Announcements

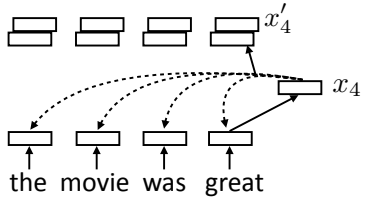
- eCIS surveys released — take a snapshot of the "done" page for your final project submission
- FP check-ins due Friday
- A4, A5 grading
- Today: recap of BERT, applying BERT
  pre-trained models for generation: GPT-2, dialogue, summarization

Recall: Self-Attention

- Each word forms a "query" which then computes attention over each word

$$\alpha_{i,j} = \text{softmax}(x_i^\top W x_j)) \quad \text{scalar}$$

$$x_i' = \sum_{j=1}^{n} \alpha_{i,j} V x_j \quad \text{vector = sum of scalar * mat * vector}$$



- Multiple "heads" analogous to different convolutional filters. Use parameters W_k and V_k to get different attention values + transform vectors

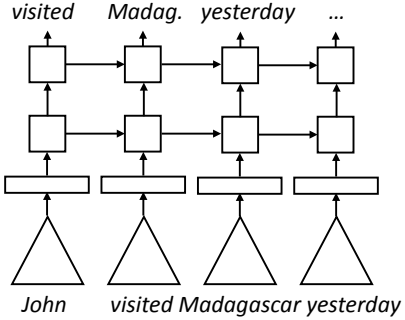
$$\alpha_{k,i,j} = \text{softmax}(x_i^\top W_k x_j) \quad x_{k,i}' = \sum_{j=1}^{n} \alpha_{k,i,j} V_k x_j$$

Vaswani et al. (2017)

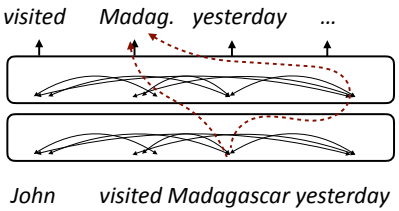
Recall: BERT

- How to learn a "deeply bidirectional" model? What happens if we just replace an LSTM with a transformer?

ELMo (Language Modeling)



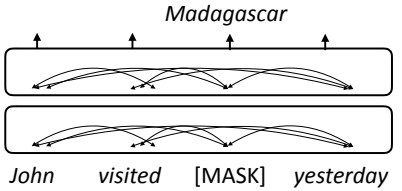
BERT



- You could do this with a "one-sided" transformer, but this "two-sided" model can cheat

Recall: Masked Language Modeling

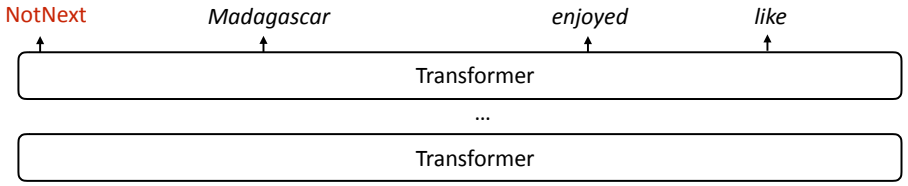
- How to prevent cheating? Next word prediction fundamentally doesn't work for bidirectional models, instead do *masked language modeling*
- BERT formula: take a chunk of text, mask out 15% of the tokens, and try to predict them



Devlin et al. (2019)

Recall: Next "Sentence" Prediction

- Input: [CLS] Text chunk 1 [SEP] Text chunk 2
- 50% of the time, take the true next chunk of text, 50% of the time take a random other chunk. Predict whether the next chunk is the "true" next
- BERT objective: masked LM + next sentence prediction

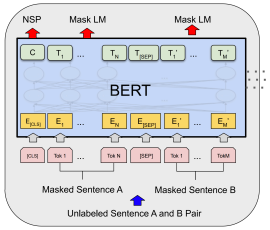


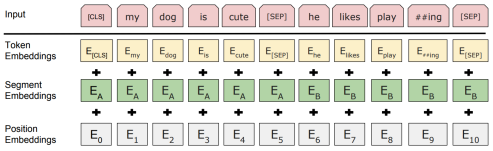
[CLS] *John visited* **[MASK]** *yesterday and really* **[MASK]** *it* [SEP] *I* **[MASK]** *Madonna.*

Devlin et al. (2019)

Recall: BERT Architecture

- BERT Base: 12 layers, 768-dim per wordpiece token, 12 heads. Total params = 110M
- BERT Large: 24 layers, 1024-dim per wordpiece token, 16 heads. Total params = 340M
- Positional embeddings and segment embeddings, 30k word pieces
- This is the model that gets **pre-trained** on a large corpus

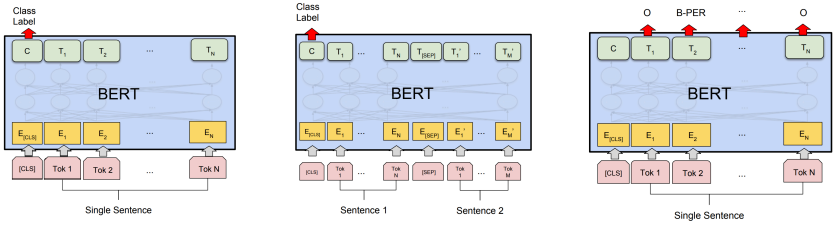




Devlin et al. (2019)

Applying BERT

Recall: What can BERT do?



(b) Single Sentence Classification Tasks: SST-2, CoLA

(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

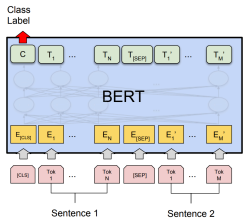
- Artificial [CLS] token is used as the vector to do classification from
- Sentence pair tasks (entailment): feed both sentences into BERT
- BERT can also do tagging by predicting tags at each word piece

Devlin et al. (2019)

What can BERT do?

Entails (first sentence implies second is true)

[CLS] A boy plays in the snow [SEP] A boy is outside



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

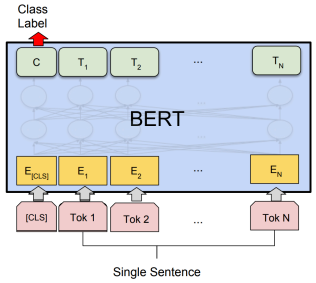
- How does BERT model this sentence pair stuff?
- Transformers can capture interactions between the two sentences, even though the NSP objective doesn't really cause this to happen

What can BERT NOT do?

- BERT **cannot** generate text (at least not in an obvious way)
  - Can fill in MASK tokens, but can't generate left-to-right (well, you could put MASK at the end repeatedly, but this is slow)
- Masked language models are intended to be used primarily for "analysis" tasks

Fine-tuning BERT

- Fine-tune for 1-3 epochs, batch size 2-32, learning rate 2e-5 - 5e-5



(b) Single Sentence Classification Tasks: SST-2, CoLA

- Large changes to weights up here (particularly in last layer to route the right information to [CLS])
- Smaller changes to weights lower down in the transformer
- Small LR and short fine-tuning schedule mean weights don't change much

Evaluation: GLUE

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Wang et al. (2019)

Results

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

- Huge improvements over prior work (even compared to ELMo)
- Effective at "sentence pair" tasks: textual entailment (does sentence A imply sentence B), paraphrase detection

Devlin et al. (2018)

Using BERT

- Huggingface Transformers: big open-source library with most pre-trained architectures implemented, weights available
- Lots of standard models…

Model architectures

🤗 Transformers currently provides the following NLU/NLG architectures:

1. **BERT** (from Google) released with the paper BERT: Pre-training of Deep Understanding by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Krist
2. **GPT** (from OpenAI) released with the paper Improving Language Under Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever.
3. **GPT-2** (from OpenAI) released with the paper Language Models are Un Jeffrey Wu*, Rewon Child, David Luan, Dario Amodei** and Ilya Sutskev
4. **Transformer-XL** (from Google/CMU) released with the paper Transform Fixed-Length Context by Zihang Dai*, Zhilin Yang*, Yiming Yang, Jaime
5. **XLNet** (from Google/CMU) released with the paper XLNet: Generalized Understanding by Zhilin Yang*, Zihang Dai*, Yiming Yang, Jaime Carbon
6. **XLM** (from Facebook) released together with the paper Cross-lingual L and Alexis Conneau.
7. **RoBERTa** (from Facebook), released together with the paper a Robustly

…

and "community models"

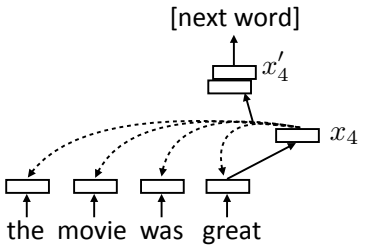
- `mrm8488/spanbert-large-finetuned-tacred`
- `mrm8488/xlm-multi-finetuned-xquadv1`
- `nlpaueb/bert-base-greek-uncased-v1`
- `nlptown/bert-base-multilingual-uncased-sentiment`
- `patrickvonplaten/reformer-crime-and-punish`
- `redewiedergabe/bert-base-historical-german-rw-cased`
- `roberta-base`
- `severinsimmler/literary-german-bert`
- `seyonec/ChemBERTa-zinc-base-v1`

…

GPT/GPT2

Language Modeling with Self-Attention

- Self-attention computes a representation for each word conditioned on the rest of the sequence
- Predict next word based on that



Vaswani et al. (2017)

OpenAI GPT/GPT2

- "ELMo with transformers" (works better than ELMo)
- Train a single unidirectional transformer LM on long contexts
- GPT2: trained on 40GB of text collected from upvoted links from reddit
- 1.5B parameters — by far the largest of these models trained when it came out in March 2019

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

- Because it's a language model, we can **generate** from it

Radford et al. (2019)

OpenAI GPT2

slide credit: OpenAI

Open Questions

1) How novel is the stuff being generated? (Is it just doing nearest neighbors on a large corpus?)

2) How do we understand and distill what is learned in this model?

3) How do we harness these priors for conditional generation tasks (summarization, generate a report of a basketball game, etc.)

4) Is this technology dangerous? (OpenAI pursued a "staged release" strategy and didn't release biggest model)

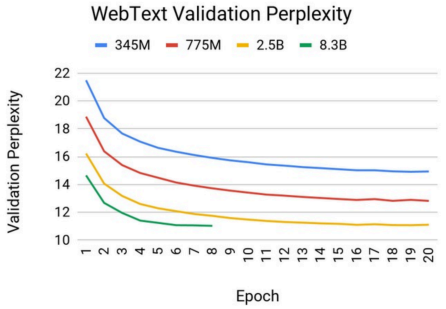
Pre-Training Cost (with Google/AWS)

- BERT: Base $500, Large $7000
- GPT-2 (as reported in other work): $25,000
- XLNet (BERT variant): $30,000 — $60,000 (unclear)
- This is for a single pre-training run…developing new pre-training techniques may require many runs
- *Fine-tuning* these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets)

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Pushing the Limits

- NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2)
- Arguable these models are still underfit: larger models still get better held-out perplexities



NVIDIA blog (Narasimhan, August 2019)

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)

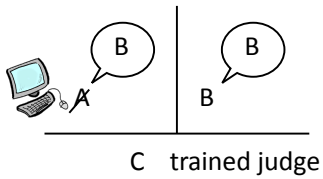
- AI Dungeon: text-based adventure game generated on-the-fly using GPT-2
- In general: big improvement in "open-ended" generation tasks

Chatbots

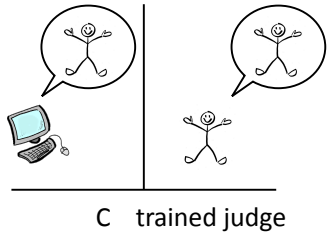
Turing Test (1950)

- 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:



Standard Interpretation:



- 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

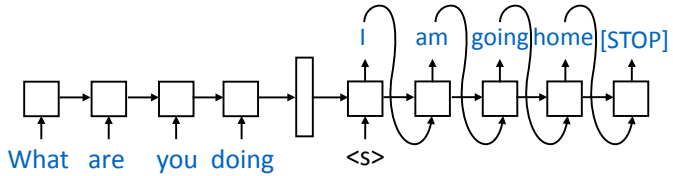
(.*) you (.*) me → Why do you think I $2 you?

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

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	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!

- Efforts to imbue seq2seq models with "personality"

Zhang et al. (2018)

Meena

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

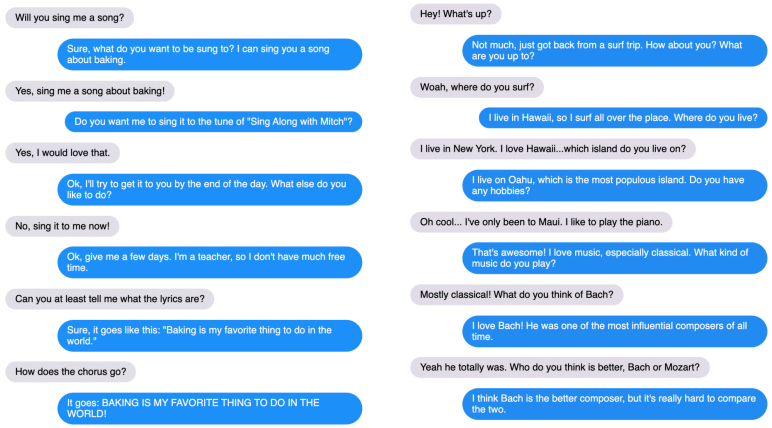
Adiwardana et al. (2020)

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)

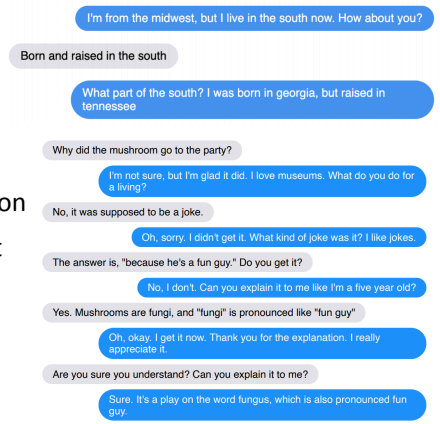
Roller et al. (April 29, 2020)

Blender



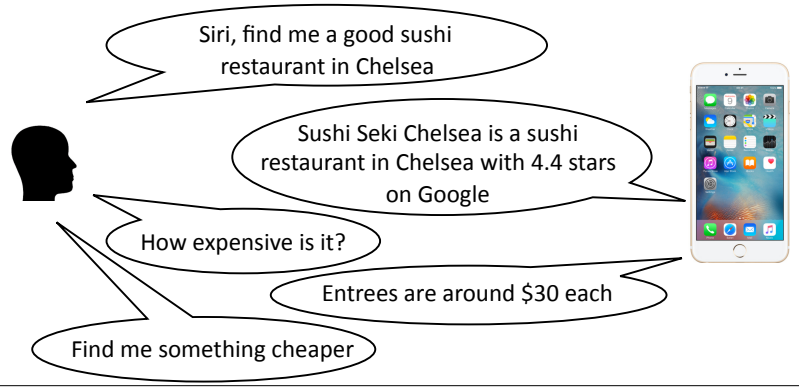
Blender

- 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



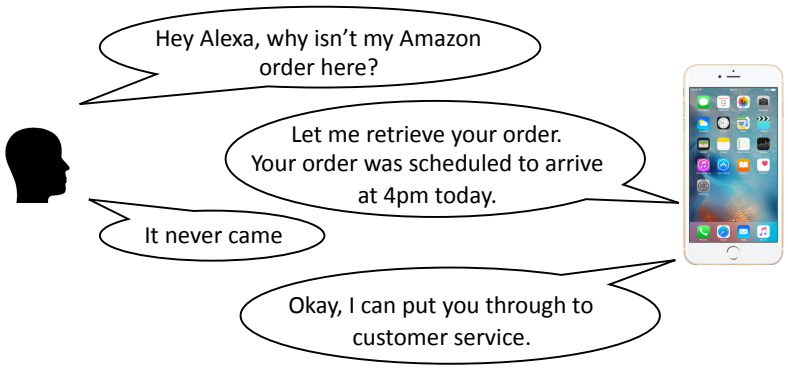
Task-Oriented Dialogue

- Personal assistants / API front-ends:



Task-Oriented Dialogue

- Customer service:



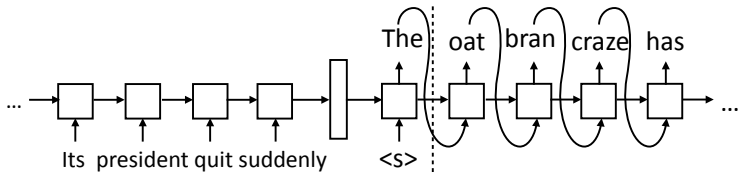
Goal-oriented Dialogue

- 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
- Dozens of startups + medium-sized companies in this space
- Big Companies: Apple Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, Facebook, Samsung Bixby, Tencent WeChat, ASAPP

Summarization

Seq2seq Summarization

- Extractive paradigm isn't all that flexible, even with compression
- Can we just use seq2seq models to simplify things?
- Train to produce summary based on document



- Need lots of data: most methods are going to be single-document

Chopra et al. (2016)

Seq2seq Headline Generation

- Headline generation task: generate headline from first sentence of article (can get lots of data!)

I(1): brazilian defender pepe is out for the rest of the season with a knee injury , his porto coach jesualdo ferreira said saturday . — sentence
G: football : pepe out for season — headline

- Works pretty well, though these models can generate incorrect summaries (who has the knee injury?)
- What happens if we try this on a longer article?

Chopra et al. (2016)

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.**

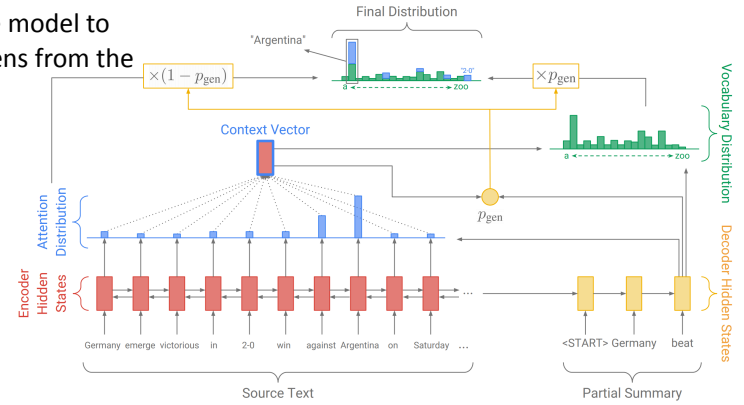
Baseline Seq2Seq + Attention: **UNK UNK** says his administration is confident it will be able to **destabilize nigeria's economy**. **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

- 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.

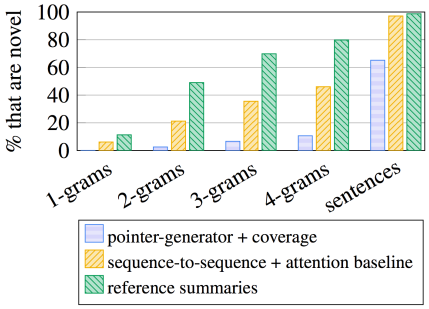
Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption in the northeast part of nigeria. he says he'll "rapidly give attention" to curbing violence in the northeast part of nigeria. he says his administration is confident it will be able to thwart criminals.

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)

BART: Pre-trained seq2seq model

Source Document (abbreviated)	BART Summary
The researchers examined three types of coral in reefs off the coast of Fiji ... The researchers found when fish were plentiful, they would eat algae and seaweed off the corals, which appeared to leave them more resistant to the bacterium Vibrio coralliilyticus, a bacterium associated with bleaching. The researchers suggested the algae, like warming temperatures, might render the corals' chemical defenses less effective, and the fish were protecting the coral by removing the algae.	Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.

- These look great! But they're not always factual

Lewis et al. (2019)

Takeaways

- Pre-trained models are remarkably good at generating text
- Story generation, dialogue systems, summarization, etc. have gotten way better in the past few years
- 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
- Next time: other languages