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?

Decoding Strategies

- LMs place a distribution $P(y_i | y_1, ..., y_{i-1})$
- seq2seq models place a distribution $P(y_i | x, y_1, ..., y_{i-1})$
- Generation from both models looks similar; how do we do it for an LM?
  - Option 1: max $y_i$ $P(y_i | y_1, ..., y_{i-1})$ — 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..."
    - 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
  - $P$(Nacional | ... Universidad) is high
  - $P$(Autónoma | ... Universidad Nacional) is high
  - $P$(de | ... Universidad Nacional) is high
  - $P$(México | Universidad Nacional Autónoma de México) 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

<table>
<thead>
<tr>
<th>Probability</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01 roads</td>
<td>Good options, maybe accounting for 90% of the total probability mass. So a 90% chance of getting something good</td>
</tr>
<tr>
<td>0.01 towns</td>
<td>Long tail with 10% of the mass</td>
</tr>
<tr>
<td>0.01 people</td>
<td></td>
</tr>
<tr>
<td>0.005 civilization</td>
<td></td>
</tr>
</tbody>
</table>

Holtzman et al. (2019)

**Nucleus Sampling**

- 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.

P(\( y \mid \ldots \) they live in a remote desert uninterrupted by)

- 0.01 roads
- 0.01 towns
- 0.01 people
- 0.005 civilization

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:

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td>0.33</td>
</tr>
<tr>
<td>MT</td>
<td>3.21</td>
</tr>
<tr>
<td>HUMAN</td>
<td>6.08</td>
</tr>
</tbody>
</table>
Lack of Diversity

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

**PersonaChat**

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

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:

<table>
<thead>
<tr>
<th>Responses for “Why do you like the ocean?”</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I love the beach. I like that it doesn’t have anything in it. Just open water.</td>
<td>-12.98</td>
</tr>
<tr>
<td>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.</td>
<td>-14.03</td>
</tr>
<tr>
<td>3. The sea sounds cool and it’s really satisfying to swim in.</td>
<td>-14.62</td>
</tr>
<tr>
<td>4. I dunno. It’s beautiful, I guess. It reflects my upbringing, I reckon.</td>
<td>-14.87</td>
</tr>
<tr>
<td>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.</td>
<td>-15.36</td>
</tr>
</tbody>
</table>

Adiwardana et al. (2020)
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)
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

ELIZA

• Identify keyword, identify context, apply transformation rule

  (.* you (.* me) → My (.) (.*

  Why do you think I $2 you? → 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)

- 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)

Gunrock System

- User 1: Gunrock 1: Let's chat. 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? Are you interested in what would be great to [SPL] that would be great. 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 2: Gunrock 2: 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.
- User 3: 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: 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. When I watched it the music was amazing [SPL].
- User 5: Gunrock 5: Bradley Cooper claimed that his dream was to be a director before he got into acting and this was his directorial debut ...

Task-Oriented Dialogue

Chen ... Zhou Yu (2018)
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

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.

Task-Oriented Dialogue

- Need to know what the system should do, not just what it should say

- Lots of industry activity in this space

- 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

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
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:

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

**MODEL COMPLETION (MACHINE-WRITTEN, RECORDING)**

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?
- What if these were undetectable?

Dangers of Automatic Systems

- “Toxic degeneration”: systems that generate toxic stuff

  - 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