Recall: BART

- BERT is good for “analysis” tasks, GPT is a good language model
- What to do for seq2seq tasks?
- Sequence-to-sequence BERT variant: permute/make/delete tokens, then predict full sequence autoregressively
- Uses the transformer encoder-decoder we discussed for MT (decoder attends to encoder)

Recall: GPT-3

- Few-shot
  In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.
This Lecture

- Chatbots: intro, rule-based
- Seq2seq chatbots
- Summarization

Chatbots: Intro, Rule-based

Turing Test (1950)

- Imitation game: A and B are locked in rooms and answer C’s questions via typewriter. Both are trying to act like B

Original Interpretation:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

C trained judge

Standard Interpretation:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

C trained judge

- The test is not “does this computer seem human-like to random people with a web browser?”

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?

  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:**

*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? sure [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:* 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:* ten

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 the music was amazing [SPL] and bradley cooper was super talented in the movie [SPL] i really like him

You’ve got me wanting to watch the movie again. Oui, 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)

Seq2seq Chatbots

- Seq2seq models

  Train a seq2seq model on subtitles data

  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>

Seq2seq models

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

<table>
<thead>
<tr>
<th>Input: What are you doing?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.86 I don’t know.</td>
</tr>
<tr>
<td>-1.03 I don’t know!</td>
</tr>
<tr>
<td>-1.06 Nothing.</td>
</tr>
<tr>
<td>-1.09 Get out of the way.</td>
</tr>
<tr>
<td>-1.10 I’m talking to you.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: What is your name?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.91 I don’t know.</td>
</tr>
<tr>
<td>-0.92 I don’t know!</td>
</tr>
<tr>
<td>-0.97 Oh, my god!</td>
</tr>
<tr>
<td>-1.55 My name is Robert.</td>
</tr>
<tr>
<td>-1.58 My name is John.</td>
</tr>
<tr>
<td>-1.59 My name’s John.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: How old are you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.79 I don’t know.</td>
</tr>
<tr>
<td>-1.06 I’m fine.</td>
</tr>
<tr>
<td>-1.17 I’m all right.</td>
</tr>
<tr>
<td>-1.17 I’m not sure.</td>
</tr>
</tbody>
</table>

Lack of Diversity

Li et al. (2016)
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 gone 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

Pre-training in Chatbots

- Initialize the parameters of this model with one of these pre-trained models (GPT-3 trained on the web), then fine-tune it on some data

Meena

- 2.6B-parameter seq2seq model (larger than GPT-2)
- 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)
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)

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 (stochastic parrot!)

Task-Oriented Dialogue

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

- Building these systems takes a ton of engineering — 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
- Tons of work in this space, much of it in industry

Summarization

- What makes a good summary?
A massive earthquake of magnitude 7.3 struck Iraq on Sunday. The epicenter was close to the Iranian border. Eight villages were damaged and six people were killed in Iran.

Summary

What makes a good summary?

- Content selection: pick the right content
  - Right content was repeated within and across documents
- Domain-specific (magnitude + epicenter of earthquakes are important)
- Generation: write the summary
  - Extraction: pick whole sentences from the summary
  - Abstraction: rewrite + reexpress content freely

---

Single-document vs. multi-document: both are valuable and interesting, we’ll focus on single-document

- Extractive models: focus on picking the right sentences from existing documents
  - Can essentially do this with a classifier, the best supervised approaches use BERT now
  - Works particularly well in the multi-document case
- Abstractive models: generate the summary text “from scratch”
Seq2seq Summarization

Original Text (truncated): Lagos, Nigeria (CNN) a day after winning Nigeria's presidency, Muhammadu Buhari told CNN's Christiane Amanpour that he plans to aggressively fight corruption that has long plagued Nigeria and go after the root of the nation’s unrest. Buhari said he’ll “rapidly give attention” to curbing violence in the northeast part of Nigeria, where the terrorist group Boko Haram operates. By cooperating with neighboring nations Chad, Cameroon and Niger, he said his administration is confident it will be able to thwart criminals and others contributing to Nigeria’s instability. For the first time in Nigeria’s history, the opposition defeated the ruling party in democratic elections. Buhari defeated incumbent Goodluck Jonathan by about 2 million votes, according to Nigeria’s independent national electoral commission. The win comes after a long history of military rule, coups and botched attempts at democracy in Africa’s most populous nation.

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize Nigeria’s economy. UNK says his administration is confident it will be able to thwart criminals and other Nigerians, he says the country has long Nigeria and Nigeria’s economy.

See et al. (2017)

CNN/Daily Mail dataset: ~300,000 articles and summaries scraped from web

Non-pretrained seq2seq model output. What’s wrong with this summary?

See et al. (2017)

Pointer-Generator Model

Allow the model to copy tokens from the source

See et al. (2017)

Seq2seq Summarization

Solutions: copy mechanism, coverage, just like in MT...

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize Nigeria’s economy. UNK says his administration is confident it will be able to thwart criminals and other Nigerians, he says the country has long Nigeria and Nigeria’s economy.

Pointer-Gen: Muhammadu Buhari says he plans to aggressively fight corruption in the northeast part of Nigeria. He says he’ll “rapidly give attention” to curbing violence in the northeast part of Nigeria. He says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: Muhammadu Buhari says he plans to aggressively fight corruption that has long plagued Nigeria. He says his administration is confident it will be able to thwart criminals. The win comes after a long history of military rule, coups and botched attempts at democracy in Africa’s most populous nation.

See et al. (2017)

Neural Abstractive Systems

Actually this model ends up copying most of the time... 40% of the generated summary sentences are direct copies from article

See et al. (2017)
Enforcing Factuality

- When doing freeform paraphrasing or summarization from a seq2seq model, how to make sure the output is factual with respect to the input?

Seven games involving Nimes were investigated after Conrad was arrested last November.

- The relation here is wrong, even though all of the words are individually right!

- How can we verify that the text predicted by our model is factual, beyond just spotting made-up words?

CHECK(text, summary)

Enforcing Factuality

- Natural language inference systems: determine when sentence A entails sentence B (implies it must be true), contradicts it, or is neutral with respect to it

<table>
<thead>
<tr>
<th>Premise</th>
<th>Entailment</th>
<th>Neutral</th>
<th>Contradiction</th>
</tr>
</thead>
<tbody>
<tr>
<td>A woman selling bamboo sticks talking to two men on a loading dock.</td>
<td>There are at least three people on a loading dock.</td>
<td>A woman is selling bamboo sticks to help provide for her family.</td>
<td>A woman is not taking money for any of her sticks.</td>
</tr>
</tbody>
</table>

CHECK(text, summary) = entails(text, summary)

Bowman et al. (2015), Gururangan et al. (2018)
**Dependency Arc Entailment**

Classifier considers embeddings of head + child + dep label

Checker is not entailed

- BERT

**Takeaways**

- LMs and seq2seq models can be used for dialogue and other applications, not just translation.
- Can build chatbots that are primarily data-driven (with these neural models) or rule-based/templated.
- AI/chatbots/dialogue are not solved problems.
- Still much more to do: these systems usually don’t have anything to say. Goal-oriented dialogue and grounded/embodied systems (e.g., a dialogue system on a robot) are much tougher to get working.

---

*Seven games involving Nimes were investigated after Conrad was arrested last November.*

CHECK(text, summary) = for all arcs. entails(text, arc)

Goyal and Durrett (2020)