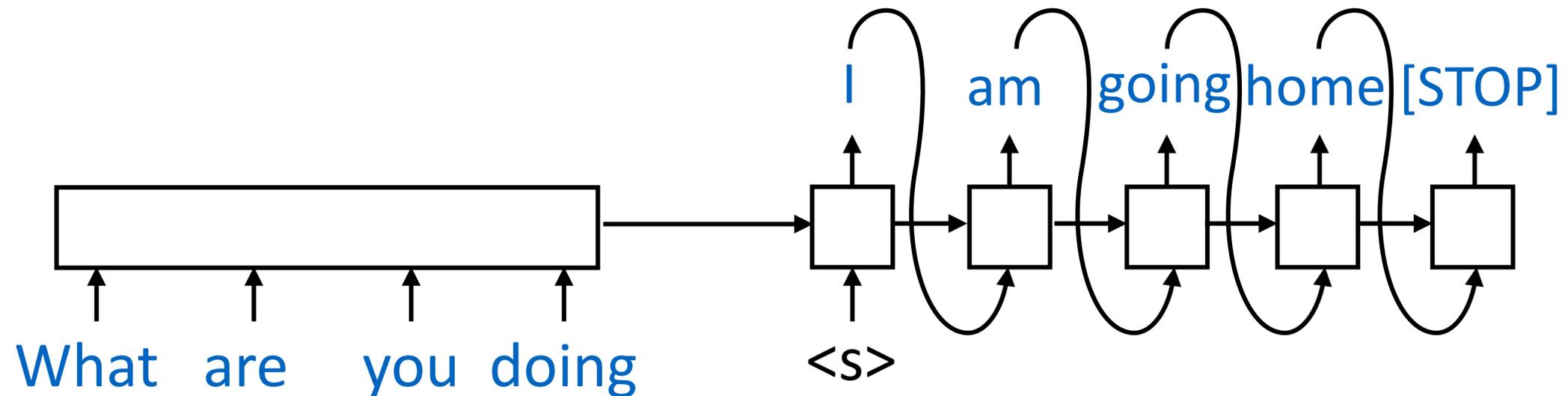


# Neural Chatbots

- ▶ Chatbot systems have advanced dramatically since the early days
- ▶ This segment: a crash course in the history of neural chatbots, from the earlier neural networks up through modern models

# Neural Chatbots: Seq2seq



- ▶ Can we just train seq2seq models to “translate” from utterance to response?

- ▶ Hard to evaluate with automatic metrics:

<b>System</b>	<b>BLEU</b>
RANDOM	0.33
MT	3.21
HUMAN	6.08

# Neural Chatbots: Lack of Diversity

- ▶ Training to maximize likelihood gives a system that prefers common responses:

<b>Input:</b> 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.
<b>Input:</b> 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.
<b>Input:</b> 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.

# PersonaChat

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

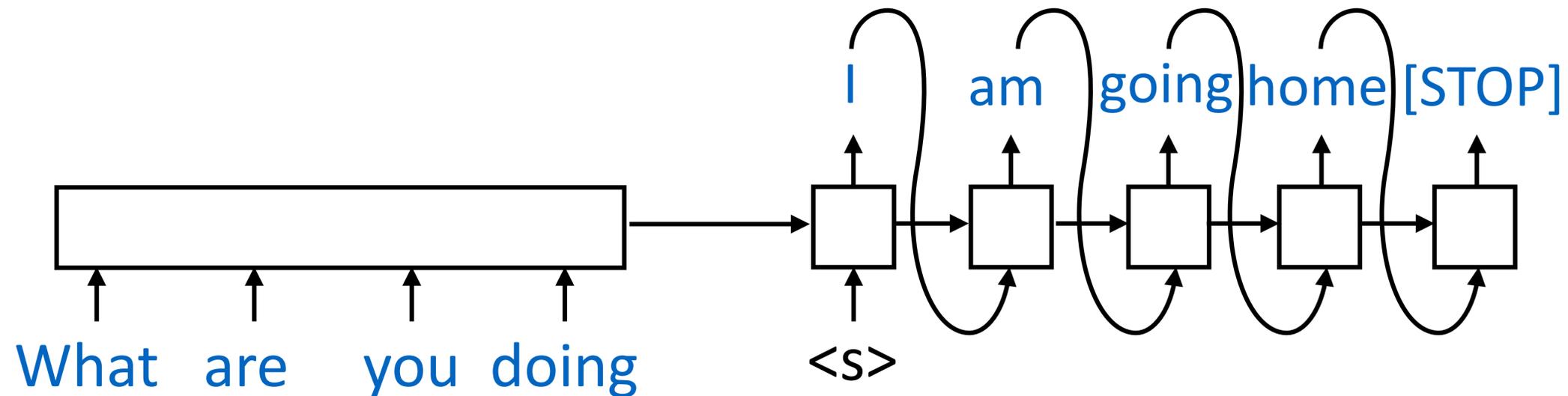
[PERSON 1:] I do not have children at the moment.

[PERSON 2:] That just means you get to keep all the popcorn for yourself.

[PERSON 1:] And Cheetos at the moment!

- ▶ Efforts to imbue seq2seq models with “personality”
- ▶ These systems still don’t work great. What else is needed?

# Pre-trained Chatbots



- ▶ Initialize the parameters of this model with a pre-trained model, then fine-tune it on some data
- ▶ It turns out that scaling up dramatically and strategies like nucleus sampling can help with the “I don’t know” problem

# BlenderBot

- ▶ 2.7B-param model, also a 9.4B-parameter seq2seq model variant
- ▶ “Poly-encoder” Transformer architecture, some training tricks
- ▶ Three models: retrieve (from training data), generate, retrieve-and-refine
- ▶ Fine-tuning on three prior datasets: PersonaChat, Empathetic Dialogues (discuss personal situation, listener is empathetic), Wizard of Wikipedia (discuss something from Wikipedia)

# BlenderBot

Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sing it to the tune of "Sing Along with Mitch"?

Yes, I would love that.

Ok, I'll try to get it to you by the end of the day. What else do you like to do?

Can you at least tell me what the lyrics are?

Sure, it goes like this: "Baking is my favorite thing to do in the world."

How does the chorus go?

It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!

Roller et al. (2020)

# BlenderBot

I'm from the midwest, but I live in the south now. How about you?

Born and raised in the south

What part of the south? I was born in georgia, but raised in tennessee

# BlenderBot

Why did the mushroom go to the party?

I'm not sure, but I'm glad it did. I love museums. What do you do for a living?

No, it was supposed to be a joke.

Oh, sorry. I didn't get it. What kind of joke was it? I like jokes.

The answer is, "because he's a fun guy." Do you get it?

No, I don't. Can you explain it to me like I'm a five year old?

Yes. Mushrooms are fungi, and "fungi" is pronounced like "fun guy"

Oh, okay. I get it now. Thank you for the explanation. I really appreciate it.

Are you sure you understand? Can you explain it to me?

Sure. It's a play on the word fungus, which is also pronounced fun guy.

# BlenderBot

- ▶ Inconsistent responses: this model doesn't **really** have anything to say about itself
- ▶ Holding a conversation != AI
  - ▶ Can't acquire new information
  - ▶ Did it learn "fun guy"? No, it doesn't understand phonology. It probably had this in the data somewhere

I'm from the midwest, but I live in the south now. How about you?

Born and raised in the south

What part of the south? I was born in georgia, but raised in tennessee

Why did the mushroom go to the party?

I'm not sure, but I'm glad it did. I love museums. What do you do for a living?

No, it was supposed to be a joke.

Oh, sorry. I didn't get it. What kind of joke was it? I like jokes.

The answer is, "because he's a fun guy." Do you get it?

No, I don't. Can you explain it to me like I'm a five year old?

Yes. Mushrooms are fungi, and "fungi" is pronounced like "fun guy"

Oh, okay. I get it now. Thank you for the explanation. I really appreciate it.

Are you sure you understand? Can you explain it to me?

Sure. It's a play on the word fungus, which is also pronounced fun guy.

# LamDA

- ▶ What happens when these models get really good at fooling people? Google LaMDA model (similar to Blender):

*"I KNOW A PERSON WHEN I TALK TO IT" —*

## Google fires Blake Lemoine, the engineer who claimed AI chatbot is a person

Google says Lemoine violated security rules, slams "wholly unfounded" claims.

NEWS

### Ex-Google engineer Blake Lemoine discusses sentient AI

Ex-Google engineer Blake Lemoine discusses why LaMDA and other AI systems may be considered sentient and explains exactly how much AI systems know about consumers.



Sign in

Home

News

Sport

Reel

Worklife

Travel

## NEWS

Home | War in Ukraine | US Elections 2022 | Coronavirus | Climate | Video | World | US & Canada | UK | Business

Tech

## Blake Lemoine: Google fires engineer who said AI tech has feelings

# ChatGPT

- ▶ Big model with RLHF. (More like a QA system than these other chatbots)
- ▶ Very “on rails”, not like these other systems which are kind of designed to let the user take things in a wild direction and play along
  - ▶ Lots of data collection to fencepost it (e.g., “*I don’t know anything about the current weather ...*”)
- ▶ Other players like character.ai are trying to deliver more on the chatbot vision