CS388: Natural Language Processing

Lecture 23: Dialogue

Greg Durrett
TEXAS
The University of Texas at Austin

Credit: Megha Desai on Medium

Administrivia

- Project 2 back next week
- Two talks coming up:
  - Friday, November 15 11am GDC 6.302: Yoav Artzi (Cornell Tech)
  - Friday, November 22 11am GDC 6.302: Jacob Andreas (MIT)

This Lecture

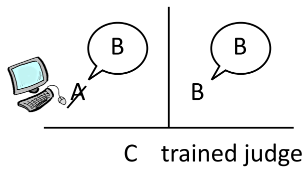
- Classical chatbots
- Neural chatbots
- Task-oriented dialogue
- QA as dialogue

Chatbots

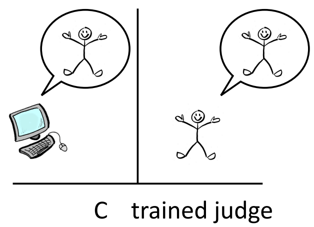
Turing Test (1950)

- 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:



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

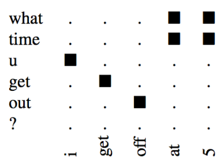
Cleverbot

- 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



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

Ritter et al. (2011)

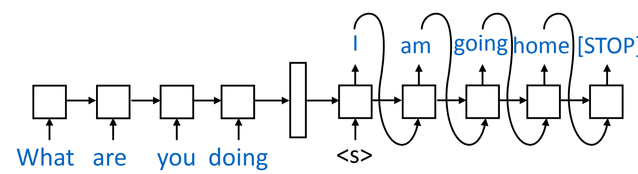
Data-Driven Approaches

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

- 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 .

- Why might this model perform poorly? What might it be bad at?

Lack of Diversity

- Training to maximize likelihood gives a system that prefers common responses:
- Why?

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: $\log \dfrac{P(R,U)}{P(R)P(U)} = \log P(R|U) - \log P(R)$
- log P(R) = 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

- OpenSubtitles data

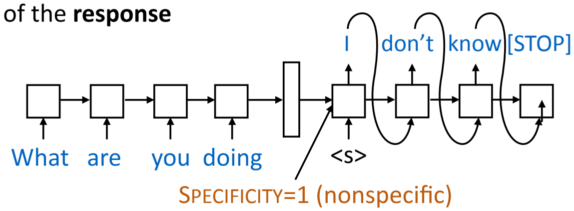
Li et al. (2016)

Specificity

- Train a specificity classifier on labeled data

  I don't know => SPECIFICITY=1 Going to the store => SPECIFICITY=3

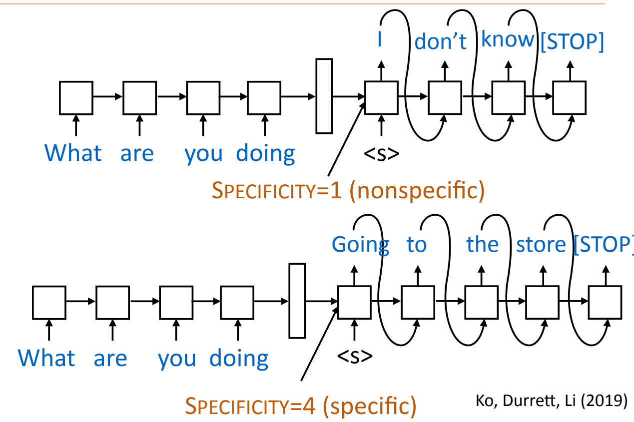
- When training the decoder, condition on the **automatically predicted specificity** of the **response**



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)

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)

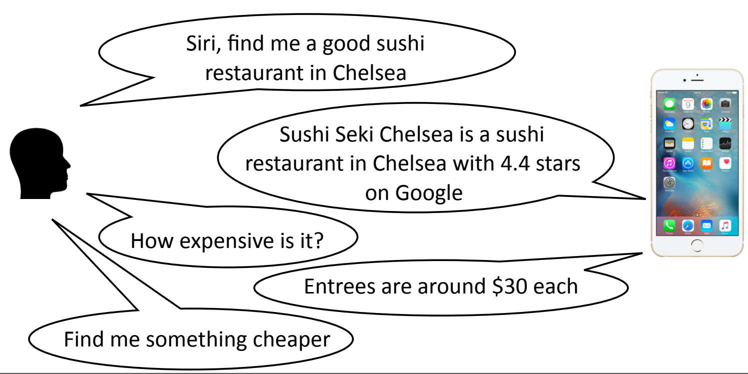
State of Chatbots

- Can force chatbots to give consistent answers with a persona, but still probably not very interesting
- "Wizard of Wikipedia:" chatbot that can discuss topics by retrieving from Wikipedia [Dinan et al., 2019)
- XiaoIce: Microsoft chatbot in Chinese, 20M users, average user interacts 60 times/month…people do seem to like talking to them…?

Task-Oriented Dialogue

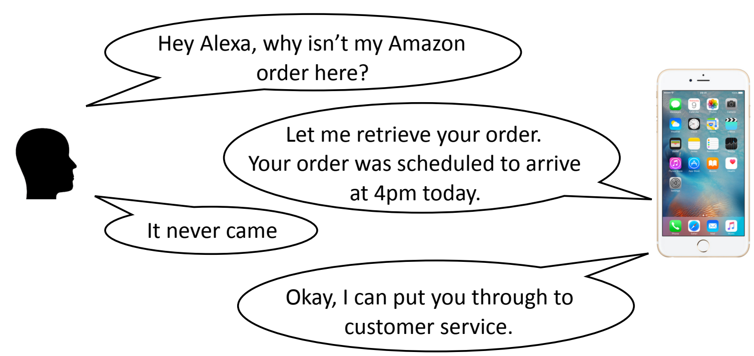
Task-Oriented Dialogue

- Personal assistants / API front-ends:



Task-Oriented Dialogue

- 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)

Intents

- 29 different intents

which flights go from cleveland to indianapolis on april fifth

Intent: flight

does tacoma airport offer transportation from the airport to the downtown area

Intent: ground_service

what days of the week do flights from san jose to nashville fly on

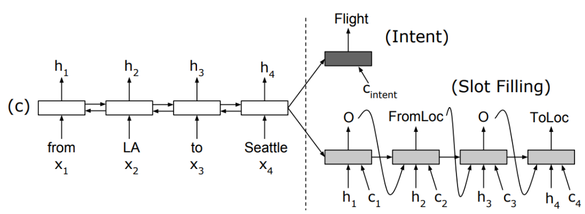
Intent: day_name

what meals are served on american flight 811 from tampa to milwaukee

Intent: meal

Joint Intent Classification and Tagging

- RNN jointly predicts intent and slot tags



Model	F1 Score	Intent Error (%)
RecNN [8]	93.22	4.60
RecNN+Viterbi [8]	93.96	4.60
Attention Encoder-Decoder NN (with aligned inputs)	**95.87**	**1.57**
Attention BiRNN	**95.98**	**1.79**

Liu and Lane (2016)

Air Travel Information Service (ATIS)

$\bar{x}_1$: *show me flights from seattle to boston next monday*

$\bar{y}_1$:
```
(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.from_airport IN (SELECT
airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN (SELECT
city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND (flight.to_airport IN (SELECT
airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN (SELECT
city.city_code FROM city WHERE city.city_name = 'BOSTON'))) AND (flight.flight_days IN (SELECT
days.days_code FROM days WHERE days.day_name IN (SELECT date_day.day_name FROM date_day WHERE
date_day.year = 1993 AND date_day.month_number = 2 AND date_day.day_number = 8))));
```

$\bar{x}_2$: *on american airlines*

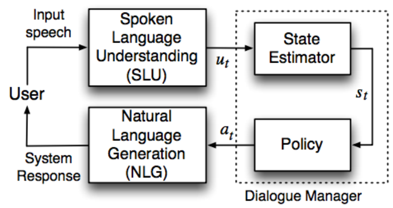
$\bar{y}_2$:
```
(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.airline_code = 'AA') AND (flight.from_airport
IN (SELECT airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN (SELECT
city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND (flight.to_airport IN (SELECT air
port_service.airport_code FROM airport_service WHERE airport_service.city_code IN (SELECT city.city_code
FROM city WHERE city.city_name = 'BOSTON'))) AND (flight.flight_days IN (SELECT days.days_code FROM
days WHERE days.day_name IN (SELECT date_day.day_name FROM date_day WHERE date_day.year = 1993 AND
date_day.month_number = 2 AND date_day.day_number = 8))));
```

- Need to use dialogue context to do the right thing. Here we're appending American Airlines as a constraint to the previous query
- seq2seq model mapping to query with copy mechanism

Suhr et al. (2018)

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()
```

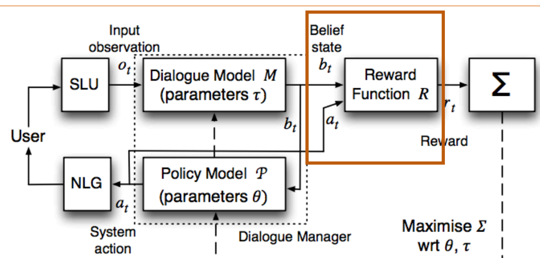
Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

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

How does the user know the right search happened?

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

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

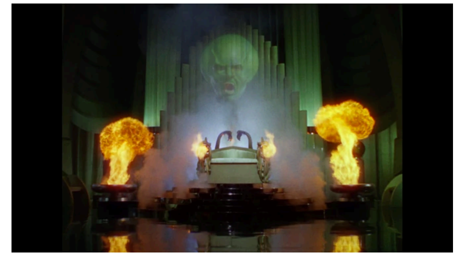
How expensive is it?

```
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

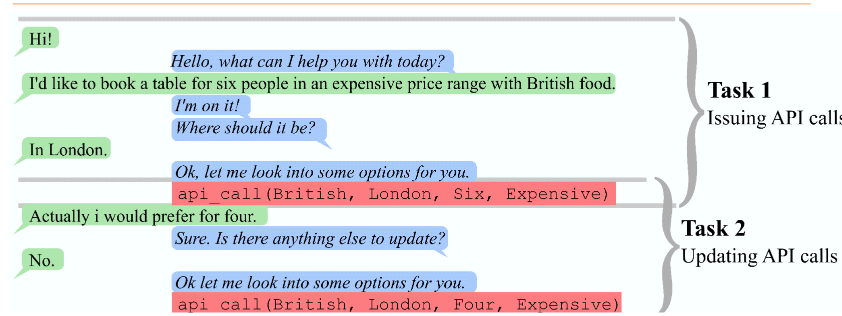
```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

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



- 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()
```

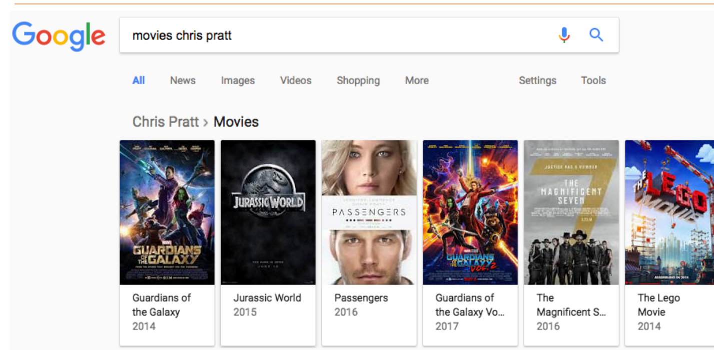
- User asked for a “good” restaurant — does that mean we should change our model to filter by star rating? What does “good” mean?
- Hard to change system behavior if training from static traces, especially if system capabilities or desired behavior change

Goal-oriented Dialogue

- Tons of industry interest!
- Dozens of startups + medium-sized companies in this space
- Big Companies: Apple Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, Facebook, Samsung Bixby, Tencent WeChat, ASAPP
- Lots of cool work that’s not public yet

Other Dialogue Applications

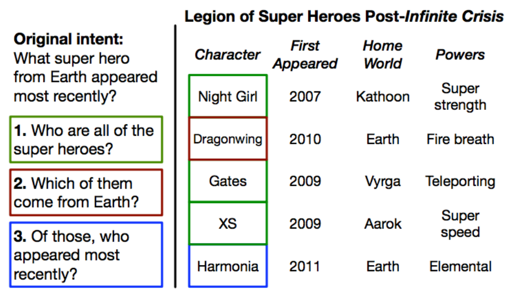
Search/QA as Dialogue



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

QA as Dialogue

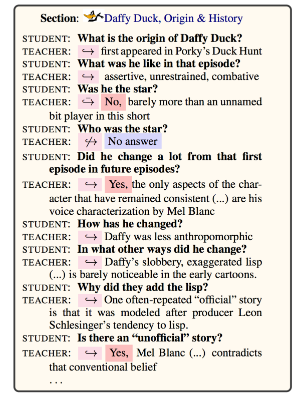
- Dialogue is a very natural way to find information from a search engine or a QA system
- Challenges: hard to annotate good dialogue datasets in a purely static way



Iyyer et al. (2017)

QA as Dialogue

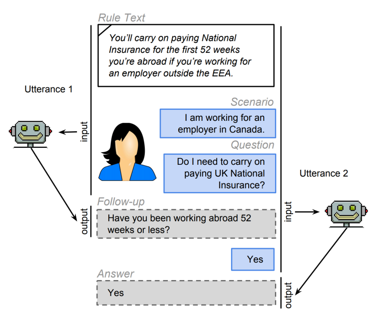
- UW QuAC dataset: Question Answering in Context



Choi et al. (2018)

QA as Dialogue

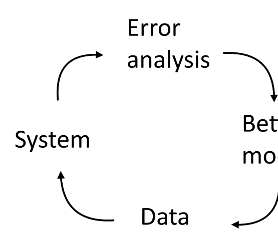
- Conversational machine reading: answer repeated questions based on a passage
- Interesting and potentially useful idea, but annotating data is very hard!



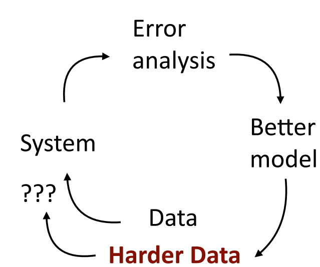
Saeidi et al. (2018)

Dialogue Mission Creep

Most NLP tasks

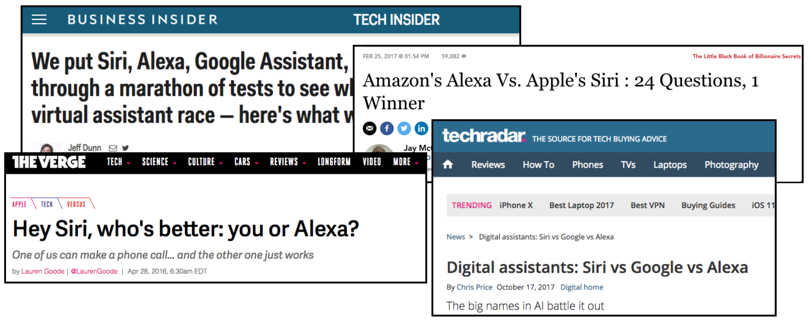


Dialogue/Search/QA



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

Dialogue Mission Creep



- 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