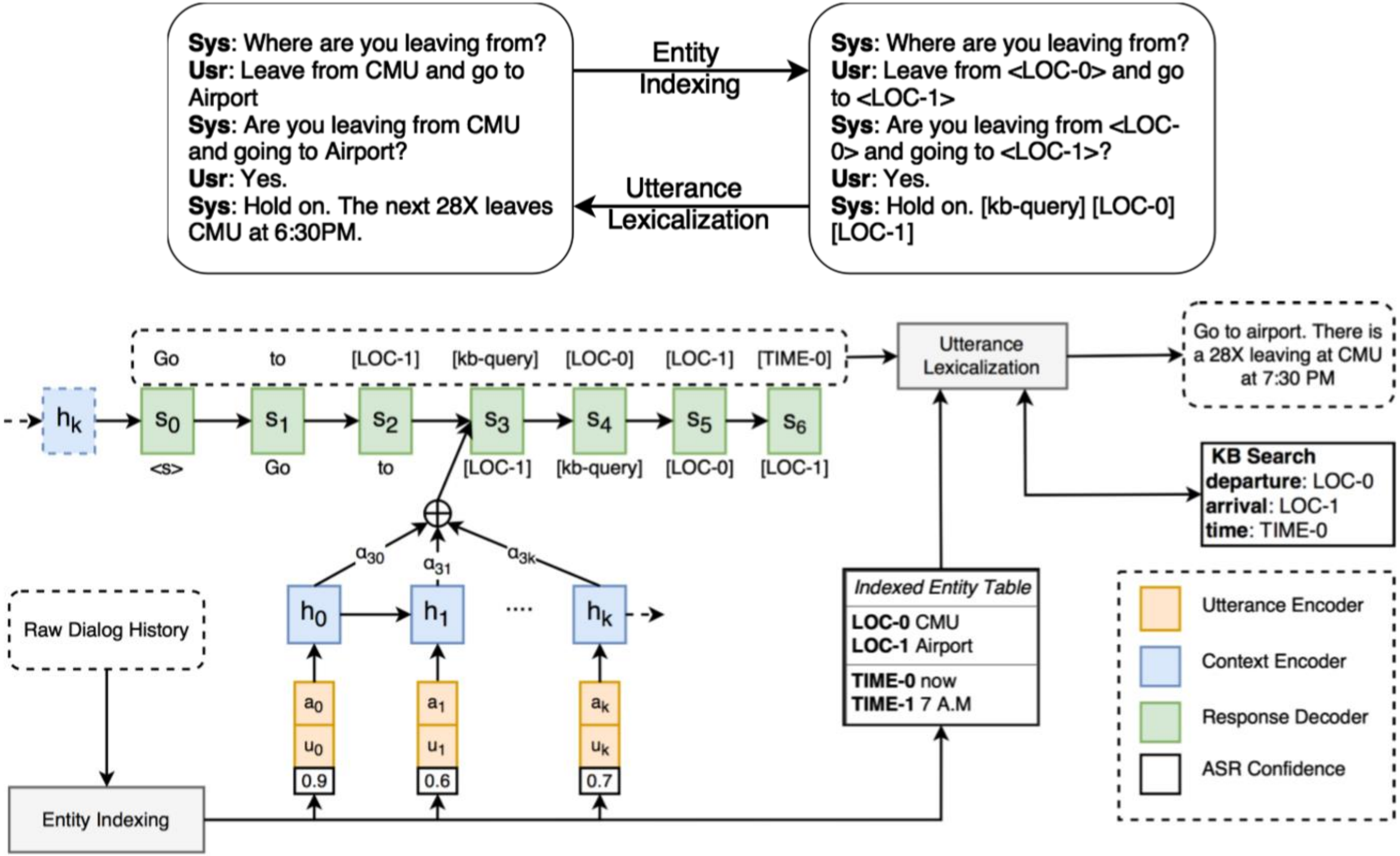
In the real world, new entities come up all the time!

In task-oriented systems, this can be disqualifying

Handling OOV entities with templatization (Zhao et al., 2017)



Handling OOV entities with templatization (Zhao et al., 2017)



Handling OOV with copy-augmented models (Eric&Manning, 2017)

Copy mechanism: add the input tokens as possible outputs in the final softmax, with probability derived from their attention scores

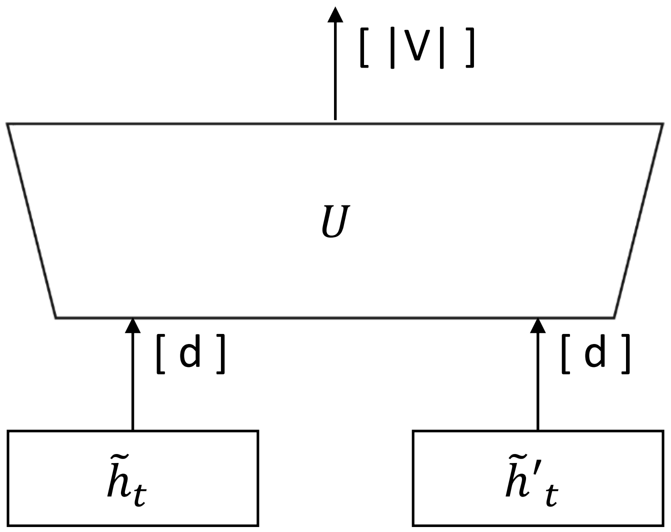
$$u_i^t = v^T \tanh\left(W_1 h_i + W_2 \tilde{h}_t\right)$$

$$a_i^t = \mathrm{softmax}\left(u_i^t\right)$$

$$\tilde{h}'_t = \sum_{i=1}^{m} a_i^t h_i$$

$$o_t = U\left[\tilde{h}_t, \tilde{h}'_t\right]$$

$$y_t = softmax(o_t)$$



Handling OOV with copy-augmented models (Eric&Manning, 2017)

Copy mechanism: add the input tokens as possible outputs in the final softmax, with probability derived from their attention scores

$$u_i^t = v^T \tanh\left(W_1 h_i + W_2 \tilde{h}_t\right)$$

$$a_i^t = \mathrm{softmax}\left(u_i^t\right)$$

$$\tilde{h}'_t = \sum_{i=1}^{m} a_i^t h_i$$

$$o_t = U\left[\tilde{h}_t, \tilde{h}'_t, a_{[1:m]}^t\right]$$

$$y_t = softmax(o_t)$$