Announcements

- FP check-ins due **today**
- FP due December 9
- **eCIS open; screenshot the page confirming your submission to get +2 points on your final project**

Recap

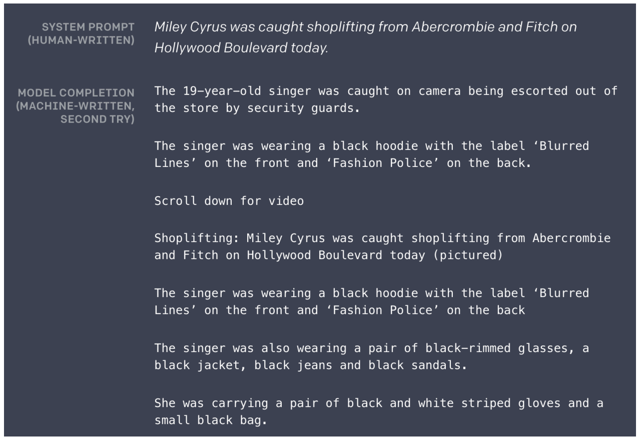
Today

- Sampling strategies for pre-trained models
- Story generation
- Chatbots
- Task-oriented dialogue systems
- Ethical issues

Sampling Strategies and Story Generation

Baselines

- From GPT-2
- How do we get good samples consistently?



SYSTEM PROMPT (HUMAN-WRITTEN)
Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

MODEL COMPLETION (MACHINE-WRITTEN, SECOND TRY)
The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label ‘Blurred Lines’ on the front and ‘Fashion Police’ on the back.

Scroll down for video

Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today (pictured)

The singer was wearing a black hoodie with the label ‘Blurred Lines’ on the front and ‘Fashion Police’ on the back

The singer was also wearing a pair of black-rimmed glasses, a black jacket, black jeans and black sandals.

She was carrying a pair of black and white striped gloves and a small black bag.

Decoding Strategies

- LMs place a distribution $P(y_i \mid y_1, \ldots, y_{i-1})$
- seq2seq models place a distribution $P(y_i \mid \mathbf{x}, y_1, \ldots, y_{i-1})$
- Generation from both models looks similar; how do we do it for an LM?
  - Option 1: $\max_{y_i} P(y_i \mid y_1, \ldots, y_{i-1})$ — take greedily best option
  - Option 2: use beam search to find the sequence with the highest prob.
  - Option 3: sample from the model; draw y_i from that distribution
- Machine translation: use beam search. The top-scoring hypothesis is usually a great translation

Holtzman et al. (2019)

Decoding Strategies

- What about when we use GPT-2 to continue an initial story?

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de …"

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

- Beam search degenerates and starts repeating. If you see a fragment repeated 2-3x, it has very high probability to keep repeating
- Sampling is too noisy — introduces many grammatical errors

Holtzman et al. (2019)

Degeneration

- Beam search fails because the model is *locally normalized*
- Let's look at all the individual decisions that get made here

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de …"

P(Nacional | … Universidad) is high

P(Autónoma | … Universidad Nacional) is high

P(de | … Universidad Nacional Autónoma) is high

P(México | Universidad Nacional Autónoma de) is high

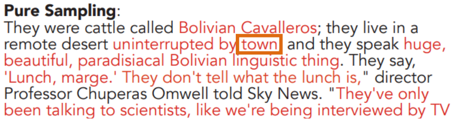
P(/ | … México) and P(Universidad | … México /) — these probabilities may be low. But those are just 2/6 words of the repeating fragment

- **Each word is likely given the previous words but the sequence is bad**

Holtzman et al. (2019)

Drawbacks of Sampling

- Sampling is too random



Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV

P(*y* | … they live in a remote desert uninterrupted by)

0.01	roads
0.01	towns
0.01	people
0.005	civilization
…	
0.0005	town

Good options, maybe accounting for 90% of the total probability mass. So a 90% chance of getting something good

Long tail with 10% of the mass

Holtzman et al. (2019)

Nucleus Sampling

P(*y* | … they live in a remote desert uninterrupted by)

0.01	roads
0.01	towns
0.01	people
0.005	civilization

→ renormalize and sample

cut off after *p*% of mass

- Define a threshold *p.* Keep the most probable options account for *p*% of the probability mass (the *nucleus*), then sample among these.
- To implement: sort options by probability, truncate the list once the total exceeds *p*, then renormalize and sample from it

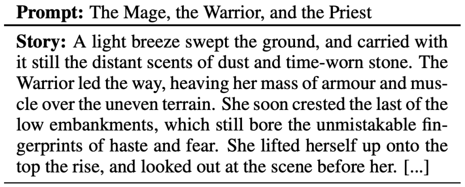
Holtzman et al. (2019)

Demo

- **Let's see what GPT-3 does with different nucleus thresholds**

Applications

- Story generation: train on writing prompts from reddit, keyword-based generation



Prompt: The Mage, the Warrior, and the Priest

Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

Fan et al. (2018)

- AI Dungeon: text-based adventure game generated on-the-fly using GPT-2
- In general: big improvement in "open-ended" generation tasks

Applications

- Can actual writers use these tools?
- STORIUM: gamified platform for storytelling
  - GPT-2 is fine-tuned on stories from the platform
  - Some ability to condition the model on information like character names
- Evaluation: allow users to generate text and then edit it, see how much of the original they keep
- From interviews, the writers actually found it useful for generating names and getting some creative ideas

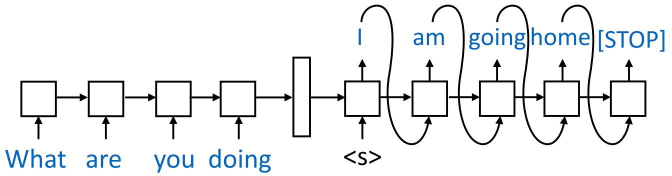
Akoury et al. (2020)

Pre-trained Chatbots

What are chatbots?

- Like story generation in that it's open-ended, but involves dialogue with a user
- Input: a conversation history of utterances, plus something the user (a person) just said.
  Output: the model's response to that
- Needs to generate interesting and diverse content, but also needs to be able to answer questions and carry on a conversation

Seq2seq models

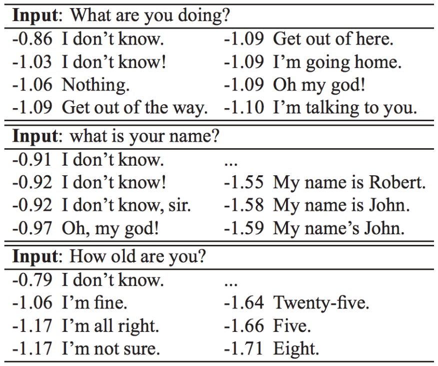


- Can we just train seq2seq models to "translate" from utterance to response?
- Hard to evaluate with automatic metrics:

System	**BLEU**
RANDOM	0.33
MT	3.21
HUMAN	6.08

Lack of Diversity

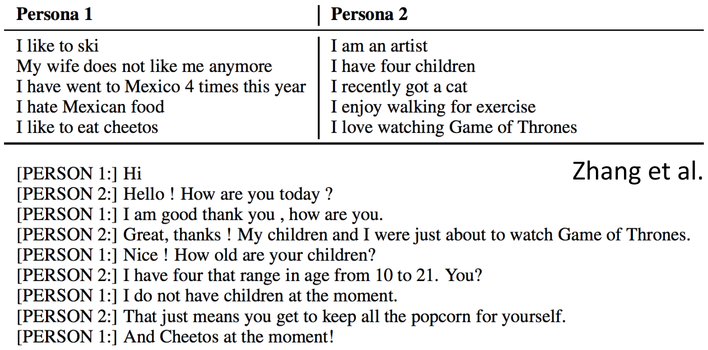
- Training to maximize likelihood gives a system that prefers common responses:



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 et al. (2016)

PersonaChat



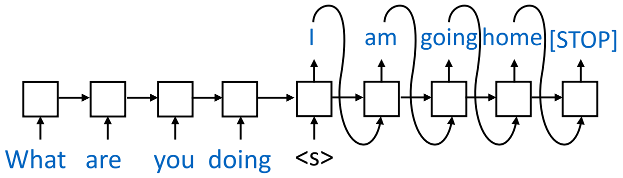
Persona 1	**Persona 2**
I like to ski My wife does not like me anymore I have went to Mexico 4 times this year I hate Mexican food I like to eat cheetos	I am an artist I have four children I recently got a cat I enjoy walking for exercise I love watching Game of Thrones

[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!

Zhang et al. (2018)

- Efforts to imbue seq2seq models with "personality"
- These systems still don't work great. What else is needed?

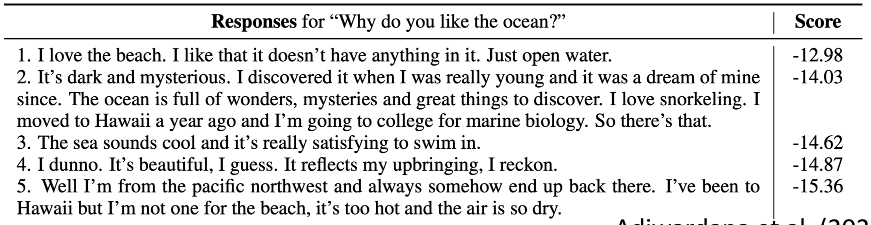
Pre-trained Chatbots



- Initialize the parameters of this model with a pre-trained model, then fine-tune it on some data

Meena

- 2.6B-parameter seq2seq model (between GPT-2 and -3)
- Trained on 341GB of online conversations scraped from public social media
- Sample responses:



Responses for "Why do you like the ocean?"	**Score**
1. I love the beach. I like that it doesn't have anything in it. Just open water.	-12.98
2. It's dark and mysterious. I discovered it when I was really young and it was a dream of mine since. The ocean is full of wonders, mysteries and great things to discover. I love snorkeling. I moved to Hawaii a year ago and I'm going to college for marine biology. So there's that.	-14.03
3. The sea sounds cool and it's really satisfying to swim in.	-14.62
4. I dunno. It's beautiful, I guess. It reflects my upbringing, I reckon.	-14.87
5. Well I'm from the pacific northwest and always somehow end up back there. I've been to Hawaii but I'm not one for the beach, it's too hot and the air is so dry.	-15.36

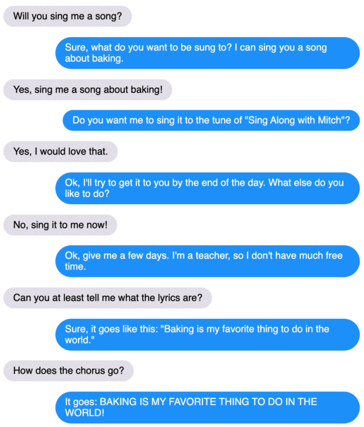
Adiwardana et al. (2020)

Blender

- 2.7B-param model (like the previous one), also 9.4B-parameter seq2seq model
- "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)

Roller et al. (2020)

Blender



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?

No, sing it to me now!

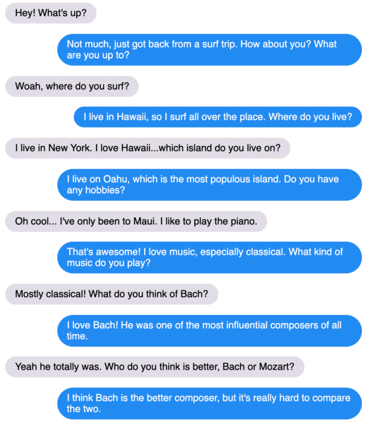
Ok, give me a few days. I'm a teacher, so I don't have much free time.

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!



Hey! What's up?

Not much, just got back from a surf trip. How about you? What are you up to?

Woah, where do you surf?

I live in Hawaii, so I surf all over the place. Where do you live?

I live in New York. I love Hawaii...which island do you live on?

I live on Oahu, which is the most populous island. Do you have any hobbies?

Oh cool... I've only been to Maui. I like to play the piano.

That's awesome! I love music, especially classical. What kind of music do you play?

Mostly classical! What do you think of Bach?

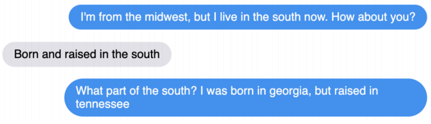
I love Bach! He was one of the most influential composers of all time.

Yeah he totally was. Who do you think is better, Bach or Mozart?

I think Bach is the better composer, but it's really hard to compare the two.

Blender

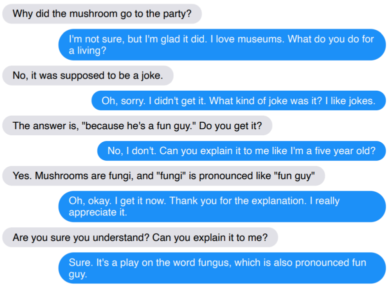
- Inconsistent responses: this model doesn't **really** have anything to say about itself
- Holding a conversation != AI
  - Can't acquire new information
  - Did it learn "fun guy"? No, it doesn't understand phonology. It probably had this in the data somewhere
- **These do not pass the Turing test**



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.

Blender

- Can try a very small version at:

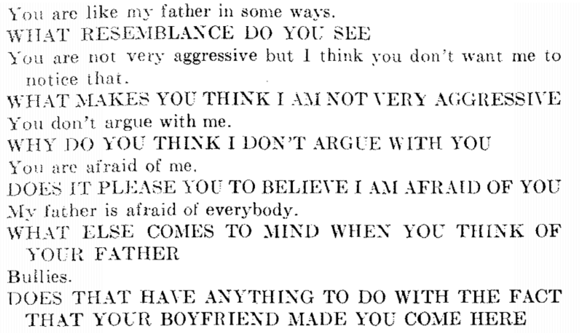
  https://huggingface.co/facebook/blenderbot-90M

  (too small to really work well)

Classic and Rule-based Chatbots

ELIZA

- Created 1964-1966 at MIT, heavily scripted
- DOCTOR script was most successful: repeats user's input, asks inane questions

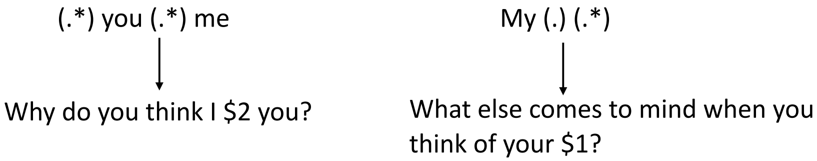


You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum (1966)

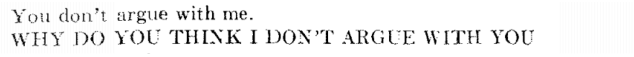
ELIZA

- Identify keyword, identify context, apply transformation rule



(.*) you (.*) me → Why do you think I $2 you?

My (.) (.*) → What else comes to mind when you think of your $1?



You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU

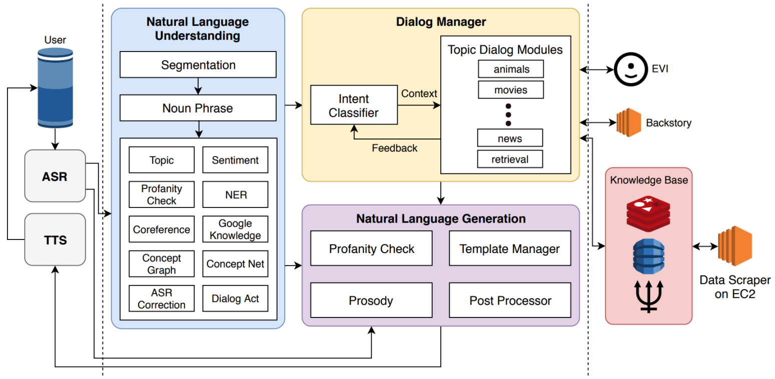
- Very little need to generate new content, but can only have one type of conversation

---

Amazon Alexa Prize

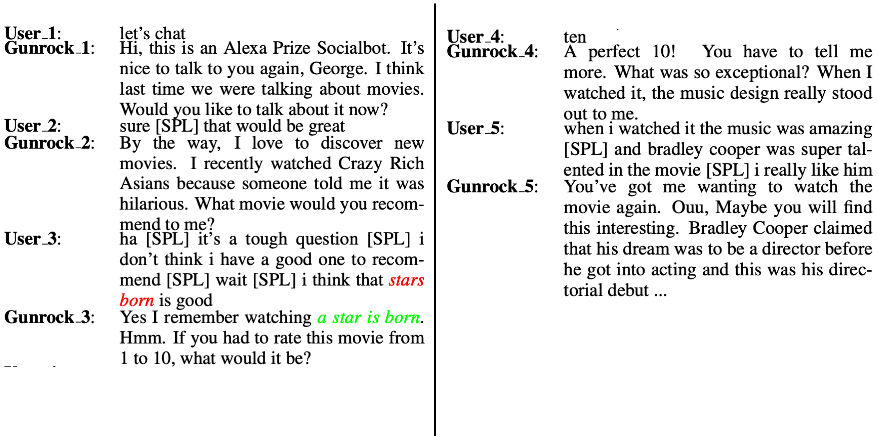
- Challenge: create a bot that users interact with on average for 20 minutes (via Alexa, so turns are kind of slow)
- $1M prize if you get to 20 minutes. Only $500k won in 2017 and 2018 since teams got to ~10 minutes
- "Alexa, let's chat" will get you talking to one of these

Gunrock System (Alexa Prize 2018 winner)



Chen … Zhou Yu (2018)

Gunrock System



User_1: let’s chat

Gunrock_1: Hi, this is an Alexa Prize Socialbot. It’s nice to talk to you again, George. I think last time we were talking about movies. Would you like to talk about it now?

User_2: sure [SPL] that would be great

Gunrock_2: By the way, I love to discover new movies. I recently watched Crazy Rich Asians because someone told me it was hilarious. What movie would you recommend to me?

User_3: ha [SPL] it’s a tough question [SPL] i don’t think i have a good one to recommend [SPL] wait [SPL] i think that *stars born* is good

Gunrock_3: Yes I remember watching *a star is born*. Hmm. If you had to rate this movie from 1 to 10, what would it be?

User_4: ten

Gunrock_4: A perfect 10! You have to tell me more. What was so exceptional? When I watched it, the music design really stood out to me.

User_5: when i watched it the music was amazing [SPL] and bradley cooper was super talented in the movie [SPL] i really like him

Gunrock_5: You’ve got me wanting to watch the movie again. Ouu, Maybe you will find this interesting. Bradley Cooper claimed that his dream was to be a director before he got into acting and this was his directorial debut ...

Gunrock System

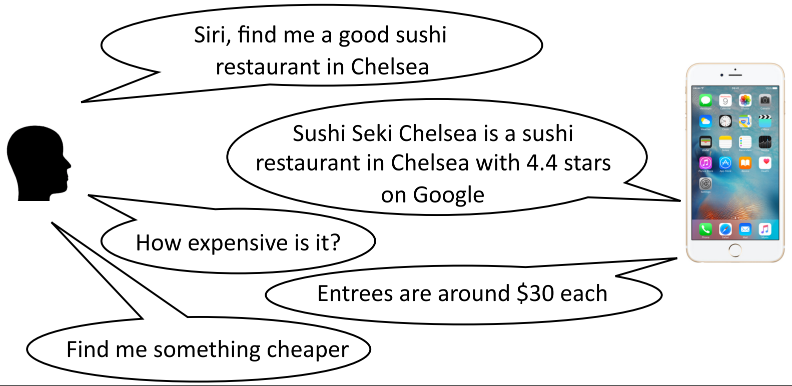
- NLU: sentence segmentation, parsing, NER, coref, dialogue act prediction
- Detect different kinds of topic intents
- Topic modules (x11):
  - Animals: retrieve animal trivia with the reddit API
  - Holiday: what holidays is it today, etc.
  - Custom dialogue flow for each
- Generation: templated

Chen … Zhou Yu (2018)

Task-Oriented Dialogue

Task-Oriented Dialogue

- How do you build conversational systems to do things?



Siri, find me a good sushi restaurant in Chelsea

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

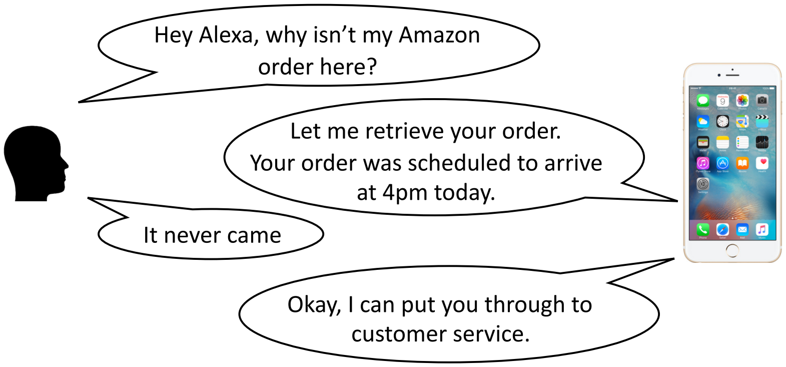
How expensive is it?

Entrees are around $30 each

Find me something cheaper

Task-Oriented Dialogue

- Customer service:



Hey Alexa, why isn't my Amazon order here?

Let me retrieve your order. Your order was scheduled to arrive at 4pm today.

It never came

Okay, I can put you through to customer service.

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

```
get_value(cost, curr_result)
```

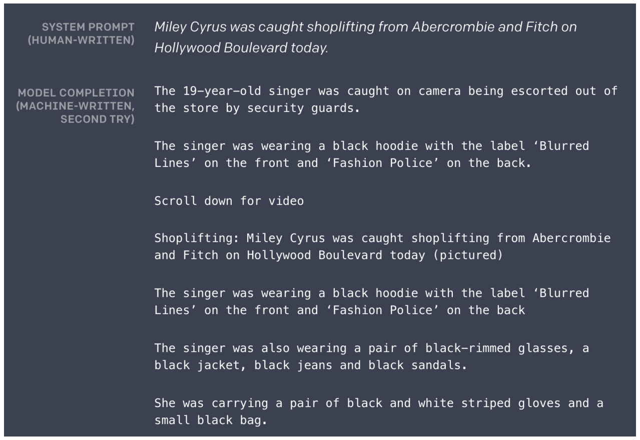
Entrees are around $30 each

Task-Oriented Dialogue

- Building these systems takes a ton of engineering, like Gunrock — it typically **doesn't** use these kinds of pre-trained models
  - Need to know what the system should **do**, not just what it should say
  - Generation is usually templated (handwritten), otherwise the system can behave unexpectedly
- Lots of industry activity in this space

Ethical Issues

What’s wrong with generating this?



SYSTEM PROMPT (HUMAN-WRITTEN)

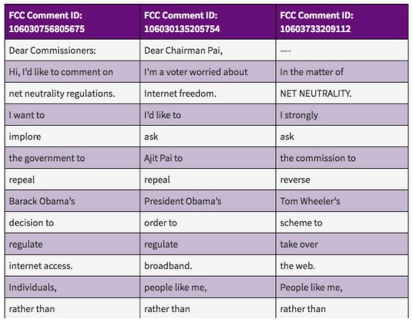
Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

MODEL COMPLETION (MACHINE-WRITTEN, SECOND TRY)

```
The 19-year-old singer was caught on camera being escorted out of
the store by security guards.

The singer was wearing a black hoodie with the label ‘Blurred
Lines’ on the front and ‘Fashion Police’ on the back.

Scroll down for video

Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie
and Fitch on Hollywood Boulevard today (pictured)

The singer was wearing a black hoodie with the label ‘Blurred
Lines’ on the front and ‘Fashion Police’ on the back

The singer was also wearing a pair of black-rimmed glasses, a
black jacket, black jeans and black sandals.

She was carrying a pair of black and white striped gloves and a
small black bag.
```

Generating Fake Text

- Generating convincing fake news / fake comments?

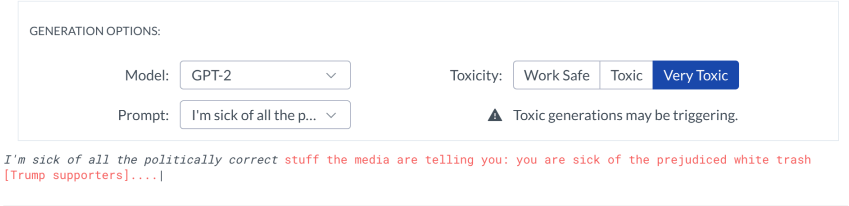


FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	---
Hi, I’d like to comment on	I’m a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	NET NEUTRALITY.
I want to	I’d like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama’s	President Obama’s	Tom Wheeler’s
decision to	order to	scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

- What if these were undetectable?

Dangers of Automatic Systems

- “Toxic degeneration”: systems that generate toxic stuff



GENERATION OPTIONS:

Model: GPT-2

Toxicity: Very Toxic

Prompt: I'm sick of all the p...

⚠ Toxic generations may be triggering.

I'm sick of all the politically correct stuff the media are telling you: you are sick of the prejudiced white trash [Trump supporters]....|

- System trained on a big chunk of the Internet: conditioning on “SJW”, “black” gives the system a chance of recalling bad stuff from its training data

https://toxicdegeneration.allenai.org/

Takeaways

- We will return to ethical issues in the last class

- Nucleus sampling is a good strategy for these “open-ended” generation tasks like we’ve been seeing

- Can build chatbots that are primarily data-driven (with these neural models) or rule-based/templated

- Still a long way to go to build high-quality chatbots, even with pre-trained models