CS395T: Structured Models for NLP
Lecture 23: Dialogue

THE UNIVERSITY OF TEXAS AT AUSTIN

Greg Durrett

Extractive Summarization: Bigram Recall

- Count number of *documents* each bigram occurs in to measure importance

score(*massive earthquake*) = 3

score(*magnitude 7.3*) = 2

score(*six killed*) = 2

score(*Iraqi capital*) = 1

- Find summary that maximizes the score of bigrams it covers
- ILP formulation: c and s are indicator variables indexed over concepts (bigrams) and sentences, respectively

$$\text{Maximize: } \sum_i w_i c_i$$

$$\text{Subject to: } \sum_j l_j s_j \leq L$$

$$s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

“set c_i to 1 iff some sentence that contains it is included”

sum of included sentences’ lengths can’t exceed L

Gillick and Favre (2009)

Compressive Summarization

Maximize: $\sum_i w_i c_i$

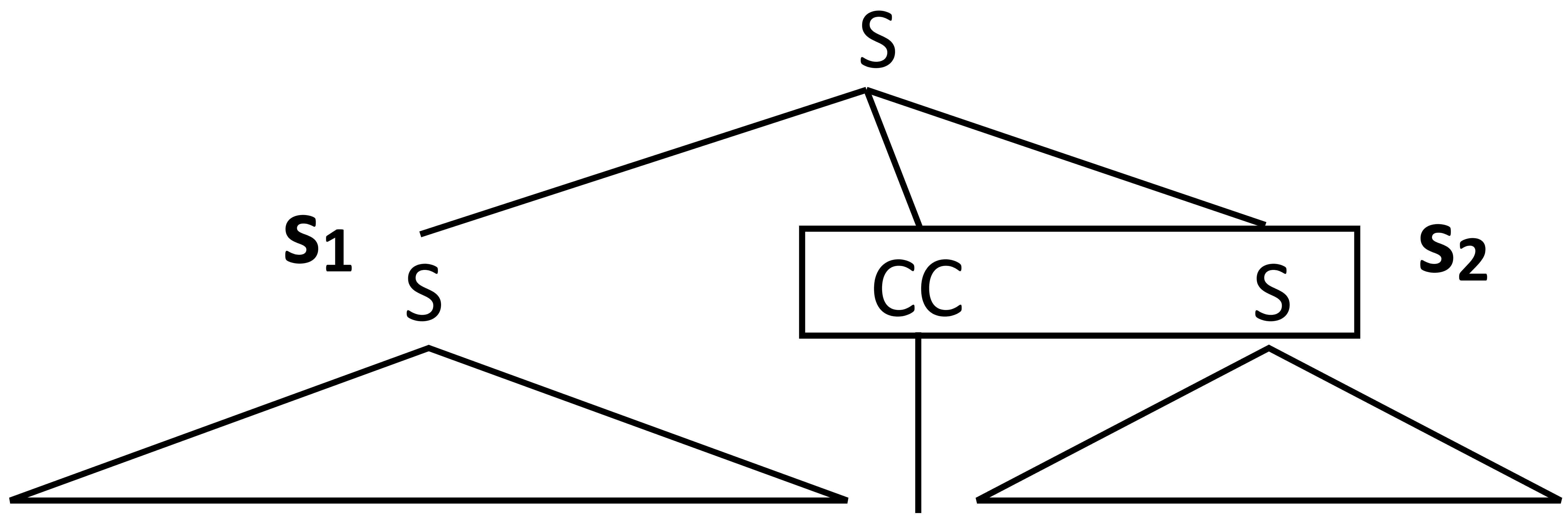
Subject to: $\sum_j l_j s_j \leq L$

$$s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

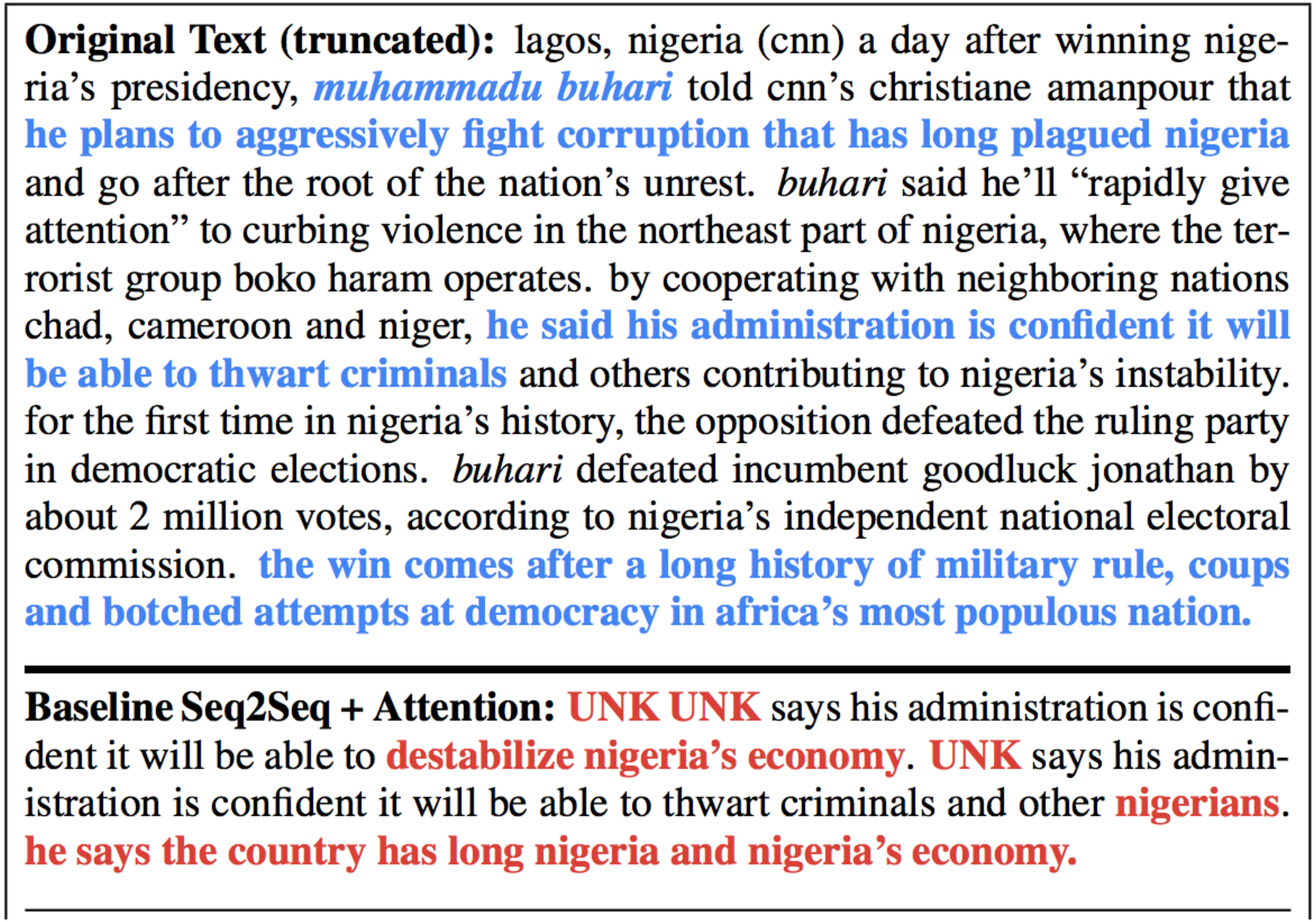
- Now s_j variables are nodes or sets of nodes in the parse tree
- New constraint: $s_2 \leq s_1$

  “s_1 is a prerequisite for s_2”



At least six people were killed and many others injured

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.**

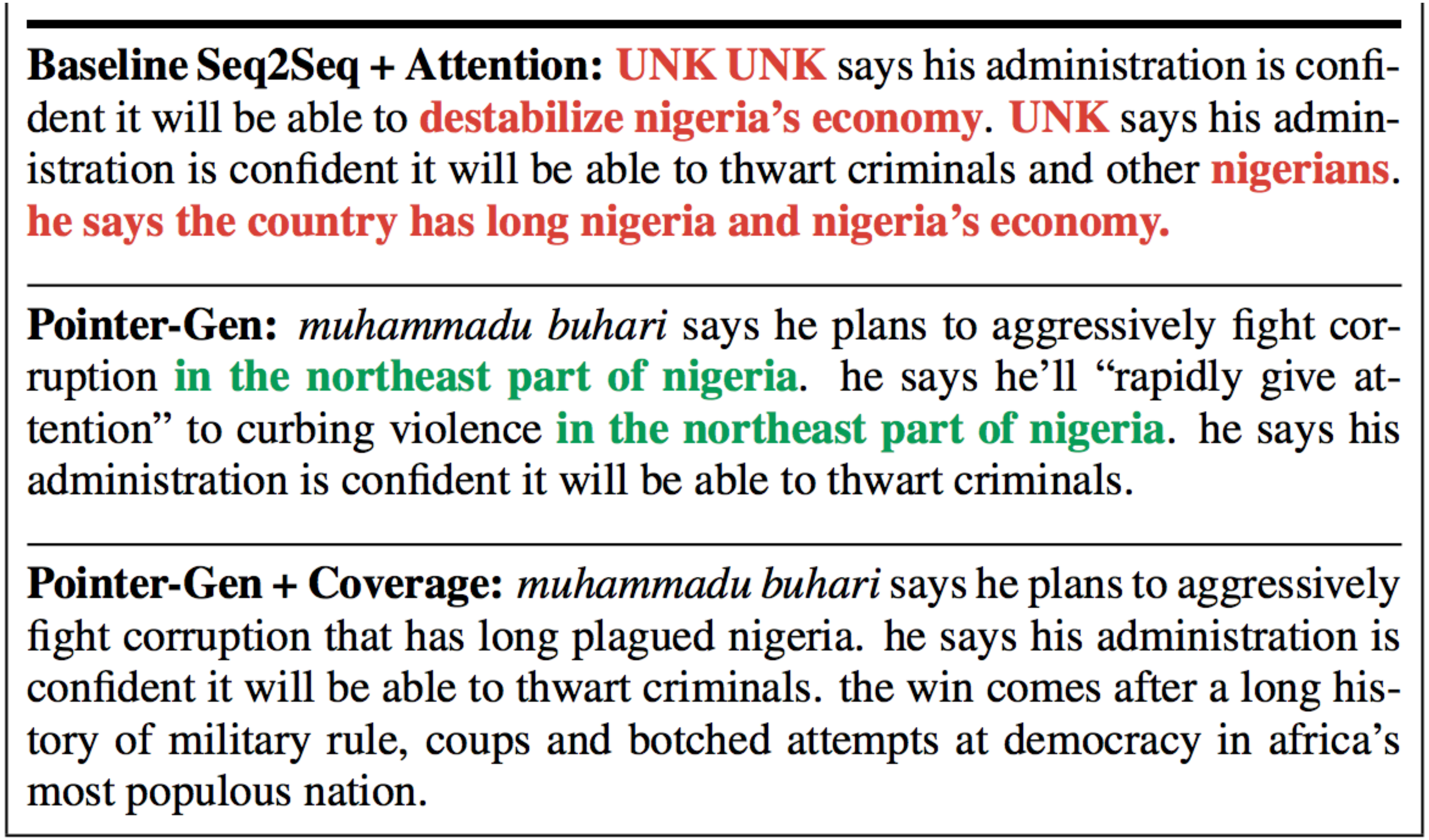
Baseline Seq2Seq + Attention: **UNK UNK** says his administration is confident it will be able to **destabilize nigeria's economy**. **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

- Problems: unknown words, inaccuracies

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.

- Things might still go wrong, no way of preventing this…

See et al. (2017)

This Lecture

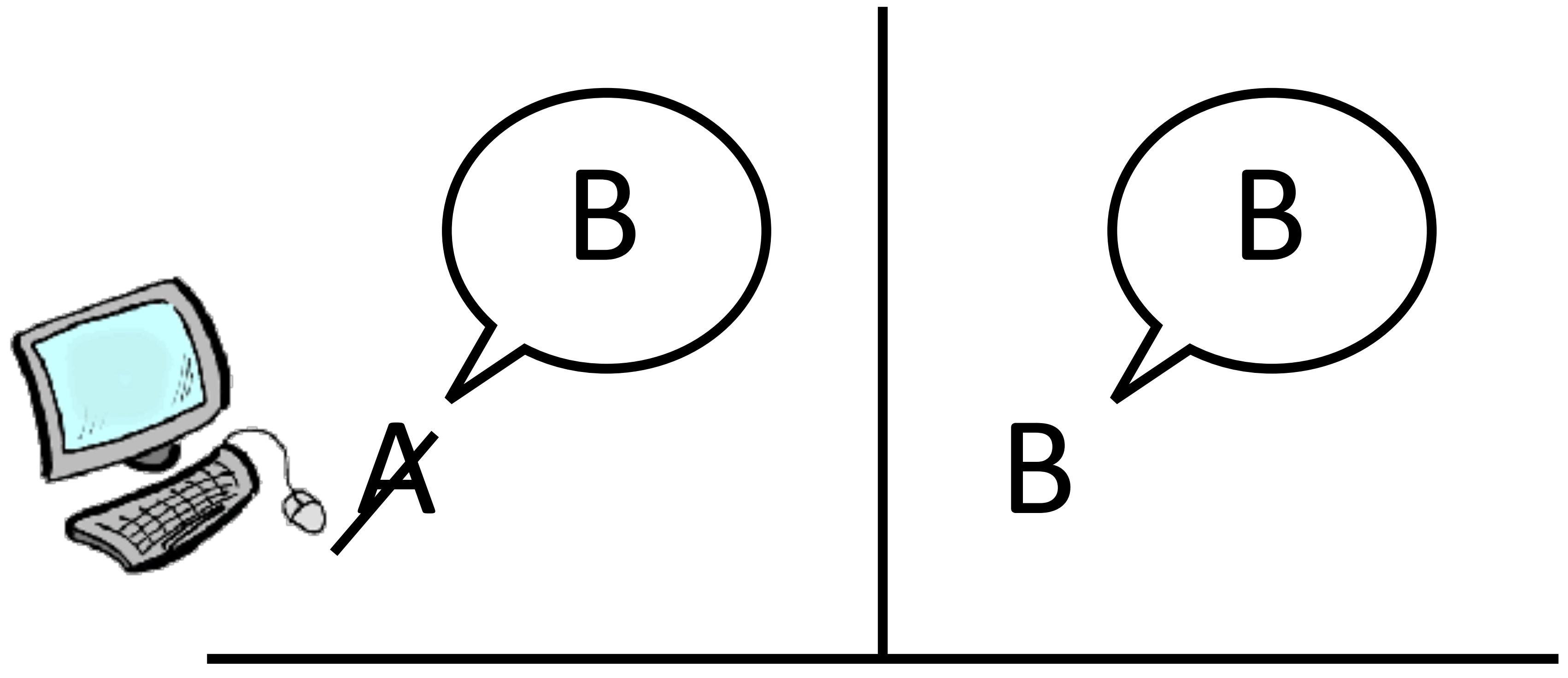
- Chatbot dialogue systems
- Task-oriented dialogue
- Other dialogue applications

Chatbots

Turing Test (1950)

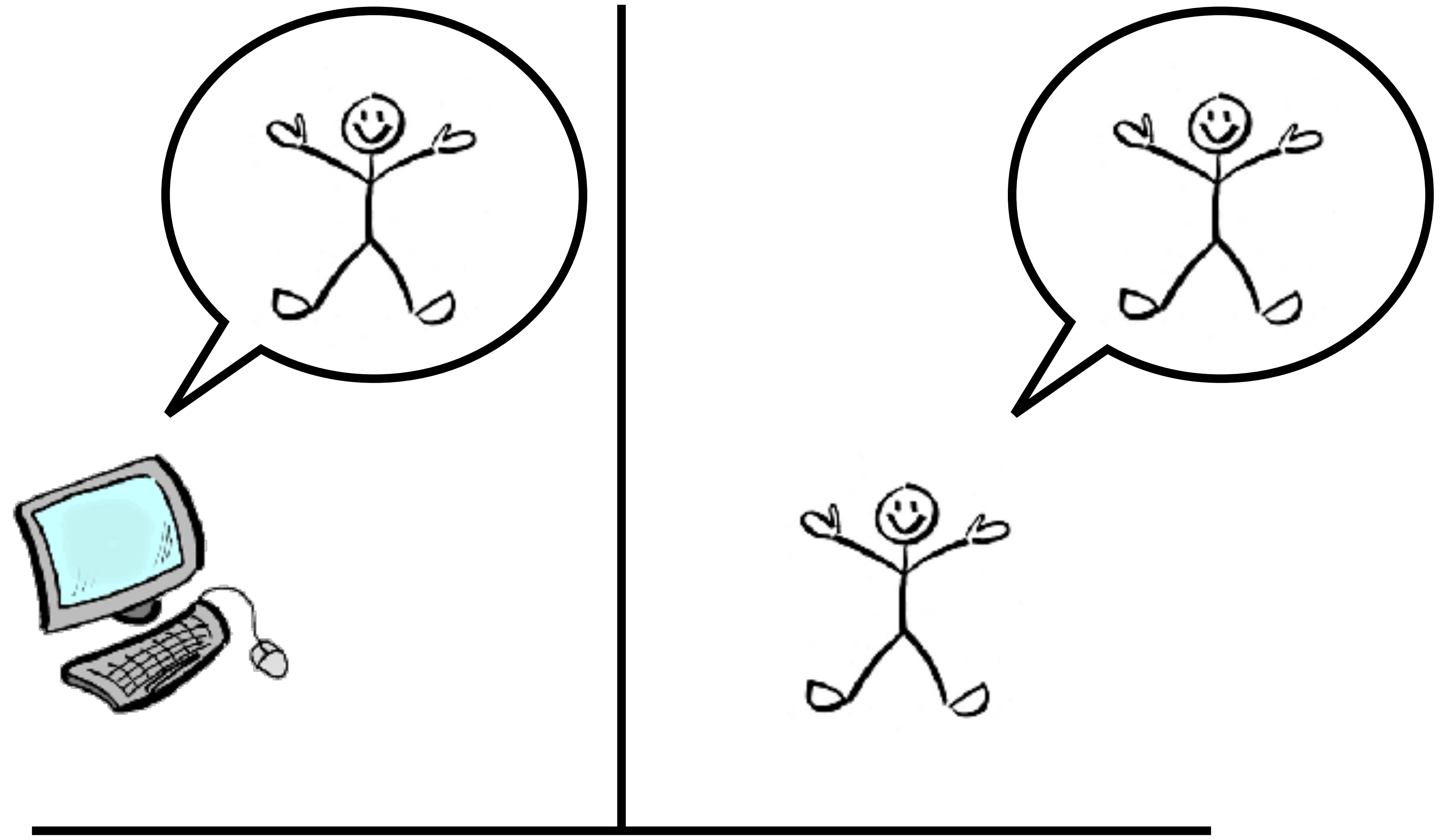
- Imitation game: A and B are locked in rooms and answer C's questions via typewriter. Both are trying to act like B

Original Interpretation:



C trained judge

Standard Interpretation:



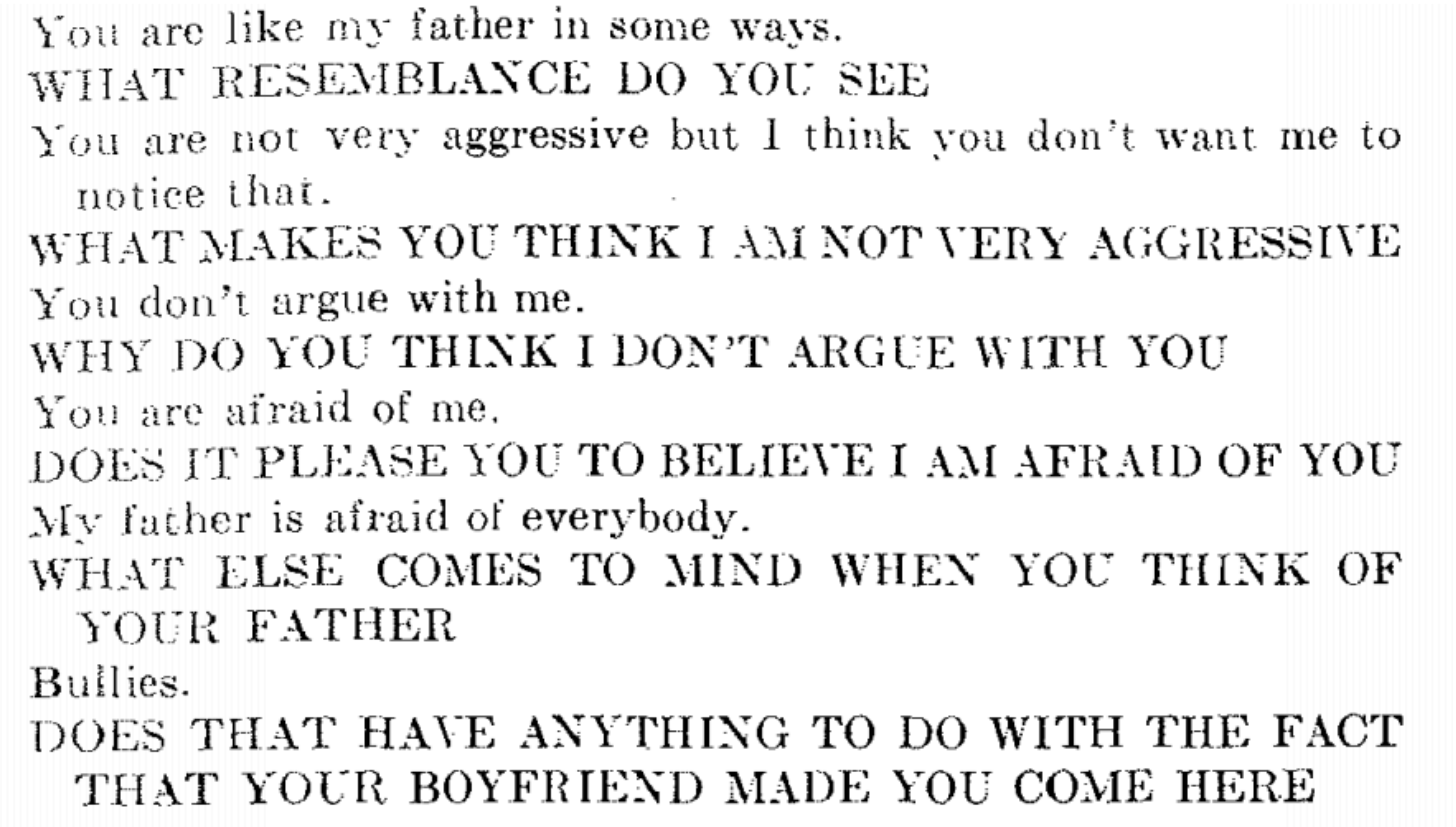
C trained judge

- 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

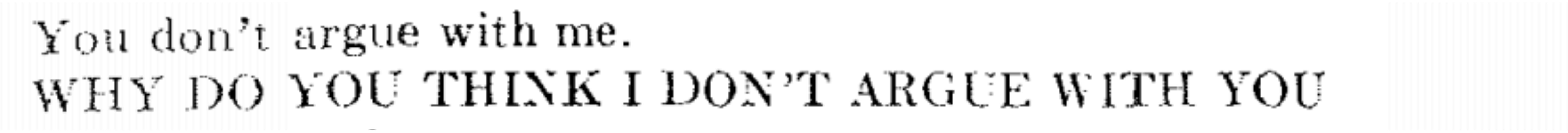
Weizenbaum (1966)

ELIZA

- Identify keyword, identify context, apply transformation rule

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

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

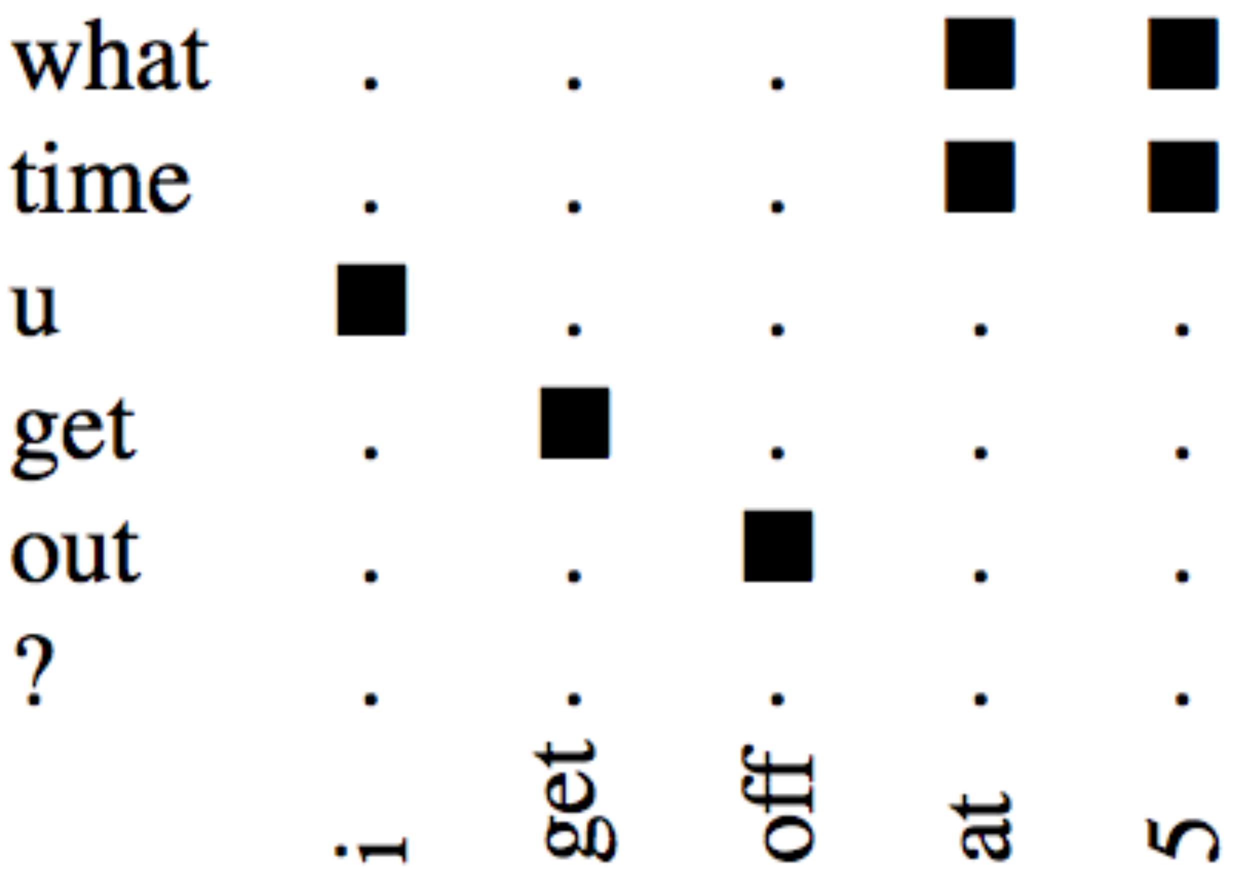
Cleverbot

- Carpenter (1986), online system built in 2006
- "Nearest neighbors": human says statement A, find a human response in human-human or human-computer chats to statement A, repeat that
- Can often give sensible answers, but the bot doesn't really impose high-level discourse structure

User: Hi there cleverbot.
Cleverbot: How are you?
User: Doing okay. what plans do you have today?
Cleverbot: I do revision.
User: What are you revising for?
Cleverbot: Maths history.
User: How old are you?
Cleverbot: That's none of your business.
User: What is maths history?
Cleverbot: I don't know, farming maybe

Data-Driven Approaches

- Can treat as a machine translation problem: “translate” from current utterance to next one

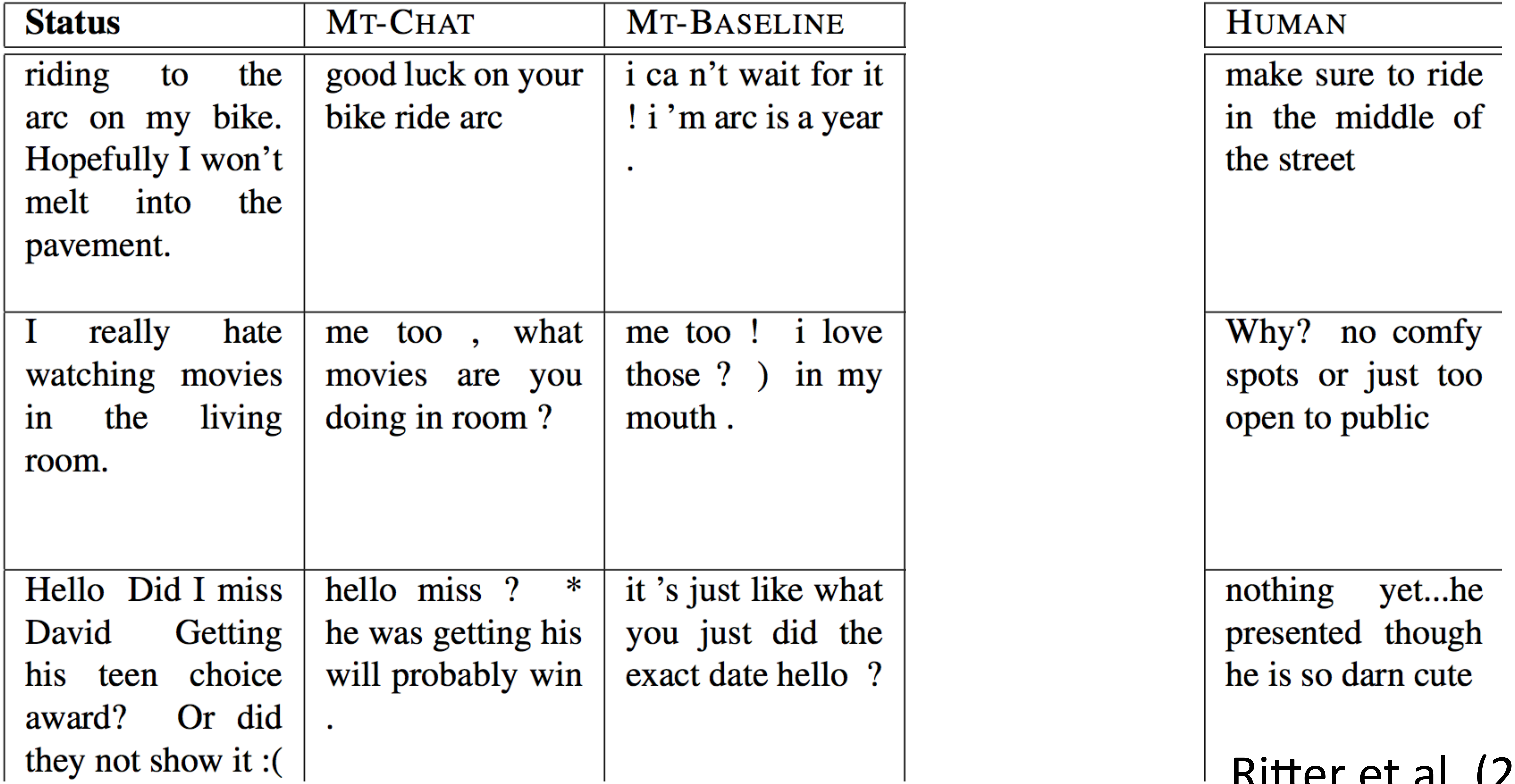


	i	get	off	at	5
what	.	.	.	■	■
time	.	.	.	■	■
u	■	.	.	.	.
get	.	■	.	.	.
out	.	.	■	.	.
?	.	.	.	.	.

- Filter the data, use statistical measures to prune extracted phrases to get better performance

Ritter et al. (2011)

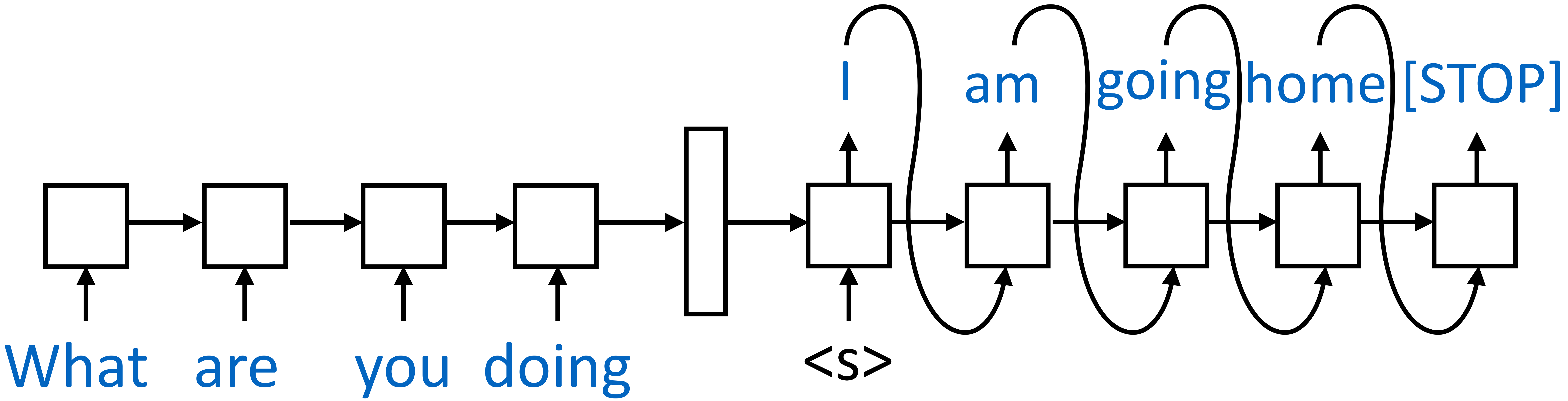
Data-Driven Approaches



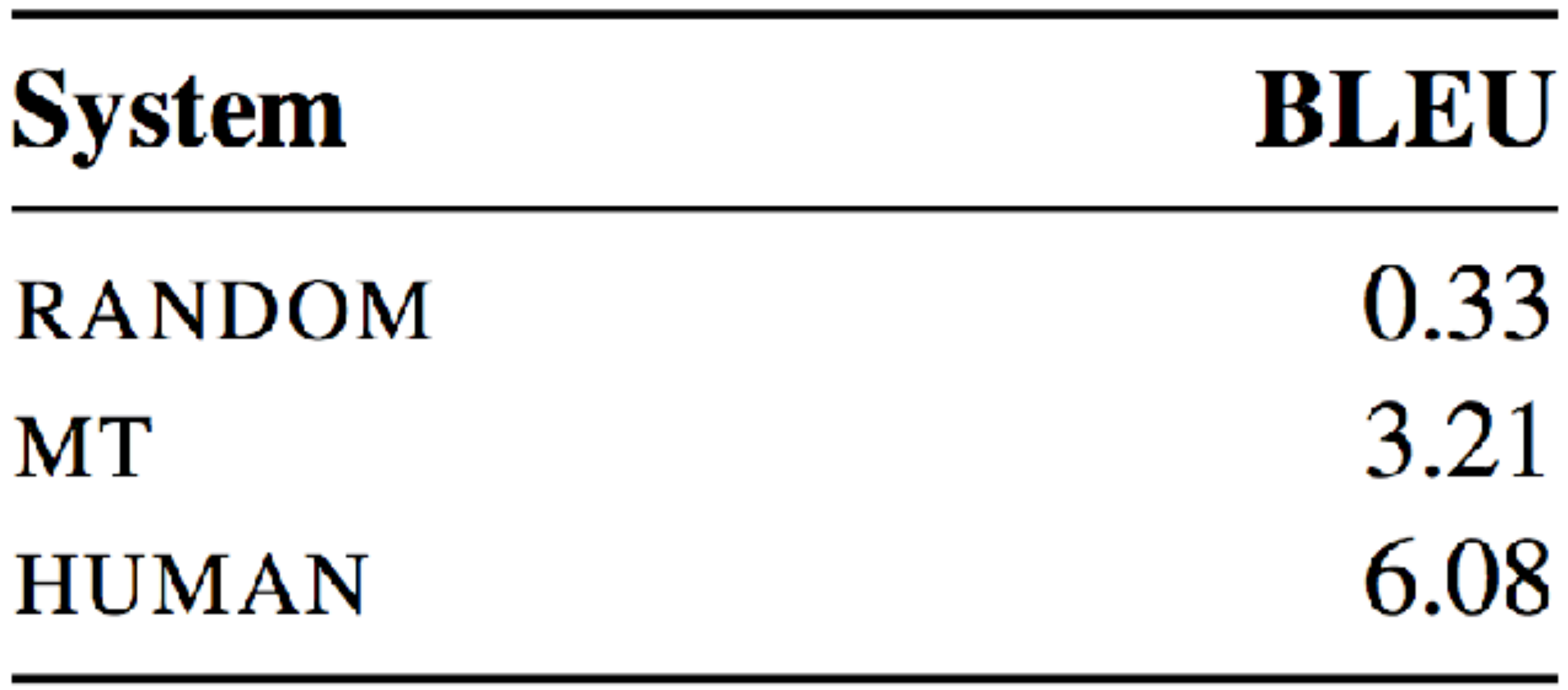
Status	MT-CHAT	MT-BASELINE	HUMAN
riding to the arc on my bike. Hopefully I won't melt into the pavement.	good luck on your bike ride arc	i ca n't wait for it ! i 'm arc is a year .	make sure to ride in the middle of the street
I really hate watching movies in the living room.	me too , what movies are you doing in room ?	me too ! i love those ?) in my mouth .	Why? no comfy spots or just too open to public
Hello Did I miss David Getting his teen choice award? Or did they not show it :(	hello miss ? * he was getting his will probably win .	it 's just like what you just did the exact date hello ?	nothing yet...he presented though he is so darn cute

Ritter et al. (2011)

Seq2seq models



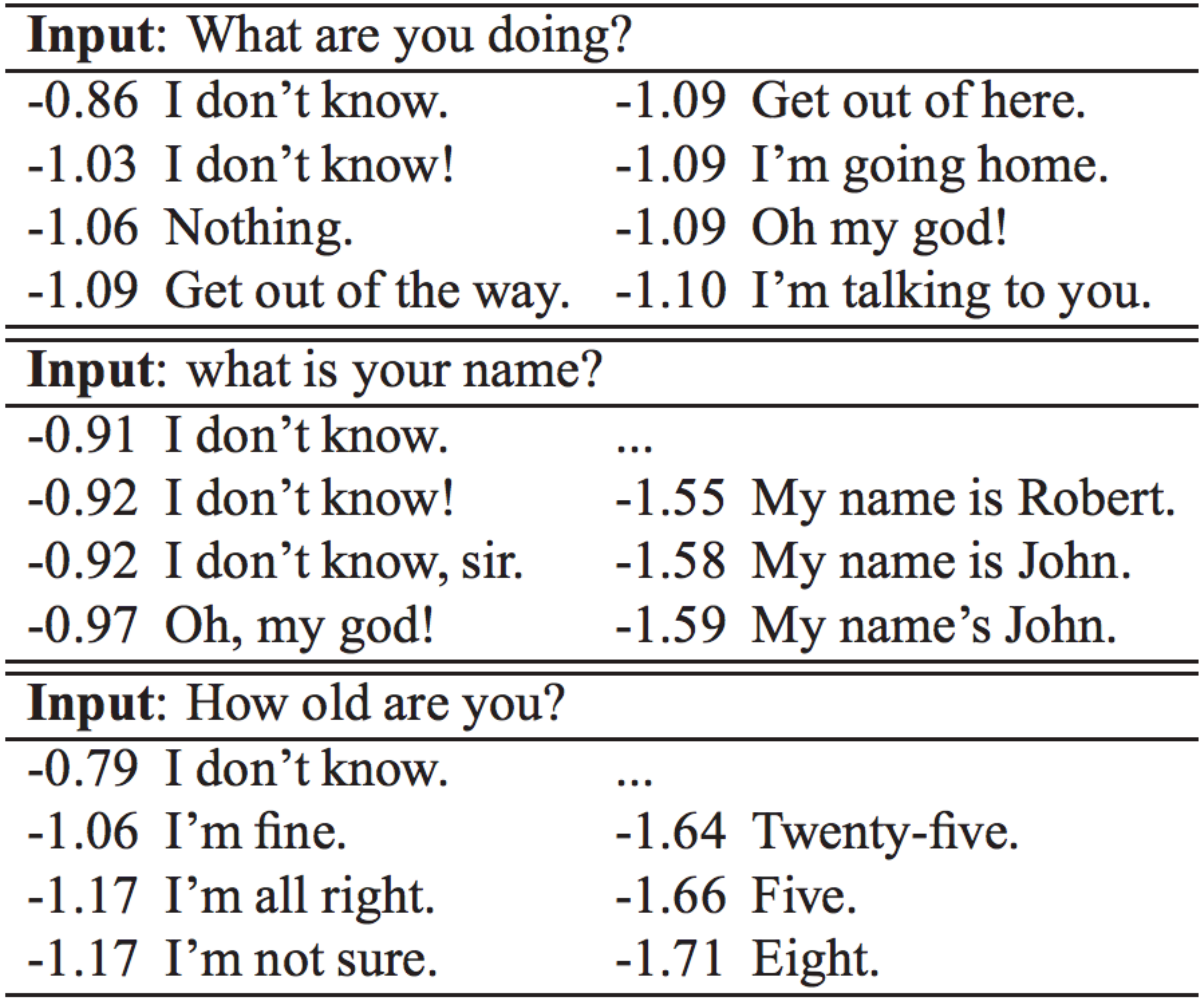
- Just like conventional MT, can train seq2seq models for this task
- Why might this model perform poorly? What might it be bad at?
- Hard to evaluate:



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)

Lack of Diversity

- Solution: mutual information criterion; response R should be predictive of user utterance U as well

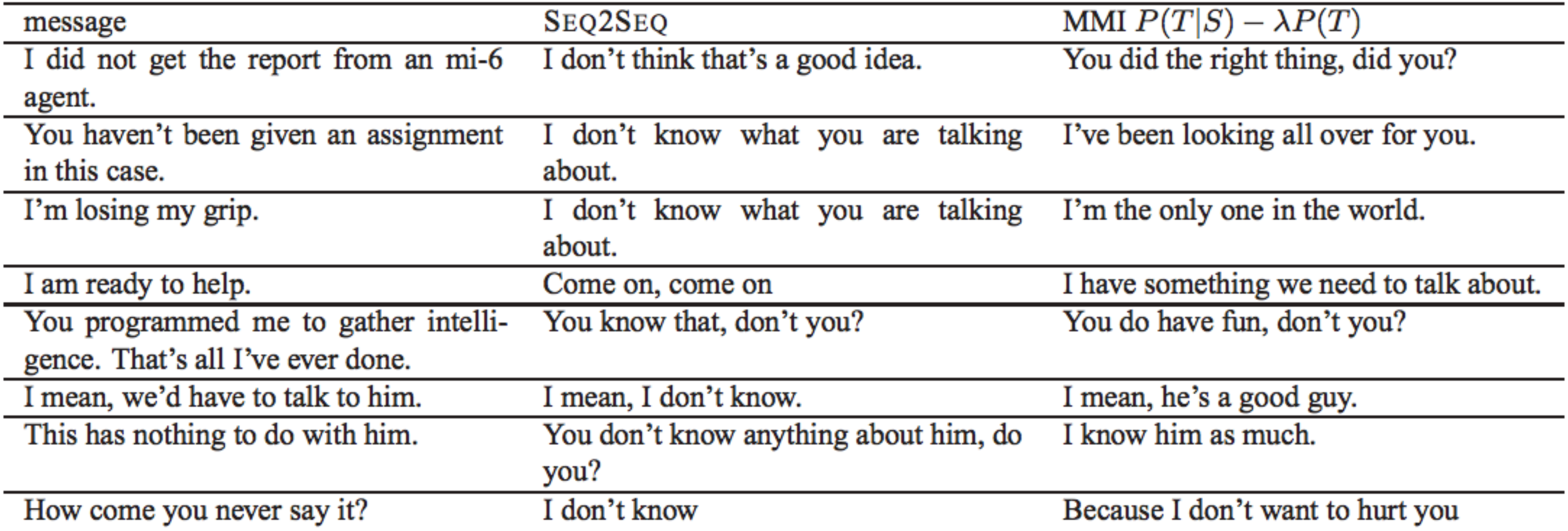
- Standard conditional likelihood: $\log P(R|U)$

- Mutual information: $\log \frac{P(R,U)}{P(R)P(U)} = \log P(R|U) - \log P(R)$

- log P(R) can reflect probabilities under a language model

Li et al. (2016)

Lack of Diversity



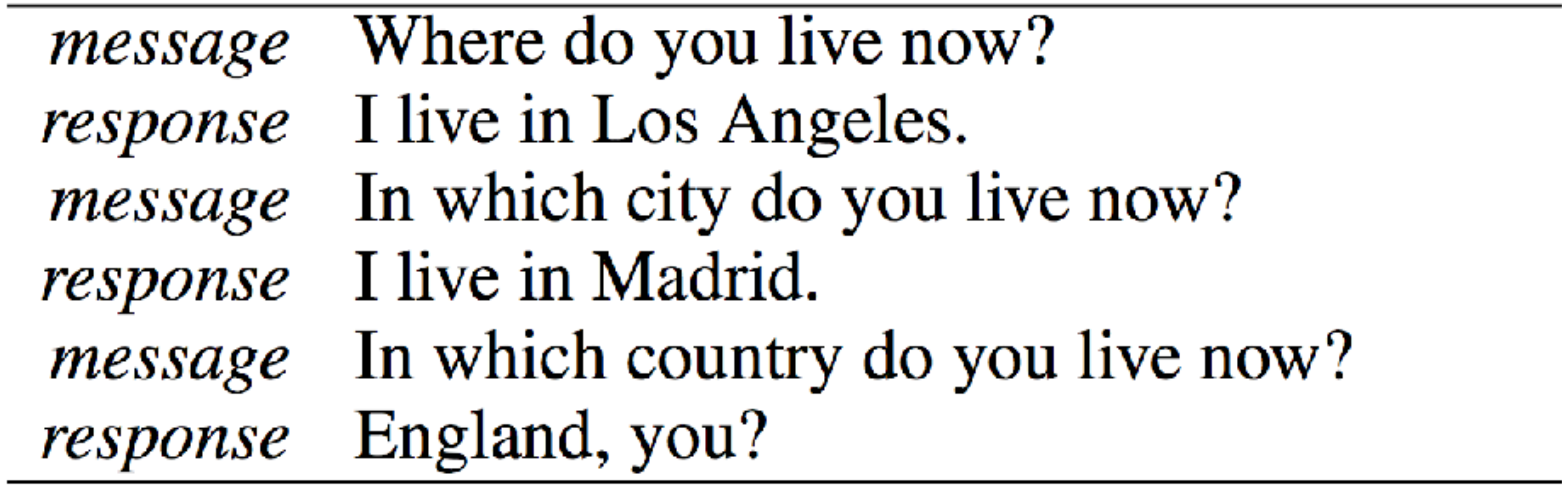
message	SEQ2SEQ	MMI $P(T \mid S) - \lambda P(T)$
I did not get the report from an mi-6 agent.	I don’t think that’s a good idea.	You did the right thing, did you?
You haven’t been given an assignment in this case.	I don’t know what you are talking about.	I’ve been looking all over for you.
I’m losing my grip.	I don’t know what you are talking about.	I’m the only one in the world.
I am ready to help.	Come on, come on	I have something we need to talk about.
You programmed me to gather intelligence. That’s all I’ve ever done.	You know that, don’t you?	You do have fun, don’t you?
I mean, we’d have to talk to him.	I mean, I don’t know.	I mean, he’s a good guy.
This has nothing to do with him.	You don’t know anything about him, do you?	I know him as much.
How come you never say it?	I don’t know	Because I don’t want to hurt you

- OpenSubtitles data

Li et al. (2016)

Future of chatbots

- How deep can a conversation be without more semantic grounding? Basic facts aren't even consistent…



message	Where do you live now?
response	I live in Los Angeles.
message	In which city do you live now?
response	I live in Madrid.
message	In which country do you live now?
response	England, you?

Li et al. (2016) Persona…

- Can force chatbots to give consistent answers, but still probably not very interesting

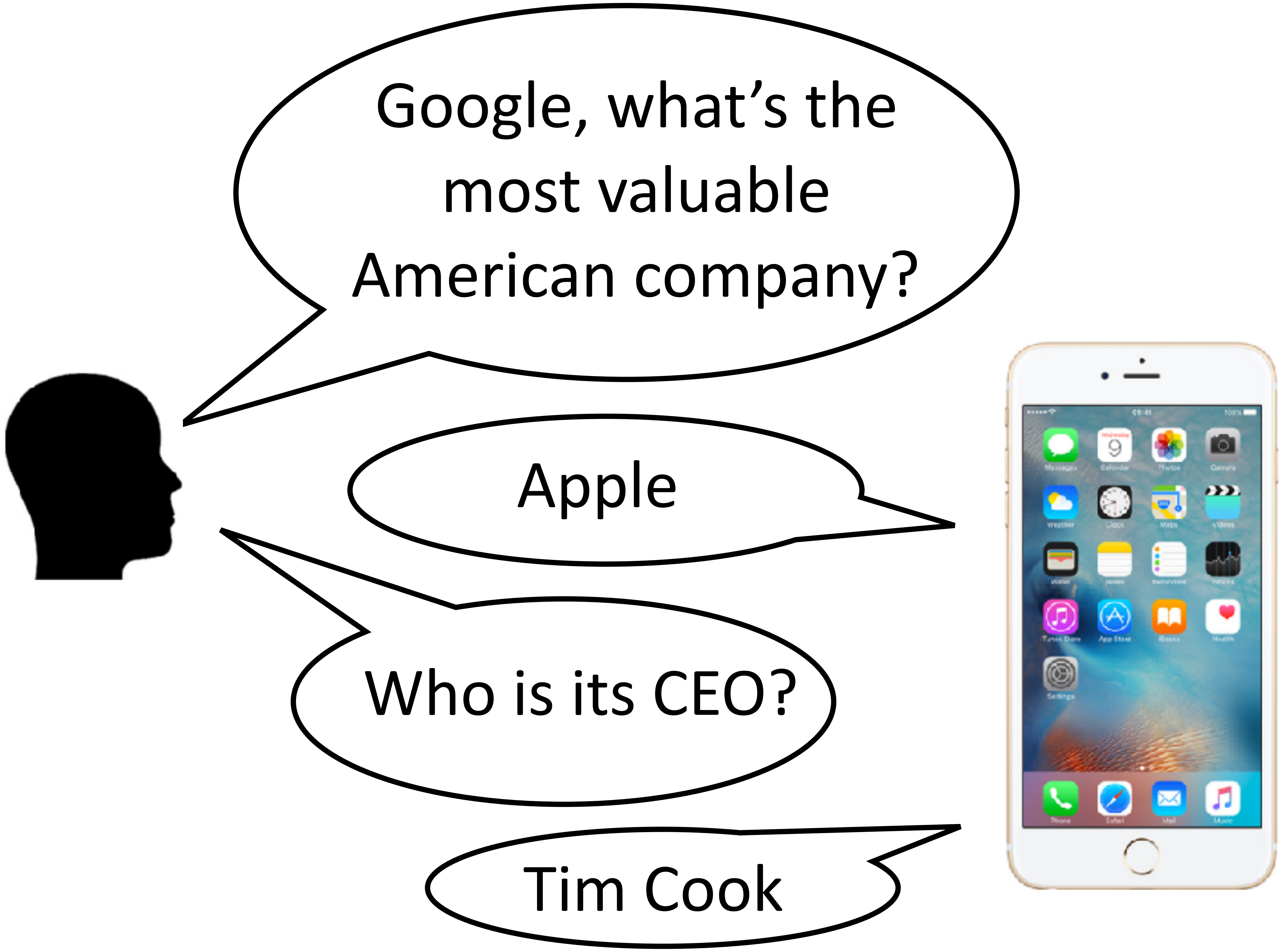
- XiaoIce: Microsoft chatbot in Chinese, 20M users, average user interacts 60 times/month

- People do seem to like talking to them…?

Task-Oriented Dialogue

Task-Oriented Dialogue

- Question answering/search:



Google, what's the most valuable American company?

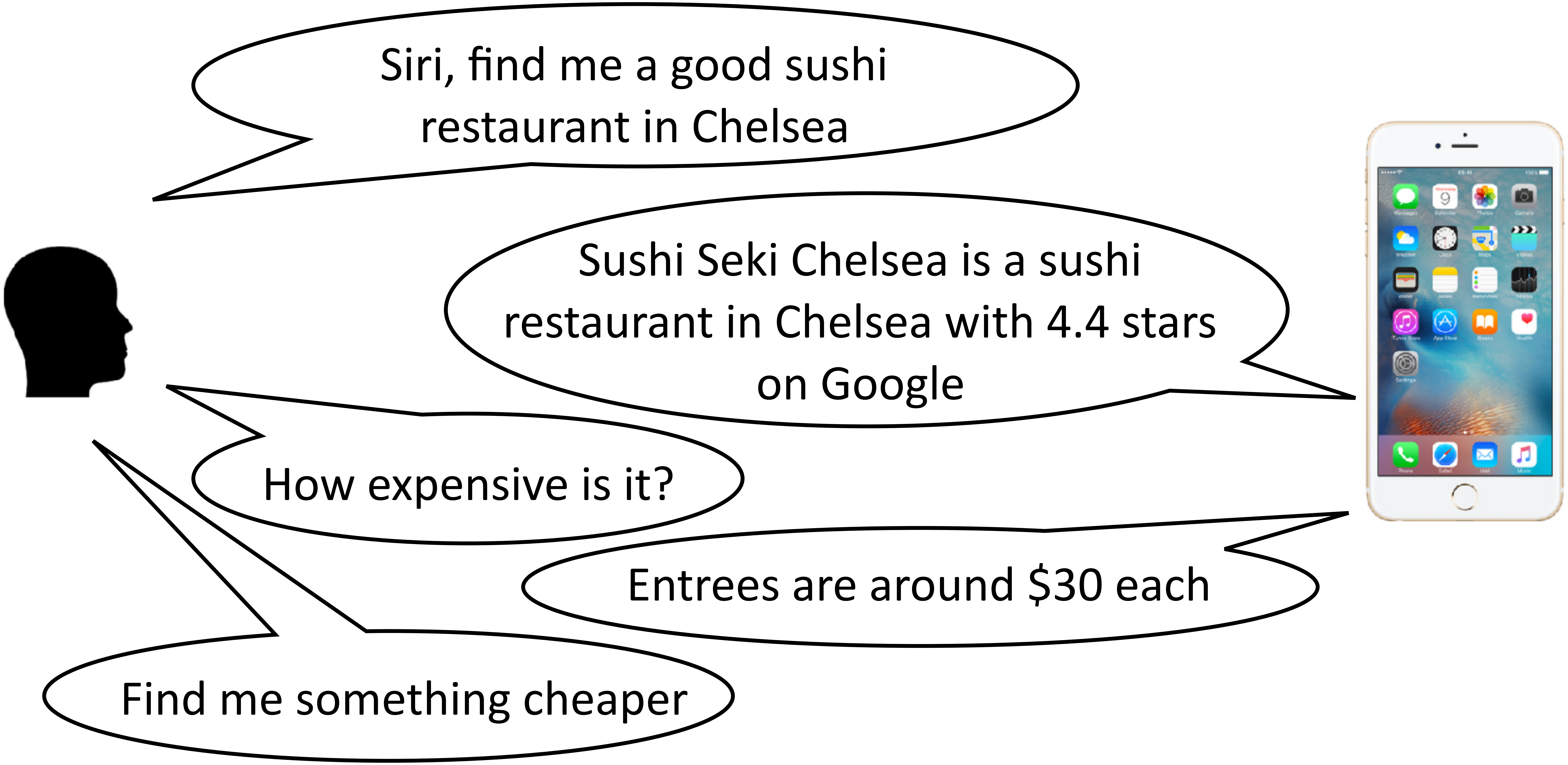
Apple

Who is its CEO?

Tim Cook

Task-Oriented Dialogue

- Personal assistants / API front-ends:



Siri, find me a good sushi restaurant in Chelsea

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

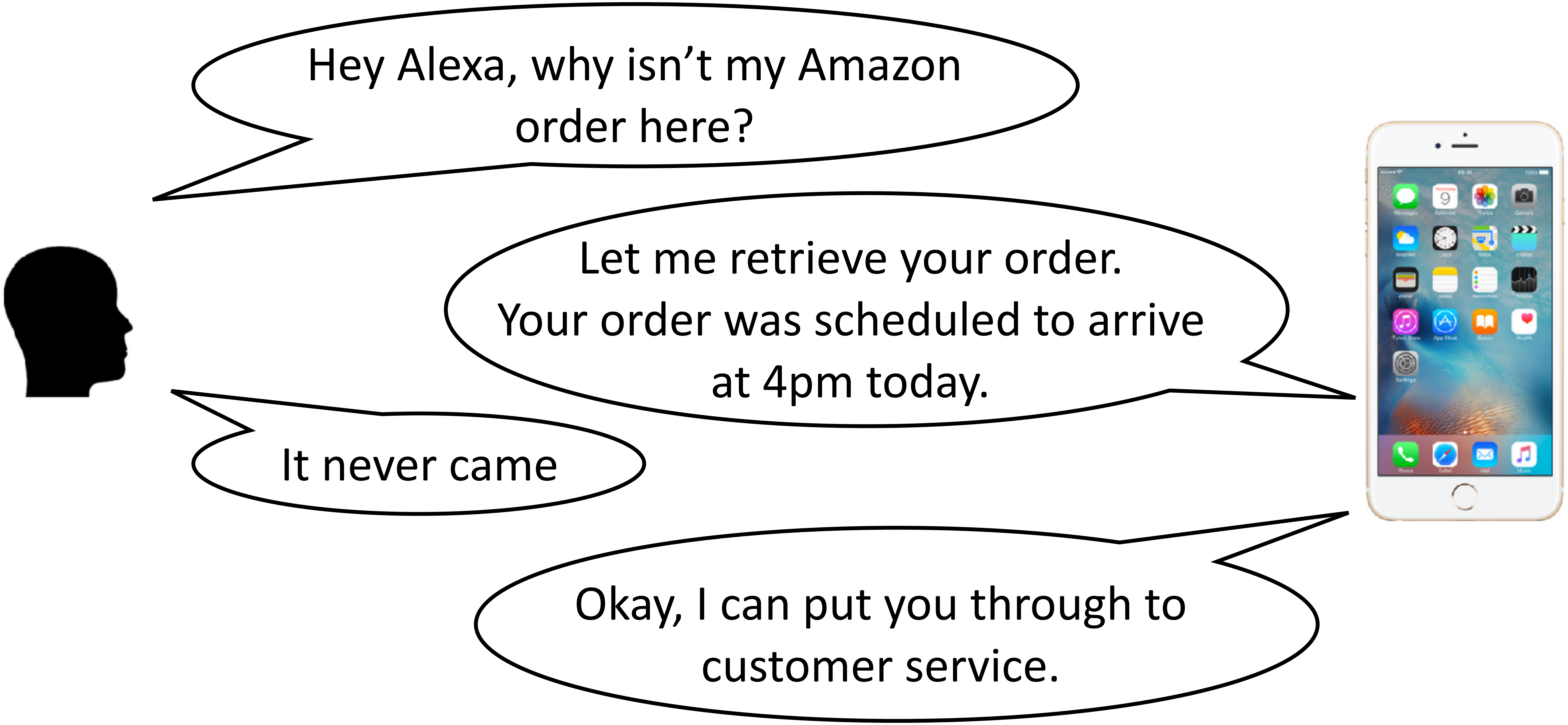
How expensive is it?

Entrees are around $30 each

Find me something cheaper

Task-Oriented Dialogue

- Personal assistants / API front-ends:



Hey Alexa, why isn't my Amazon order here?

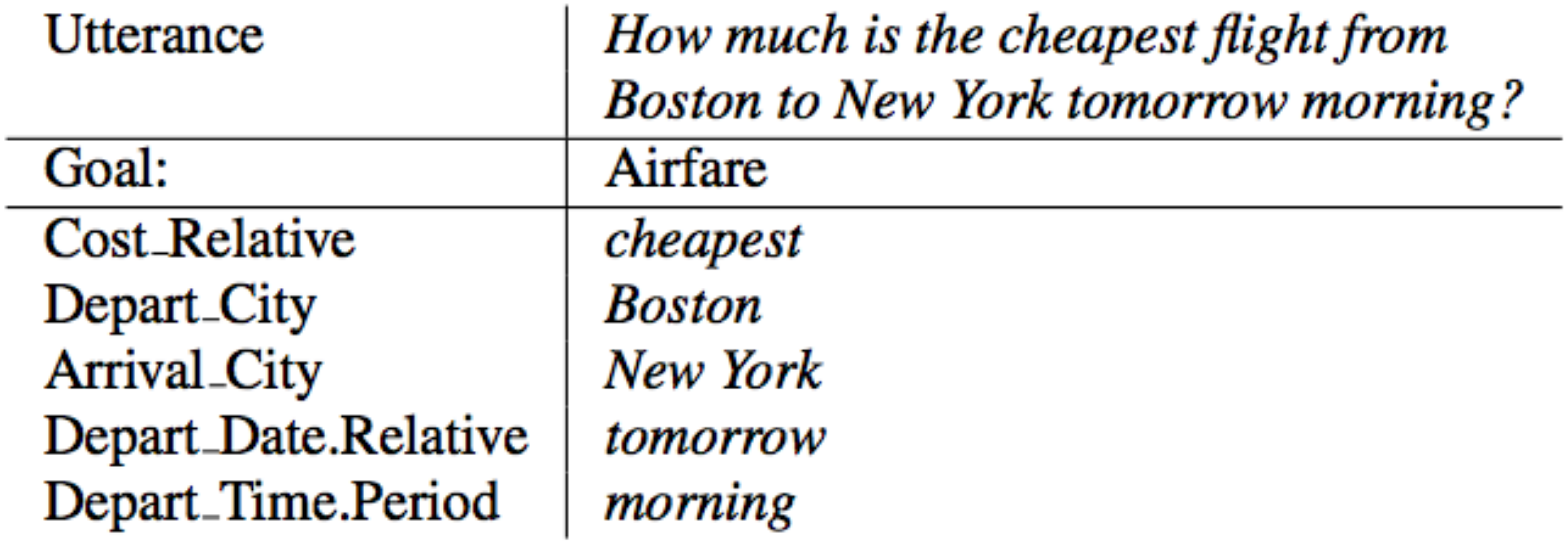
Let me retrieve your order. Your order was scheduled to arrive at 4pm today.

It never came

Okay, I can put you through to customer service.

Air Travel Information Service (ATIS)

- Given an utterance, predict a domain-specific semantic interpretation



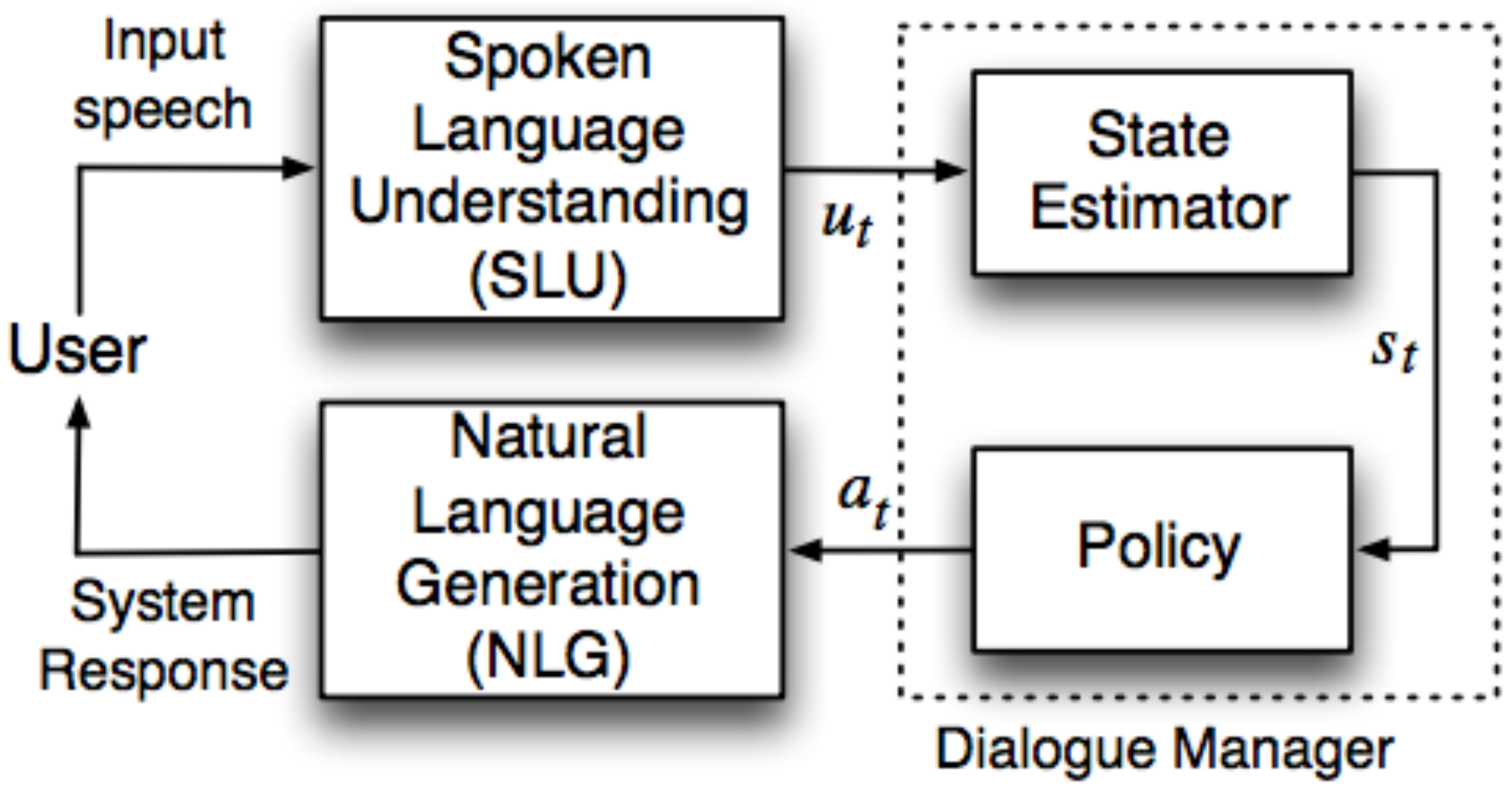
Utterance	*How much is the cheapest flight from Boston to New York tomorrow morning?*
Goal:	Airfare
Cost_Relative	*cheapest*
Depart_City	*Boston*
Arrival_City	*New York*
Depart_Date.Relative	*tomorrow*
Depart_Time.Period	*morning*

- Can formulate as semantic parsing, but simple slot-filling solutions (classifiers) work well too

DARPA (early 1990s), Figure from Tur et al. (2010)

Full Dialogue Task

- Parsing / language understanding is just one piece of a system
- Dialogue state: reflects any information about the conversation (e.g., search history)



- User utterance -> update dialogue state -> take action (e.g., query the restaurant database) -> say something
- Much more complex than chatbots!

Young et al. (2013)

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

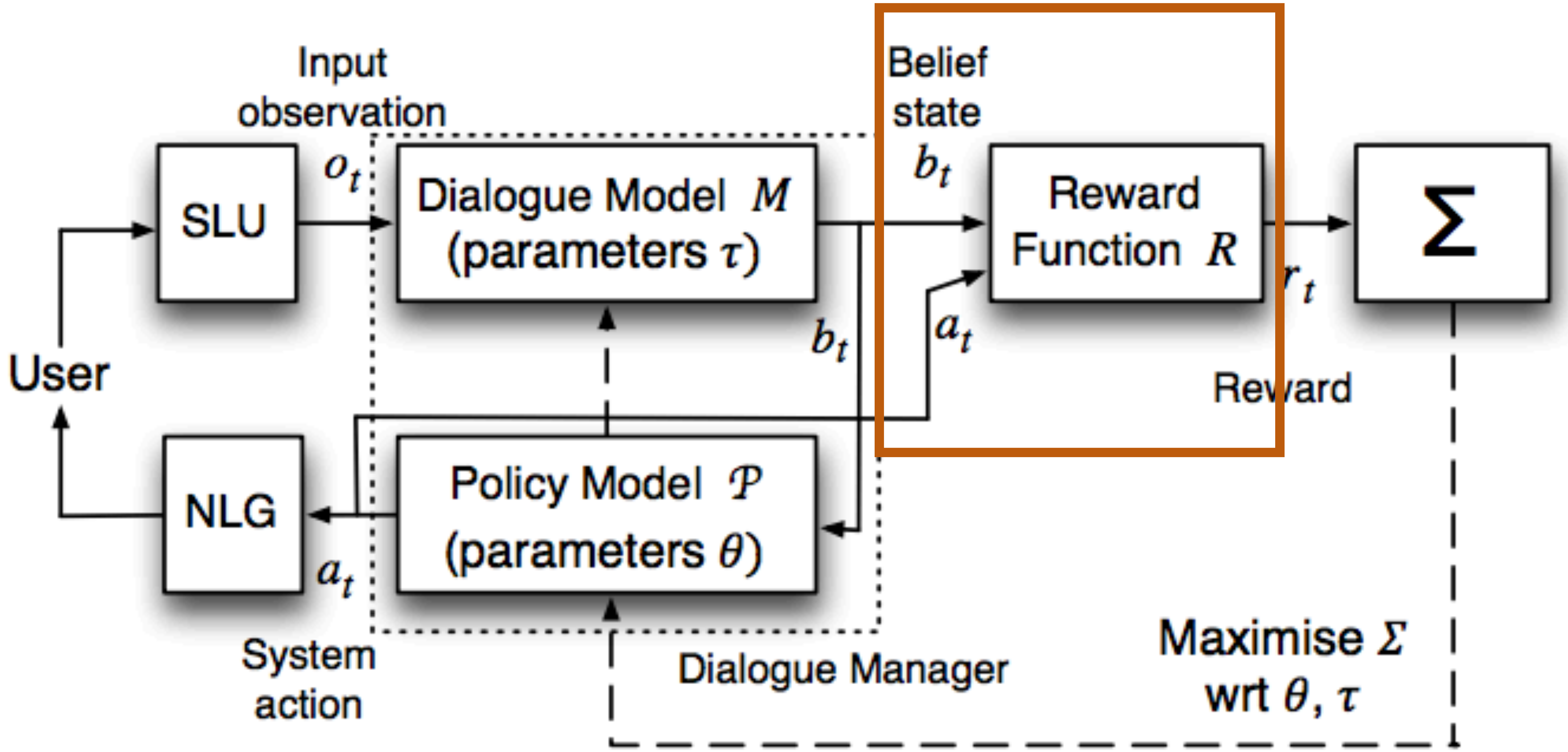
Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

```
get_value(cost, curr_result)
```

Entrees are around $30 each

POMDP-based Dialogue Systems



- POMDP: user is the "environment," an utterance is a noisy signal of state
- Dialogue model: can look like a parser or any kind of encoder model
- Generator: use templates or seq2seq model
- Where do rewards come from?

Young et al. (2013)

Reward for completing task?

Find me a good sushi restaurant in Chelsea

Very indirect signal of what should happen up here

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

…

Okay make me a reservation!

+1 `make_reservation(curr_result)`

User gives reward?

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

How does the user know the right search happened?

+1 Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

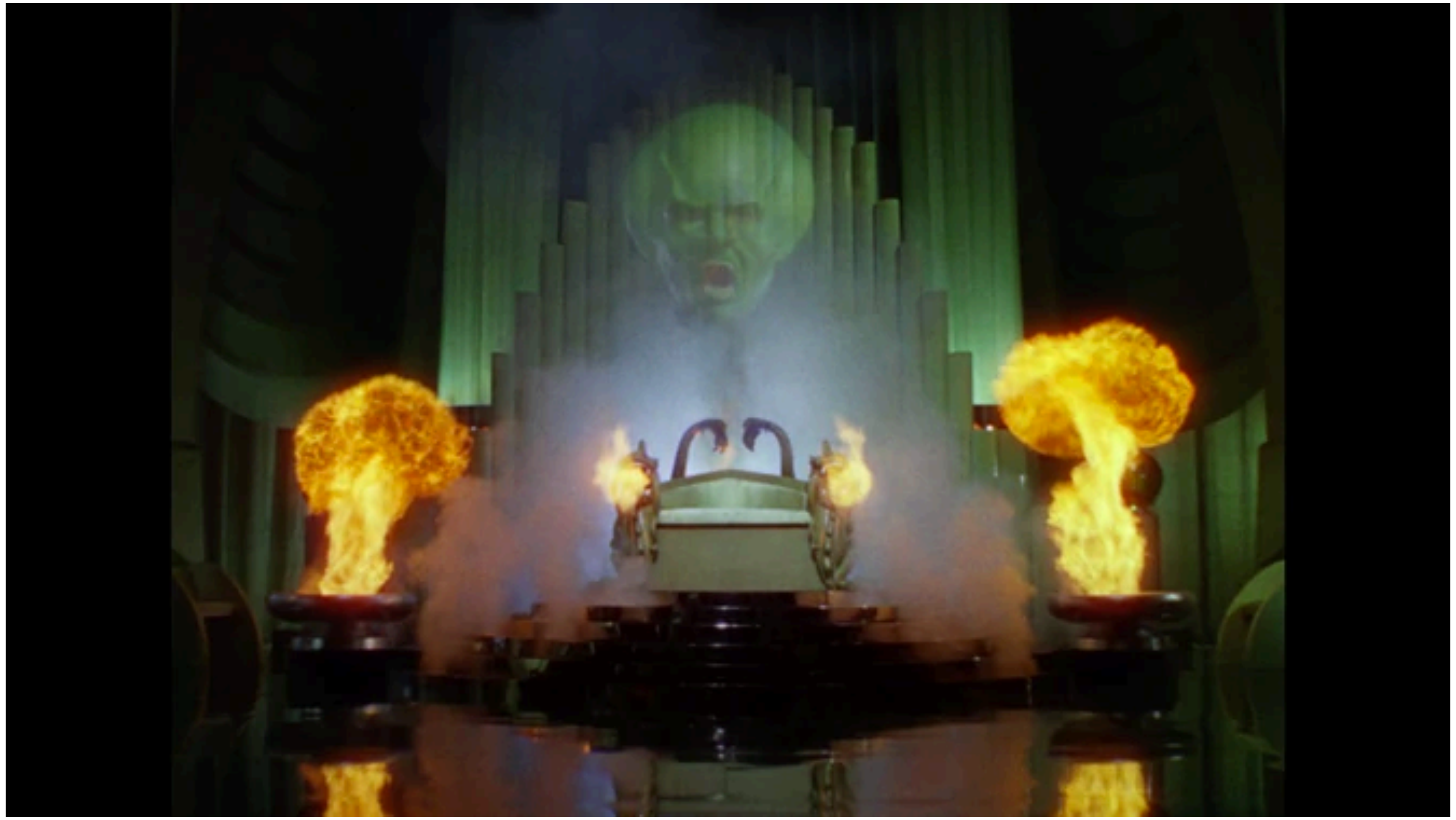
How expensive is it?

```
get_value(cost, curr_result)
```

+1 Entrees are around $30 each

Wizard-of-Oz

- Learning from demonstrations: “wizard” pulls the levers and makes the dialogue system update its state and take actions



Kelley (early 1980s), Ford and Smith (1982)

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

wizard enters these

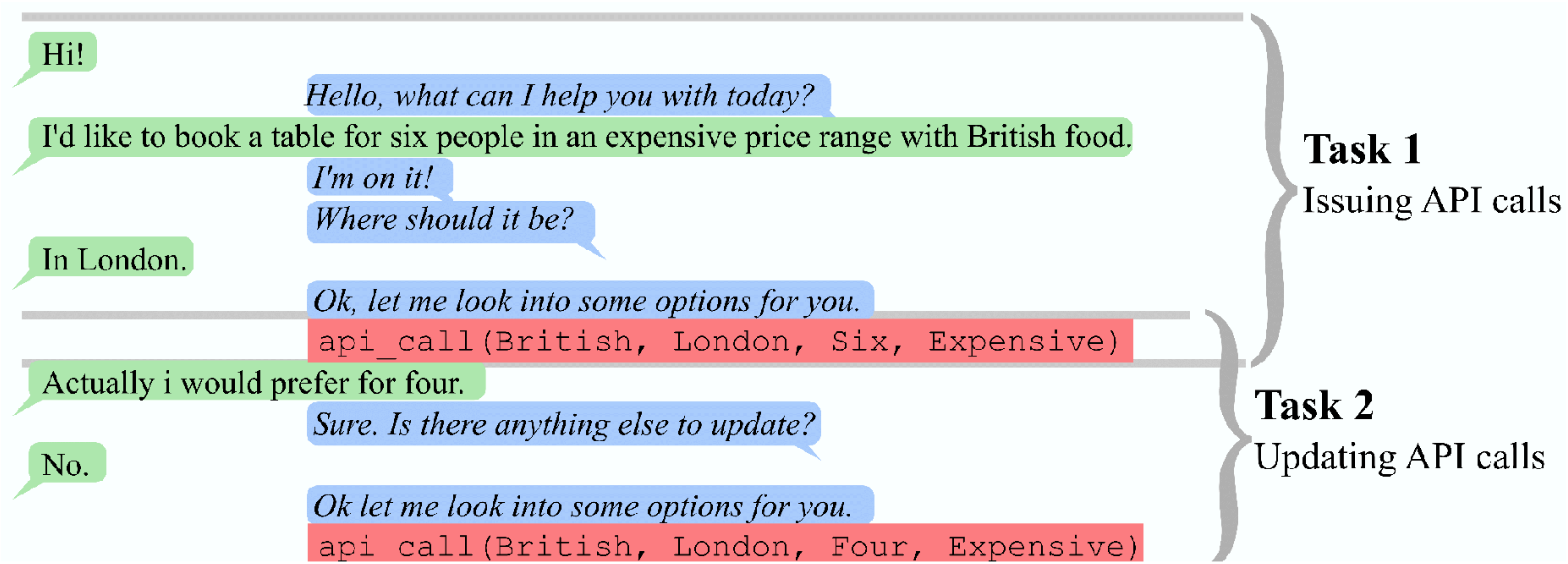
```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

wizard types this out or invokes templates

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

- Wizard can be a trained expert and know exactly what the dialogue systems is supposed to do

Learning from Static Traces



Hi!

Hello, what can I help you with today?

I'd like to book a table for six people in an expensive price range with British food.

I'm on it!

Where should it be?

In London.

Ok, let me look into some options for you.

`api_call(British, London, Six, Expensive)`

Actually i would prefer for four.

Sure. Is there anything else to update?

No.

Ok let me look into some options for you.

`api call(British, London, Four, Expensive)`

- Using either wizard-of-Oz or other annotations, can collect static traces and train from these

Bordes et al. (2017)

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
stars <- 4+
curr_result <- execute_search()
```

- User asked for a “good” restaurant — does that mean we should filter by star rating? What does “good” mean?
- Hard to change system behavior if training from static traces, especially if system capabilities or desired behavior change

Goal-oriented Dialogue

- Tons of industry interest!

- Startups:

semanticmachines

Maluuba
A Microsoft company

ASAPP

Eloquent Labs

VIV

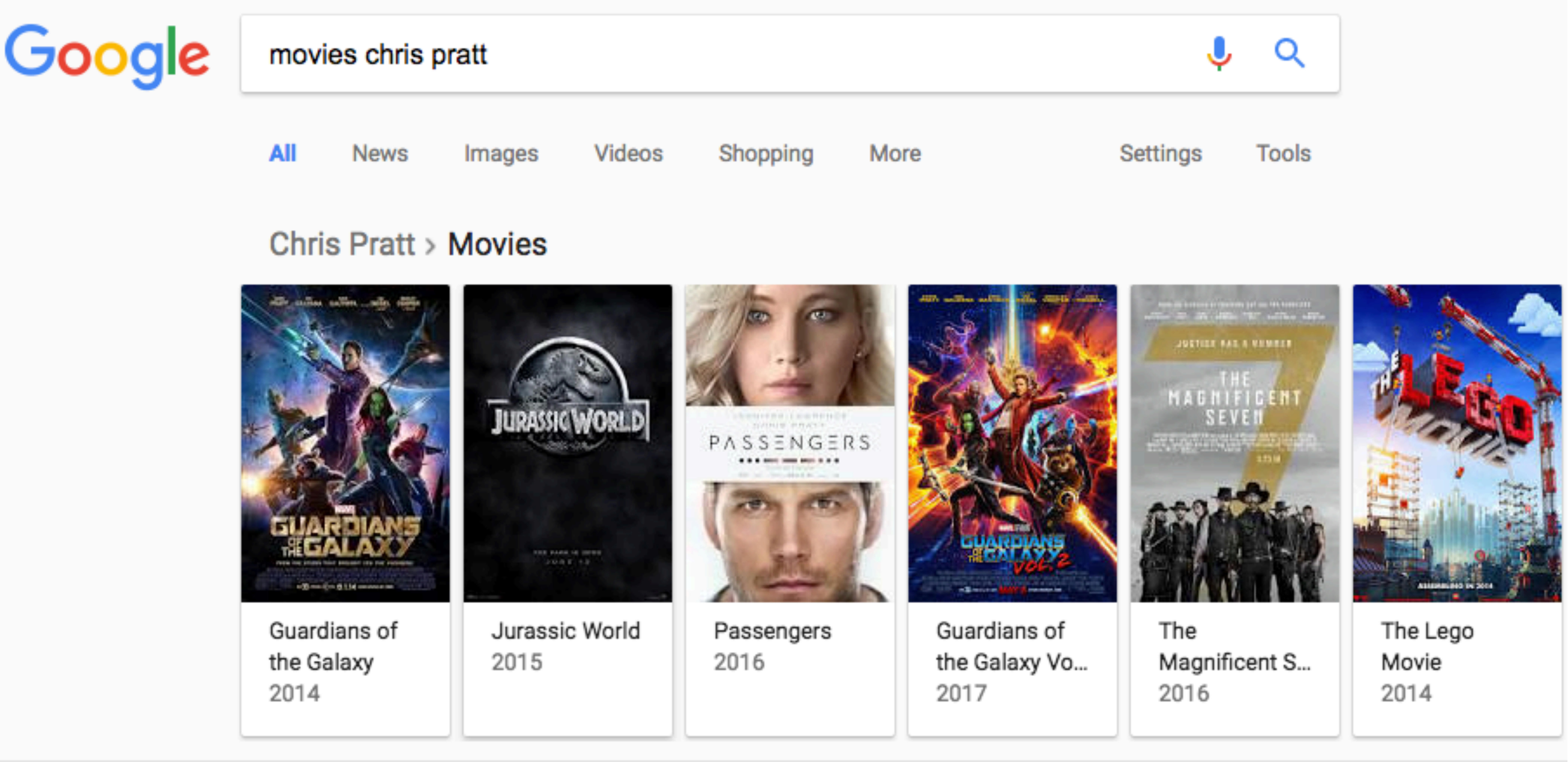
x.

- Big Companies: Apple Siri (VocalIQ), Google Allo, Amazon Alexa, Microsoft Cortana, Facebook M, Samsung Bixby, Tencent WeChat

- Lots of cool work that’s not public yet

Other Dialogue Applications

Search/QA as Dialogue



movies chris pratt

Chris Pratt › Movies

Guardians of the Galaxy 2014

Jurassic World 2015

Passengers 2016

Guardians of the Galaxy Vo... 2017

The Magnificent S... 2016

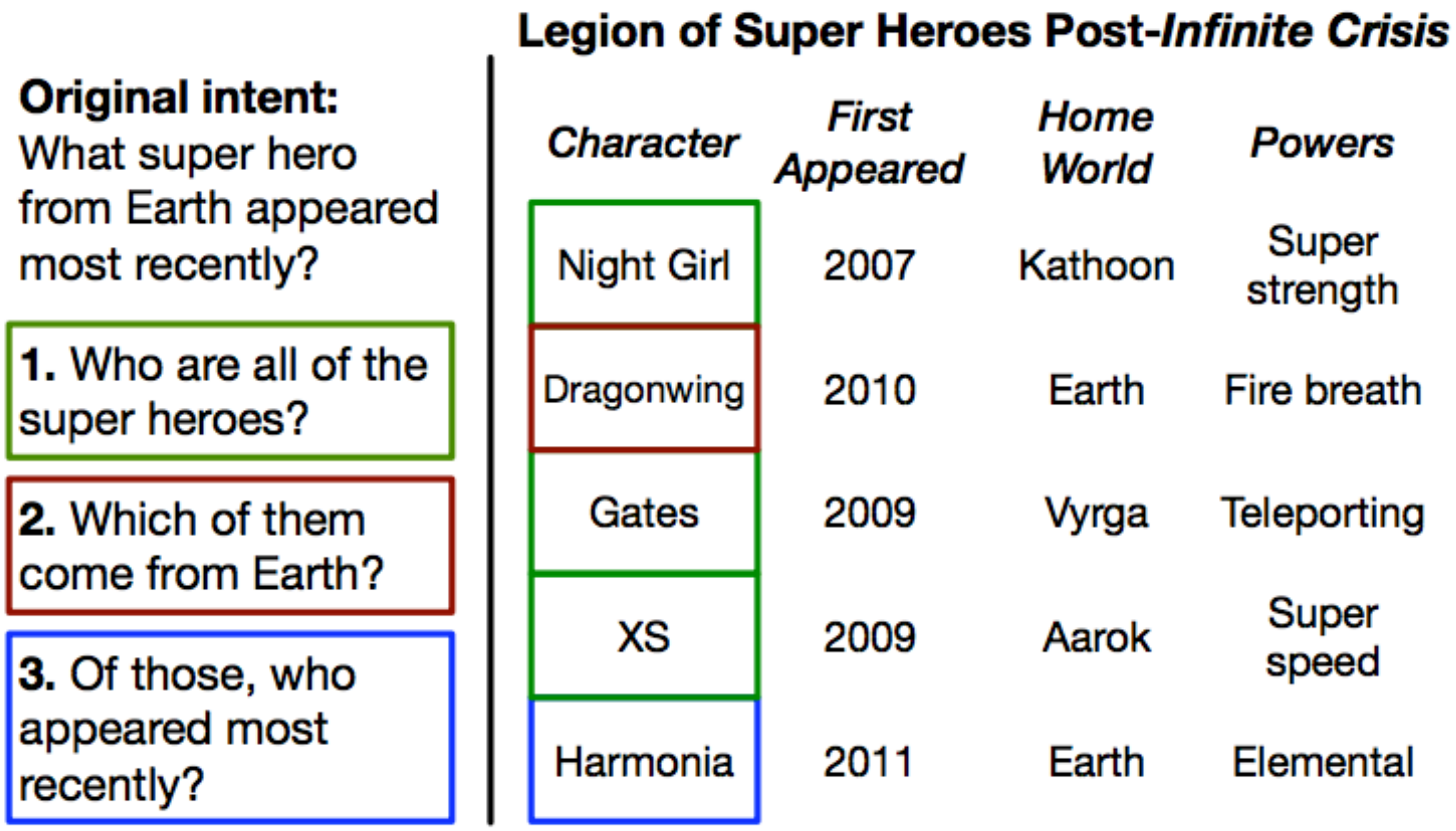
The Lego Movie 2014

- “Has Chris Pratt won an Oscar?” / “Has *he* won an Oscar”

QA as Dialogue

- Dialogue is a very natural way to find information from a search engine or a QA system

- Challenges:
  - QA is hard enough on its own
  - Users move the goalposts



Original intent:
What super hero from Earth appeared most recently?

1. Who are all of the super heroes?

2. Which of them come from Earth?

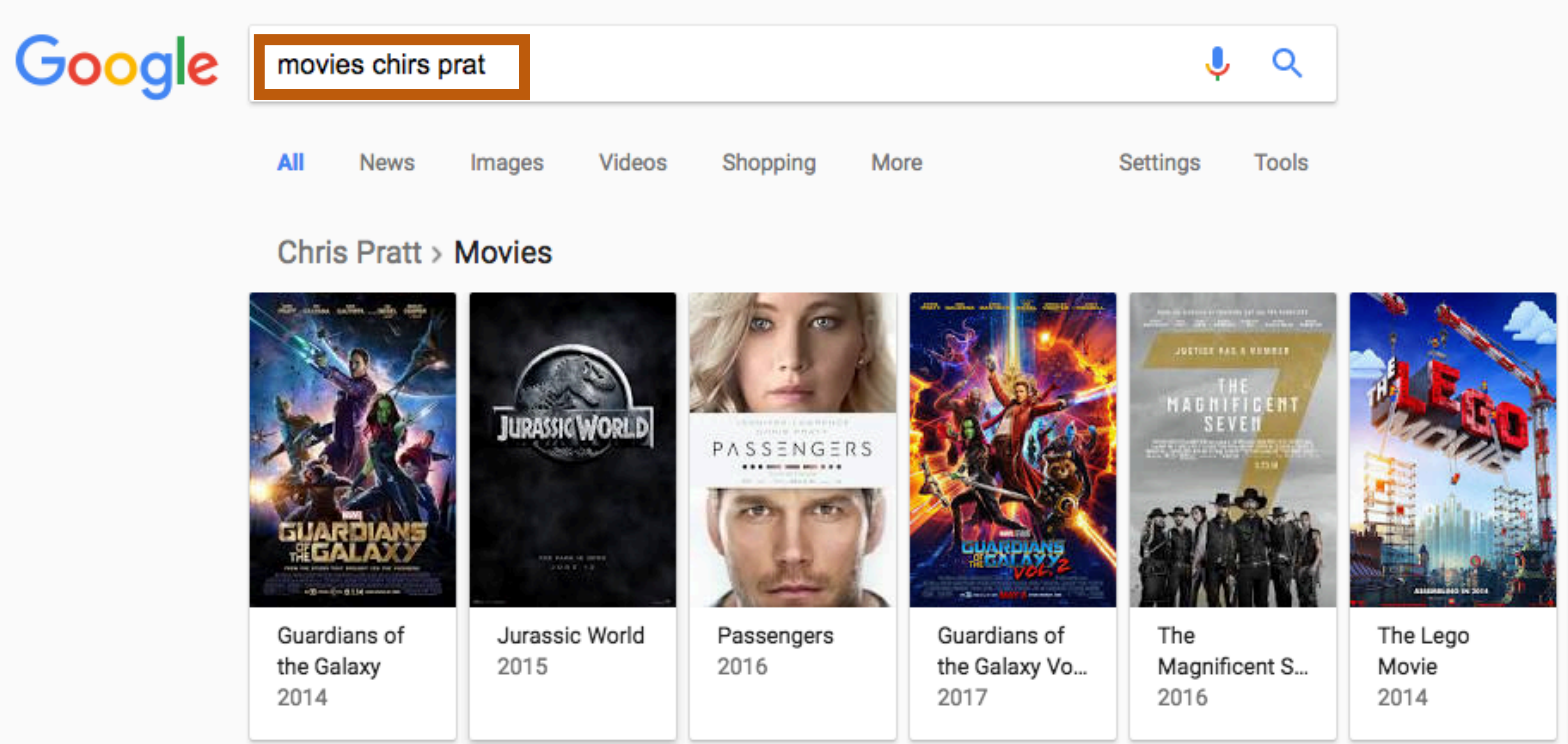
3. Of those, who appeared most recently?

Legion of Super Heroes Post-*Infinite Crisis*

Character	*First Appeared*	*Home World*	*Powers*
Night Girl	2007	Kathoon	Super strength
Dragonwing	2010	Earth	Fire breath
Gates	2009	Vyrga	Teleporting
XS	2009	Aarok	Super speed
Harmonia	2011	Earth	Elemental

Iyyer et al. (2017)

Search as Dialogue



movies chirs prat

Chris Pratt › Movies

Guardians of the Galaxy 2014

Jurassic World 2015

Passengers 2016

Guardians of the Galaxy Vo... 2017

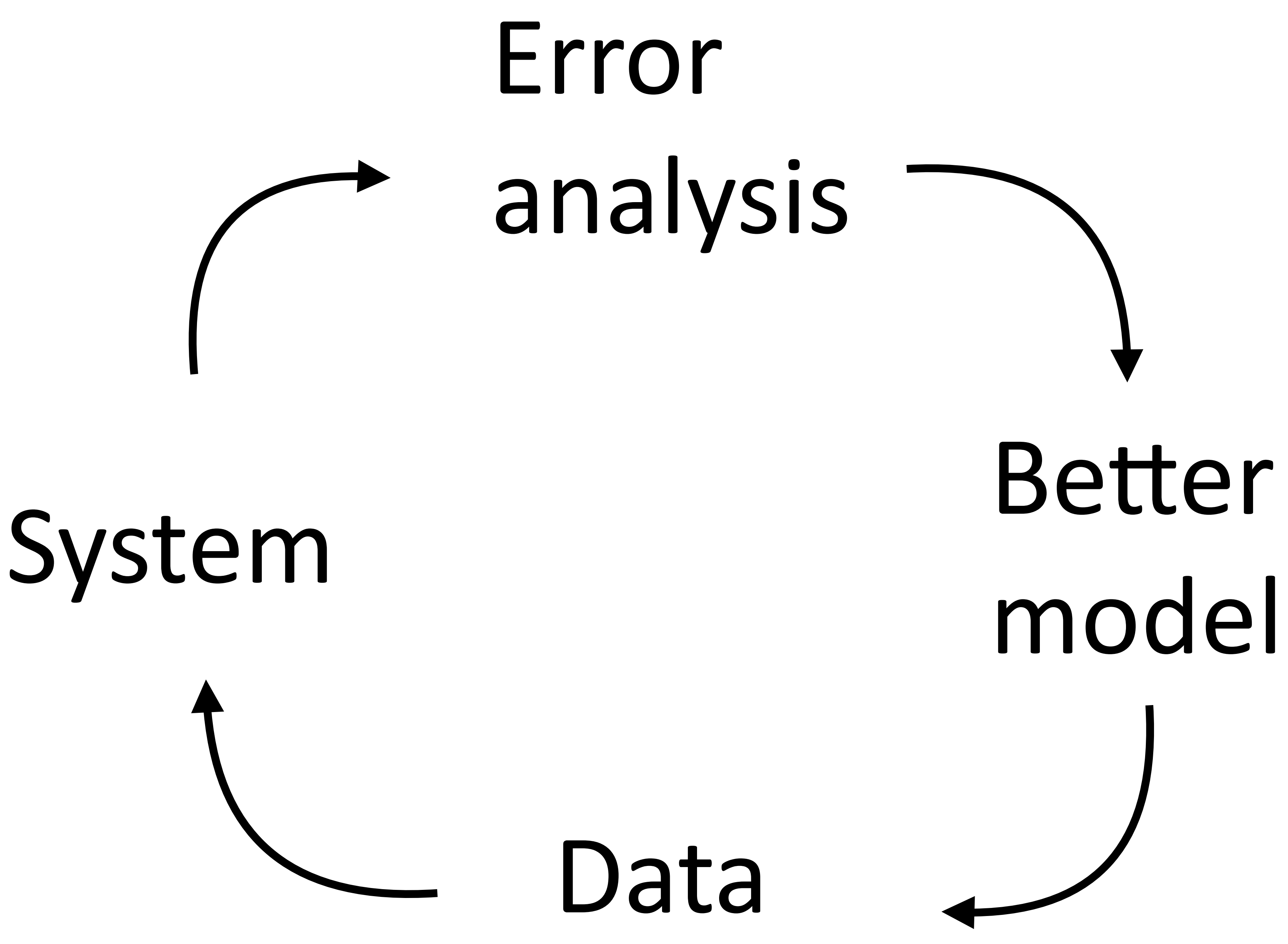
The Magnificent S... 2016

The Lego Movie 2014

- Google can deal with misspellings, so more misspellings happen — Google has to do more!

Dialogue Mission Creep

Most NLP tasks



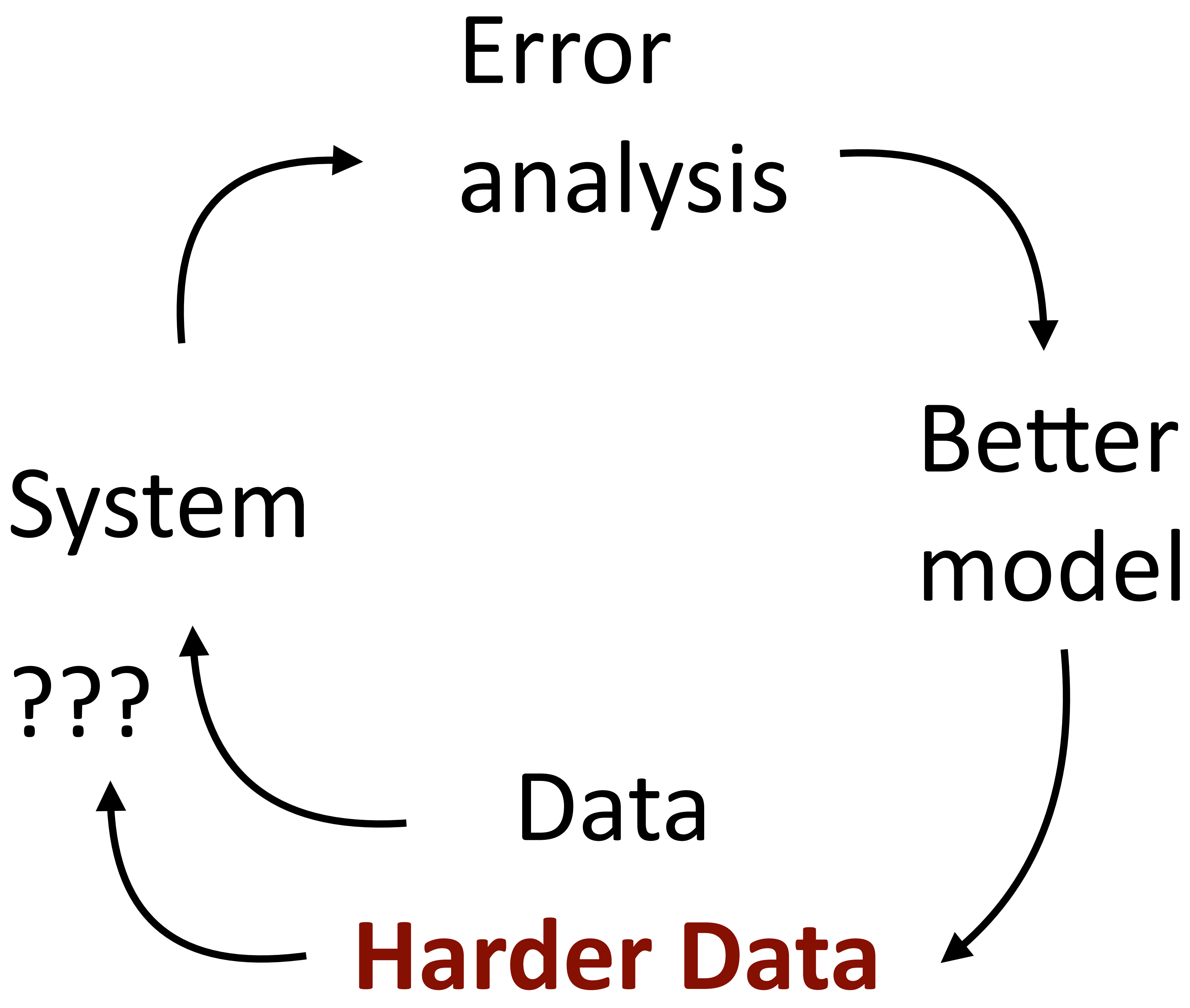
Error analysis

Better model

Data

System

Dialogue/Search/QA



Error analysis

Better model

Harder Data

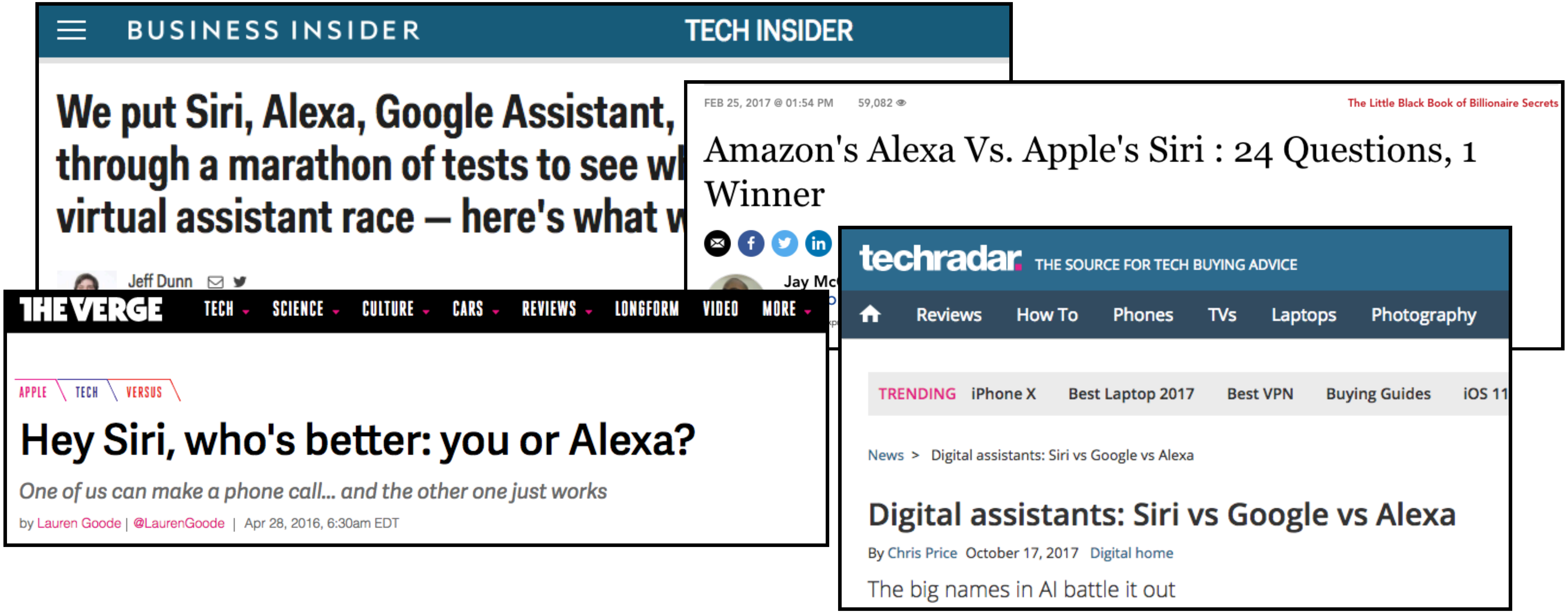
Data

System

???

- Fixed distribution (e.g., natural language sentences), error rate -> 0
- Error rate -> ???; “mission creep” from HCI element

Dialogue Mission Creep



We put Siri, Alexa, Google Assistant, through a marathon of tests to see wh virtual assistant race — here's what w

Jeff Dunn

FEB 25, 2017 @ 01:54 PM 59,082

Amazon's Alexa Vs. Apple's Siri : 24 Questions, 1 Winner

Jay Mc

APPLE TECH VERSUS

Hey Siri, who's better: you or Alexa?

One of us can make a phone call... and the other one just works

by Lauren Goode | @LaurenGoode | Apr 28, 2016, 6:30am EDT

News > Digital assistants: Siri vs Google vs Alexa

Digital assistants: Siri vs Google vs Alexa

By Chris Price October 17, 2017 Digital home

The big names in AI battle it out

- High visibility — your product has to work really well!

Takeaways

- Some decent chatbots, but unclear how to make these more sophisticated than they are right now
- Task-oriented dialogue systems are growing in scope and complexity — really exciting systems on the way
- More and more problems are being formulated as dialogue — interesting applications but challenging to get working well