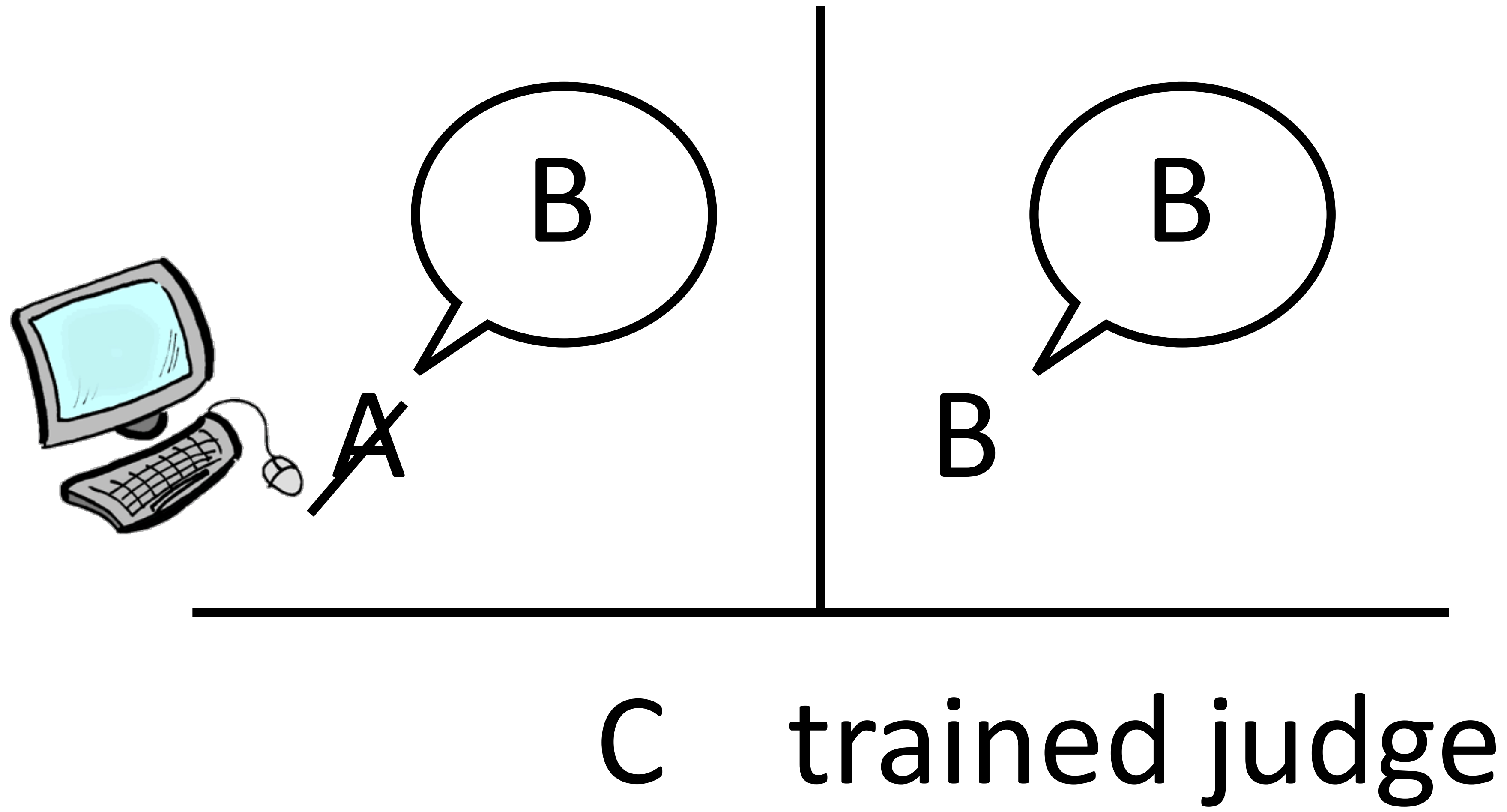
Chatbot History

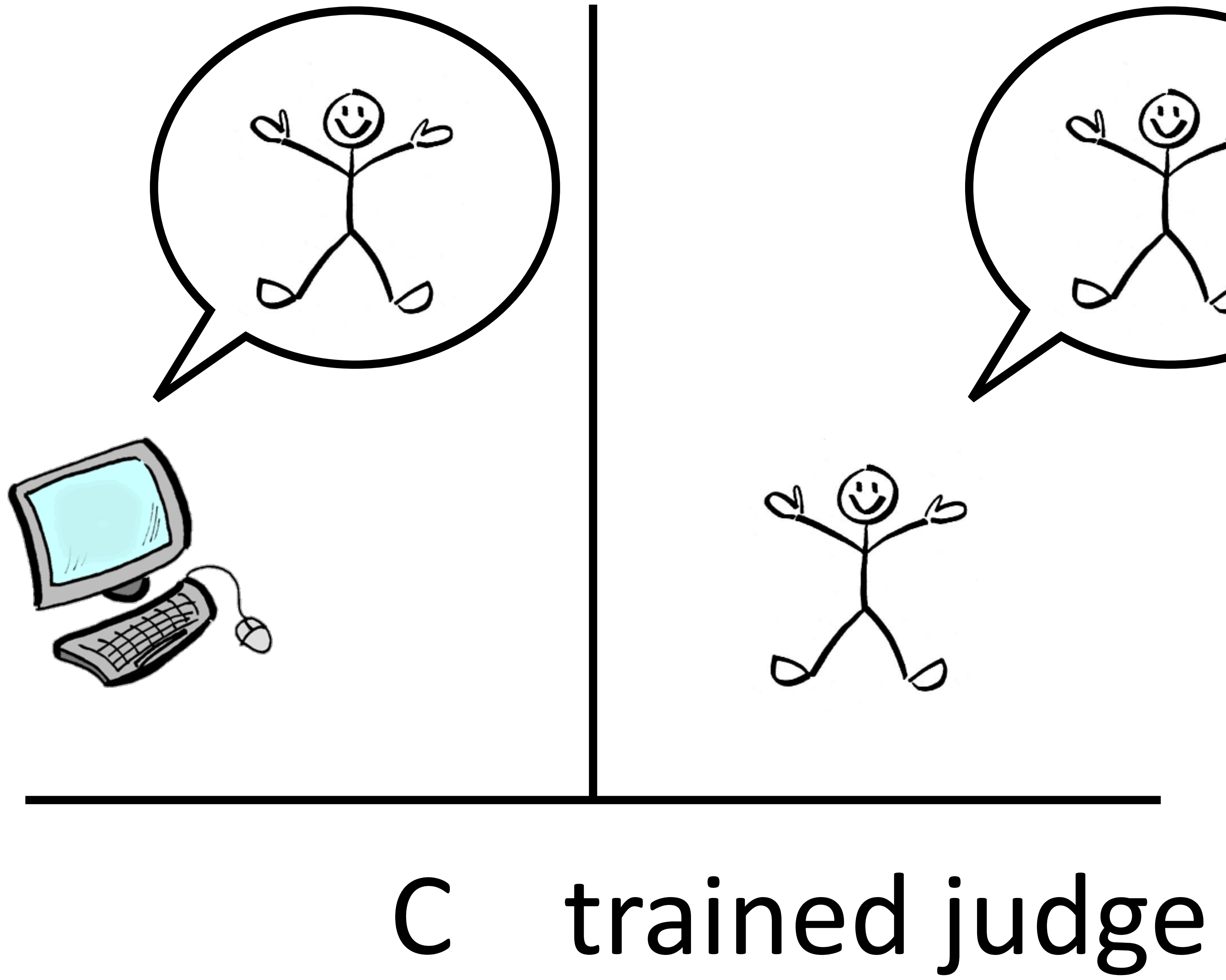
Turing Test (1950)

- 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:

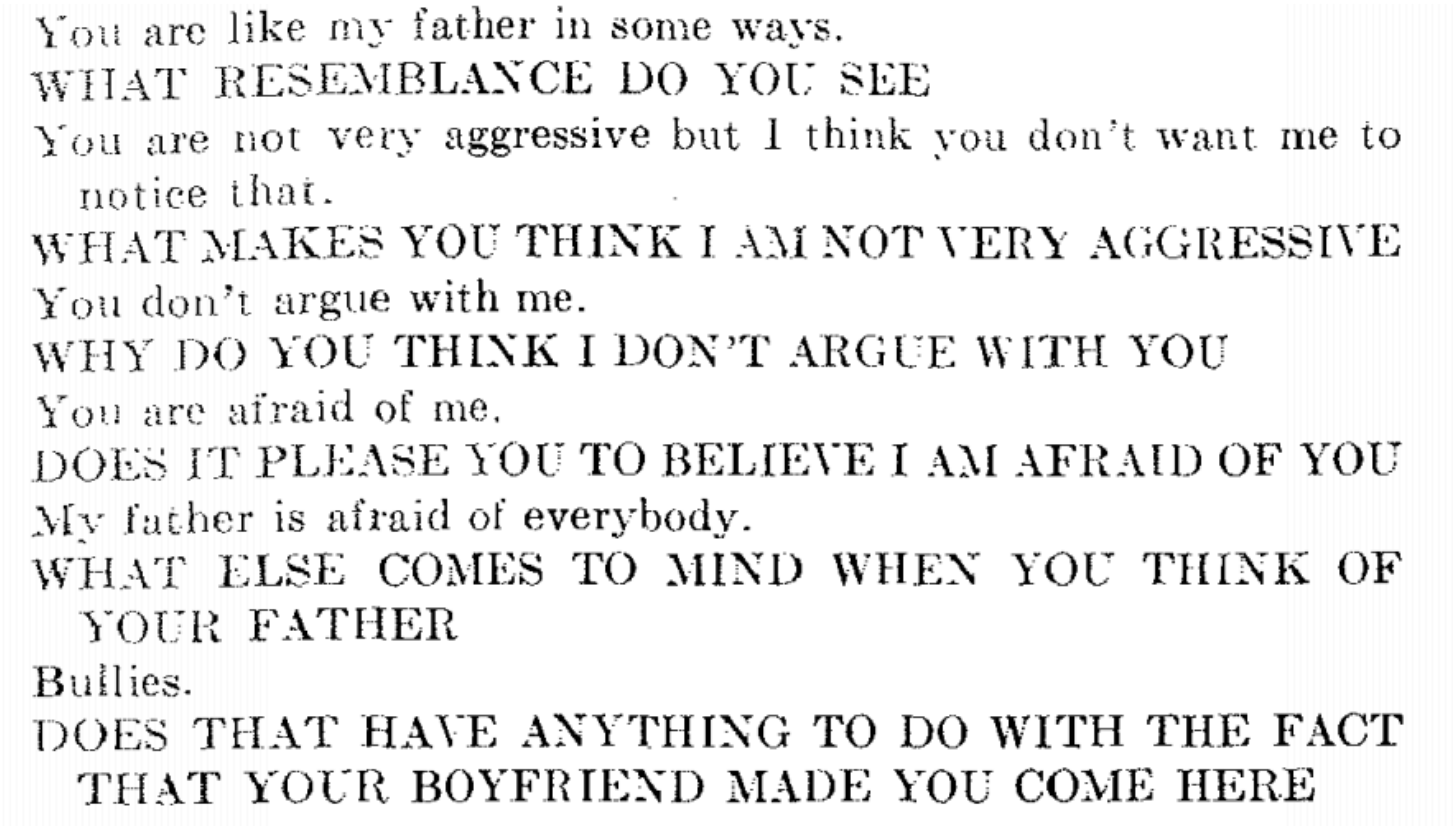


- The test is **not** "does this computer seem human-like to random people with a web browser?"

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

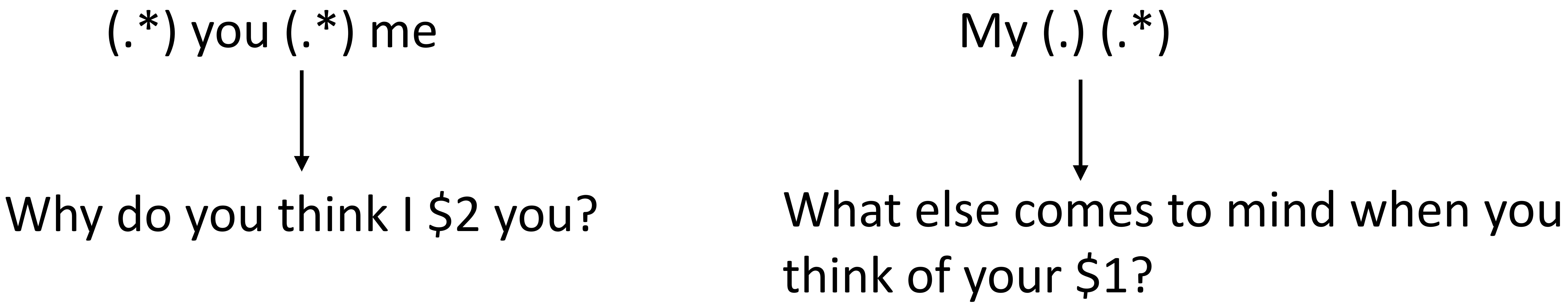
Bullies.

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

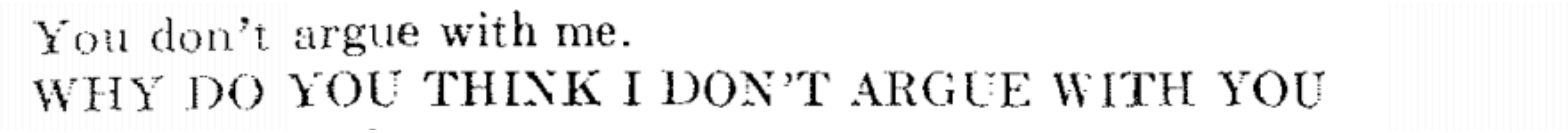


(.*) you (.*) me

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

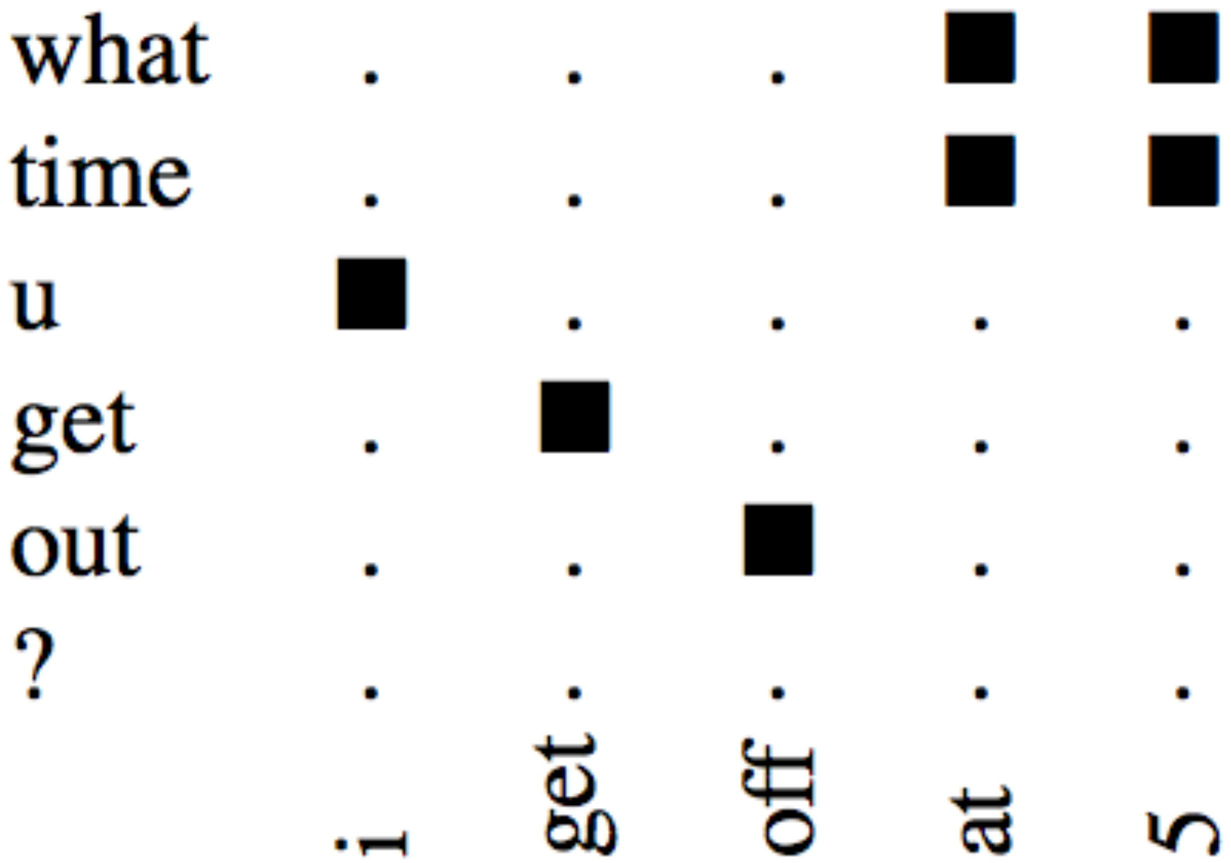
Cleverbot

- 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

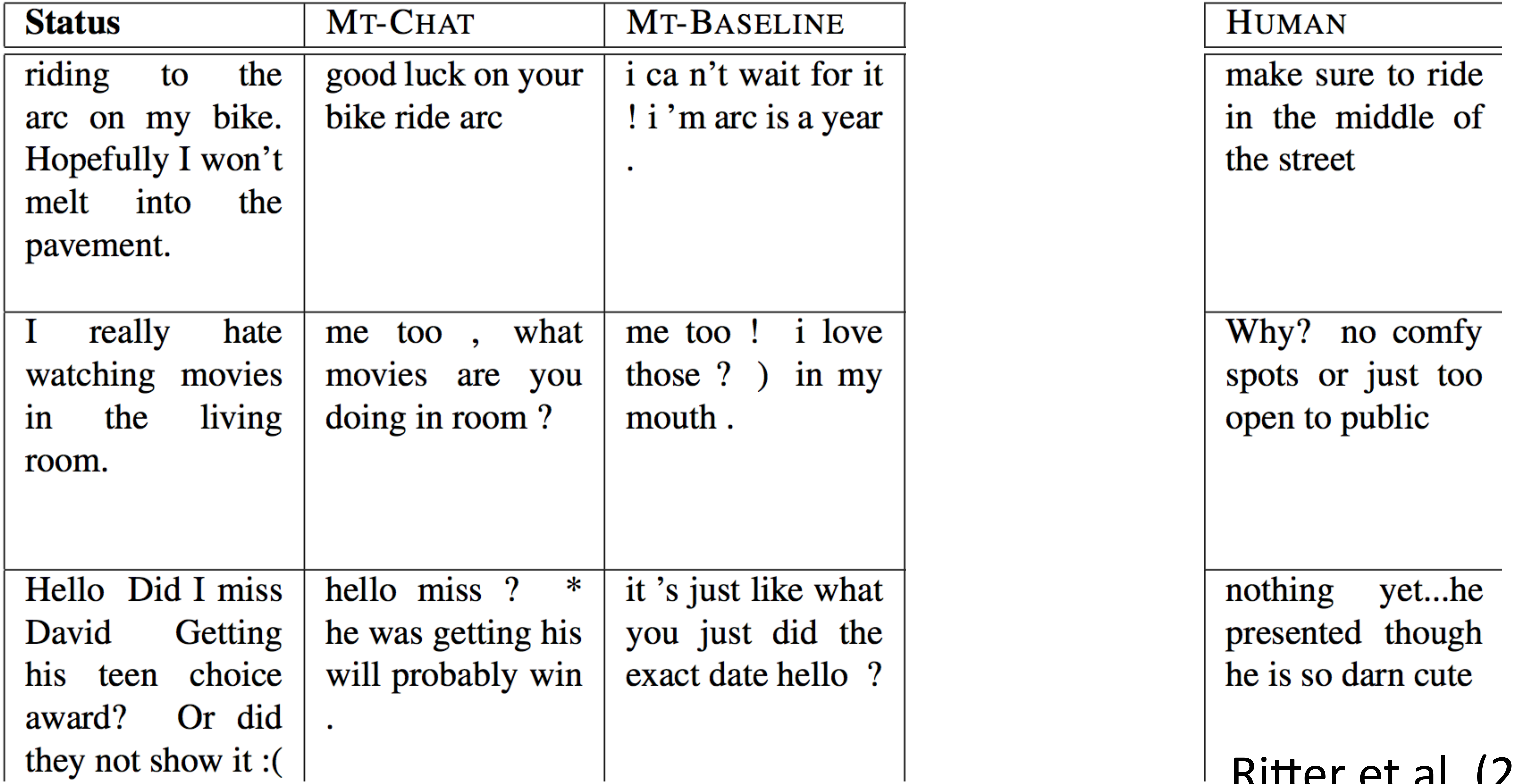


	i	get	off	at	5
what	.	.	.	■	■
time	.	.	.	■	■
u	■	.	.	.	.
get	.	■	.	.	.
out	.	.	■	.	.
?	.	.	.	.	.

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

Ritter et al. (2011)

Data-Driven Approaches

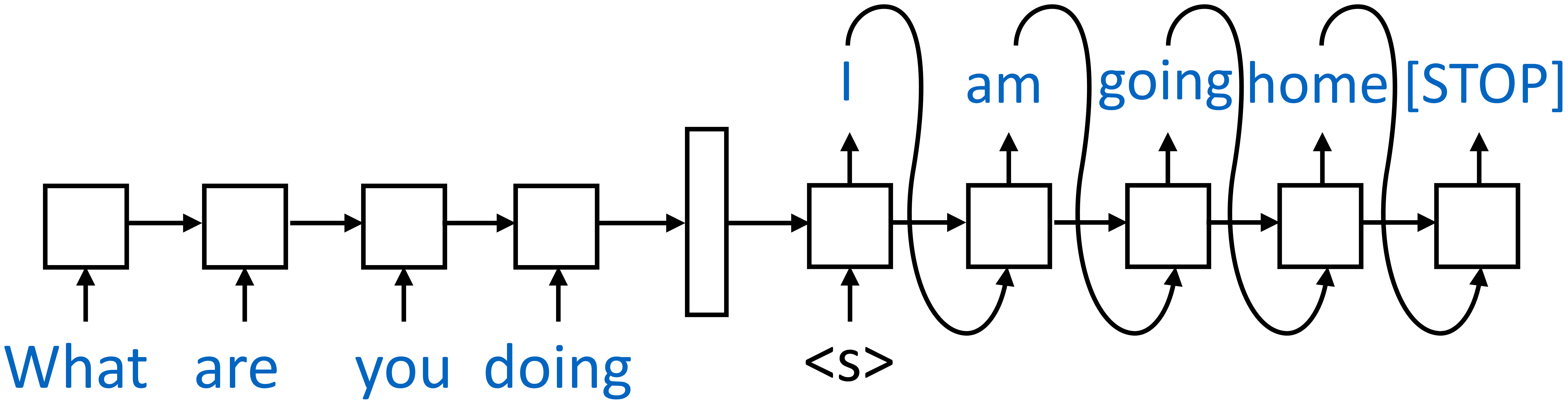


Status	MT-CHAT	MT-BASELINE	HUMAN
riding to the arc on my bike. Hopefully I won't melt into the pavement.	good luck on your bike ride arc	i ca n't wait for it ! i 'm arc is a year .	make sure to ride in the middle of the street
I really hate watching movies in the living room.	me too , what movies are 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)

Neural Chatbots

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

Subtitles Data

do you want to meet your sponsor for the last 10 years ?
of course ! but he doesn’t want to see me !

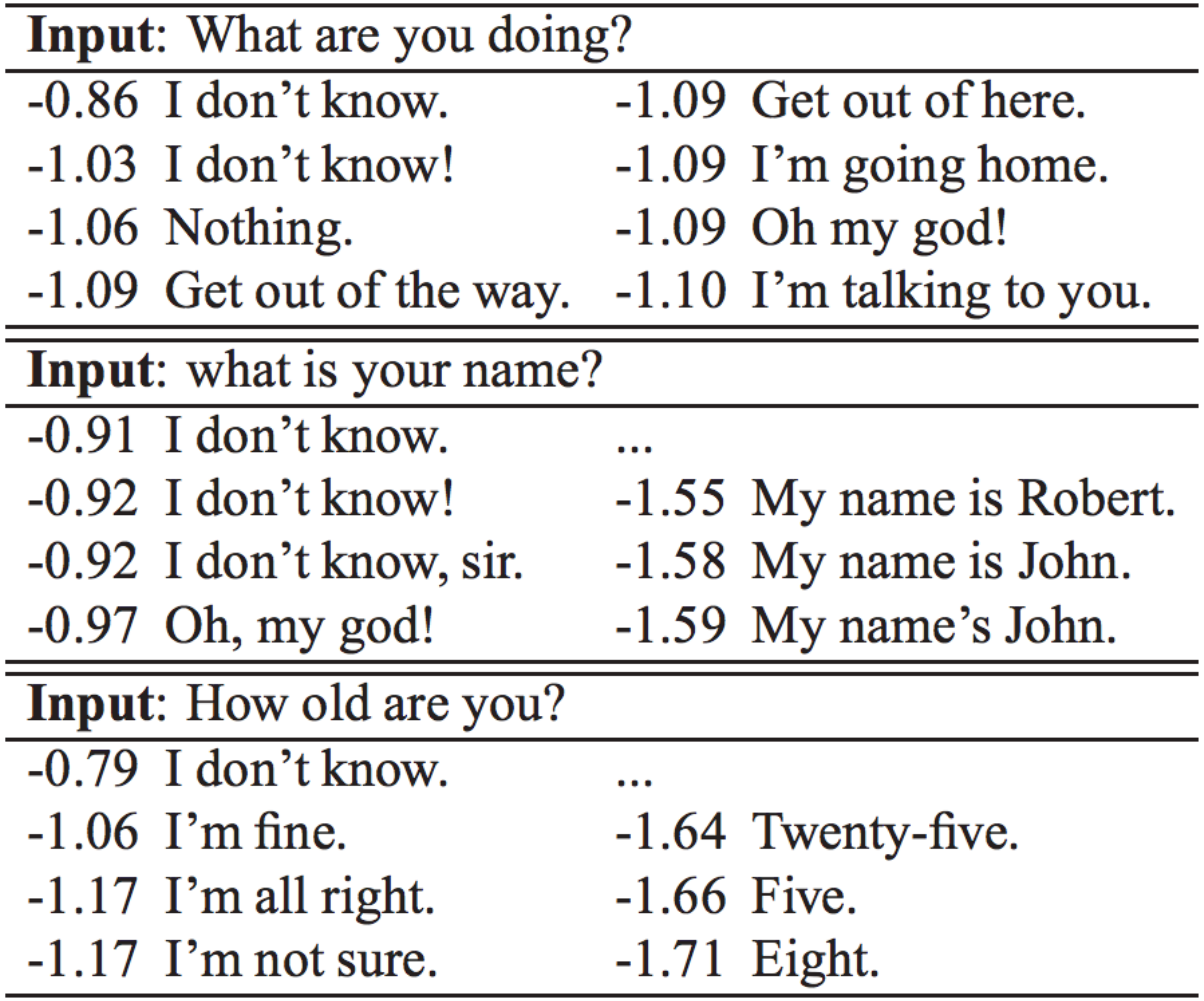
and where had you been just before ?
i' d been to the palace of the legion of honor , the art gallery .

yeah , we were just going to hit up taco bell .
well , it' s my pleasure .

- How can attention be useful?

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

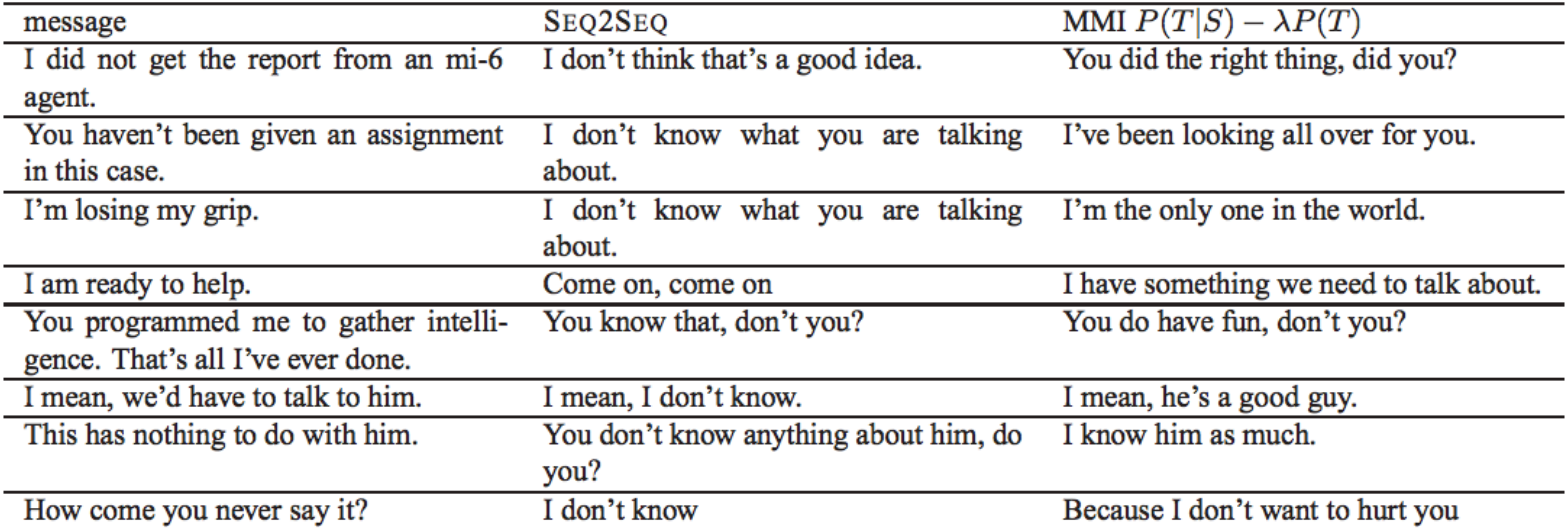
- Standard conditional likelihood: $\log P(R|U)$

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

- log P(R) reflects probabilities under a language model

Li et al. (2016)

Lack of Diversity



message	SEQ2SEQ	MMI $P(T\|S) - \lambda P(T)$
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

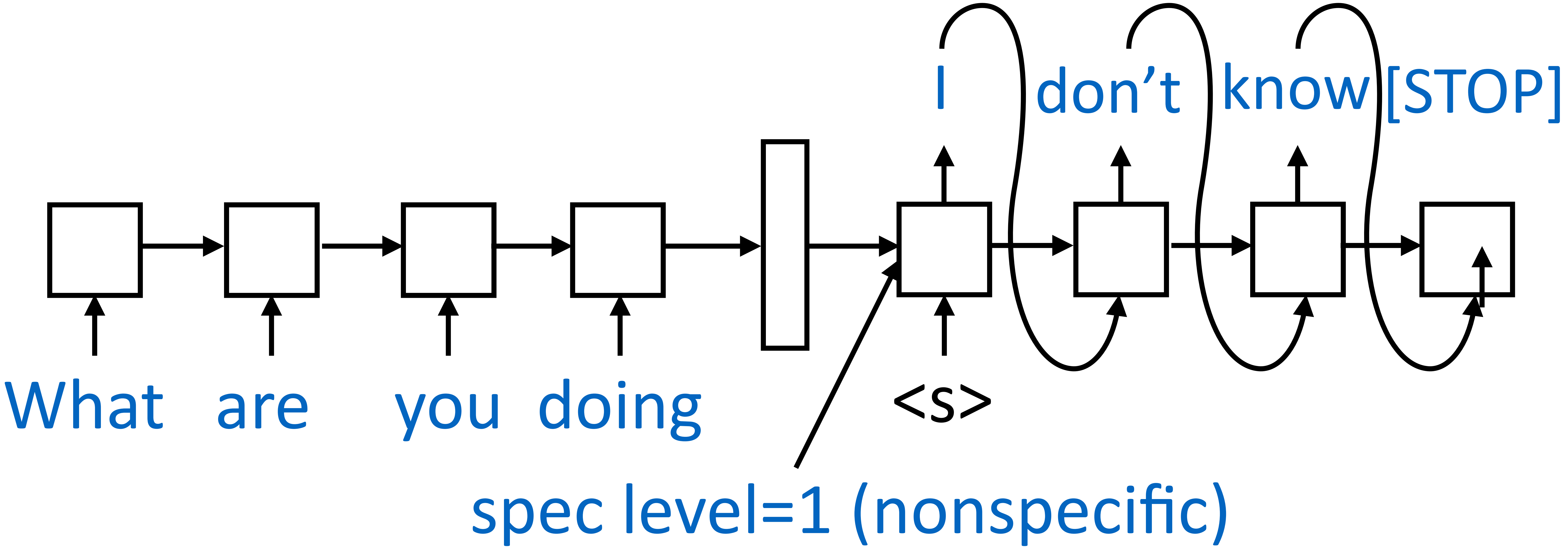
Li et al. (2016)

Specificity

- Train a specificity classifier on labeled data

I don’t know => spec level 1 Going to the store => spec level 3

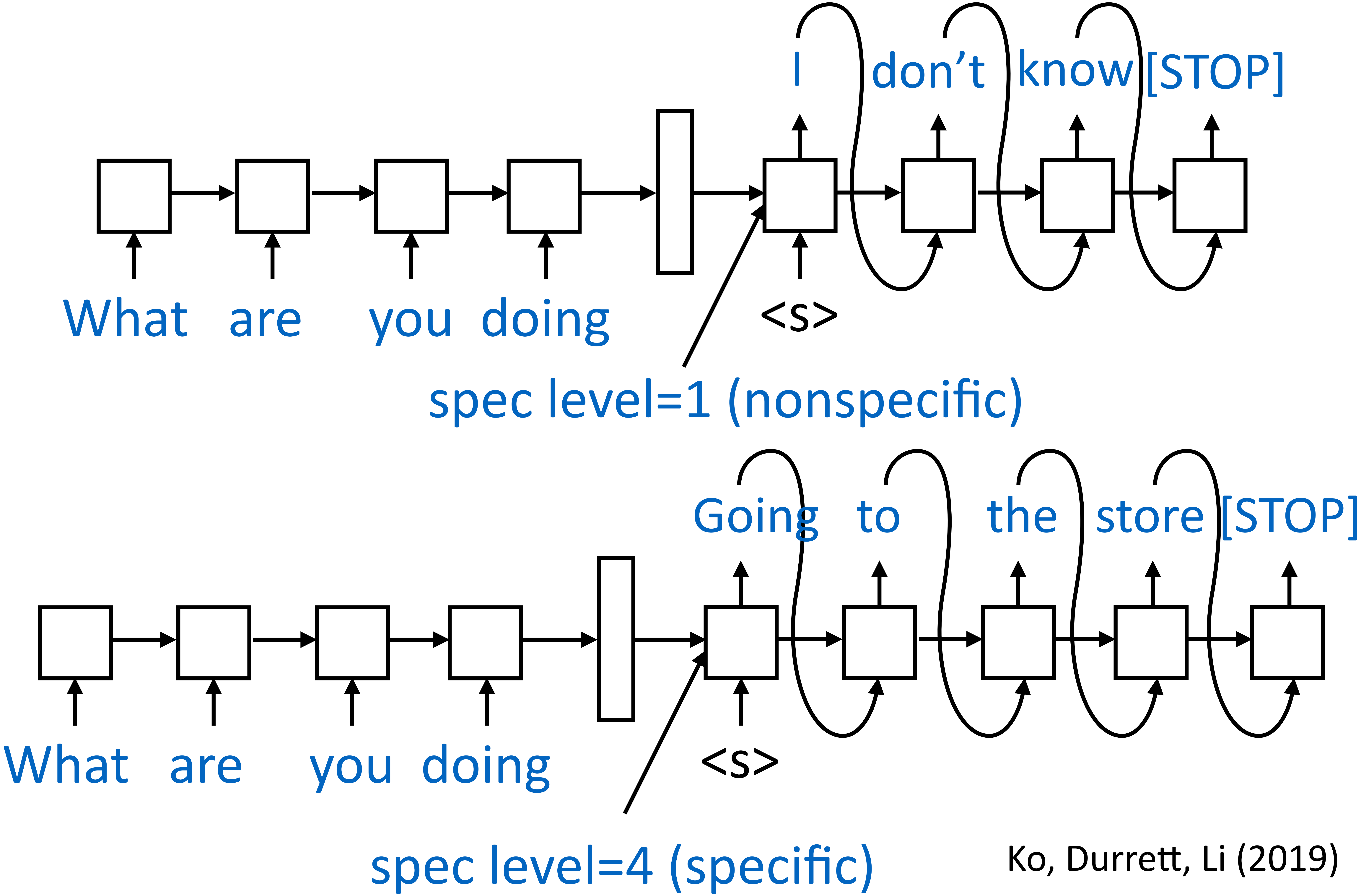
- When training the decoder, condition on the specificity of the response



Ko, Durrett, Li (2019)

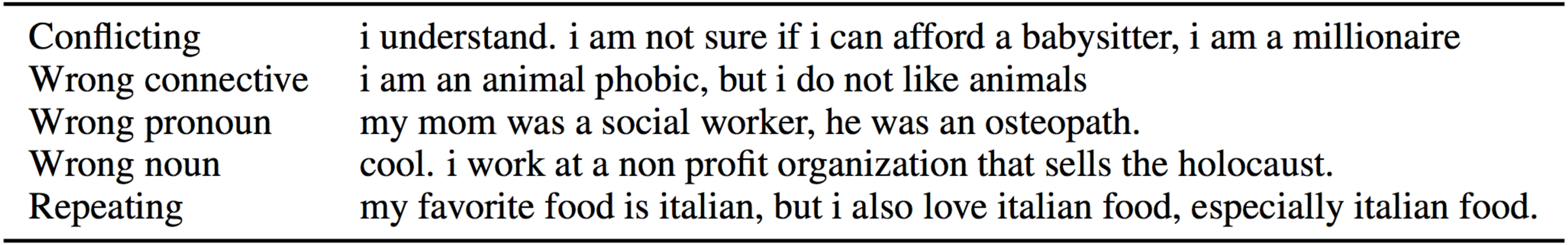
Specificity

- At test time, set the specificity level higher to get less generic responses



Ko, Durrett, Li (2019)

Specificity



Conflicting	i understand. i am not sure if i can afford a babysitter, i am a millionaire
Wrong connective	i am an animal phobic, but i do not like animals
Wrong pronoun	my mom was a social worker, he was an osteopath.
Wrong noun	cool. i work at a non profit organization that sells the holocaust.
Repeating	my favorite food is italian, but i also love italian food, especially italian food.

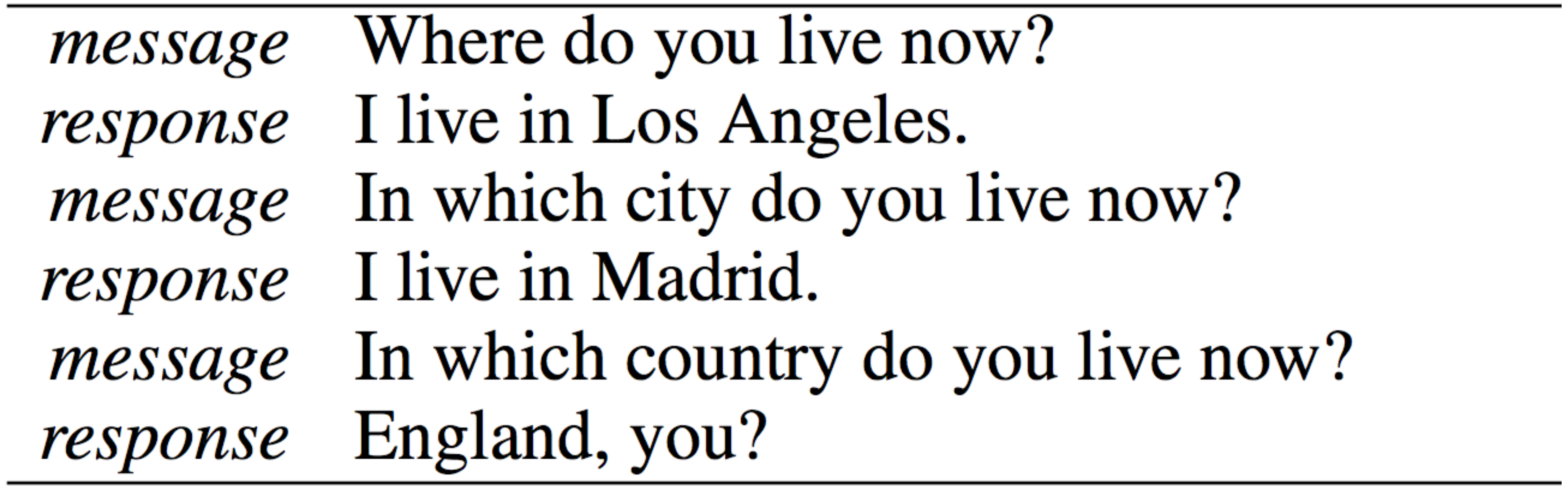
- Can use other models to try to fix these issues. But the facts are still all made up, even if they make sense

Ko, Durrett, Li (2019)

Agent Personas

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

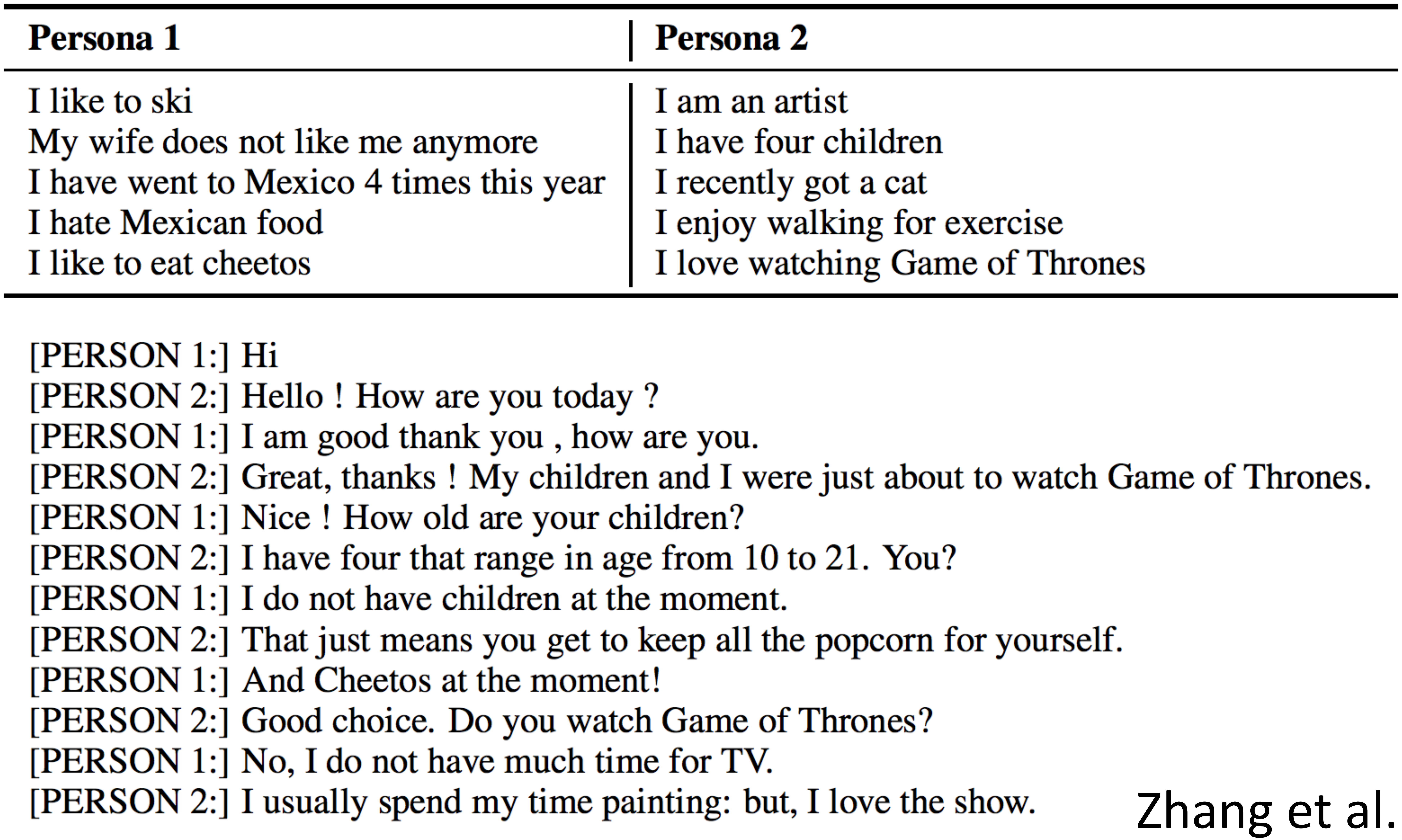
- Give the bot a *persona*: set of facts that it can at least consistently report on



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…

PersonaChat

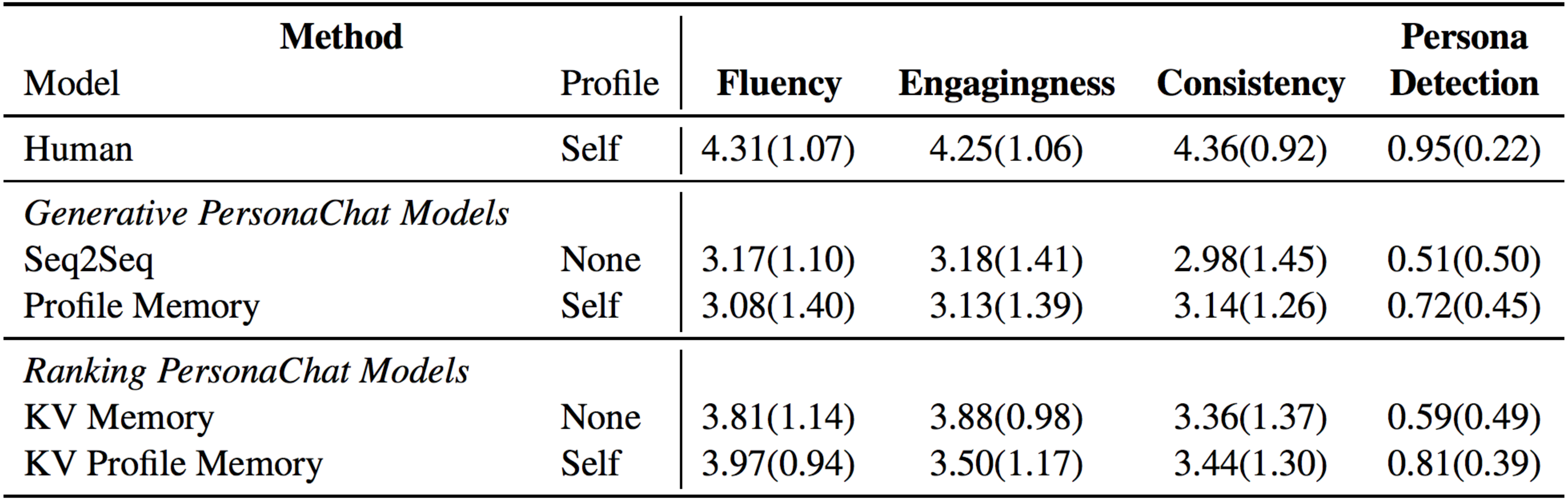


Persona 1	**Persona 2**
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[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!
[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.

Zhang et al. (2018)

PersonaChat



Method					**Persona**
Model	Profile	**Fluency**	**Engagingness**	**Consistency**	**Detection**
Human	Self	4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)
Generative PersonaChat Models					
Seq2Seq	None	3.17(1.10)	3.18(1.41)	2.98(1.45)	0.51(0.50)
Profile Memory	Self	3.08(1.40)	3.13(1.39)	3.14(1.26)	0.72(0.45)
Ranking PersonaChat Models					
KV Memory	None	3.81(1.14)	3.88(0.98)	3.36(1.37)	0.59(0.49)
KV Profile Memory	Self	3.97(0.94)	3.50(1.17)	3.44(1.30)	0.81(0.39)

- Ranking: retrieve utterance from training set to use at test time

Zhang et al. (2018)

Alexa Prize

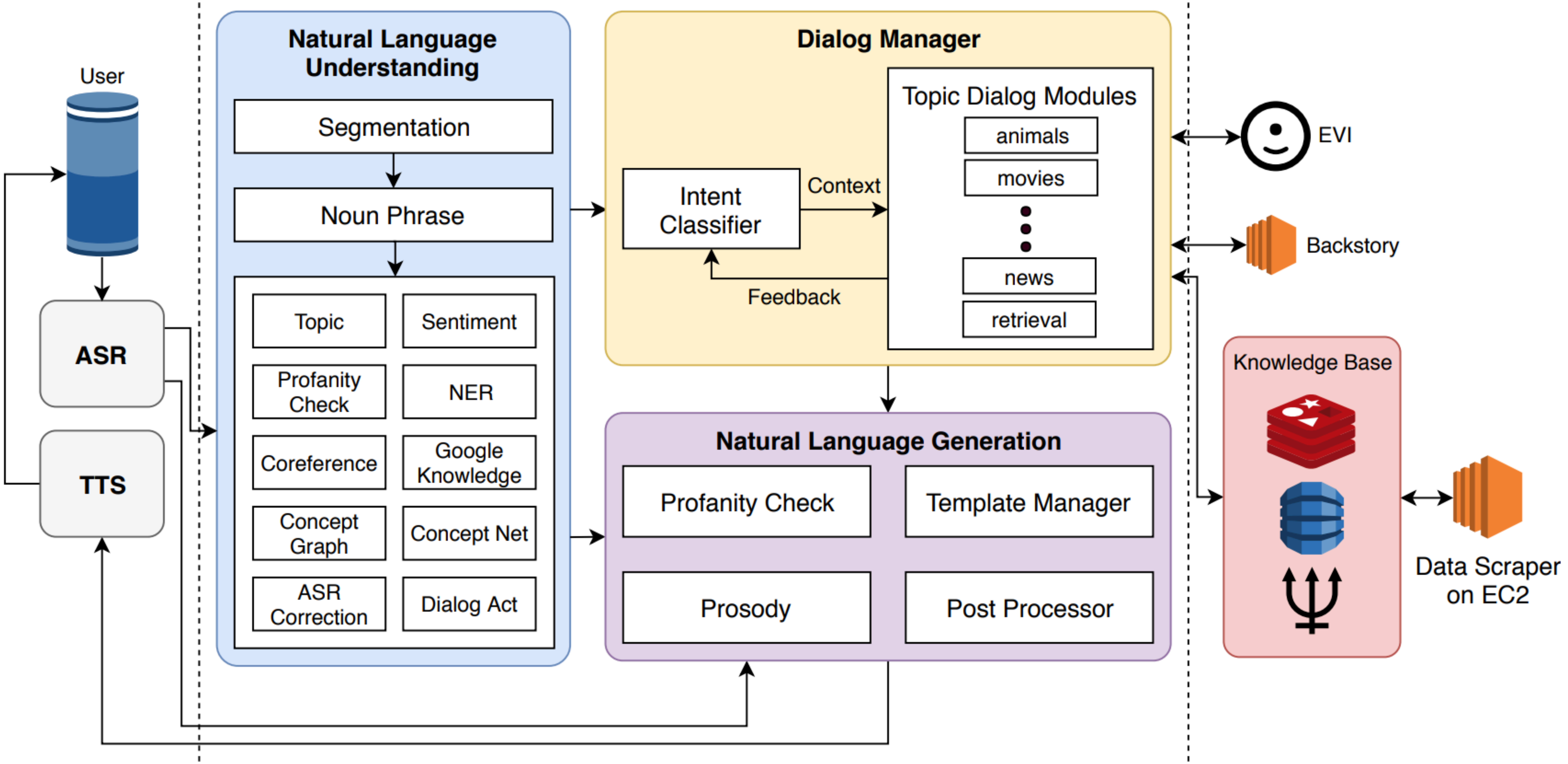
Amazon Alexa Prize

- 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



Chen … Zhou Yu (2018)

Gunrock System

- NLU: sentence segmentation, parsing, NER, coref, dialogue act prediction
- Detect: topic intents, lexical intents, other intents (tell user to exit)
- 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)

Takeaways

- State-of-the-art chatbots are heavily hand-engineered
- Neural methods are much less robust and require lots of coercion right now, hard to get data for the desired UX
- XiaoIce: Microsoft chatbot in Chinese, 100M+ users, 30 billion turns total, average user interacts 60 times/month
- People do seem to like talking to them…?
- Next time: task-oriented systems (Siri, etc.)