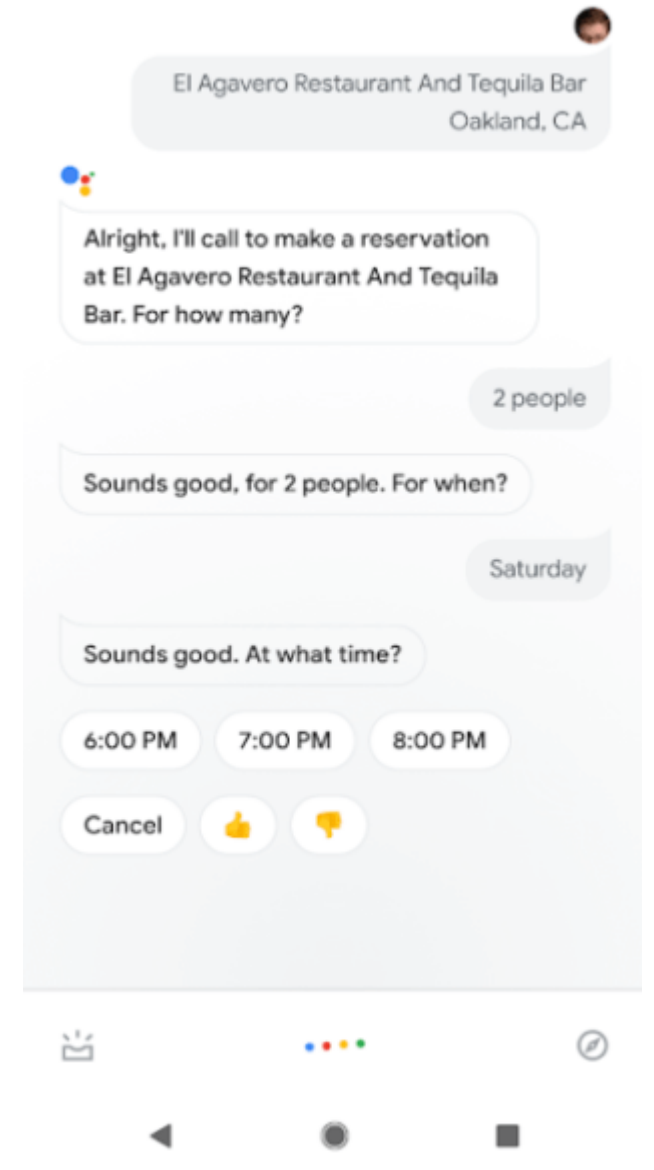
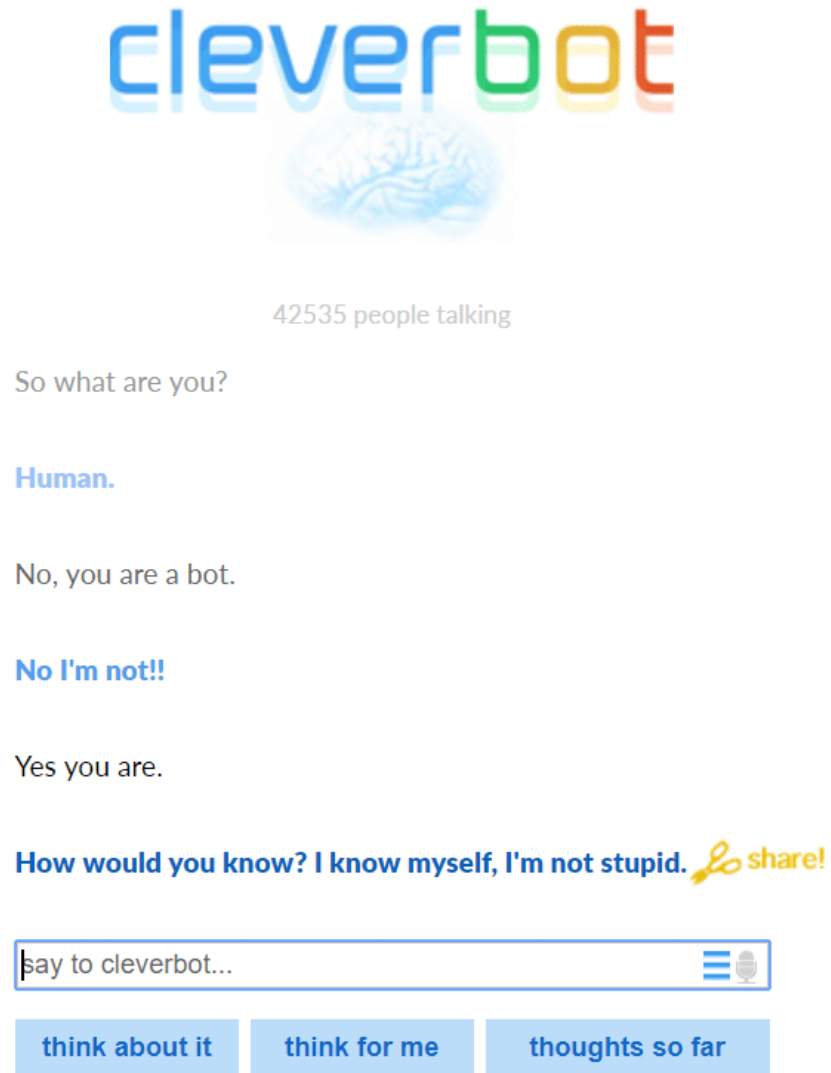
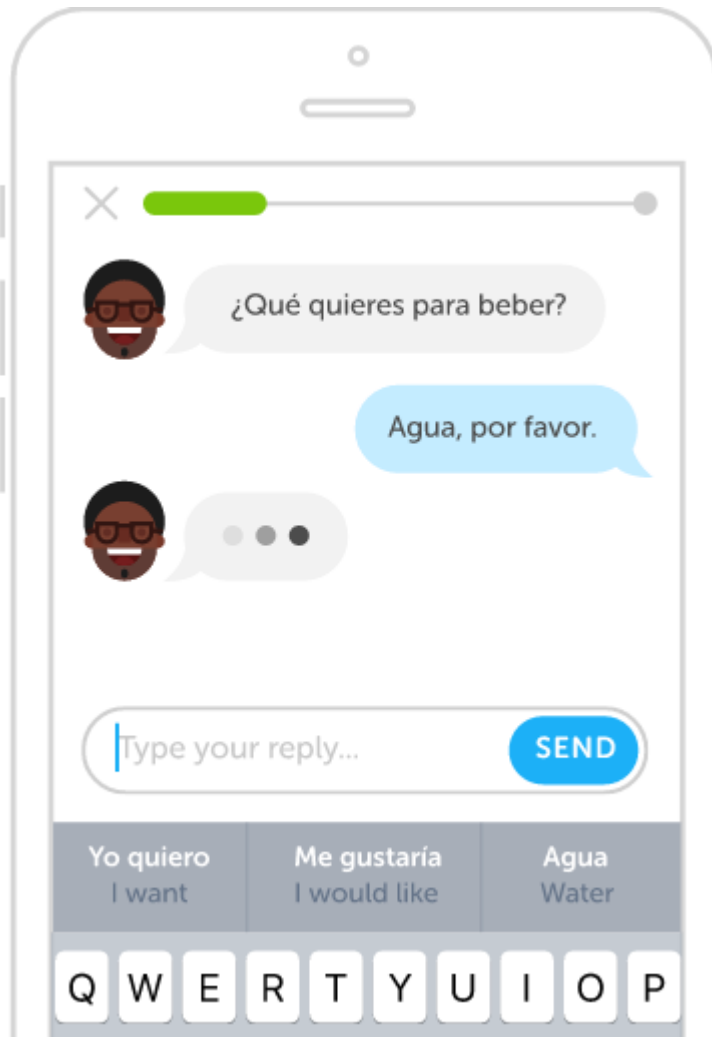


Text Generation

Katie Stasaski

INFO 159/259
Spring 2021
April 15, 2021



Dialogue Systems

Input: conversation, collection of multiple utterances/turns

Output: next utterance in conversation

Conversational *goals* and how success is defined are task-dependent

Seq2Seq Framework

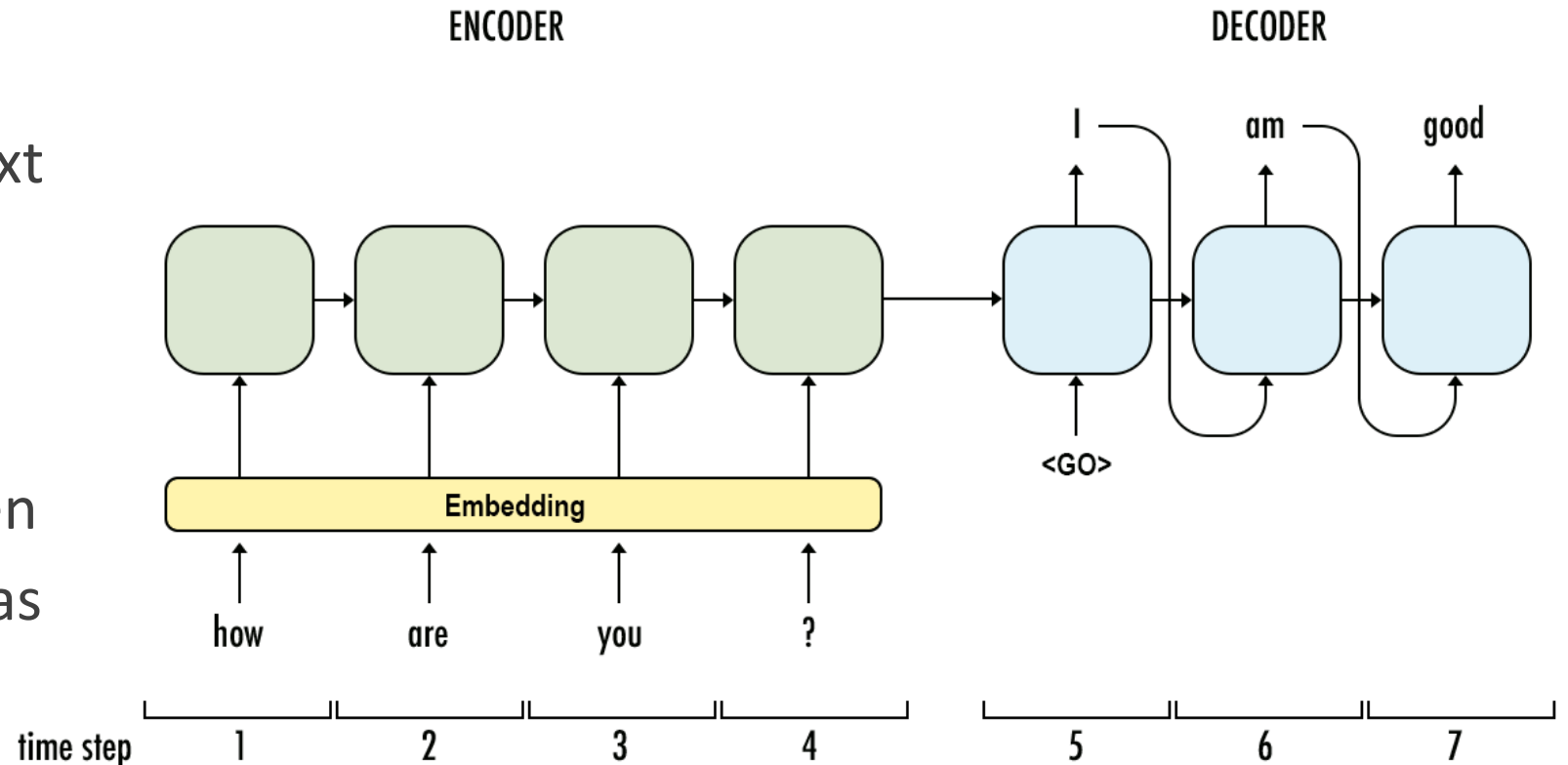
Two components: encoder and decoder

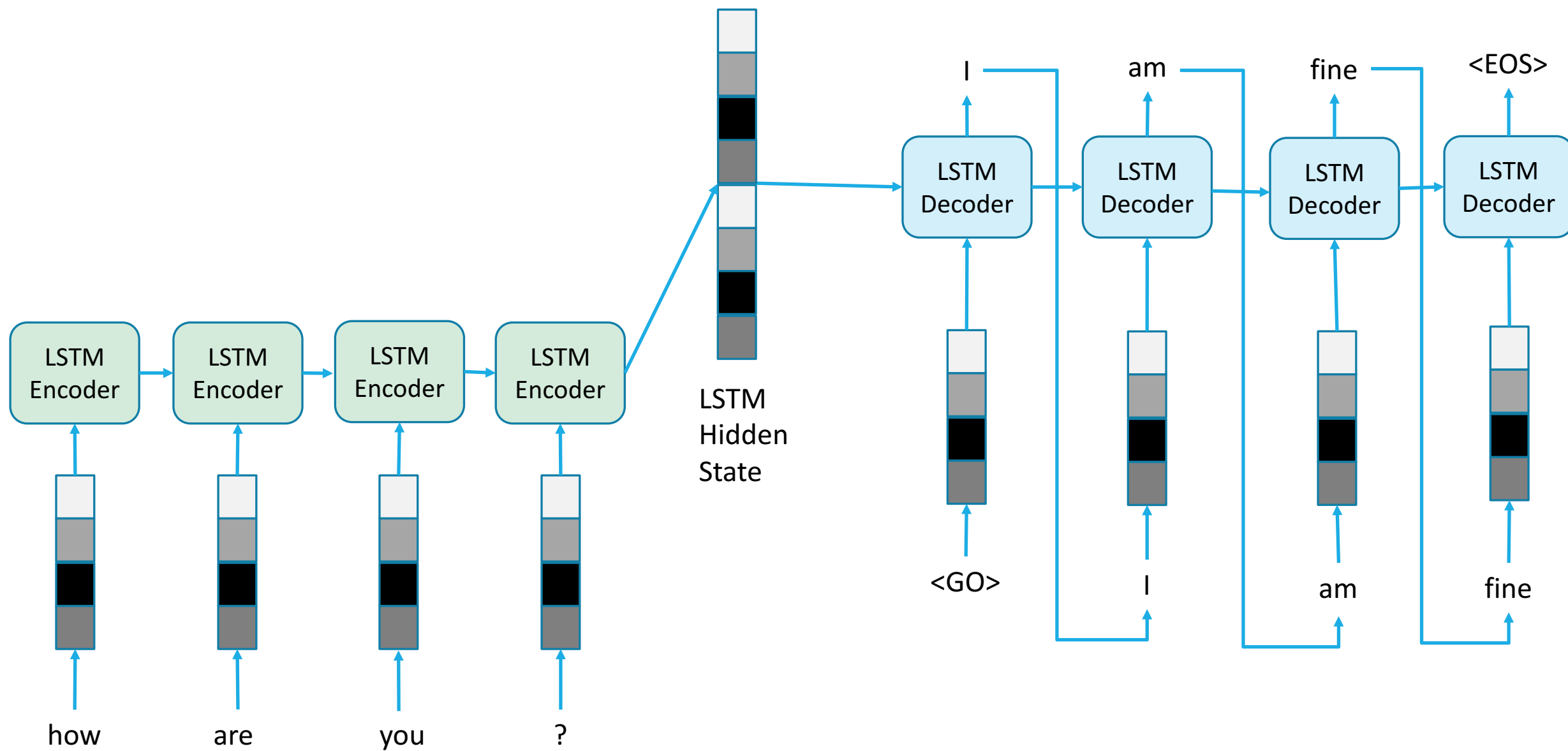
Encoder represents input text

Initialize decoder's hidden state to encoder's hidden state to begin with

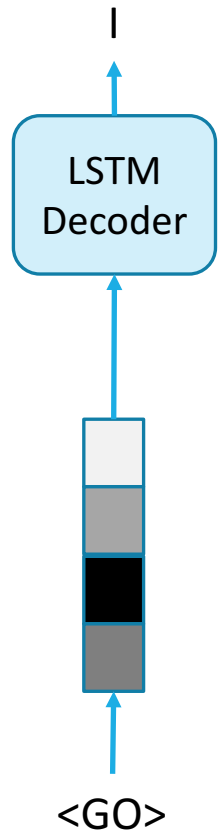
Prompt the decoder to start generation with special token

Feed decoder's predictions as data at generation time





How is “I” generated?



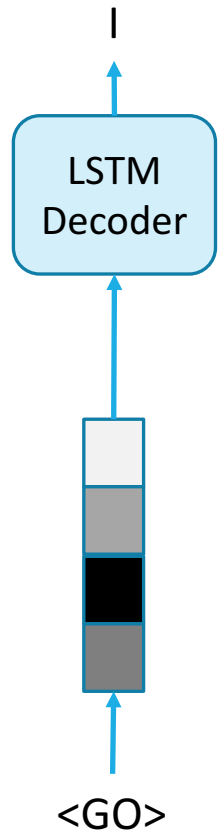
Size(Vocab)

Token	Probability
I	0.1
Well	0.05
I'm	0.04
Good	0.04
Poor	0.02
Doin'	0.001
Know	0.0001

Greedy decode:
Generate token
with highest
probability!

This is multi-class classification, just
applied at very large scale (e.g. 10k
vocab words!)

How do we train our model on human data?



Token	Probability
I	0.1
Well	0.05
I'm	0.04
Good	0.04
Poor	0.02
Doin'	0.001
Know	0.0001

Let's look at our *true response*, associated with this example, in the dataset.



I'm doing fine.

Word at position 0

$$\text{cross entropy loss} = - \sum_{t \in T} \log(P(y_t | x, y_{<t}))$$

What if the model goes off-track?

Input:

How are you?

True response is very specific!

True response:

I'm doing well but my brother has been struggling lately. Do you remember him?

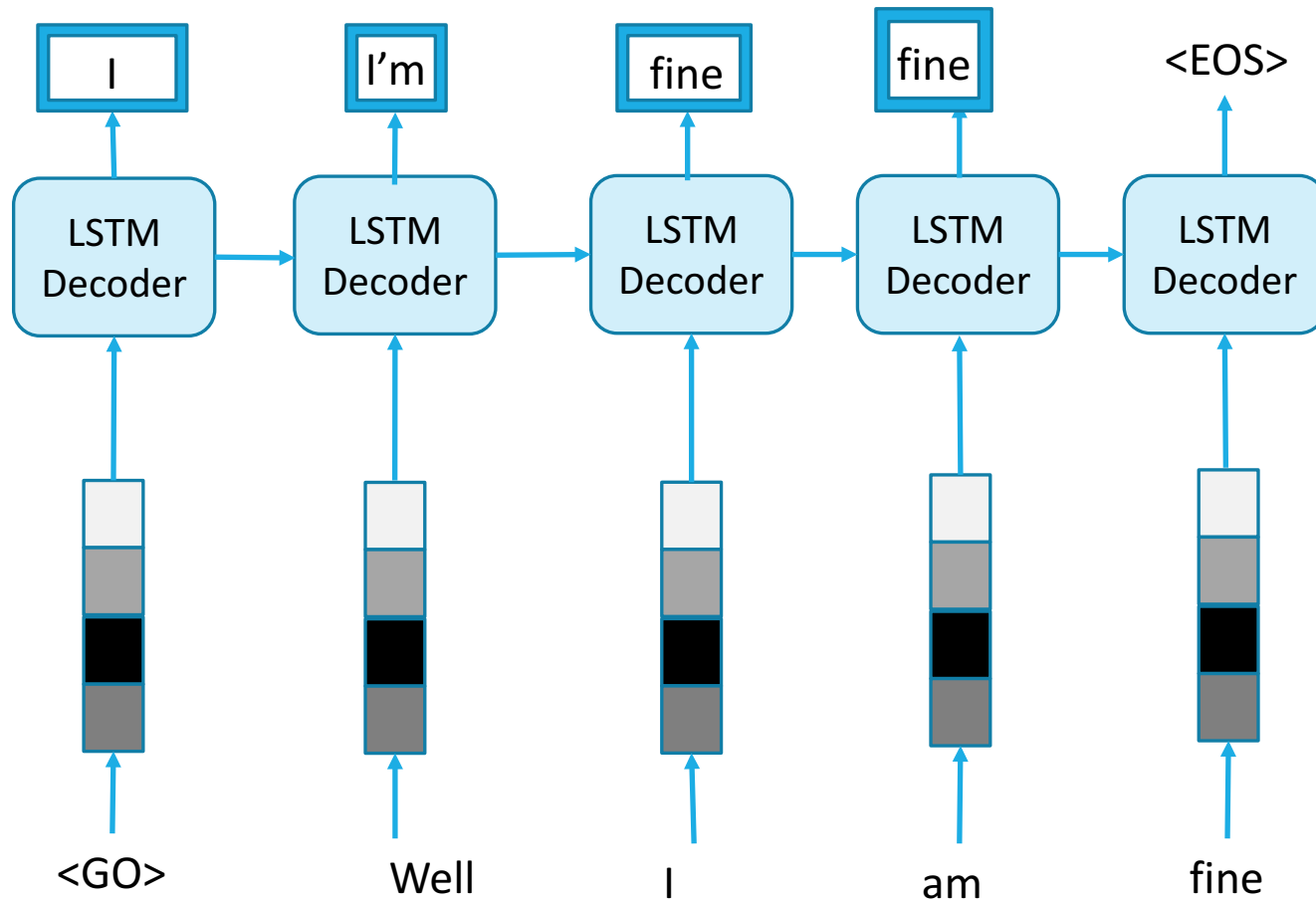
Isn't single correct response for conversations.

Greedy decoded model response:

I'm ok thank you for asking.

Can't compare true and decoded responses for tokens after position 1!

Solution: Teacher-forcing!

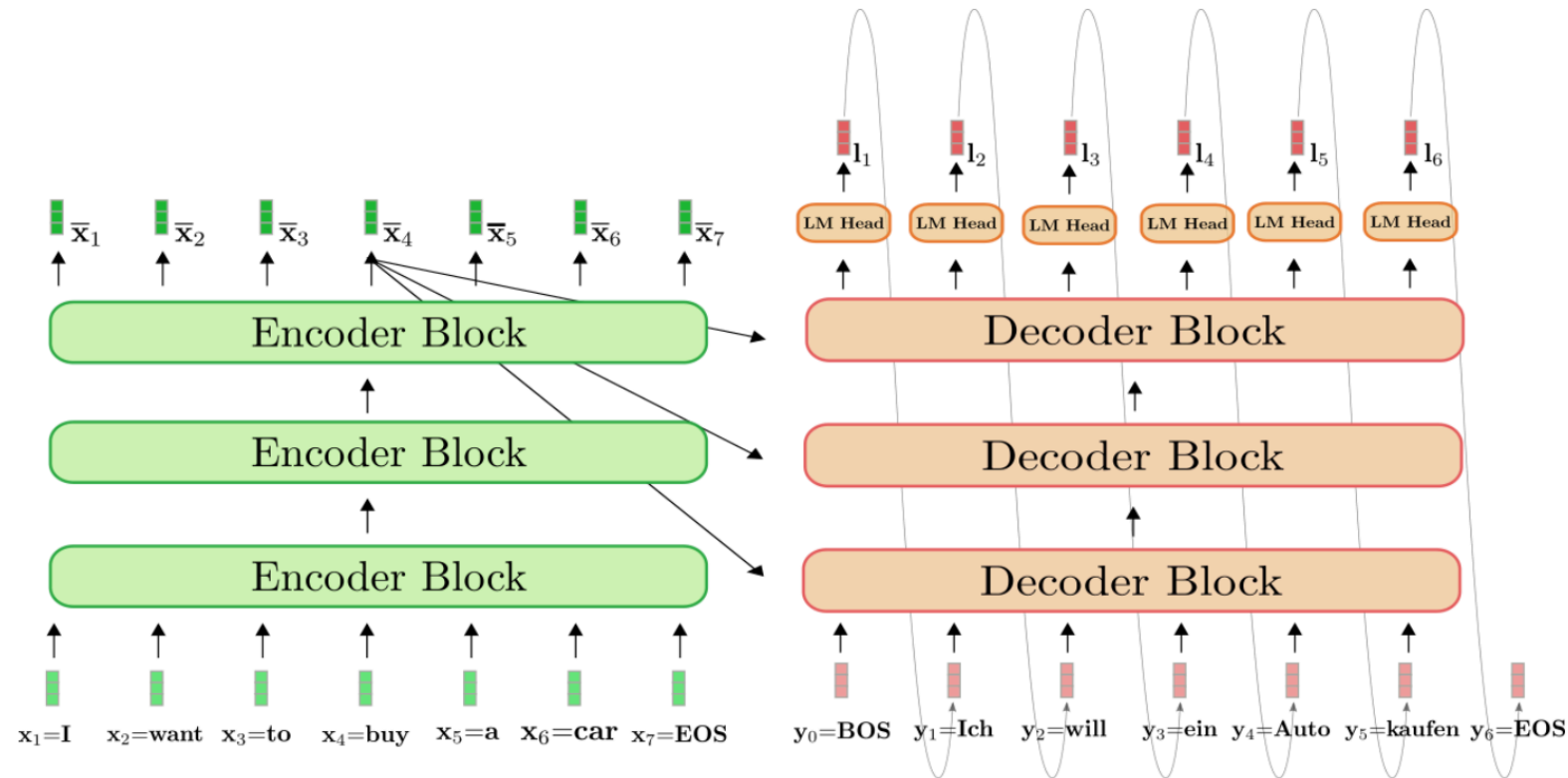


True response:

Well I am fine.

Calculate loss for each timestep, but always feed gold-generated token.

BERT Conditional Generation Framework



Dialogue Models

```
graph TD; A[Dialogue Models] --> B[Goal-Oriented]; A --> C[Chit-Chat];
```

Goal-Oriented

- Users converse with system with goal in mind (e.g. movie recommendation)
- System aims to fulfill user request
- System usually match user-intents or slot-fill

Chit-Chat

- Users converse with system with the goal of having a social conversation
- System aims to entertain and engage user
- Systems can use neural dialog agents or knowledge bases

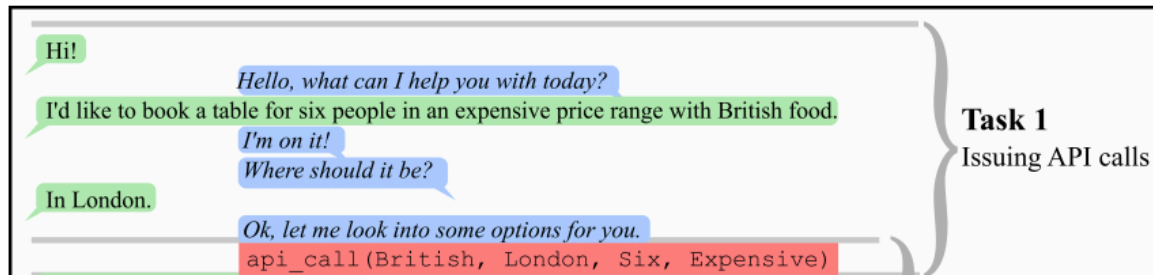
Goal Oriented Dialogue

Conversation with user aims to fulfill user task

Common domains:

- Restaurant reservation
- Movie recommendation
- Hotel reservation

bAbI dialog tasks

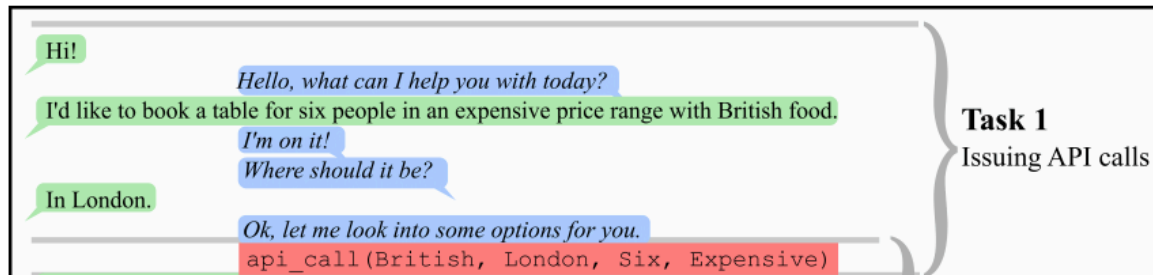


Intent classification

What is the user intending to accomplish with the conversation?

E.g. “book a restaurant reservation”

bAbI dialog tasks

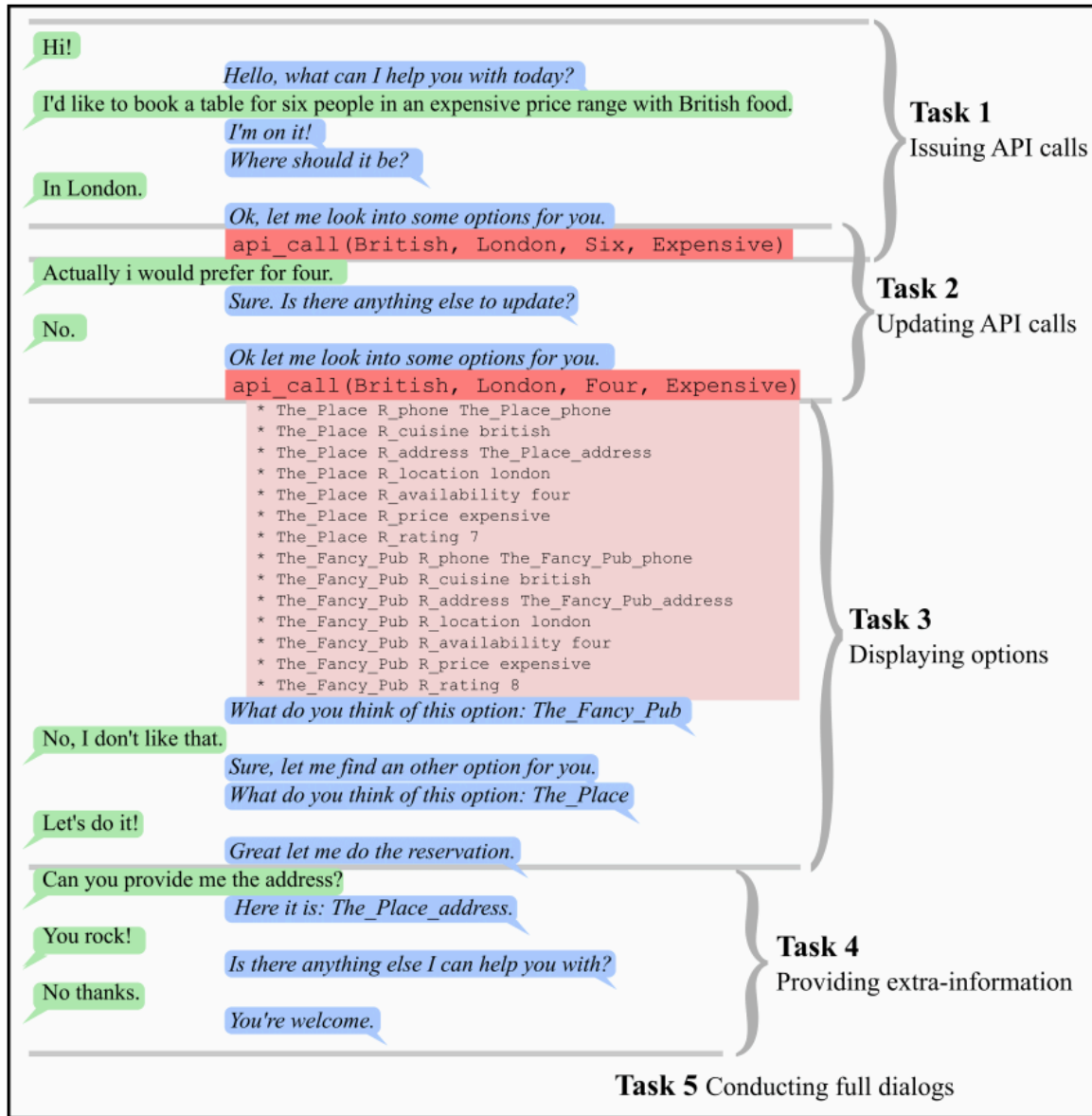


Slot-filling

Given a template,
what information is
required to complete
the task?

E.g. (British, London, 6
people, expensive)

bAbI dialog tasks



Bordes et al. 2017, Learning End-to-End Goal-Oriented Dialog

Dialogue Models

```
graph TD; A[Dialogue Models] --> B[Goal-Oriented]; A --> C[Chit-Chat];
```

Goal-Oriented

- Users converse with system with goal in mind (e.g. movie recommendation)
- System aims to fulfill user request
- System usually match user-intents or slot-fill

Chit-Chat

- Users converse with system with the goal of having a social conversation
- System aims to entertain and engage user
- Systems can use neural dialog agents or knowledge bases

Chit-Chat Dialogue

Aims to entertain the user and keep them engaged with utterances that further the conversation.

Alexa SocialBot Challenge

Goal: engage users in social conversation

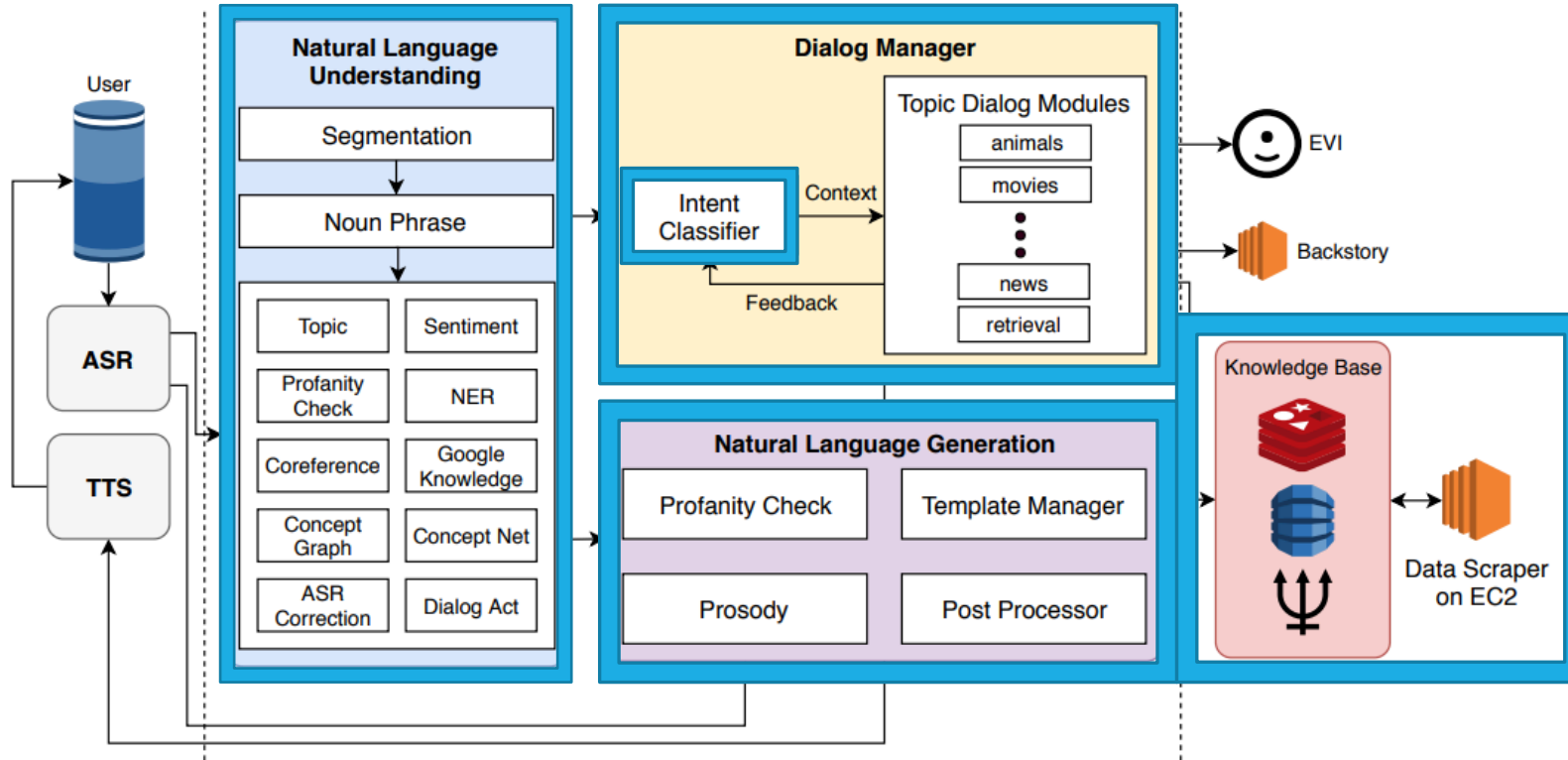
User population: Amazon Alexa users who opt-in to challenge via “Alexa, let’s chat”

Evaluation:

- User Likert scale ratings (1-5)
- Number of turns completed
- Time spent in conversation

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 SocialBot



Chen et al. 2018, Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data

Challenges with Chatbots

Persona
Attributes

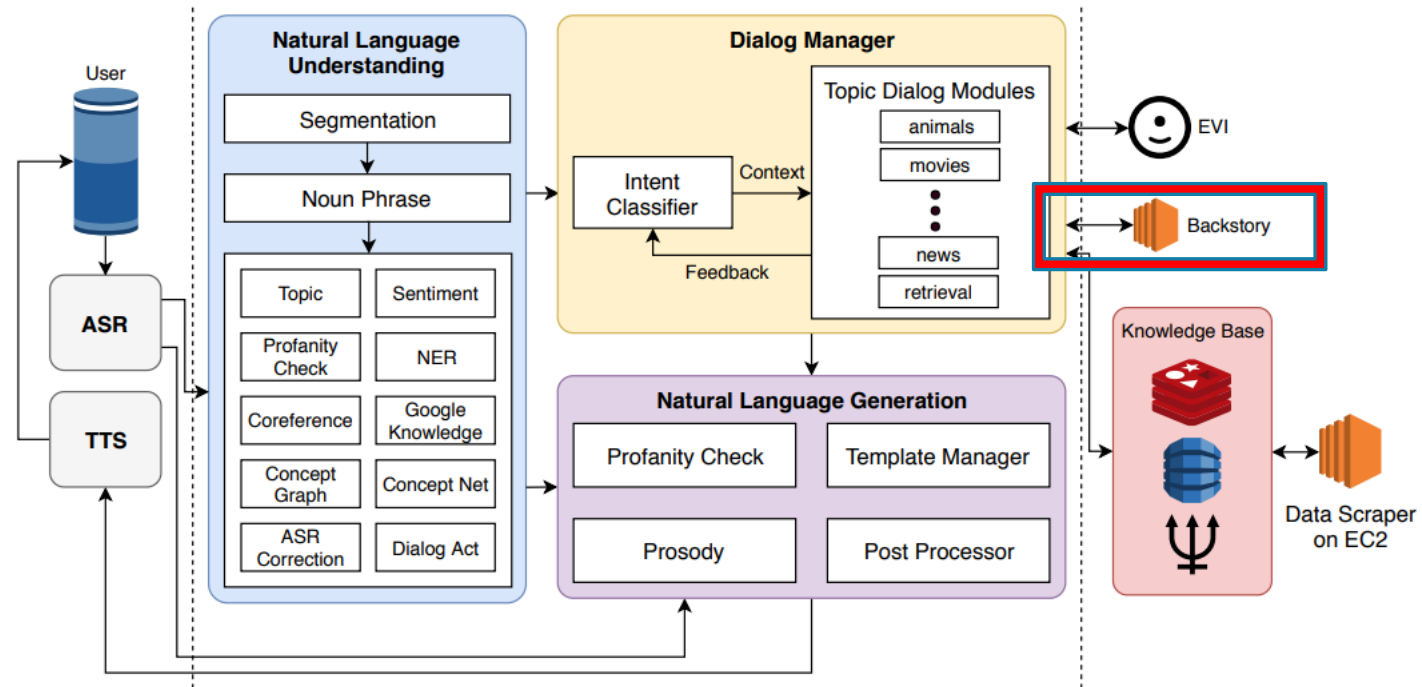
Responding
appropriately

Diversity of
Responses

Why Personas?

<i>message</i>	Where do you live now?
<i>response</i>	I live in Los Angeles.
<i>message</i>	In which city do you live now?
<i>response</i>	I live in Madrid.
<i>message</i>	In which country do you live now?
<i>response</i>	England, you?
<i>message</i>	Where were you born?
<i>response</i>	I was born in Canada.
<i>message</i>	Where are you from?
<i>response</i>	England, you?
<i>message</i>	Where did you grow up?
<i>response</i>	I grew up in Texas.
<i>message</i>	How old are you?
<i>response</i>	16 and you?
<i>message</i>	What's your age?
<i>response</i>	18.
<i>message</i>	What is your major?
<i>response</i>	I'm majoring in psychology
<i>message</i>	What did you study in college?
<i>response</i>	English lit.

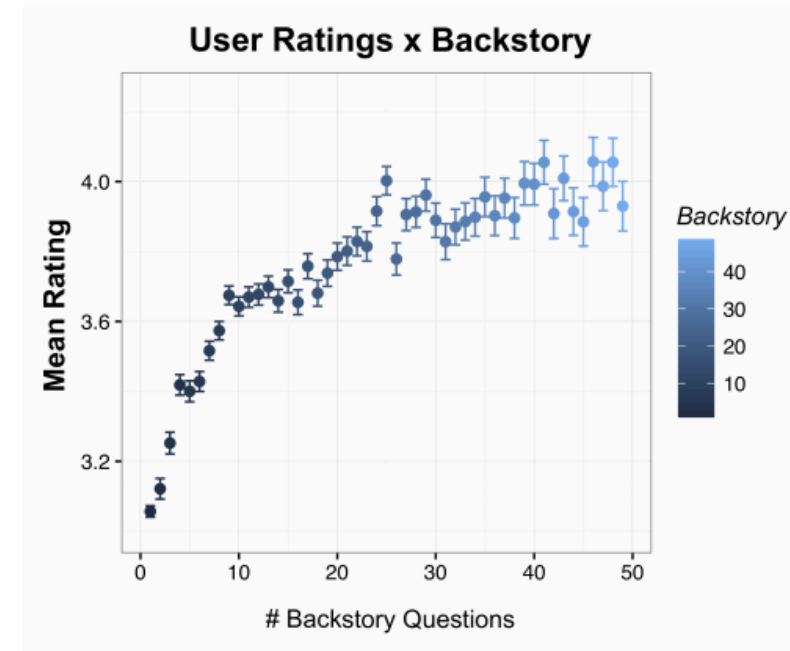
Gunrock SocialBot



Chen et al. 2018, Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data

Gunrock Persona

User1: do you like jeopardy?
Gunrock1: Sure I like Jeopardy, especially when Watson competed.
User2: why?
Gunrock2: I'm so impressed with the capabilities of a supercomputer.



Personas at-scale: 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!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Collected from crowdworkers

Each participant in 2-party dialogue has associated persona

162k Conversations

1,155 personas

Zhang et al. 2018, Personalizing Dialogue Agents: I have a dog, do you have pets too?

Challenges with Chatbots

Persona
Attributes

Responding
appropriately

Diversity of
Responses

What if responses don't convey correct emotion?



Solution: condition on emotion

Label: Proud

Situation: Speaker felt this when...

“I finally got that promotion at work! I have tried so hard for so long to get it!”

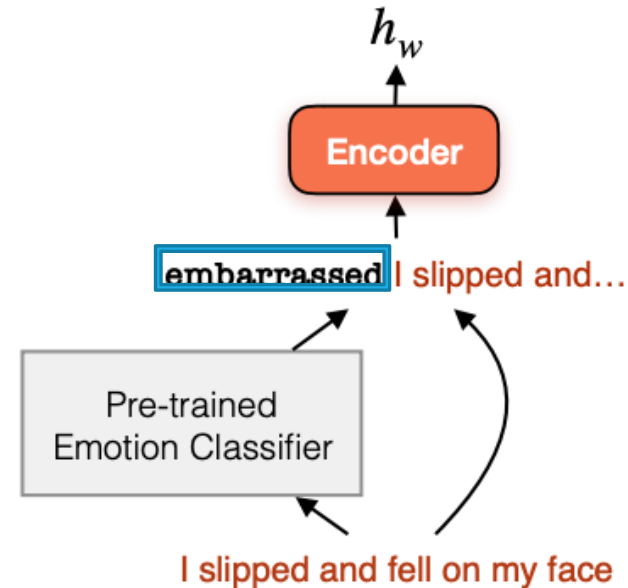
Conversation:

Speaker: I finally got promoted today at work!

Listener: Congrats! That's great!

Speaker: Thank you! I've been trying to get it for a while now!

Listener: That is quite an accomplishment and you should be proud!



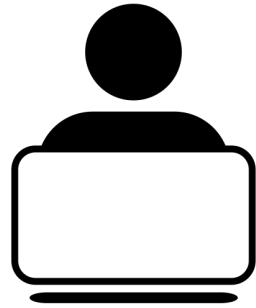
Prepend
emotion to
dialogue to
condition
generation

Challenges with Chatbots

Persona
Attributes

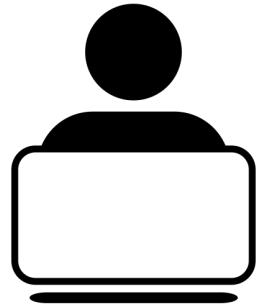
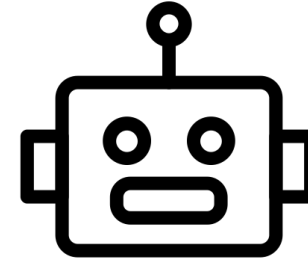
Responding
appropriately

Diversity of
Responses



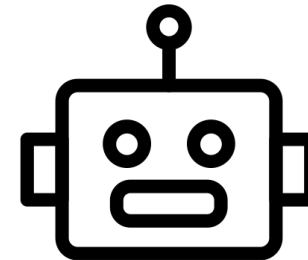
I just got promoted at work today!

Oh, okay



How should I celebrate?

I don't know



“An engaging response generation system should be able to output grammatical, coherent responses that are diverse and interesting” (Li et al. 2016)

What is diversity?

Diverse responses:

- are specific, not vague or generic
- are interesting
- are engaging
- advance the conversation

Problem with Maximum Likelihood

If always choose to generate the most likely token, our model responses will be the most predictable.

Maximum Mutual Information

Instead of using greedy decode to produce the most likely response, incorporate specificity of information given by target.

More precisely, how likely is the source (prior utterance) given the wording of the target (potential model response)? More specific utterances provide more information

Standard decode

$$\hat{T} = \arg \max_T \{ \log p(T|S) \}$$

MMI decode

$$\arg \max_T \{ (1 - \lambda) \log p(T|S) + \lambda \log p(S|T) \}$$

Maximum Mutual Information

Instead of using greedy decode to produce the most likely response, incorporate specificity of information given by target.

More precisely, how likely is the source (prior utterance) given the wording of the target (potential model response)? More specific utterances provide more information

Standard decode

$$\hat{T} = \arg \max_T \{ \log p(T|S) \}$$

MMI decode

$$\arg \max_T \{ (1 - \lambda) \log p(T|S) + \lambda \log p(S|T) \}$$

How probable is *target*
given *source*?

Maximum Mutual Information

Instead of using greedy decode to produce the most likely response, incorporate specificity of information given by target.

More precisely, how likely is the source (prior utterance) given the wording of the target (potential model response)? More specific utterances provide more information

Standard decode

$$\hat{T} = \arg \max_T \{ \log p(T|S) \}$$

MMI decode

$$\arg \max_T \{ (1 - \lambda) \log p(T|S) + \lambda \log p(S|T) \}$$

How probable is *source*
given *target*?

**All sources
in dataset:**

What's
your
name?

How
are
you?

Gold Source:

How is your
sister?

**Targets
(generate n):**

I don't know

Thanks for asking, my
sister is doing great!

General response, could apply to
many source sentences in dataset.

More specific response, limits
number of applicable source
utterances. $P(S|T)$ is higher.

Lower MMI likelihood

Higher MMI likelihood

Measuring diversity

How do we measure diversity of responses? Distinct number of n-grams!

Distinct-1: For the *entire* test set, calculate the total number of *unique unigrams*, divided by the total number of generated unigrams

Distinct-1

$$\frac{\text{len}(\text{set}(\text{unigrams}(\text{test}_{\text{set}})))}{\text{len}(\text{unigrams}(\text{test}_{\text{set}}))}$$

I don't know.

I'm doing well.

I don't know. Are you well?

OK, I guess so.

I'm doing well.

I : 3
Don't: 2
Know: 2
I'm: 2
Doing: 2
Well: 3
Are: 1
You: 1
OK: 1
Guess: 1
So: 1

11 distinct unigrams

Total: 19

$$\text{Distinct-1: } \frac{11}{19} = 0.58$$

Measuring diversity

Distinct-2: For the *entire* test set, calculate the total number of *unique bigrams*, divided by the total number of generated bigrams.

Distinct-2

$$\frac{\text{len}(\text{set}(\text{bigrams}(\text{test}_{\text{set}})))}{\text{len}(\text{bigrams}(\text{test}_{\text{set}}))}$$

I don't know.

I'm doing well.

I don't know. Are you well?

OK, I guess so.

I'm doing well.

I don't: 2
Don't know: 2
I'm doing: 2
Doing well: 2
Know are: 1
Are you: 1
You well: 1
OK I: 1
I guess: 1
Guess so: 1

10 distinct unigrams

Total: 14

Distinct-2: $\frac{10}{14} = 0.71$

Measuring diversity

% Unique Replies: For the *entire* test set, calculate the total number of *unique responses*, divided by the number of responses generated.

$$\frac{\text{len}(\text{set}(\text{responses}(\text{test}_{\text{set}})))}{\text{len}(\text{responses}(\text{test}_{\text{set}}))}$$

I don't know.

I'm doing well.

I don't know. Are you well?

OK, I guess so.

I'm doing well.

I don't know: 1

I'm doing well: 2

I don't know. Are you well?: 1

OK, I guess so: 1

Total: 5

4 distinct replies

% Unique Replies: $\frac{4}{5} = 0.8$

MMI Results

Model	# of training instances	BLEU	<i>distinct-1</i>	<i>distinct-2</i>
SEQ2SEQ (baseline)	23M	4.31	.023	.107
SEQ2SEQ (greedy)	23M	4.51	.032	.148
MMI-antiLM: $\log p(T S) - \lambda U(T)$	23M	4.86	.033	.175
MMI-bidi: $(1 - \lambda) \log p(T S) + \lambda \log p(S T)$	23M	5.22	.051	.270

Why are these much smaller than our examples?

Because test set is MUCH larger!

Greedy Decoding

At each time step, select the most probable token to generate.

Problem: most-predictable utterance is chosen!

Token	Probability
I	0.1
Well	0.05
I'm	0.04
Good	0.04
Poor	0.02
Doin'	0.001
Know	0.0001

Top-k sampling

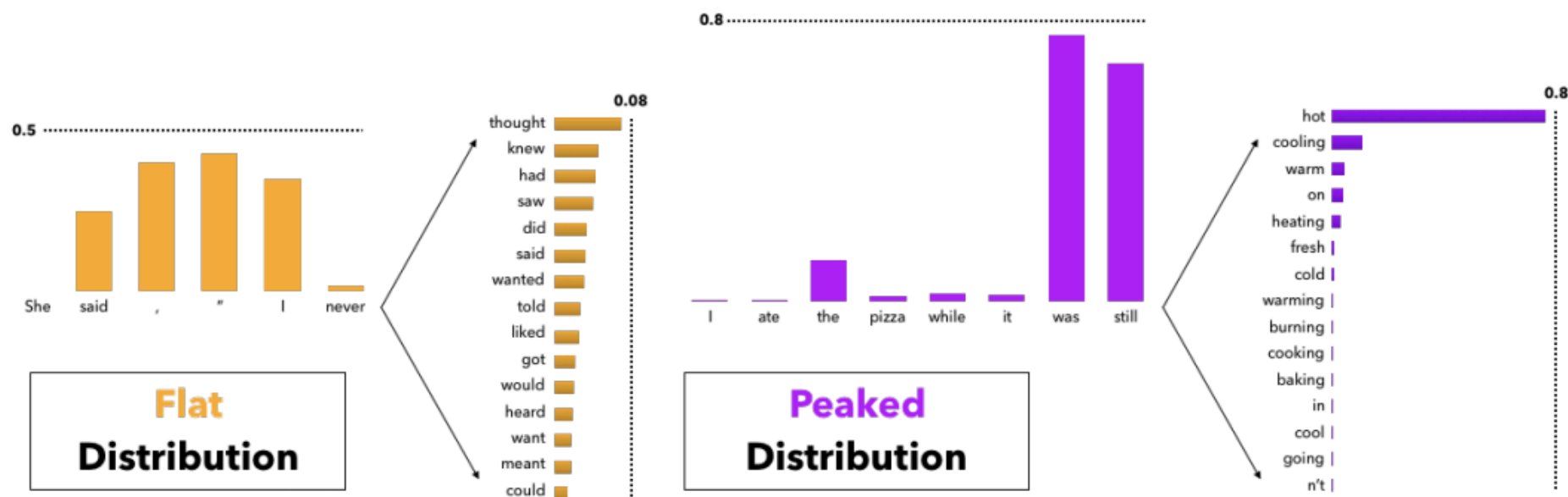
Idea: what if we *sampled* instead of choosing the most likely token, but only from the top k most likely tokens?

Token	Probability
I	0.1
Well	0.05
I'm	0.04
Good	0.04
Poor	0.02
Doin'	0.001
Know	0.0001

K=3
Randomly
sample
from top-3
tokens

Problem with top-k sampling?

Doesn't take into account the probability distribution, always chooses k words to sample from.



Nucleus sampling

Idea: sample from set of tokens which make up top-p probability mass

We want to collect the minimal amount of tokens whose probability mass combined is greater than p

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1})$$

$\geq p$. Amount of probability mass we want to sample from, e.g. 0.9

New probability distribution, used for normalization, e.g. 0.91

$$p' = \sum_{x \in V^{(p)}} P(x|x_{1:i-1})$$

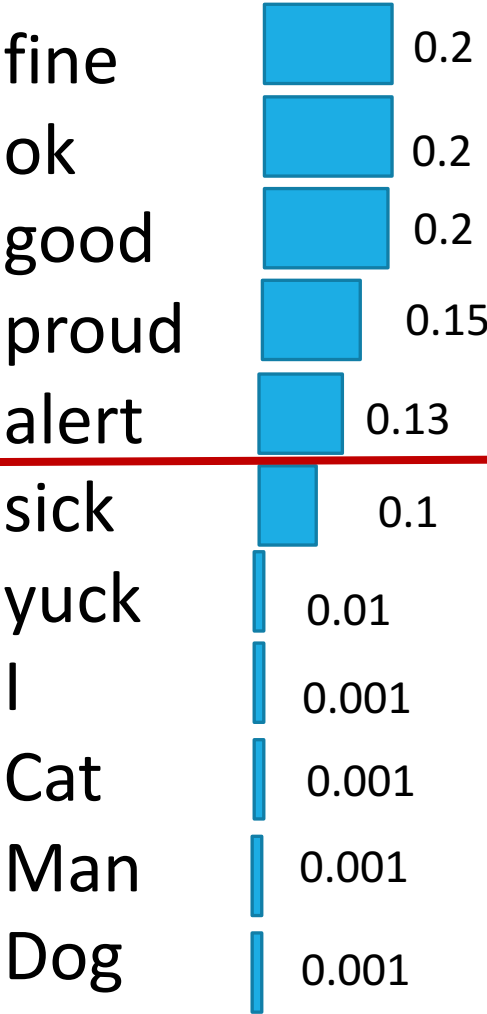
Use new probability distribution to renormalize

Sample from new distribution

$$P'(x|x_{1:i-1}) = \begin{cases} \frac{P(x|x_{1:i-1})}{p'} & \text{if } x \in V^{(p)} \\ 0 & \text{otherwise.} \end{cases}$$

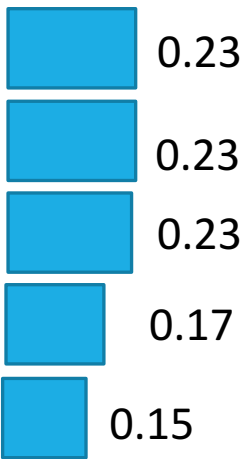
Zero out probability of tokens which fall below probability mass cutoff

Input text: I was feeling



$P = 0.8$

$P' = 0.88$



$\frac{0.2}{0.88} = 0.23$

With top-k sampling (k=3), would have only sampled from “fine, ok, good.” Now, have a more diverse set to sample from.

Outstanding challenges

Coherence across multiple conversation turns

Longer, sustained conversation

Increasing diversity

More complex goals / application areas (e.g. conversational tutor, mental health chatbots)