

# CS388: Natural Language Processing

## Lecture 20: Dialogue and Summarization

Greg Durrett



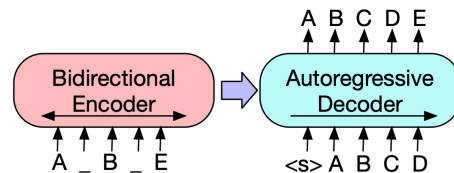
## Administrivia

- ▶ Kenton Lee talk next week in Eunsol Choi's class
- ▶ FP presentation days posted



## 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)



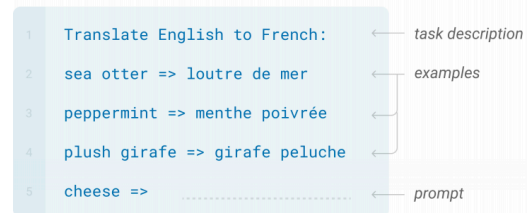
Lewis et al. (2019)



## 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.



Brown et al. (2020)



## This Lecture

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

Bender, Gebru, McMillan-Major, Shmitchell (2021)

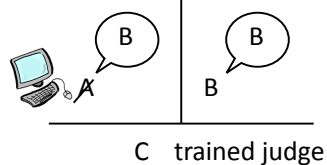
## Chatbots: Intro, Rule-based



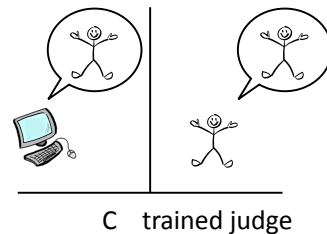
## Turing Test (1950)

- ▶ 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:



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?

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

My (.\*) (.\*)

What else comes to mind when you think of your \$1?

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

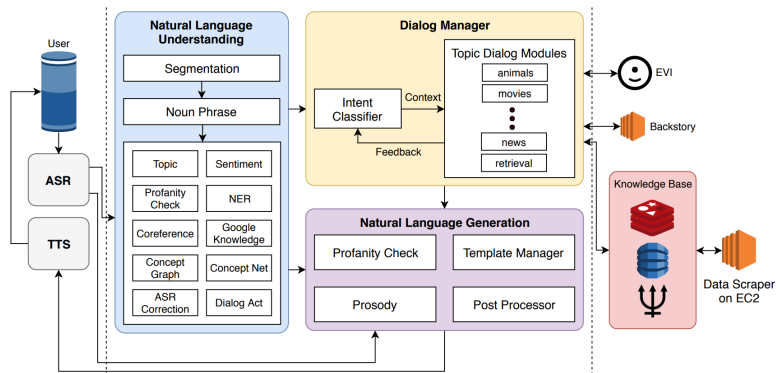


## Amazon Alexa Prize

- ▶ 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

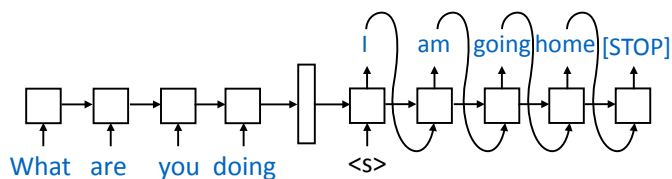
- ▶ NLU: sentence segmentation, parsing, NER, coref, dialogue act prediction
- ▶ Detect different kinds of topic intents
- ▶ Topic modules (x11):
  - ▶ Animals: retrieve animal trivia with the reddit API
  - ▶ Holiday: what holidays is it today, etc.
  - ▶ Custom dialogue flow for each
- ▶ Generation: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

Chen ... Zhou Yu (2018)

## Seq2seq Chatbots



## Seq2seq models



- ▶ Train a seq2seq model on subtitles data

- ▶ 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:

<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.

Li et al. (2016)



## PersonaChat

### Persona 1

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

### Persona 2

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!

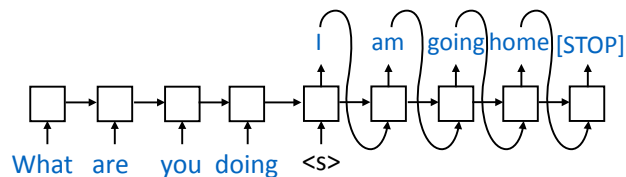
Zhang et al. (2018)

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

## Pre-trained Chatbots



## 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:

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



## 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 (stochastic parrot!)



## Task-Oriented Dialogue



## 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



## Summarization

- ▶ What makes a good summary?



## Summarization

BAGHDAD/ERBIL, Iraq (Reuters) - A strong earthquake hit large parts of northern Iraq and the capital Baghdad on Sunday, and also caused damage in villages across the border in Iran where state TV said at least six people had been killed.

There were no immediate reports of casualties in Iraq after the quake, whose epicenter was in Penjwin, in Sulaimaniyah province which is in the semi-autonomous Kurdistan region very close to the Iranian border, according to an Iraqi meteorology official.

But eight villages were damaged in Iran and at least six people were killed and many others injured in the border town of Qasr-e Shirin in Iran, Iranian state TV said.

The US Geological Survey said the quake measured a magnitude of 7.3, while an Iraqi meteorology official put its magnitude at 6.5 according to preliminary information.

Many residents in the Iraqi capital Baghdad rushed out of houses and tall buildings in panic.

...



## Summarization

Indian Express — **A massive earthquake of magnitude 7.3 struck Iraq on Sunday,** 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

The quake has been felt in several Iranian cities and **eight villages have been damaged.** Electricity has also been disrupted at many places, suggest few TV reports.

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.



## What makes a good summary?

Summary

A strong earthquake of magnitude 7.3 struck Iraq and Iran on Sunday. The epicenter was close to the Iranian border. Eight villages were damaged and six people were killed in Iran.

- ▶ 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



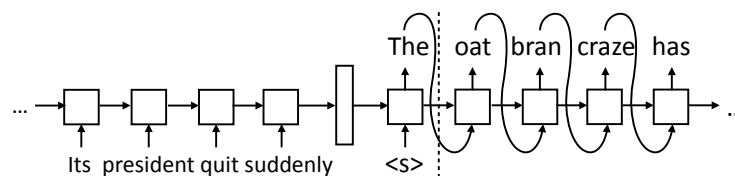
## Summarization Landscape

- ▶ 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

- ▶ Train to produce summary based on document



- ▶ Works okay for generating very short summaries like headlines, but what happens on longer articles?

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 amapour 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.**

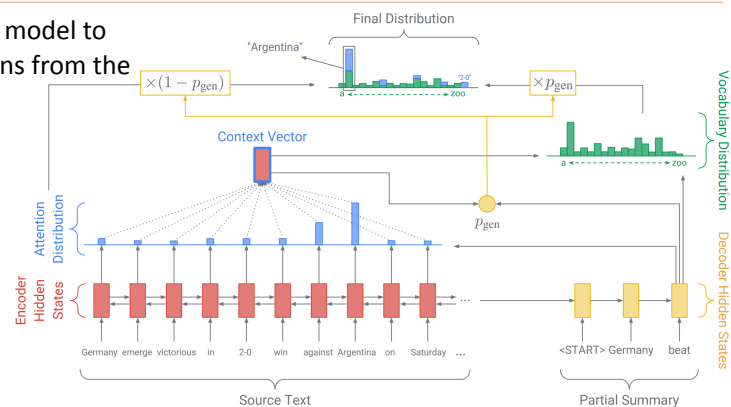
- ▶ CNN/Daily Mail dataset: ~300,000 articles and summaries scraped from web
- ▶ **Non-pretrained** seq2seq model output. What's wrong with this summary?

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.

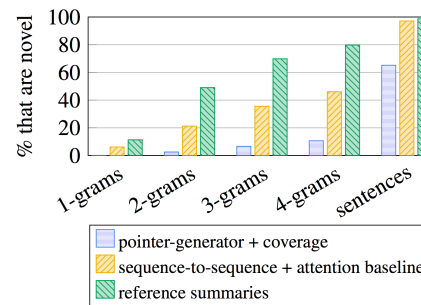
**Pointer-Gen + Coverage:** *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's 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)

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.

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."

### BART Summary

Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.

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)



## 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. seq2seq → Seven games involving Nimes were 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

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

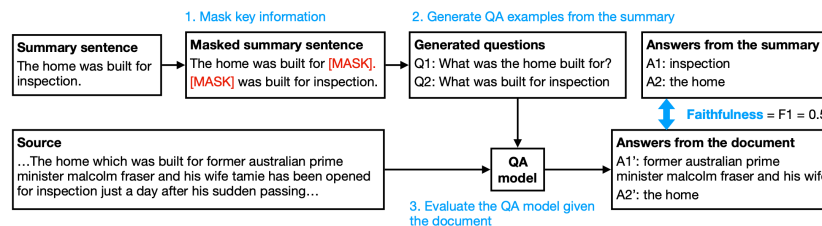
CHECK(text, summary) = entails(text, summary)

Bowman et al. (2015), Gururangan et al. (2018)



## Enforcing Factuality

- ▶ Question generation approaches Alex Wang et al. (2020), Durmus et al. (2020), Deutsch et al. (2020)



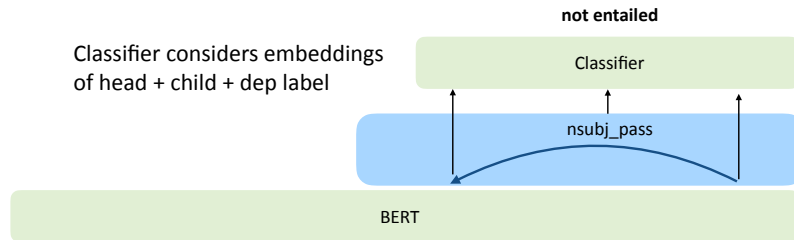
CHECK(text, summary) = { generate questions, answer questions from text, answer questions from summary + compare }

- ▶ What can go wrong? Answer selection, question generation, question answering...



## Dependency Arc Entailment

Classifier considers embeddings of head + child + dep label



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

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

Goyal and Durrett (2020)



## 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/emodied systems (e.g., a dialogue system on a robot) are much tougher to get working