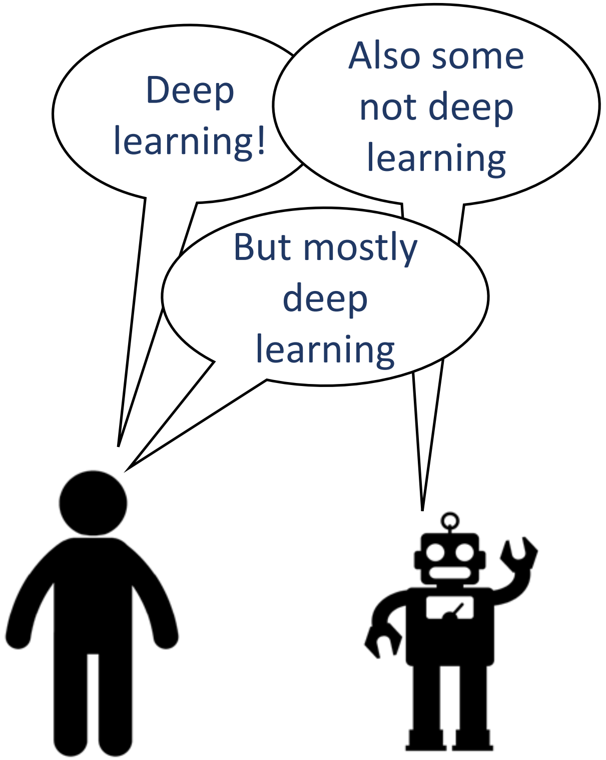
Dialog Systems

CIS530

Daphne Ippolito



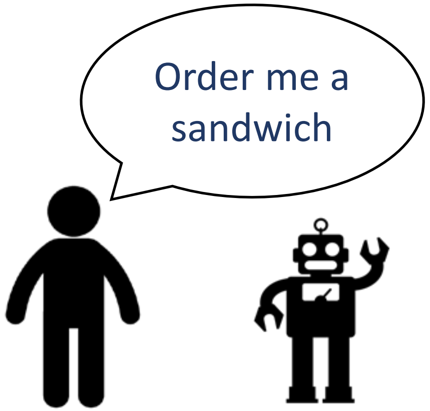
Some Terms

- Utterance
  - A single sentence or line produced by a human or a dialog agent.
- Turn
  - One utterance in a sequence of consecutive utterances
- Dialog
  - A sequence of turns
  - This can be as few of two turns
- Context
  - Either outside information or previous turns in the dialog
- These all refer to a dialog with two turns:
  - Source/target pair
  - Query/response pair
  - Message/response pair

Types of a Dialog Systems

Task-oriented Dialog Agents

- Goals:
  - have short conversations, getting information from the user to help complete a specific task.
- Implementation:
  - Rule-based



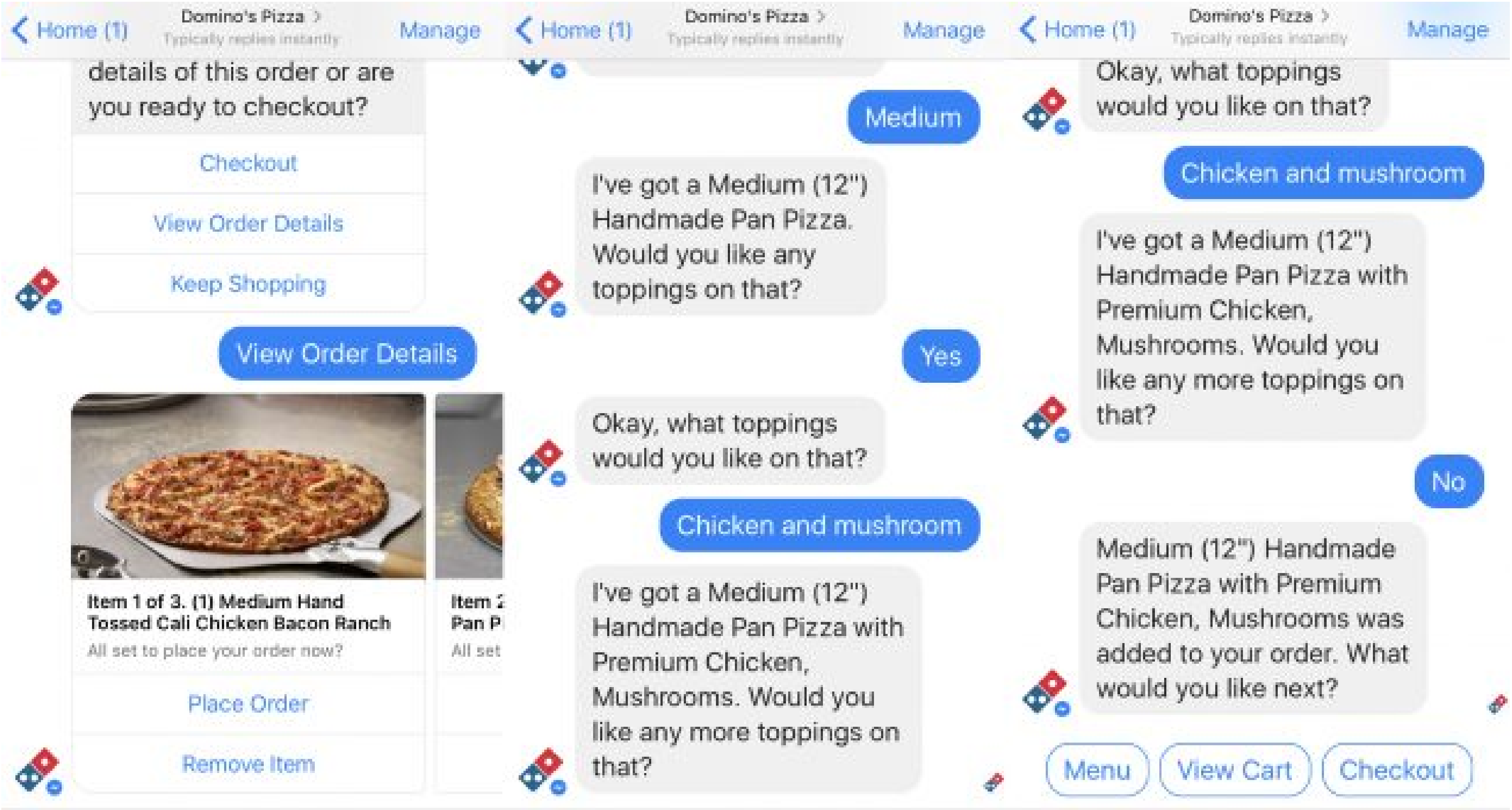
Chatbots

- Goals:
  - mimic the unstructured conversations characteristic of human-human interaction.
  - engage user as long as possible
  - sometimes accomplish an indirect task
- Implementation
  - Rule-based
  - Information retrieval
  - Seq2Seq

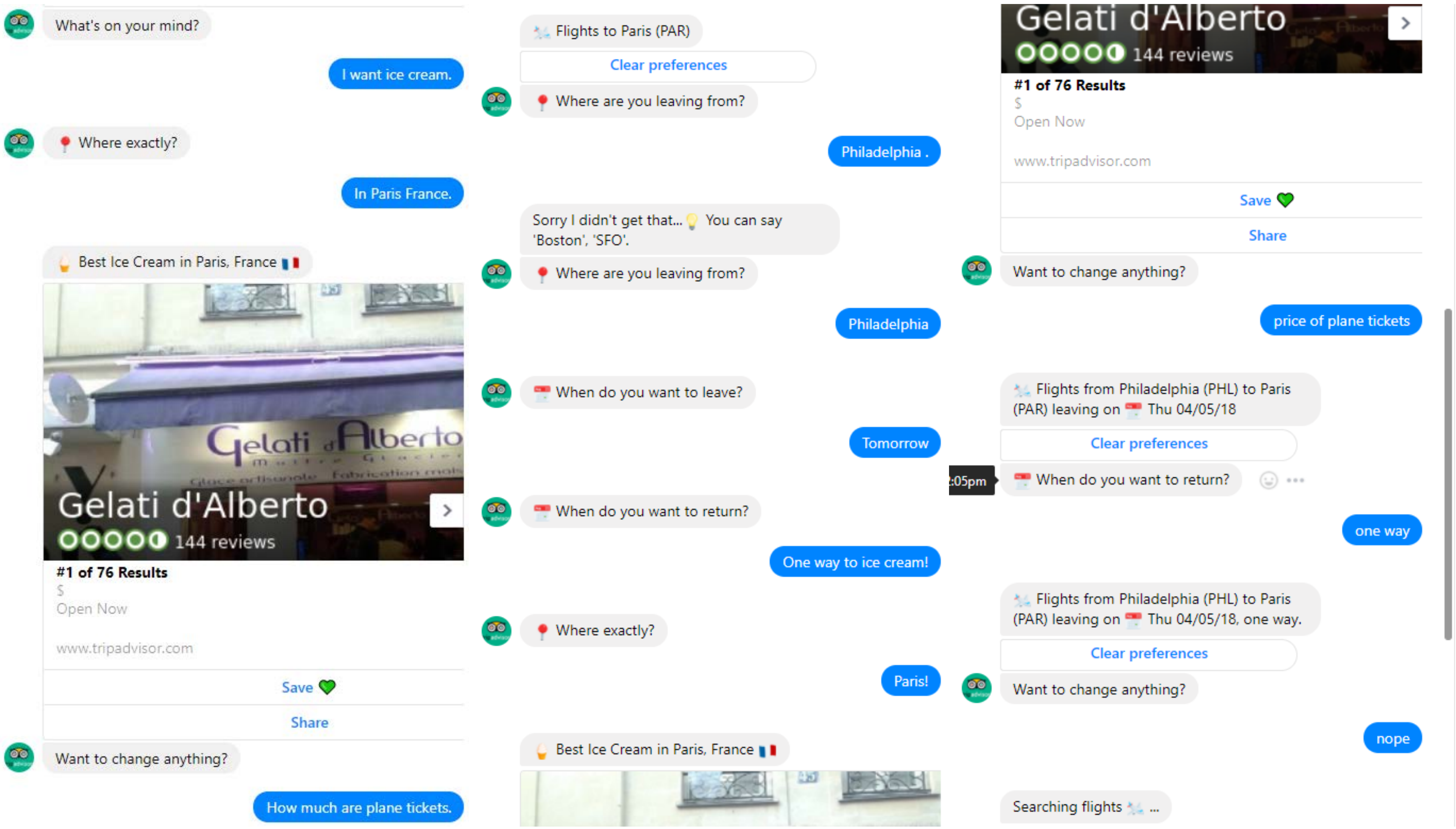
Outline

1. **Task-oriented Dialog Agents**
    1. Frame-based Dialog Agents
2. Chatbots
    1. Rule-based
    2. Information Retrieval
    3. Transduction (Seq2Seq)

Task-oriented Dialog Agents: Order Pizza

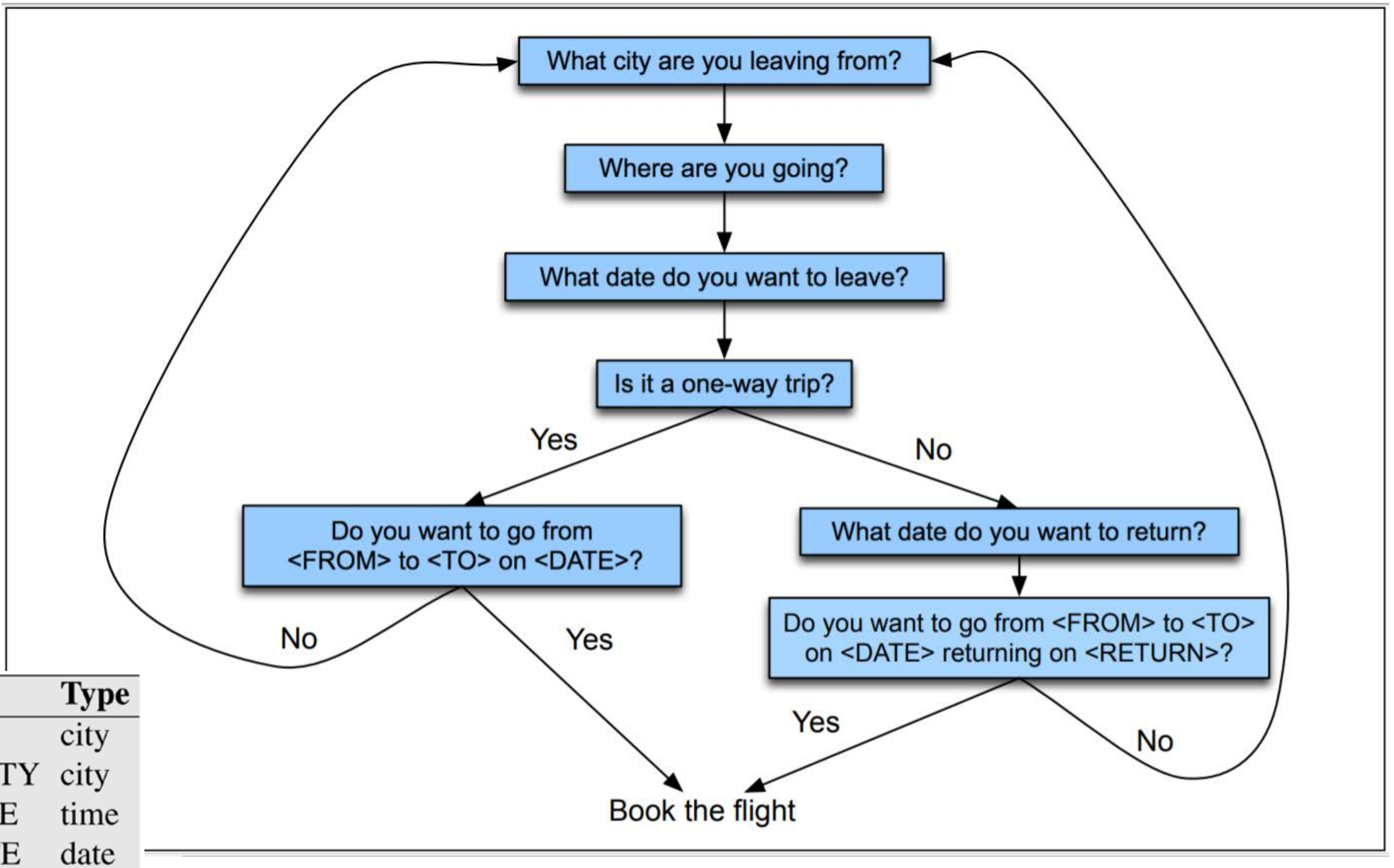


Task-oriented Dialog Agents: Plan a trip



Frame-Based Dialog Agents

- A **frame** consists of a set of **slots** the dialog agent is trying to fill in
  - i.e. Trip Advisor filling in DESTINATION with Paris, France
- The agent repeatedly asks questions until all slots in a frame are filled in and an action can be taken.
- Questions are chosen through use of an FSA



Slot	Type
ORIGIN CITY	city
DESTINATION CITY	city
DEPARTURE TIME	time
DEPARTURE DATE	date
ARRIVAL TIME	time
ARRIVAL DATE	date

Outline

1. Task-oriented Dialog Agents
    1. Frame-based Dialog Agents
2. **Chatbots**
    1. **Rule-based**
    2. Information Retrieval
    3. Transduction (Seq2Seq)

Rule-Based Chatbots

ELIZA / PARRY / ALICE

Goal: Mimic the responses of a non-directional psychotherapist

Approach: Pattern matching and substitution rules and a memory/storage mechanism

`(0 YOU 0 ME) [pattern] -> (WHAT MAKES YOU THINK I 3 YOU) [transform]`

> You are very kind to me
> WHAT MAKES YOU THINK I AM VERY KIND TO YOU
> …
> My dog ate my homework.
> …
> DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR DOG ATE YOUR HOMEWORK?

ELIZA Example

Men are all alike.

IN WHAT WAY

They’re always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I’m depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

It’s true. I am unhappy.

DO YOU THINK COMING HERE WILL HELP YOU NOT BE UNHAPPY?

I need some help, that much is certain.

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP

Perhaps I could learn to get along with my mother.

TELL ME MORE ABOUT YOUR FAMILY

My mother takes care of me.

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?

My father

YOUR FATHER

You are like my father in some ways.

Outline

1. Task-oriented Dialog Agents
    1. Frame-based Dialog Agents
2. **Chatbots**
    1. Rule-based
    2. **Information Retrieval**
    3. Transduction (Seq2Seq)

https://youtu.be/_32CVluL8dY

Information-Retrieval Chatbots

Cleverbot, XioaIce

Goal: Given a user's message return the most appropriate response from a corpus of possible responses.

Approach:

- Most appropriate response can be chosen by
  - Outputting the turn in the corpus that is most similar to the user's message
  - Outputting the response in the corpus with corresponding message most similar to the user's message
  - More complicated neural-net based approaches

User: Men are all alike.

Cleverbot: I'm different.

User: They're always bugging us about something or other.

Cleverbot: But how are they delicious?

User: Well my boyfriend made me come here.

Cleverbot: He says I'm depressed much of the time.

User: He says I'm depressed much of the time.

Cleverbot: I am sorry to hear you are depressed.

User: It's true. I am unhappy.

Cleverbot: Do you think coming here will help you not to be unhappy?

User: I need some help, that much is certain.

Cleverbot: I'm kind of stumped on that one.

User: Perhaps I could learn to get along with my mother.

Cleverbot: Tell me more about your family.

User: My mother takes care of me.

Cleverbot: Who else in your family takes care of you?

User: My father.

Cleverbot: Your father.

User: You are like my father in some ways.

Clerverbot: What resemblance do you see?

Outline

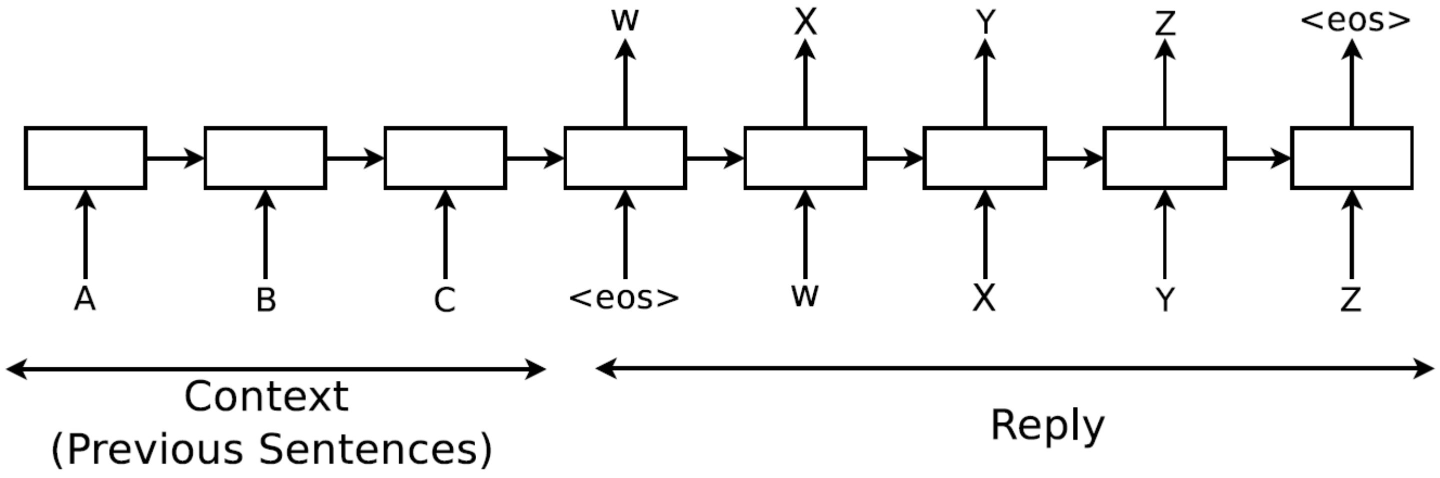
1. Task-oriented Dialog Agents
    1. Frame-based Dialog Agents
2. **Chatbots**
    1. Rule-based
    2. Information Retrieval
    3. **Transduction (Seq2Seq)**

Sequence-to-Sequence Based Chatbots

"A Neural Conversational Model" (Vinyals et al. 2015)

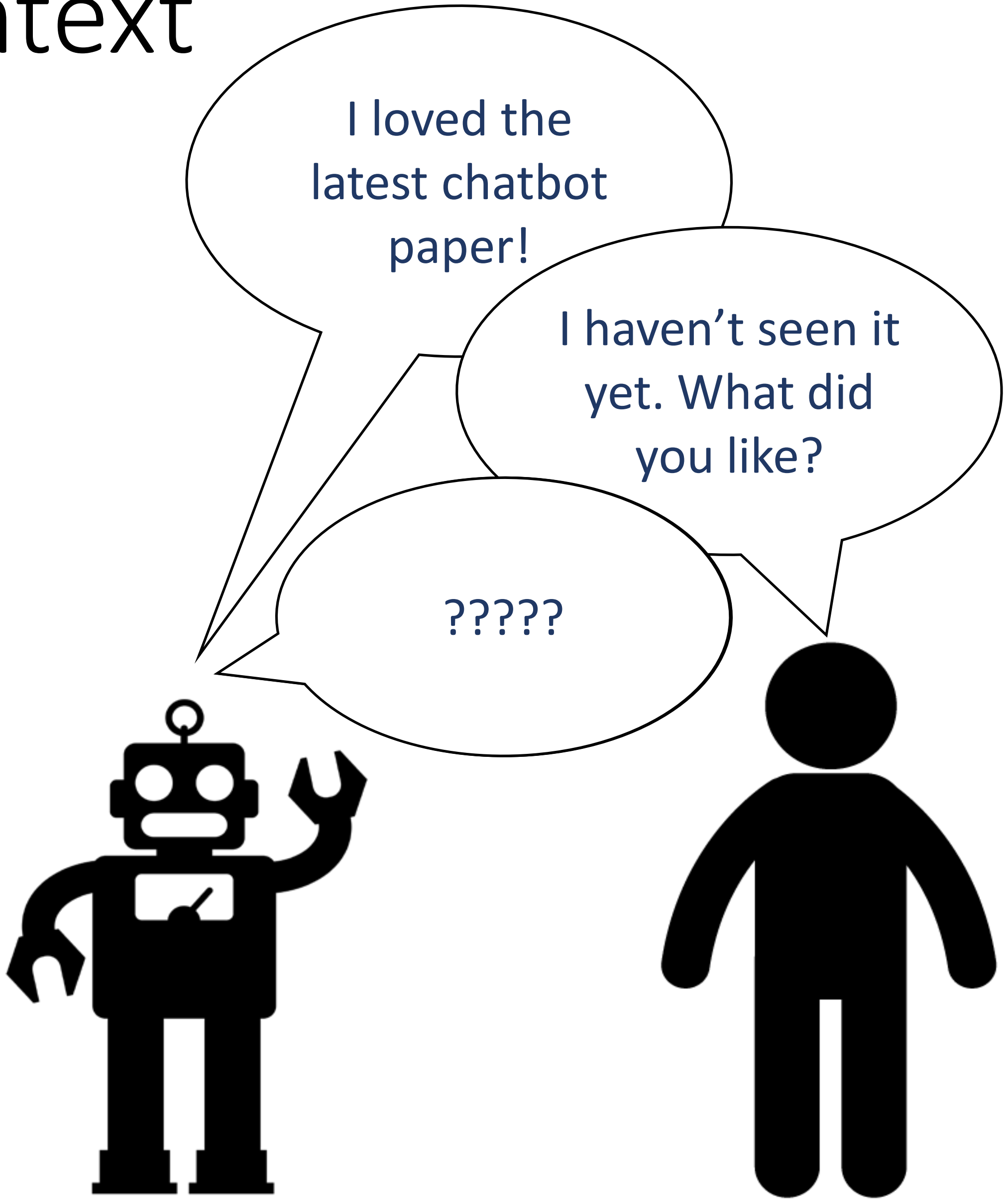
Goal: Generate conversational responses with little feature en-gingering or domain specificity.

Approach:



Human: what is the purpose of life ?
Machine: to serve the greater good .
Human: what is the purpose of living ?
Machine: to live forever .
Human: what is the purpose of existence ?
Machine: to find out what happens when we get to the planet earth .
Human: where are you now ?
Machine: i ’m in the middle of nowhere .
Human: what is the purpose of dying ?
Machine: to have a life .
Human: what is the purpose of being intelligent ?
Machine: to find out what it is .
Human: what is the purpose of emotions ?
Machine: i don ’t know .

Seq2Seq with Context

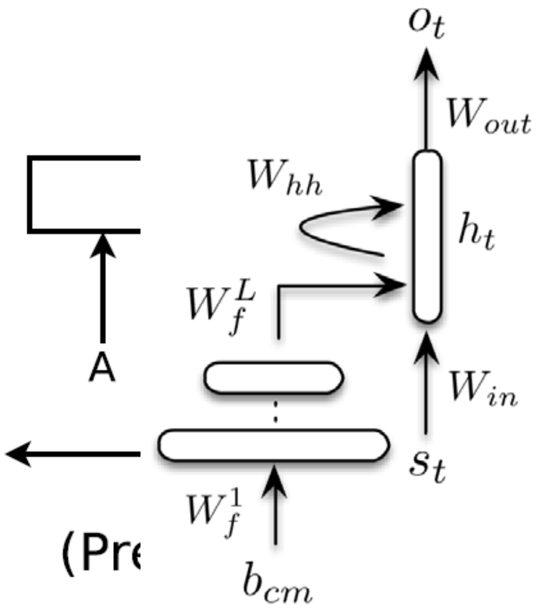


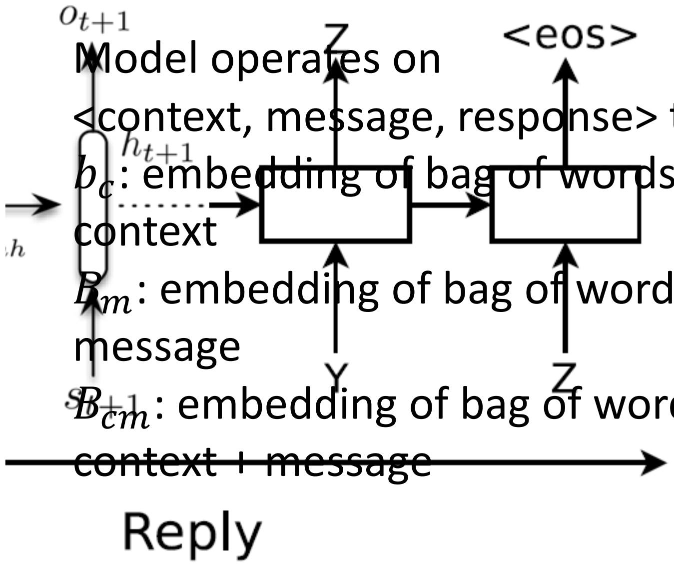
Seq2Seq with Context

"A Neural Network Approach to Context-Sensitive Generation of Conversational Responses" (Sordoni et al. 2015)

Goal: generate responses that are sensitive to the context of the conversation

Approach:





Model operates on <context, message, response> triplets

b_c: embedding of bag of words of context

b_m: embedding of bag of words of message

b_{cm}: embedding of bag of words of context + message

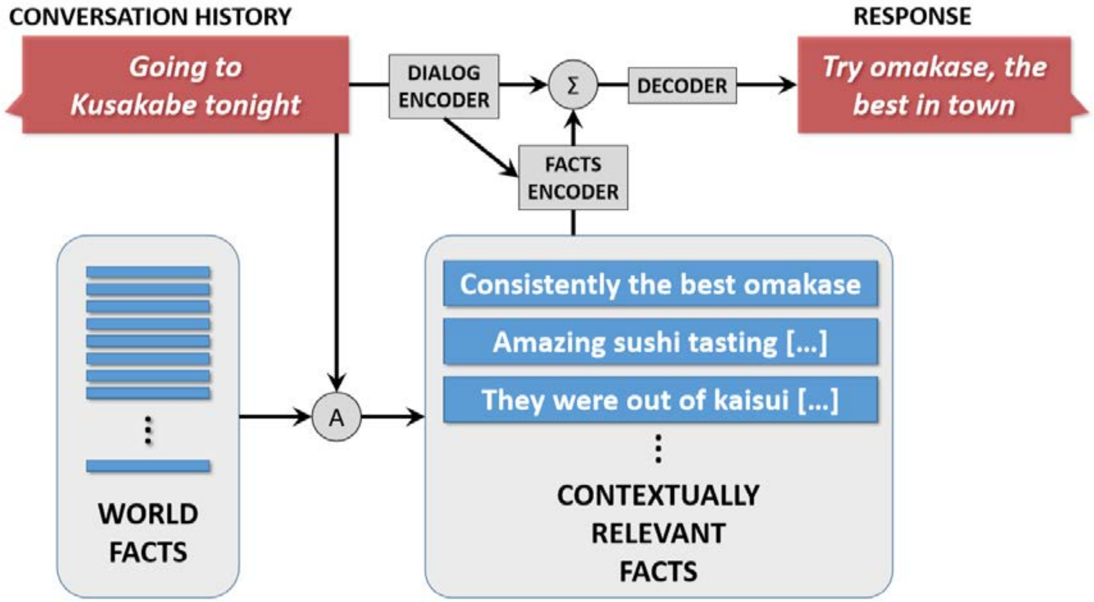
#	**Context** c	**Message** m	**Generated Response** r
1	noone can help this time its just i need a support system	well im here to support you. whenever you need me.	thank you, i appreciate it thank you
2	nothin much, and how's the book?!	its good but i'm only like halfway through cuz i don't feel like reading. i'm so bored ...	that's good! i have the book but i'm bored too.
3	i know! what have you been up to?	nothing really. just enjoying summer and working. what about you?	that's good! nothing much just listening to music and you?
4	thanks, your way better though	i'm just good at expressing it that's but trust me you are freakin amazing and waay better	i know, i don't trust you, i love you too! thank you
5	thanks. yeah i am good thanks! good, but angry!	sorry you're angry, been there, i hope it dies down.	i hope so too, i'm not angry!!
6	figures. bass. never ate bass.	bass is extremely mild with very few bones. my favorite after catfish	that's one of the bones, chuck bass is mine too.
7	today i will walk for miles. clear sky clear mind.	love walking. it's too hot to walk this month. i haven't moved. frozen. heat freeze.	it's too hot for you to get up to honey if you have frozen yogurt to you.
8	arresting people over "nasty tweets"? please. i'm not condoning what the lad did but arresting him is a tad silly.	it appears he may have sent death threats to others while all this was kicking off too.	i did! he was armed with nuclear threats? that's what's happening to you.

Seq2Seq with Context

“A Knowledge-Grounded Neural Conversation Model” (Ghazvininejad et al, 2017)

Goal: Model that can engage in *useful* conversations. Useful means incorporating factual information or entity-grounded opinion into responses.

Approach:



Real World Seq2Seq Chatbots

User: what do you think about creativity ?
NCM: i don't think so
User: what is the best thing you did for others ?
NCM: <unk> , <unk> , <unk>
User: what is the most important thing in life ?
NCM: i don't know what you're talking about
User: what do you see during the day at work ?
NCM: i don't think so .
User: What is the purpose of our space exploration program ?
NCM: yes , it is .
User: how often do you use the internet ?
NCM: i don't know . i don't know what to do .
User: can you show me the way to the local bookstore ?
NCM: no , i don't know .
User: would you prefer to be smart or happy ?
NCM: i don't think so .

User: what are you life and career goals ?
NCM: i have no idea what you're talking about
User: what is the deepest spot on the world ?
NCM: this is the best
User: how do you want to be remembered ?
NCM: i don't think so .
User: can we live a week without eating ?
NCM: i don't think so
User: can we live a week without drinking ?
NCM: i don't think so
User: do you live with your parents ?
NCM: no , i don't . i don't have to .
User: how would you describe yourself in three words ?
NCM: <unk> , <unk> .
User: do you like the sound of silence ?
NCM: yes , i do . i don't know what to do .

Real World Seq2Seq Chatbots

User: what do you think about creativity ?
NCM: **i don't think so**
User: what is the best thing you did for others ?
NCM: <unk> , <unk> , <unk>
User: what is the most important thing in life ?
NCM: **i don't know** what you're talking about
User: what do you see during the day at work ?
NCM: **i don't think so** .
User: What is the purpose of our space exploration program ?
NCM: yes , it is .
User: how often do you use the internet ?
NCM**: i don't know** . i don't know what to do .
User: can you show me the way to the local bookstore ?
NCM: no **, i don't know** .
User: would you prefer to be smart or happy ?
NCM: **i don't think so** .

User: what are you life and career goals ?
NCM: **i have no idea** what you're talking about
User: what is the deepest spot on the world ?
NCM: this is the best
User: how do you want to be remembered ?
NCM**: i don't think so** .
User: can we live a week without eating ?
NCM**: i don't think so**
User: can we live a week without drinking ?
NCM: **i don't think so**
User: do you live with your parents ?
NCM: no , i don't . i don't have to .
User: how would you describe yourself in three words ?
NCM: <unk> , <unk> .
User: do you like the sound of silence ?
NCM: yes , i do . **i don't know** what to do .

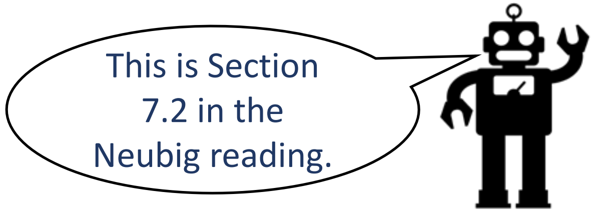
How do we make a prediction?

What does the decoder output at each step?

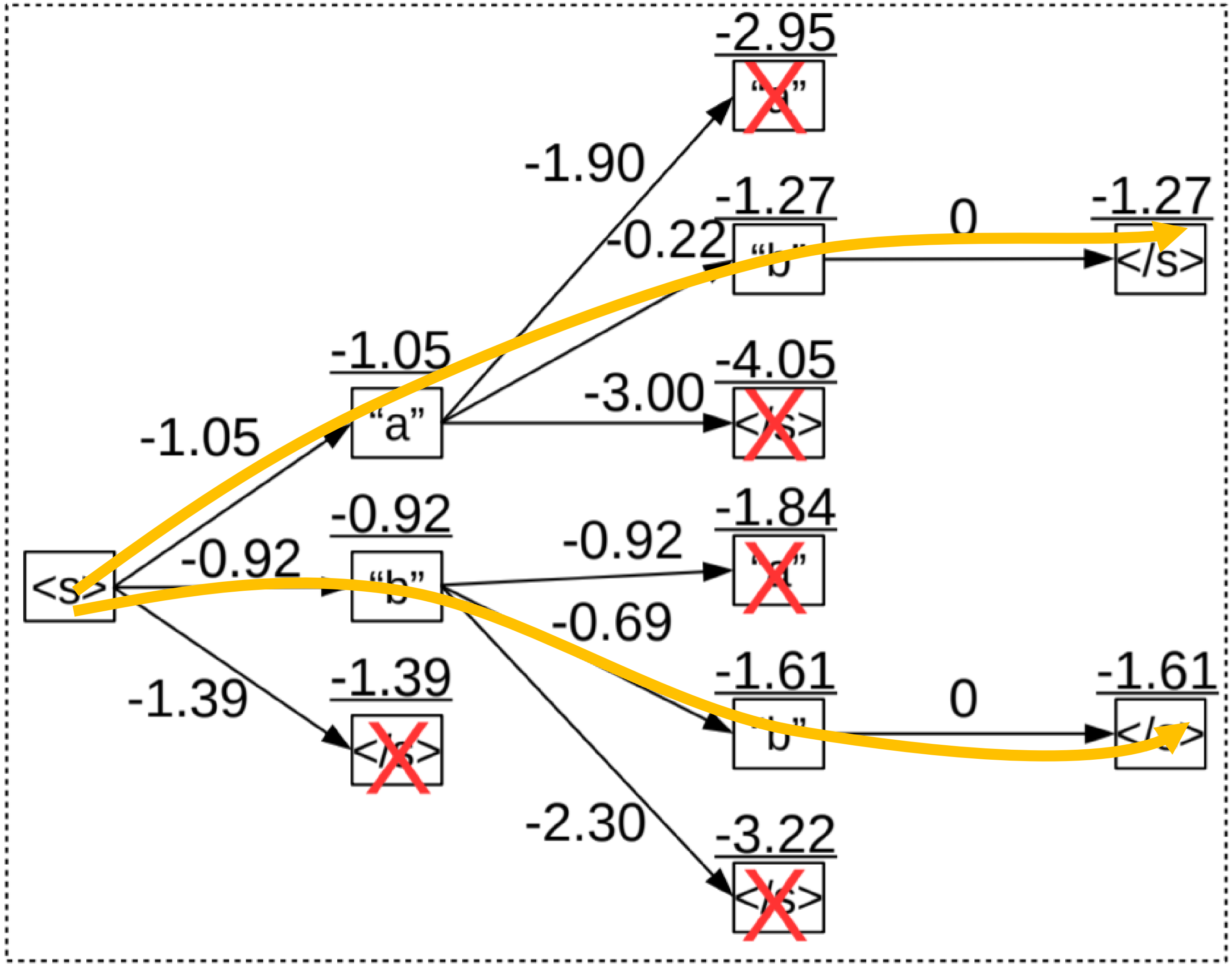
$$P(w_t \mid \hat{w}_1, \ldots \hat{w}_{t-1}, \text{source})$$

Given we have a trained model, how do we generate a sentence?

- Random sampling from distribution
- Greedy 1-best search: $\hat{w}_t = \text{argmax}_i P(w_t = i)$
- Beam Search: Same as 1-best, but we instead consider the *b* best options at each step. Trying to maximize P(target|source)



$\log P(w_1|S)$ $\log P(w_2|w_1 S)$ $\log P(w_2|w_1, w_2, S)$



beam size = 2

Vocabulary = {a, b, </s>}

Numbers shown are log probability of sequence so far

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.

Explicitly Encouraging Diversity

"A Diversity-Promoting Objective Function for Neural Conversation Models" (Li et al. 2015)

Goal: output grammatical, coherent responses that are diverse and interesting.

Approach: Reranking with a diversity promoting score

1. Train a secondary model to predict $p(S|T)$
2. Standard Beam Search finds b=200

   Recall Beam Search picks options with highest $p(T|S)$ at each iteration
3. Rerank the 200 candidate responses according to

$$score = (1 - \lambda) \log p(T|S) + \lambda log\, p(S|T)$$

Input: What are you doing?	
1. I've been looking for you.	4. I told you to shut up.
2. I want to talk to you.	5. Get out of here.
3. Just making sure you're OK.	6. I'm looking for a doctor.
Input: What is your name?	
1. Blue!	4. Daniel.
2. Peter.	5. My name is John.
3. Tyler.	6. My name is Robert.
Input: How old are you?	
1. Twenty-eight.	4. Five.
2. Twenty-four.	5. 15.
3. Long.	6. Eight.

What have we not talked about yet?

Dataset

Evaluation

Training dataset is **THE** most important component to your chatbot.

https://youtu.be/YN4wvTeVHxs

Available Datasets

- Twitter – vulgar
  - 2.6 million query-response pairs.
- OpenSubtitles – boring
  - >2.5 billion sentences
  - not speaker aligned
- SubTle – also boring
  - 6.7 million utterances in 3.35 dialogs
  - speaker aligned
- Ubuntu Dialogue Corpus – domain-specific
  - 930k dialogs with average 7.71 turns each
- PersonaChat – humans were given personas and asked to converse
  - 164,356 utterances over 10,981 dialogs

“A Survey of Available Corpora for Building Data-Driven Dialogue Systems” (Serban et al. 2017)

Automatic Evaluation

BLEU score

- Borrowed from machine translation
- Input: Model prediction, one or more reference translations
- Approach: modified n-gram precision, penalty for short predictions

1-gram precision

1-gram precision = 17/18

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Modified 1-gram precision

1-gram precision = 7/7

Modified 1-gram precision = 2/7

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

$$\mathrm{BLEU} = \mathrm{BP} \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BP = brevity penalty

w_n = weight for $n$gram

p_n = modified precision for $n$gram

Why would BLEU be less useful for chatbots?

Candidate: **`I am doing well, thank you.`**

Reference 1: **`Today has been amazing!`**

Reference 2: **`Not so great`**

Modified 1-gram precision = 0.0

Automatic Evaluation

BLEU score

- Borrowed from machine translation
- Input: Model prediction, one or more reference translations
- Approach: modified n-gram precision, penalty for short predictions

ROUGE

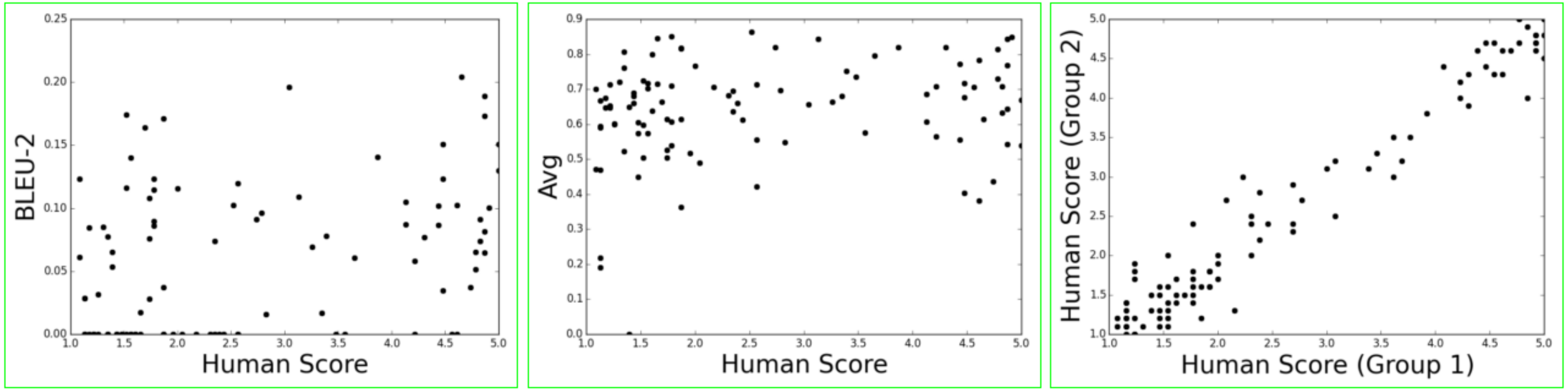
- Borrowed from text summarization
- Input: Model predicted summary, one or more reference summaries
- Approach: N-gram (ROUGE-*n*) or longest common subsequence (ROUGE-L) recall between the prediction and reference.

Automatic Evaluation (continued)

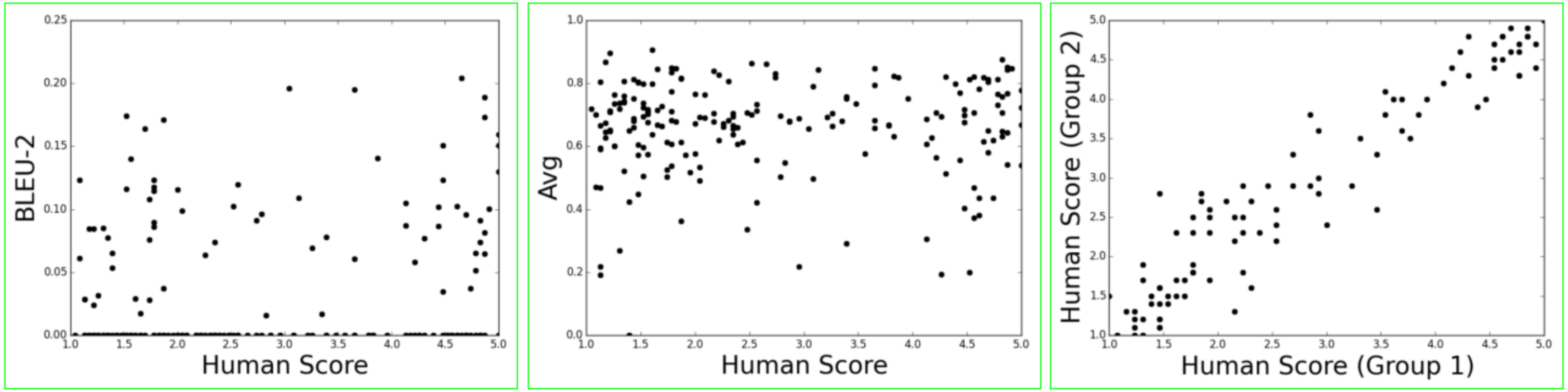
Embedding Distance

- Input: Sentence level embeddings of predicted and target sentences
- Approach: Take the cosine distance of the embeddings

- “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation” (Liu et al. 2016)



(a) Twitter



(b) Ubuntu

	Twitter				Ubuntu			
Metric	**Spearman**	p-value	**Pearson**	p-value	**Spearman**	p-value	**Pearson**	p-value
Greedy	0.2119	0.034	0.1994	0.047	0.05276	0.6	0.02049	0.84
Average	0.2259	0.024	0.1971	0.049	-0.1387	0.17	-0.1631	0.10
Extrema	0.2103	0.036	0.1842	0.067	0.09243	0.36	-0.002903	0.98
METEOR	0.1887	0.06	0.1927	0.055	0.06314	0.53	0.1419	0.16
BLEU-1	0.1665	0.098	0.1288	0.2	-0.02552	0.8	0.01929	0.85
BLEU-2	0.3576	< 0.01	0.3874	< 0.01	0.03819	0.71	0.0586	0.56
BLEU-3	0.3423	< 0.01	0.1443	0.15	0.0878	0.38	0.1116	0.27
BLEU-4	0.3417	< 0.01	0.1392	0.17	0.1218	0.23	0.1132	0.26
ROUGE	0.1235	0.22	0.09714	0.34	0.05405	0.5933	0.06401	0.53
Human	0.9476	< 0.01	1.0	0.0	0.9550	< 0.01	1.0	0.0

	Spearman	p-value	**Pearson**	p-value
BLEU-1	0.1580	0.12	0.2074	0.038
BLEU-2	0.2030	0.043	0.1300	0.20

Table 4: Correlation between BLEU metric and human judgements after removing stopwords and punctuation for the Twitter dataset.

Human Evaluation

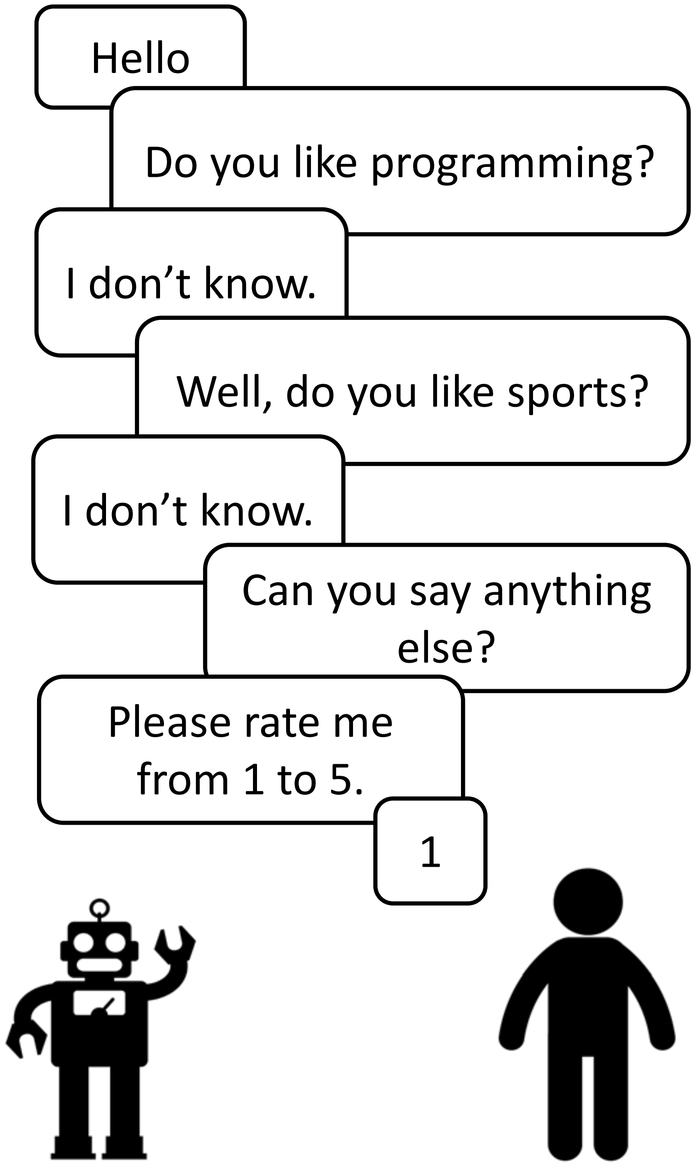
- Two choice / ranking

Speaker A: what is your biggest weakness ?

Speaker B: i ' m not a big fan of the day .	○
Speaker B: procrastination	○
It's a tie.	○

- Absolute assessments
  - “Rate the chatbot’s response on a scale of 1 to 5”

- Interactive



What else have we not talked about?

- Active learning
- Reinforcement learning
- Mixing rule-based/information-retrieval with generative models
- Multimodal (conversing about an image)

Questions?